STUDY OF THE KEY DETERMINANTS OF STATISTICAL POWER IN LARGE SCALE GENETIC ASSOCIATION STUDIES

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester

By

Amadou Gaye BSc MSc

Department of Health Sciences

University of Leicester

2012

ABSTRACT

STUDY OF THE KEY DETERMINANTS OF STATISTICAL POWER IN LARGE SCALE GENETIC ASSOCIATION STUDIES

AMADOU GAYE

A large number of participants is often required by association studies investigating the causal mechanisms of complex diseases because of the generally weak causal effects involved in these conditions. The large sample sizes necessary for adequately powered analyses are mainly achieved by large studies. This can be an expensive undertaking and it is important that the correct sample size is identified. But, the analysis of the statistical power of large consortia and major biobanks demands that a number of complicating issues are taken into proper account. This includes the impact of unmeasured aetiological determinants and the quality of measurement of both outcome and explanatory variables. Conventional methods to analyse power use closed-form solutions that are not flexible enough to allow for these elements to be taken easily into account and this results in a potentially substantial overestimation of the actual power.

In this thesis, I describe the radical rebuilding of an existing power calculator known as ESPRESSO to develop and implement the ESPRESSO-forte algorithm. ESPRESSO-forte is intended as a comprehensive study simulation platform aimed at supporting the design of large scale association studies and biobanks. I then applied the newly developed software to two real world scientific problems: (1) to assess the power of a large multi-provincial Canadian cohort for the study of quantitative traits; and (2) to estimate the impact of the particular standard operating procedures that were applied to the collecting and processing of biosamples in UK Biobank, on the likely power of future nested case-control studies.

Some analyses now explore the role of copy-number variants (CNVs) in disease. I evaluated the accuracy of CNVs genotypes measured on four SNP genotyping platforms to inform future studies that plan to use existing SNP intensity data to measure CNVs or carry de novo CNV measurements from SNP genotyping platforms.

ACKNOWLEDGEMENT

I would firstly thank Professor Paul R. Burton and Professor Martin D. Tobin for offering me the opportunity to undertake this doctoral work. Thanks go to Paul, Martin and Dr Louise V. Wain for their supervision and encouragement throughout the thesis; I cannot put money on all what I learned from them.

I am pleased to thank my wife Myriam for her patience and support over the past years. I am grateful to my mum and dad who never went to school but yet emphasized to me the importance of education and knowledge at a young age.

I would like to express my gratitude to Boubacar Khouma and Dr Djabir Yahya who respectively sponsored me and supported me at some critical moments of my educational career. I would probably have not pursued higher education without the help of Boubacar.

Lastly I would like to thank the students and staff of the department who made my stay pleasant. I did really appreciate the time I spent with all of them.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

1. GENERAL INTRODUCTION

1.1. BACKGROUND AND AIM

According to the World Health Organisation, common chronic diseases such as asthma, diabetes, cancer and cardiovascular diseases are expected to account for 57% of the burden of all diseases by 2020 (**1**). Common diseases are caused by complex interactions between genetic determinants as well as between genetic and environmental factors (**2, 3**). They represent an important public health issue which is expected to worsen in the future. It is hence important to develop and improve methods that would help to understand the genetic mechanisms of these diseases for a better prevention, diagnosis and treatment.

The field of genetic epidemiology has raised hope for the understanding of the pathogenesis of common diseases (**4-6**) by means of genetic association studies. There are, however, some technical challenges to solve for an optimal use of these studies. One of the limitations of genetic association studies in the investigation of the genetic basis of common diseases is the lack of power: the ability to detect the *true* genetic determinants of a disease (**7, 8**).

This PhD project aims to extend our knowledge of the key determinants that influence the power of genetic association studies in order to help us to design studies that take full account of their impact. The work has three principal elements:

a. Explore and understand the biology, epidemiology and statistical characteristics of the factors that most critically determine the statistical power of contemporary genetic association studies.

b. Contribute to the extension of the study simulation platform known as ESPRESSO that has explicitly been designed to enable the impact of the key power determining factors to be studied and to be taken into full account in designing large scale biobanks and genetic association studies.

c. Apply the extended ESPRESSO version to answer relevant scientific questions that have been identified as being critical in these regards.

The work undertaken in this PhD required the understanding and application of some key concepts in genetics and genetic epidemiology. The remainder of this chapter is a recap of those fundamental concepts, a review of some previous studies that have influenced the field and a brief introduction to the method development and analyses carried out in the thesis.

1.2. GENETICS

1.2.1. DNA, GENOME AND GENES

Deoxyribonucleic acid (DNA) is the nucleic acid that carries genetic information in humans. This information is stored as a code of four chemical bases or nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA contains about 3200 million nucleotide pairs (A paired with T and G paired with C) maintained by a sugar phosphate backbone forming a double helix (Figure 1). DNA can replicate identically by opening up and using each strand of the double helix as a template for a new complementary strand. The entire genetic complement of an individual is referred

to as the genome. The human genome consists of 23 pairs of chromosomes: 22 paired autosomes, one X-chromosome and one Y-chromosome for males, and 22 paired autosomes and two X-chromosomes for females. Since they have two sets of homologous chromosomes, humans are diploid.

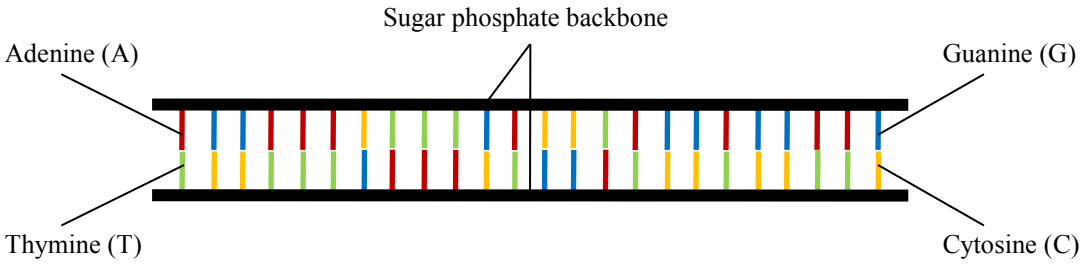


Figure 1: Simplified view of DNA structure.

Genes are segments of chromosomes coding, mainly, for proteins (polypeptide chains). Gene expression is the process that leads to a functional product, usually a protein. The gene is transcribed into messenger RNA (mRNA), a single stranded ribonucleic acid similar to DNA with the base uracil (U) instead of thymine (T) and the sugar ribose instead of deoxyribose, and following some degree of post-transcriptional processing, the mature mRNA is ultimately translated into an amino acid chain (Figure 2). It is the DNA sequence that determines the mRNA sequence which then determines the amino acid sequence of the protein. Most genes are made up of coding segments (exons), which are ultimately translated into protein, and non-coding segments (introns) spliced out during transcription. Gene expression is mainly regulated by the control of the rate of transcription. Transcription can be up-regulated by a regulatory region named an enhancer that can be located thousands of base pairs upstream of the regulated gene. There are other regulatory elements termed silencers that can suppress or down-regulate gene expression.

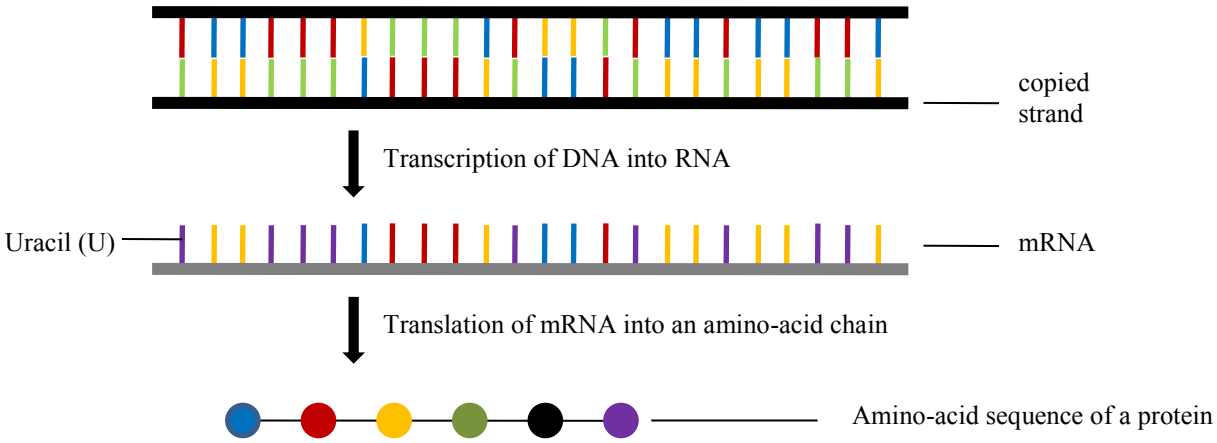


Figure 2: Overview of gene expression.

Each human genome is unique partly due to genetic recombination and mutations affecting the genome.

1.2.2. GENETIC RECOMBINATION AT MEIOSIS

Gametes are haploid human cells. The process by which gametes are formed is termed meiosis. Meiosis is a cell division process which results in the production of four haploid cells (gametes) from one diploid cell. In the first stage of meiosis the chromosomes in the diploid cell are duplicated to obtain a tetrad of chromosomes. Homologous chromosomes pair-up and the cell undergoes a division resulting in two diploid cells. In the second stage of meiosis, each of these diploid cells divides to produces two haploid cells, resulting in four haploid gametes. When homologous chromosomes pair-up in meiosis they exchange (crossover) some segments so that the resulting gametes are a mixture of segments from the initial chromatids; this re-composition of chromosomes during meiosis is one form of genetic recombination. Crossovers do not occur at an even rate across the genome. There are some regions of the genome, termed recombination hotspots (**9**), where a crossover is more likely to occur and other regions where crossover is less likely to occur.

1.2.3. GENETIC POLYMORPHISMS AND LINKAGE DISEQUILIBRIUM

Genetic polymorphisms are mutations caused by errors occurring during DNA replication or by external agents (mutagens) acting on the DNA. A mutation that occurs within a gene can alter the product of gene expression whilst a mutation occurring within a regulatory region can affect the regulation of gene expression; in both cases there could be some functional consequences for the individual. Mutations that cause a change in the amino acid chain of the protein are termed non-synonymous mutations and mutations that do not cause a change in the amino acid chain of the protein are synonymous mutations, also termed silent mutations. A polymorphism which consists of the change of one single nucleotide (one base) is termed a Single Nucleotide Polymorphism (SNP). Some polymorphisms are genomic structural variations resulting from genomic rearrangements involving anywhere between a small number of nucleotides to megabase regions (**10**). One form of structural variation is copy-number variation (CNV) which is a duplication or deletion of a genomic region of hundreds to millions of base pairs. There are other types of polymorphism, but my work focuses on SNP and CNV polymorphisms which are covered in the next two sub-sections.

DNA sequences at a specific chromosomal position (locus) that vary between different copies of the same chromosome are termed alleles. Two or more alleles are observed at a polymorphic locus; an initial allele termed the wild type and the (new) mutant allele(s). The genotype of an individual at a particular locus is the combination of the alleles on each of the two homologous chromosomes. The individual is homozygous if both chromosomes present the same allele at that locus and he/she is heterozygous if the alleles are different.

1.2.3.1. Single Nucleotide Polymorphism

A single nucleotide polymorphism (SNP) results from an ancestral mutation that replaces one nucleotide by another. Most SNPs are biallelic; the frequency of the less common allele is termed minor allele frequency (MAF).There are approximately 10 million SNPs in the human genome (**6**). The HapMap project (**11**) has generated a large database of well characterized SNPs of varying minor allele frequencies. SNP variation is well characterised (e.g. HapMap) and SNPs are relatively straightforward to accurately genotype on a large scale; this makes SNPs suitable polymorphisms for use in genetic association studies.

SNP genotyping

SNP genotyping is the identification of the alleles at a SNP locus within one individual. The genotyping can be done using methods based on oligonucleotide hybridisation analysis. In such analysis a short single stranded oligonucleotide of about 50bp is synthesized and allowed to hybridise (pair up) with a single stranded DNA sequence (target DNA, DNA of the individual being genotyped) under stringent conditions that allow only for the formation of a fully paired hybrid. If both strands are fully complementary, the formed hybrid is stable but if two nucleotides cannot pair up at every position because they are not complementary (i.e. they contain different alleles) then the hybrid is not stable (Figure 3). This way the two alleles of a SNP can be identified since one forms a stable hybrid and the other not.

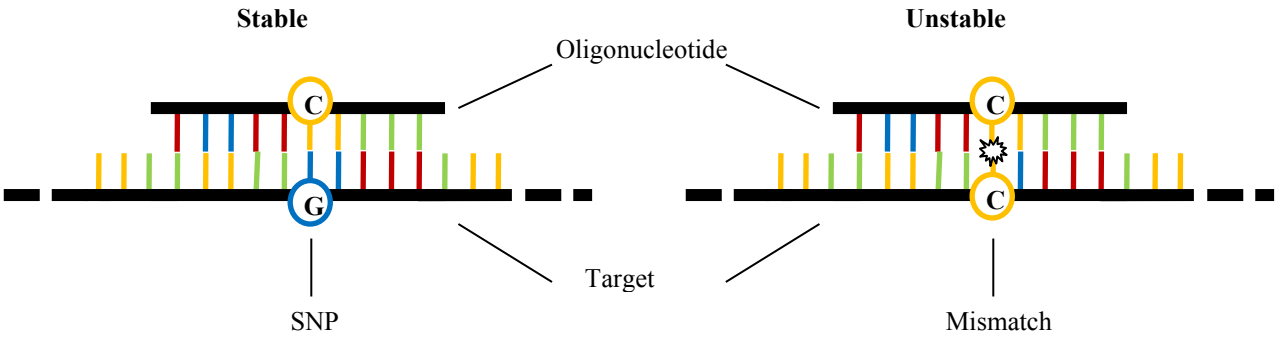


Figure 3: Overview of the oligonucleotide hybridization method.

The raw intensities from each allele can be represented graphically; in Figure 4, the y-axis is assigned to the red intensity (allele A) and the x-axis is assigned to the green intensity (allele C). Figure 4a shows the expected positions of three individuals (an individual homozygous for the *A* allele, a heterozygous individual and an individual homozygous for the C allele). Figure 4b shows the same plot for a population of individuals who have been assayed on the same array; individuals with the same underlying genotype will tend to cluster together. Genotype calling algorithms use these clusters to assign individuals to genotypes. Due to possible genotyping errors, poor DNA quality (sample quality) and calling algorithm errors some genotypes will not be determined precisely causing overlaps between clusters, outlying genotypes and undetermined genotypes (individuals that failed to genotype).

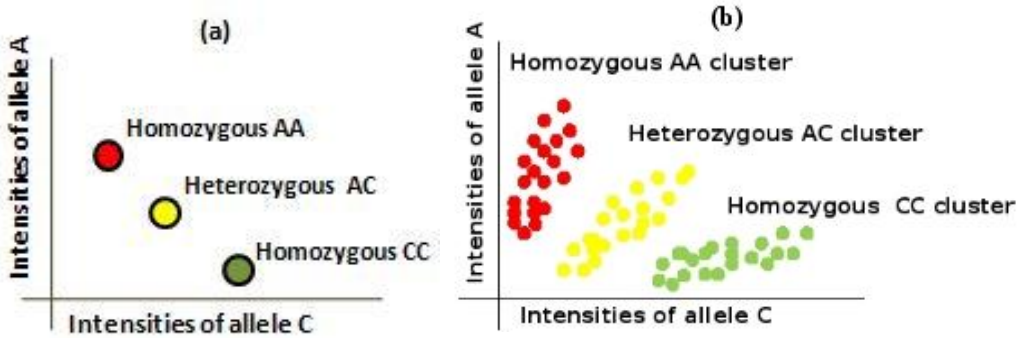


Figure 4: Graphical representation of genotype calls for one SNP.
The calls are for three individuals (a) or several individuals (b).

1.2.3.2. Copy-Number Variation

Copy Number Variation (CNV) is a form of genomic structural variation where DNA segments of hundreds to millions of base pairs are deleted or duplicated (**12**). CNVs can be simple deletions or duplications as well as multiallelic variation at a given locus involving both duplication and deletion.

It is not fully understood yet how all CNVs originate but some mechanisms have been suggested. CNVs can arise from non-allelic homologous recombination (NAHR) when, at meiosis (see section 1.2.2), crossover between chromosomes does not occur at the same homologous position on both chromosomes (non-symmetric – Figure 5) (**10**). This results in a gain or loss of segments of DNA. NAHR occurs because there is a similarity of sequence on two non-homologous positions of the chromosomes where the crossover took place. This phenomenon can also occur between two daughter chromatids (interchromatid NAHR) or within one chromatid which then forms a loop and crossover with itself (intrachromatid NAHR).

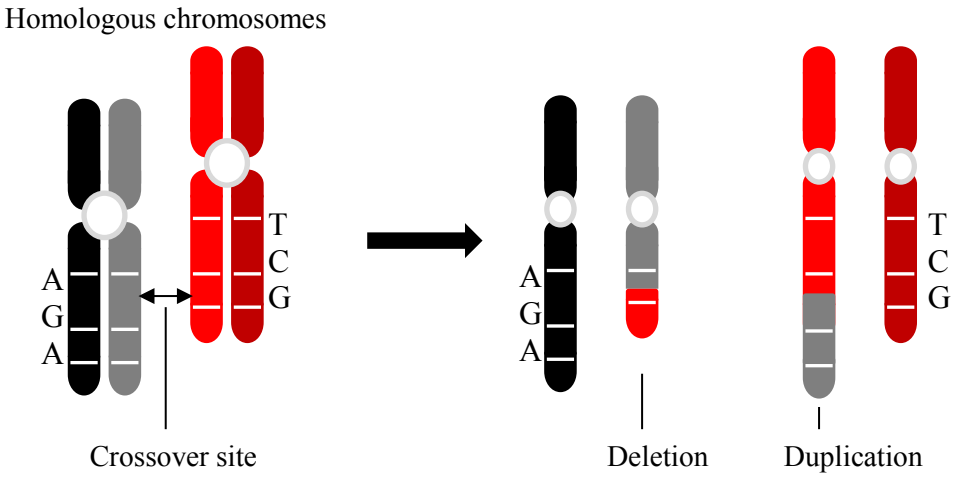


Figure 5: Genomic rearrangement due to interchromosal NAHR.

Non-homologous end joining (NHEJ) during DNA repair (Figure 6) can also lead to CNV. NHEJ is a mechanism used to repair DNA double-strand breaks which otherwise

could result in cell death (**13**). The two broken ends are joined without using a short homologous template and the new DNA segment may be elongated or shortened due to addition or loss of some nucleotides (**10**).

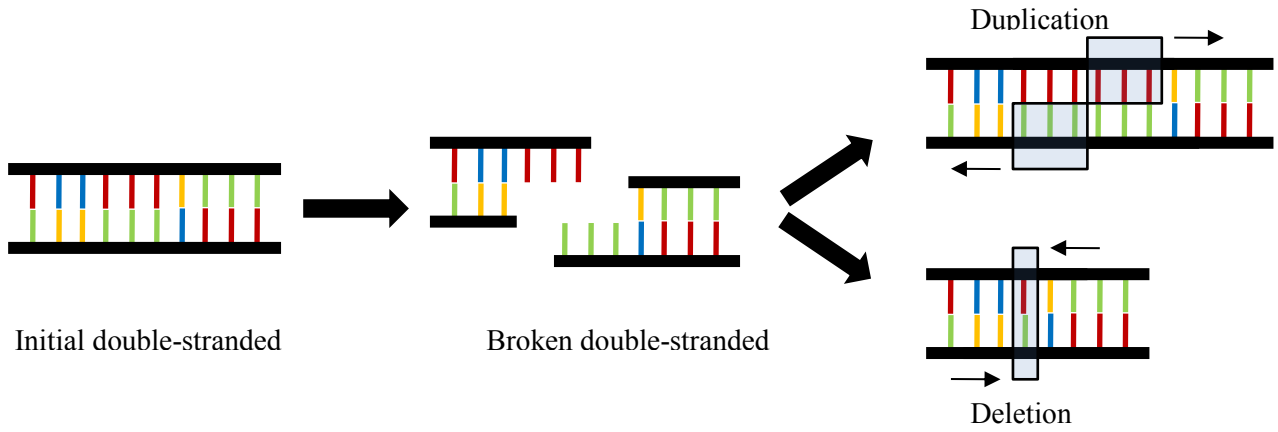


Figure 6: Genomic rearrangement due to NHEJ.

Array based Comparative Genome Hybridization (aCGH) and SNP genotyping arrays are two of the methods used to detect CNVs.

CNV detection and genotyping

In array based Comparative Genome Hybridization, a target DNA and a reference DNA are labelled differently with a fluorescent dye; they are then mixed and allowed to hybridize with probes disposed as spots on an array (Figure 7). The array is then scanned and the fluorescence of each spot recorded as an intensity which reflects the concentration of the spotted probe. The fluorescent ratio between target and reference DNA is finally calculated for each spot and plotted relative to the position of the probes on the genome.

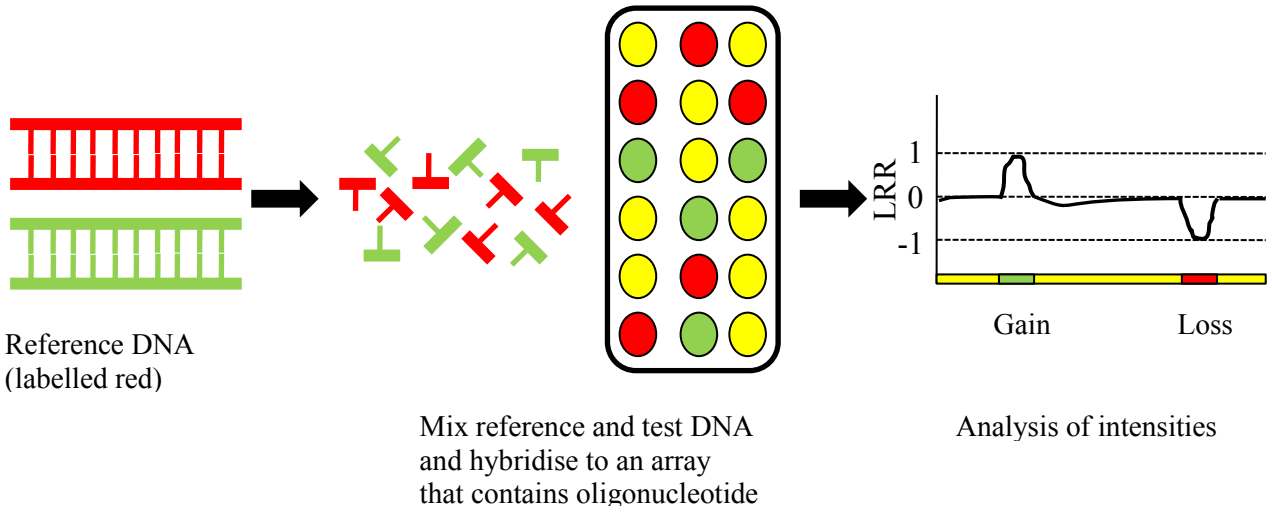


Figure 7: Overview of array-comparative genomic hybridization method.
The plot of the fluorescent ratios shows the regions of the genome where there is a gain or loss of DNA segment(s).

For each probe a logarithmic ratio of test DNA over reference DNA termed Log R Ratio (LRR), is computed and the values are normalized to have a mean of zero over the array. This ratio allows us to identify locations where the test DNA and the reference DNA have unequal copy number. Typically, if the reference DNA has two copies at a particular location whilst the test DNA has only one copy at that location, the log R ratio will be equal to -1.A log R ratio of 0 indicates two copies for both the test and the reference DNA. A log R ratio of 1means the copy number of the test DNA is twice the copy number of the reference DNA. Array-CGH enables the identification of regions of test and reference DNA that differ in copy number but cannot tell if the deletion or duplication occurred in the test or reference sample. For example, if LRR = -1, it is not possible to tell if what is observed corresponds to a deletion in the test DNA or a duplication in the reference DNA. Array-CGH is also limited in allowing one to determine the copy number as this must be inferred from the intensity ratio which can be very noisy. The resolution of array-CGH depends on the probe lengths; if probes have, for example, a length of 500 base-pairs (bp), the length of CNVs shorter than

500bp will be difficult to determine as it is impossible to tell how much shorter than the probe they are.

CNV genotyping can also be carried out using SNP genotyping array data. Some SNP arrays contain both SNP probes and CNV probes (CNVIs); CNVIs are non-polymorphic probes which, whilst not measuring sequence variation, are useful for regions likely to contain CNVs(**14**).

CNV genotyping consists of the identification of the number of diploid copies carried by an individual (copy number status). Several algorithms have been developed for calling CNVs and determining the copy number statuses of individuals; details of how the algorithm CNVtools works are covered in section 5.2.2.

One of the early limitations (prior to 2006) in the use of SNP arrays to study CNVs was their low resolution; it was not possible to fit very large numbers of SNPs. Also SNPs located within CNV regions were generally excluded from SNP arrays after failing a Hardy-Weinberg Equilibrium (HWE) test because CNV duplications or deletions cause a deviation of the SNP genotypes from HWE (the test principle states that if individuals in a population are randomly mating and there are no significant evolutionary forces acting; the frequencies of the genotypes present in the population and the frequencies of the alleles are stable over generations). A general limitation is related to the way to interpret overlapping CNVs and the ascertainment of CNV boundaries. A calling algorithm may detect overlapping CNV segments across individuals, and in this case it is difficult to determine if the detected overlapping regions actually represent signals from the same underlying CNV or if they are separate CNVs; thus there is a risk of misclassification (**15**). It is important to measure precisely the copy number level of an individual at a given CNV locus in order to investigate a genetic association without a

serious risk of bias. The analysis in chapter 5 is an investigation of the level of accuracy when copy numbers are measured using intensity data from SNP genotyping arrays.

Recent CNV maps

The recently developed CNV maps have improved the characterisation of CNVs and the on-going improvements allow one to undertake CNV association studies without carrying out CNV detection and characterisation at the same time. McCarroll *et al.*(**16**) used the Affymetrix *SNP Array 6.0* to analyse simultaneously SNPs and CNVs and develop a human CNV map. In the study carried out by McCarroll *et al.*, approximately half of the observed CNVs were present in many unrelated individual at a resolution of approximately 2kb. A large majority of the detected CNVs were inherited. The results also indicated that the contribution of some CNVs to disease can be captured by analysing the association between disease and SNPs that are correlated with CNVs. Conrad *et al.*(**17**)used a high density aCGH approach and developed a map of CNVs greater than ~500bp in length using data from two populations (European and West African). They detected ~11700 copy number variants; 77% of the common CNVs (CNVs with a frequency > 0.05) were tagged by SNPs. The study found 30 CNV loci that correlate with SNPs associated with complex traits, which means that these CNV loci may have some influence on complex traits. According to Conrad *et al.* most (80-90%) of the CNVs greater than 1kb in length have now been identified.

1.2.3.3. Linkage Disequilibrium

Linkage Disequilibrium (LD) is the dependence of alleles located at different positions in the genome at a population level, or the measure of correlation between alleles at population level (**18**). If two alleles at two different loci are in strong LD then each of the alleles may be used to predict the other.

LD calculation

If we consider two biallelic loci; locus A, with alleles ***A*** and ***a***, and locus C, with alleles ***C*** and ***c***, the combination of alleles that could occur together and their probabilities is summarized in Table 1 which is similar to an example presented by R.C. Lewontin (**19**).

Alleles	**C**	**c**
A	P_{AC}	P_{Ac}
a	P_{Ca}	P_{ac}

Table 1: Possible allelic combinations given two biallelic loci.
*P_{AC} = Probability of **A** and **C** occurring together. P_{Ac} = Probability of **A** and **c** occurring together. P_{Ca} = Probability of **C** and **a** occurring together. P_{ac} =Probability of **a** and **c** occurring together.*

The Pearson correlation coefficient r and Lewontin's coefficients D and D' are usually used to compute the magnitude of a linkage disequilibrium; $r^2 = 1$ denotes a perfect correlation and $r^2 = 0$ denotes the absence of correlation. The Lewontin's D coefficient is given by(**20**):

$$D = (P_{AC} \times P_{ac}) - (P_{Ac} \times P_{Ca})$$

The Lewontin's D' coefficient is a normalized version of D because D depends on alleles frequencies. If $D \geq 0$, D' is given by $D' = \frac{D}{D_{max}}$ where D_{max} is the smallest value between $(P_A \times P_a)$ and $(P_C \times P_c)$, the products of the probabilities of the occurrence of the alleles. If $D < 0$, D' is given by $D' = \frac{D}{D_{min}}$ where D_{min} is the largest value between $-(P_A \times P_C)$ and $-(P_a \times P_c)$ (**19, 20**).D' varies between -1 and 1, $|D'| = 1$ denotes maximum linkage (**21**). D' is inflated for small samples and not accurate for rare alleles (**22**). The square of the correlation coefficient, r^2,is given by:

$$r^2 = \frac{D^2}{P_A \times P_a \times P_C \times P_c}$$

Alleles at distinct loci, on the same chromosome, which are in high LD, tend to be inherited together as a block; such a set of alleles represents a haplotype. Haplotype usually refers to a set of SNP alleles but can also include other polymorphisms. Regions with low recombination exhibit higher LD than average (with extended haplotype blocks).Within a haplotype, information about the other SNPs in the haplotype block can be captured by one or more SNPs termed a tag SNP(s). Tag SNPs are very useful for genetic association studies because they reduce the number of SNPs that it is necessary to genotype in order to achieve efficient coverage of the genome. The aim of the HapMap project (**11**) was to capture most of the genetic variation in the human genome by identifying haplotypes and creating a map of haplotypes where tag SNPs would facilitate association studies between common genetic variants and disease. The 1000 Genomes Project is establishing a more extensive database of human genetic variation (common and rare sequence variants and, where detected, structural variations) by sequencing the genomes of many individuals from different populations. This project will also help to determine the relative allelic frequencies of the detected genetic variants across several populations.

1.2.4. GENETIC POLYMORPHISMS AND DISEASE

The change in the quality or quantity of the product of gene expression caused by variants within genes or within regulatory regions may explain why some polymorphisms play a role in the onset of diseases as well as influencing continuous traits. Both SNPs and CNVs have now been associated with a number of diseases or continuous traits. For example, the SNP rs1801282 in the *PPARG* gene was found to be associated with type 2 diabetes (**23, 24**). Two SNPs, rs10516526 and rs2571445, located

respectively in the genes *GSTCD* and *TNS1* were found to be associated with the trait FEV_1 - forced expiratory volume in 1 second (**25**).

CNVs may lead to functional consequences by affecting gene dosage, the number of functional copies of a gene, which may determine the amount of functional protein produced. They could also disrupt regulatory regions of genes. For example, a common CNV lying about 20kb upstream of the *IRGM* gene is strongly associated with Crohn's disease and suspected to play a role in the disease by altering the regulation of *IRGM* (**26**). A CNV on chromosome 17 is believed to be associated with the rate of AIDS progression. This CNV is subject to a varying number of duplications; a higher copy number of the gene*CCL3L1* is believed to be associated with a decreased susceptibility to HIV-1 infection (**27**). The effect of CNVs could have other clinical implications such as speeding up or slowing down the way drugs are metabolised in the body and hence inducing toxicity in the event of high concentration of a drug or reducing drug efficacy if the concentration of the drug is too low. Examples of this are deletions and duplications affecting the expression of *CYP2A* and *CYP2D* genes that are involved in drug metabolism (**28**). Understanding relationships between CNVs and disease has so far been limited by inaccuracies in measuring the number of copies within individuals. Accordingly, the analysis in chapter 5 specifically investigates the inferential impact of inaccuracy in the number of copies measured from SNP genotyping platform data.

Studying associations between polymorphic sites (genetic variants) and traits can lead to the identification of genes involved in the pathways of disease. This is an aim pursued by genetic epidemiologists through methods such as genetic association studies.

1.3. GENETIC EPIDEMIOLOGY

Genetic epidemiology has been defined as *"a science which deals with the aetiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations"* (**29**). Whilst traditional epidemiology explores environmental (in a wide sense) and social factors influencing the occurrence and development of disease, genetic epidemiology addresses the genetic causes of disease and investigates interactions between genetic determinants and the joint effect of genetic and environmental determinants of disease.

Although genetic epidemiology and classical epidemiology share many methods; genetic epidemiology presents unique challenges that require analytical approaches different to those of traditional epidemiology. For example in association studies, genetic epidemiology can use biological information such as LD to infer the existence of an association without directly measuring a causal variant. This is achieved by establishing an indirect association between another variant that is in LD with the causal one, and can also be demonstrated to be correlated with the disease. In contrast to studies in classical epidemiology, genetic association studies are much less subject to confounding by lifestyle factors because genotype is randomly assigned at conception; this concept is fundamental to analyses based on Mendelian randomization (**30-32**).

Complex disorders are caused by multiple genetic and environmental determinants interacting in complex ways. The small effect size of common genetic variants involved in complex disorders constitutes a constraint for many genetic epidemiology methods (**2**).

In earlier years, linkage analysis was the main method used in genetic epidemiology to map the location of genes related to a trait. Linkage analysis is based on the principle

that shorter haplotypes tend to not be broken down by recombination and are hence inherited intact. If a genetic marker is inherited together with a disease more often than it can be expected by chance this may indicate that the variant involved in the disease is located near the genetic marker. Linkage is a family-based approach where results are derived from relationships *within* a family and not between families; results from several families are then combined to obtain an overall result. Linkage analysis is useful for mapping genes that cause a large increase in risk; but linkage is in general not sufficiently powerful to investigate complex disorders because the genes involved in such disorders have usually small effect sizes. Genetic association studies provide more power than linkage analysis in the search of genes associated with complex traits (**33**); association studies are considered in more details in the next section.

1.3.1. GENETIC ASSOCIATION STUDIES

1.3.1.1. Definition and generalities

Association can formally be defined in a number of ways. It describes a statistical relationship where the distribution of the values taken by one variable depends on the values taken by another variable, so that the variables are not independent from one another (**34**). If two variables A and B are associated then there is a function E(B|A) termed the regression function of B on A which gives the expected value of B given a known value of A. The curve obtained when the regression function is drawn with respect to A is the regression curve (**35**). The nature of prediction errors around the regression curve reflects the distribution of the dependent variable. This may be a normal, Poisson, binomial or some other distribution (see section 2.4 on generalized linear modelling [GLM] in chapter 2). The values of A allow for the prediction of the values of B. The extent to which the knowledge of A enhances the prediction of B

represents the magnitude or strength of the association. In other words the magnitude of an association reflects the level of interdependence between the associated variables.

Association studies explore the relationship between an outcome variable, termed the dependent variable, and one or more other variables, termed independent variables or covariates, to identify an association and determine its magnitude. The independent variables may be of any type, while the dependent variable is typically one of the exponential family distributions (**36**). The analyses in this thesis focus mainly on binary and normally distributed dependent and on binary or quantitative independent variables.

In traditional epidemiology, association studies are used to investigate relationship between exposures, the independent variables, and a disease or disease-related trait, the dependent variable; the exposures are typically life-style or environmental factors. Genetic association studies resemble association studies in classical epidemiology. As well as life-style/environmental factors they also explore relationships between genomic variants - equivalent to exposures in classical epidemiology - and phenotypic variation with the aim to detect association between genetic variants and the phenotypic trait (**18**). The availability of the large HapMap database of polymorphisms combined with the continuous fall of the cost of genotyping and the development of genotyping platforms with dense network of markers (hundreds of thousands to one million markers) has enabled association studies using whole-genome scans termed genome-wide association studies (GWAS).

1.3.1.2. Sampling and interpretation of findings

My work focuses on individual-based genetic association studies; in such studies the data are from individuals expected to be unrelated rather than from families. This type of study enables us to achieve the large sample sizes that are more appropriate for the

investigation of the genetic causes of complex traits (**37**). Individual-based studies may have to deal with the issue of population heterogeneity or population admixture: different ethnicity or groups of different geographical locations may have different prevalence for the same disease and different frequencies for the alleles linked to the disease; such population stratification can cause spurious association, between genetic variants and disease, which in fact just reflects the difference of ancestry (**38**).

As the cost of genotyping becomes more affordable, more variants can now be included in studies. This increase in the number of variants to test for association increases the burden of multiple testing which is an issue especially when a very large number of variants is tested as in genome-wide association studies (**6**). It is then necessary to set a significance threshold that takes into account the number of variants. If the number of equivalent genetic variants that have some influence in human traits is about 10^6 then a significance threshold of 5×10^{-8} (p.value of 0.05 corrected for 10^6 tests) would be an appropriate benchmark for single association studies when there is no pre-existing biological evidence for the influence of individual SNPs (**39**). A less stringent significance threshold can be set for association studies involving genes with some prior evidence of biological effect. On the other hand if a variant reaches genome-wide significance (p.value $< 5\times10^{-8}$) then it may be worth following it up even if there is no biological evidence linking it to the trait of interest as, if artefactual association has been ruled out, it might just be that its involvement in the pathway is unknown as yet. When interpreting findings eventual bias resulting for example from genotyping errors should be taken into account as it reduces the power of the study. The impact of bias on power is covered in more details in the next section.

1.3.2. POWER IN GENETIC ASSOCIATION STUDIES

The design stage is probably one of the most important steps of a study because poor design cannot be retrieved whilst a well-designed study provides opportunities for different analytic approaches. One of the most crucial decisions to make at the design stage is to determine the sample size required to achieve adequate statistical power. If the sample size is underestimated, the study might have a low precision i.e. not enough power and might hence fail to provide a reliable answer. On the other hand, if the sample size is too large, resources may be wasted.

1.3.2.1. Definition

This section addresses some of the principles underpinning the concept of statistical power.

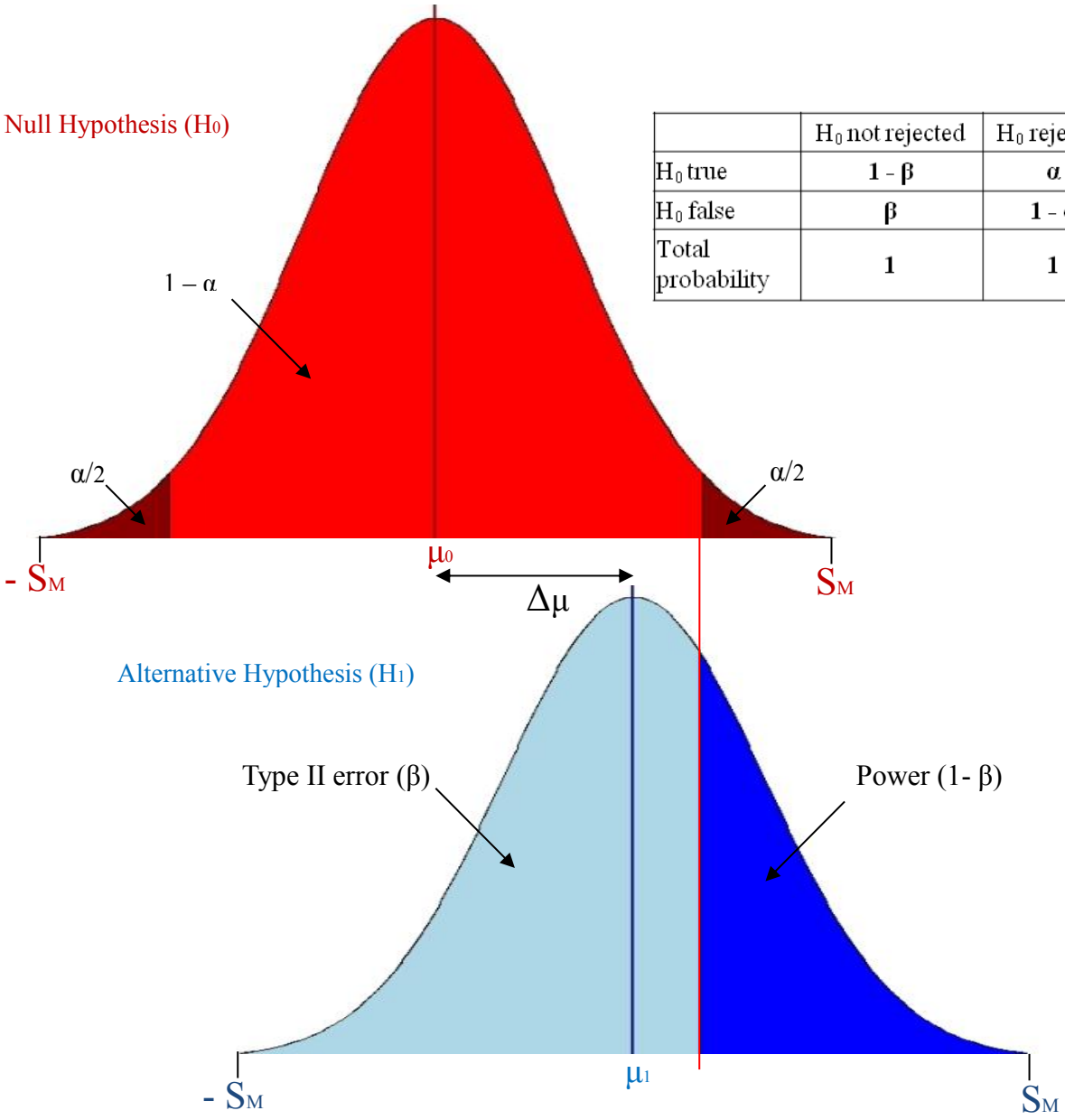
Confidence interval

Due to variation occurring by pure chance when samples are drawn from a population, the observed mean value is just an estimate of the true mean value. Different samples drawn from the same population would yield different estimates. The true mean is not therefore known but the range of values within which it is likely to lie, termed the confidence interval (CI), can be established by using the observed data to compute a lower and upper bound. Under the framework of classical frequentist inference, a confidence interval must necessarily be defined as the long-run probability of an observable event. For example, *"if new data were repeatedly to be sampled, the same model applied and a series of 95% confidence intervals calculated, 19 out of 20 such intervals would include the true value of the parameter being estimated"*(**40**). But this definition is so opaque that many contemporary statisticians instead adopt a pseudo-Bayesian interpretation. For example: *"95% confidence intervals are commonly*

*understood to represent a range of values within which one may be 95% certain that the true value of whatever one is estimating really lies. [This] is valid as a statement of Bayesian posterior probability, provided that the prior distribution that represents pre-existing beliefs is uniform, which means flat, on the scale of the main outcome variable" **(41)***. It is this latter, pseudo-Bayesian interpretation that is used in this thesis.

Hypothesis testing

In drawing formal inferences based on a conventional hypothesis test, a null hypothesis (H_0) typically refers to the assumption that there is no difference between two (or more) groups being compared *e.g.* they have the same true mean (**42**), or more generally the same value for a specified test statistic. The p.value is then defined as the probability that a particular study/analysis as specified will generate a result for the test statistic as extreme, or more extreme, than the result actually observed *given that the null hypothesis is true* (**43**). A type I error occurs when the null hypothesis is rejected whilst it is true; the probability of the occurrence of a type I error is denoted by α. A type II error occurs when the null hypothesis is not rejected whilst it is false; the probability of the occurrence of a type II error is denoted by β. Statistical power is the ability of a test to reject the null hypothesis when it is false *i.e.* the probability that a test does not commit a type II error (β); it follows that power = 1 – β (see Figure 8). If for example a study is investigating an association between a certain factor and a disease or continuous trait, the power is the ability of the study to detect the association if that association truly exists.



	H_0 not rejected	H_0 rejected
H_0 true	**1 - β**	**α**
H_0 false	**β**	**1 - α**
Total probability	**1**	**1**

Figure 8: Illustration of type I error, type II error and power in a two-tailed test. The red and blue curves are respectively the sampling distributions under the null hypothesis and under the alternative hypothesis. α is the probability of the occurrence of a type I error and β is the probability of the occurrence of a type II error. μ_0 and μ_1 are respectively the mean under the null hypothesis and the mean under the alternative hypothesis and $\Delta\mu$ is the mean difference. S_M is the standard error of the mean.

1.3.2.2. Some factors affecting power

There are a number of factors that influence power in any association study and others that are specific to genetic association studies. Power depends mainly on effect size, type 1 error, and sample size (**44-47**). When a key variable is categorical (*e.g.* binary)

the power is influenced not only by the total sample size but also by the size of individual categories – the number of observations in the rarest category typically dominates the statistical power. This is true both for outcomes and exposure variables and it means, for example, that in a genetic association study, power depends critically on the frequency of the allele conferring disease susceptibility at each locus. This may be (though is not universally) the rarer allele; the population frequency of the rarer allele (the proportion of chromosomes where the locus exhibits the rarer allele) is the minor allele frequency (MAF). The power of a genetic association is also dependent on the strength of linkage disequilibrium (LD) between a causal variant and a genotyped variant (**6**) and on the particular genetic model that applies (**7, 8, 48**). These concepts have important parallels in the non-genetic setting (*i.e.* the quality of a proxy measure for any quantity of interest and the particular bio-clinical model that may apply in any setting) but they must always be considered in the genetic setting. The next sub-sections explore the relationships between effect size, type I error, sample size and power. The effect of imperfect linkage disequilibrium between causal and observed genetic variant on power is investigated in section 2.8.

Effect size

Effect size is the strength of the relationship between two variables. It is for example the magnitude of the difference between two groups being compared (exposed vs. unexposed, treated vs. not treated) expressed as $\Delta\mu$, mean difference, in Figure 8. In epidemiological studies the effect size is often measured as a relative risk or odds ratio (see section 2.3). Here it is relevant to note that when an analysis is undertaken using logistic regression, the logarithm of the odds ratio may be regarded as a mean difference on the scale of log-odds. Effect size can also be understood as a measure of the effect of an independent variable on a dependent variable. If a study is investigating an

association between a genetic variant and a trait the effect size is how much the genetic factor influences or decreases the risk (or odds) of the trait occurring (for a binary trait) or by how much it increases or decreases the level of the trait (for a quantitative trait).

The power required to detect an association decreases with decreasing effect size of the causal variable. If the effect size (here the mean difference $\Delta\mu$) decreases, there is a shift to the left of the blue region in Figure 9, and the area that represents the power decreases. This decrease in power with decreasing effect size means simply - and intuitively - that it is more difficult to detect a difference between two groups if that difference is small. If an effect size is negligible, a study would have great difficulty detecting the underlying association even if that study is very large.

If the effect size is unknown, power calculations are undertaken based on estimating the power of the study to detect an effect size that is scientifically relevant. If practical considerations mean that a study cannot be as large as is ideally required to detect the *minimum* effect size that is scientifically (or bio-clinically) relevant one may end up concluding, for example, that the proposed study will have adequate power (*e.g.* 83%) to detect the true effect if it corresponds to an odds ratio as large as 1.75, but inadequate power (*e.g.* 29%) if the corresponding odds ratio is as low as 1.30. Crucially, once a study has been undertaken, if no association has been detected, it does not mean that there is *no* causal effect. It may only be inferred that, after adjusting for potential bias and confounding, any real effect of the putative causal factor is likely to be smaller than the minimum effect size that the study was adequately powered to detect.

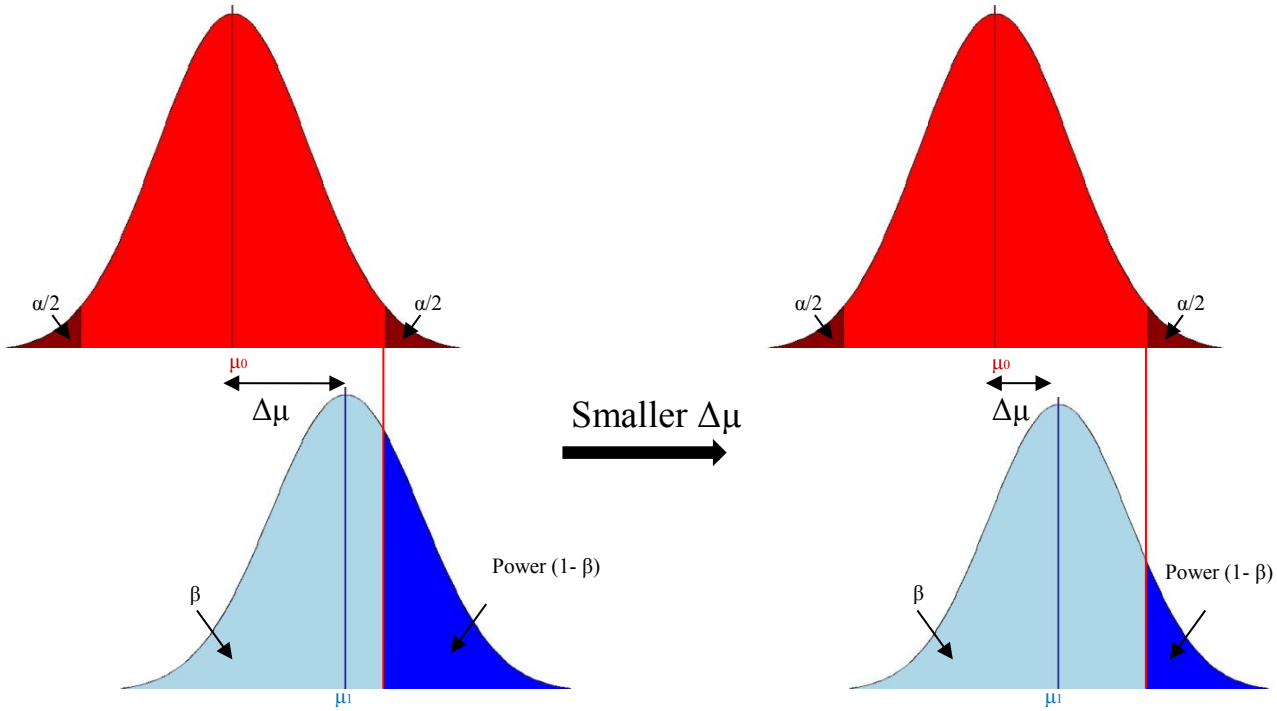


Figure 9: Illustration of how a decrease in effect size causes a decrease of power. The area that represents the power shrinks when Δμ becomes smaller. α is the probability of the occurrence of a type I error and β is the probability of the occurrence of a type II error. μ_0 and μ_1 are respectively the mean under the null hypothesis and the mean under the alternative hypothesis and Δμ is the mean difference. S_M is the standard error of the mean.

Sample size

As already mentioned, if many random samples are drawn from a population, the mean of each sample is an estimate of the true mean. Furthermore, the mean of those sample means is also a (better) estimate of the true mean. The magnitude of observed (stochastic) variation of the sample means from sample to sample may be quantified by the standard error of the mean S_M (often abbreviated to SEM or SE) which provides an estimate of the true variation of sample means. Crucially, as a sample size becomes larger, its mean becomes a more precise estimate of the true mean because the standard deviation of the mean becomes smaller as reflected in Figure 10 and in the equation below:

$$S_M = \frac{\sigma_S}{N}$$

Where σ_S is the estimated standard deviation of individual measurements in the population, and N is the number of observations in the sample.

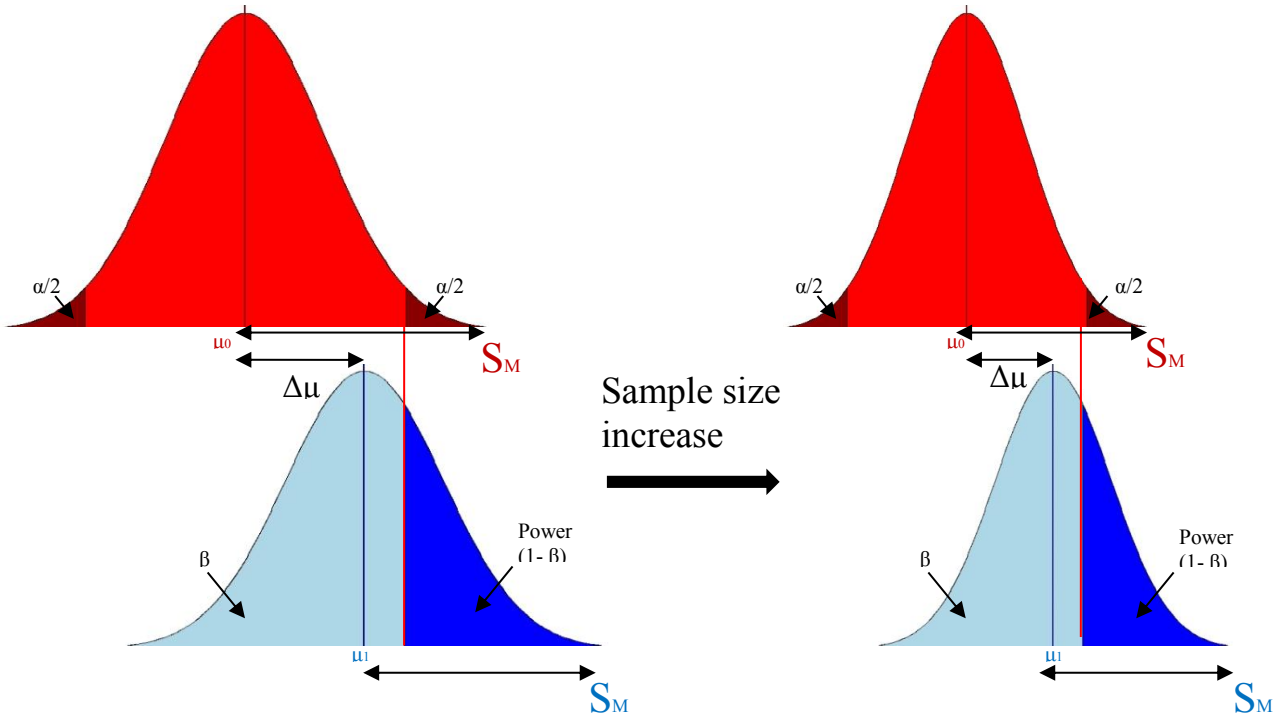


Figure 10: Illustration of how power increases following sample size increase. The standard deviation of the mean, S_M, becomes smaller whilst the effect size, $\Delta\mu$, remains unchanged. α is the probability of the occurrence of a type I error and β is the probability of the occurrence of a type II error. μ_0 and μ_1 are respectively the mean under the null hypothesis and the mean under the alternative hypothesis and $\Delta\mu$ is the mean difference. S_M is the standard error of the mean.

Type 1 error

From the graph in Figure 11, it can be seen that if the type I error (α)decreases the type II error (β) increases and since power is equal to (1 – β) it follows that if α becomes more stringent the statistical power decreases and the sample size required increases. As a corollary, if α is increased (more false positives allowed) β decreases and it follows that the power increases. However increasing power by just allowing more false

positives (using a relaxed p.value) is not a good solution because that would mean having to accept many spurious associations involving numerous genetic variants that are not real and this impedes the development of bioscience.

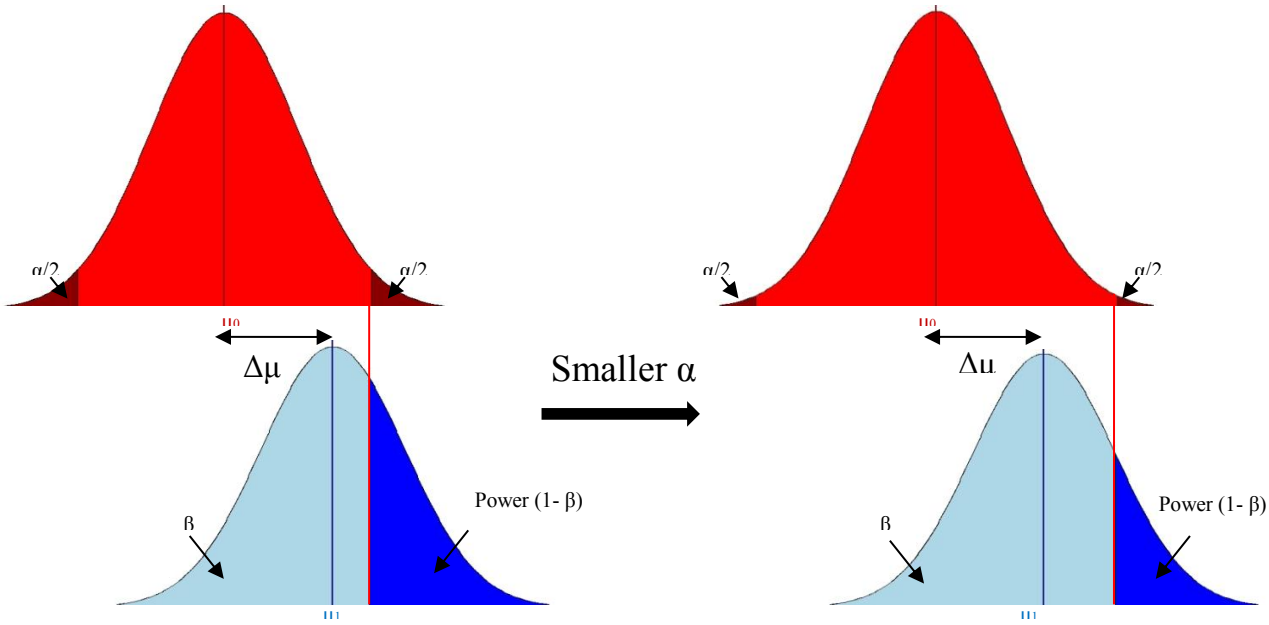


Figure 11: Illustration of the influence of type I error on power.
Allowing for a smaller value of alpha corresponds to choosing a more stringent significance threshold which decreases the power. α is the probability of the occurrence of a type I error and β is the probability of the occurrence of a type II error. μ_0 and μ_1 are respectively the mean under the null hypothesis and the mean under the alternative hypothesis and $\Delta\mu$ is the mean difference. S_M is the standard error of the mean.

1.4. OUTLINE OF THE METHOD DEVELOPMENT AND THE ANALYSES IN THE THESIS

This thesis contains six chapters; the first and the last chapters are respectively a general introduction and conclusions/recommendations. The second chapter describes the development of an algorithm for power analysis and sample size estimation. The aim of this development is to make available a tool that allows for many of the bio-clinical and study design factors that in reality affect statistical power in genetic association studies,

but that, in reality, are typically ignored by traditional power calculation tools, to be taken into account. Such a tool would enable us to achieve a more realistic estimation of the sample size required for adequate power in any given study. The algorithm developed could be of use to investigators to analyse power and estimate sample size in nested or standalone association studies. It could allow funding bodies to verify the power claimed by applicants – the funding of studies that are seriously underpowered is not viewed as an appropriate use of resources. As outlined in chapters 3, 4 and 5 such a tool would also be used in answering a range of important scientific questions.

The third chapter is an analysis to estimate the minimal sample size required by a pre-existing cohort study to enable the investigation of a set of quantitative traits. One of the goals of the cohort is to establish a well powered platform which permits biomedical researches of quantitative traits. The cohort is still recruiting participants and the eventual sample size is yet to be definitively determined. This has important strategic and financial implications and it is therefore important to know the size that the cohort should achieve if it is to enable appropriately powered studies of the targeted quantitative traits.

The fourth chapter explores the impact of some of the protocols used by UK Biobank to process bio-samples once they had been obtained. Assessment errors, introduced through poor assessment or physical measurement or because of inconsistent/or inappropriate standard operating procedures for collecting, processing, storing and/or analyzing bio-samples, can have a severely negative impact on statistical power (**49**). Thus, it is important to quantify such errors and to understand how they might impact on statistical power, particularly in association studies nested in biobanks and large bio-repositories.

The fifth chapter evaluates how accurately and precisely copy number variants can be called from recent and older SNP genotyping platforms and the impact of inaccuracy in CNV calling on the power of genetic association studies investigating the role of CNVs on the onset of diseases. CNV genotyping can potentially be carried out using SNP platforms and SNP intensity data can be used in CNV association studies. It is therefore important to estimate the nature of errors in data from these platforms to inform investigators about the accuracy and precision of the CNVs they might want to include in their studies and the appropriate design of these studies.

CHAPTER 2

2. ESPRESSO-FORTE, ALGORITHM FOR MORE REALISTIC POWER ANALYSIS AND SAMPLE SIZE ESTIMATION

The initial version of this algorithm was developed by Professor Paul Burton. The development of the current version, named ESPRESSO-forte, was a key component of my thesis work. It is the new functionalities that I built in, including the possibility to analyse quantitative variables, which made possible the analyses carried out in sections 2.8 and 2.9 and in chapters 3 and 4. The purpose of section 2.2, which follows the introduction to this chapter, is to explain my contribution by highlighting the differences between the original version of ESPRESSO and the version I developed as part of this thesis.

2.1. INTRODUCTION

A critical question to answer at the design stage of large scale studies and biobanks is what sample size is required to achieve adequate statistical power. For a large study aimed at exploring weak effects, the answer can have major implications for funding and resources. But this answer depends on a number of key issues such as the study design (longitudinal, cross-sectional, individual or family based), the class of analysis, the time required to observed sufficient number of events and the quality of the measurements, which in some cases depend on the number of repeated measurements that can be afforded. Assessment error in outcome and in explanatory variables can

substantially reduce the power of association studies (**49**).Conventional approaches to estimate the sample size required to achieve adequate power often fail to take into account some complex elements (**7**) such as the sensitivity and specificity of the assessment of binary outcomes and explanatory variables or the reliability of the assessment of quantitative outcomes and explanatory variables. A failure to include these elements in power analyses at the design stage of a study may result in a serious over-estimation of its true statistical power and a research platform that is critically underpowered when it comes to analysis.

ESPRESSO (Estimating Sample-size and Power in R by Exploring Simulated Study Outcomes) was developed to take into account the key issues and complex elements mentioned in the preceding paragraph and hence allow for more realistic power analysis and sample size calculation in genetic association studies. It is a simulation based approach, written in the R programming language (**50**) that supports power and sample size calculations for stand-alone studies and analyses nested in cohort studies. The simulation based approach, as opposed to the closed form solutions offered by conventional power and sample size calculators, allows for the high flexibility required to include the complex elements already mentioned.

The ESPRESSO algorithm can be used by researchers involved in designing and setting up studies to investigate the genetic and environmental basis of complex traits. In particular, it enables those designing large cohorts and biobanks to better estimate the sample size required to achieve adequate power. ESPRESSO also allows funding bodies to verify the statistical power calculations put forward by researchers in their grant applications, thereby helping to ensure that resources are not wasted in underpowered studies. Although well conducted underpowered studies may be useful for meta-analyses in the absence of serious reporting and publication biases, the primary aim of

an association study is not, generally, to generate results to be included in meta-analysis studies. The algorithm can also be used to explore specific scientific questions relevant to the design and set up of large-scale association studies and biobanks. For example, it may be used to help design protocols and standard operating procedures (SOPs) that enable appropriate account to be taken of errors introduced during the collection, transport and/or storage of samples. SOPs that entirely eradicate or minimise such errors are typically very expensive, in time, resources or invasiveness for participants. A sensible balance must therefore be struck between optimisation of SOPs and avoidance of approaches that are so time or resource intensive that the resultant opportunity cost - not being able to do other things instead – is simply too high. In this regard, chapter 4, details the exploration of the power implications of the SOP adopted by UK Biobank protocol that determined the distribution of delay time between the collection of bio-samples (blood and urine) and final processing and cryo-storage.

The aim of this project is to re-construct and re-program the original - published (**7**)-version of ESPRESSO to: (1) make it more user friendly and intuitive to use; (2) allow the algorithm to deal with quantitative variables (both outcomes and explanatories); and (3) extend the range of biomedical scenarios that can be investigated. The updated version of ESPRESSO has been called ESPRESSO-forte.

2.2. ORIGINAL AND NEW VERSION OF ESPRESSO

2.2.1. RECONSTRUCTING ESPRESSO

The original version of the algorithm was written in a procedural way where each line of code consisted in the execution of a command that produced an output required by one or more of the next commands. This programming paradigm was fine for a concise

project as the early version of ESPRESSO, but as the length and the complexity of the original script increased, it became less easy to debug (to find and fix errors) and maintain the programme. So it was necessary to re-write the original script in a modular way to simplify it and work around the difficulties mentioned above; I did this by writing functions which are easier to use than several interdependent lines of instructions. These modifications of the original script were sufficient for the purpose of the web based version under the Public Population Project in Genomics (P^3G) website: http://www.p3gobservatory.org/powercalculator.htm. However, to make the programme accessible to a wider audience and allow for researchers proficient in the R programming language to use the algorithm as written or modify some functions in a way that suits their analyses, I built the algorithm as an R package (an R library) which is now available for download from the Comprehensive R Archive Network (CRAN) repository of contributed packages: http://cran.r-project.org/web/packages/ESPRESSO/index.html.

2.2.2. EXTENDING ESPRESSO

The original version of ESPRESSO allowed for the fitting of two explanatory variables (a genetic exposure and a binary environmental exposure) and one binary outcome. In the new version two genetic and two environmental exposures can be analysed simultaneously (under a main effect model) and the environmental exposure can also be quantitative. The main purpose of extending the number of explanatory variables that can be fitted was to allow for the investigation of quantitative traits which are becoming increasingly important as biomarkers are being more and more measured to assess disease (binary outcome) status.

Table 2 shows the classes of analyses that were possible in the original version of ESPRESSO (cells shaded light green and light blue) and those that are now possible with the new version (cells shaded dark green and dark blue). In addition, whereas the original version of ESPRESSO focussed primarily on binary disease outcomes, ESPRESSO-forte allows for outcomes to be either binary traits or quantitative measures.

	Additive genetic variant (GA)	**Binary genetic variant(GB)**	**Quantitative environmental exposure (EQ)**	**Binary environmental exposure (EB)**
Additive genetic variant (GA)	**GA×GA**			
Binary genetic variant (GB)	**GB×GA**	**GB×GB**		
Quantitative environmental exposure (EQ)	**EQ×GA**	**EQ×GB**	**EQ×EQ**	
Binary environmental exposure (EB)	**EB×GA**	**EB×GB**	**EB×EQ**	**EB×EB**

Table 2: Overview of the original ESPRESSO algorithm and the newly developed one. This table shows the genetic and environmental main effects scenarios that could be investigated under the original version of ESPRESSO (light green) and under the new version (dark green); and the interaction scenarios that can be analysed under the original version (light blue) and under the new version (dark blue).

In working with binary outcomes, although the ESPRESSO algorithm was actually developed for the purpose of supporting design of cohort studies, the power calculations undertaken were based on the power of the nested case-control (or more often nested case-cohort) such cohorts can support. This is because, in most realistic settings, it is the nested case-control studies that a cohort can support that represent the power-limiting feature of the cohort design. If a cohort study is large enough to support typical nested case-control studies, it will almost certainly have plenty of power to provide for other

sub-study designs such as exposure-based studies (including genotype-based studies). Given the fundamental role of nested case-control analyses in the ESPRESSO algorithm, the next two sections cover some fundamental aspects of case-control design and regression analysis that are relevant to the context of ESPRESSO-forte. These will hopefully help readers to understand the architecture and logic of the algorithm.

When ESPRESSO-forte is used to explore power for analyses involving a quantitative outcome (*e.g.* measured systolic blood pressure), the required analysis is not formally a case-control analysis but rather an association analysis (usually regression based) that is carried out on a sample of participants (possibly all) from the cohort study.

2.3. CASE-CONTROL STUDIES

A case-control study compares a group of affected persons, cases, to a group of unaffected persons, controls, to investigate possible association between one or more exposures (explanatory variables) and a binary outcome. The exposures can include particular personal attributes (e.g. genetic variants in ESPRESSO-forte) or environmental factors or interactions between them. If the frequency of the exposure is higher in the case group then an association between the exposure and the outcome may be inferred. Association does not necessarily imply causation; the explanatory variable may or may not cause or causally influence the outcome variable. For example, an association might be observed because of confounding consequent upon a third variable. A confounder is a variable associated with both the outcome and the explanatory variable (Figure 13) but not on a causal pathway linking the two together. The effect of the explanatory variable is mixed with the effect of the confounder therefore the estimated effect of the explanatory variable is distorted. A confounder may be adjusted for by techniques such as matching, stratification, multivariate adjustment

(**51**), if it is known; if it is not known randomization can be used as a solution such as in randomized controlled trials where subjects are randomly allocated to treatment groups.

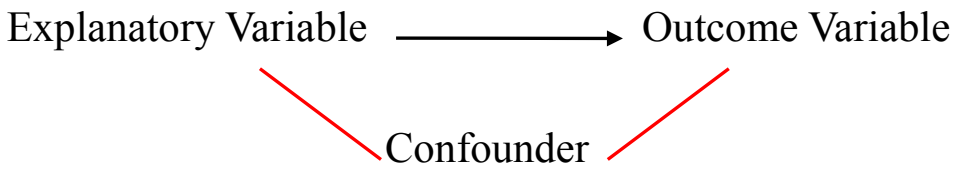


Figure 12: Graphical illustration of a confounder.
A confounder is related to both independent and dependent variable but is not in the causal pathway.

To ensure that the estimate of a true association is not masked or distorted by some bias, it is essential, as for any study design, to include only cases with strong evidence of disease to prevent misclassification (assigning an individual to a category or class he/she does not belong to). Controls should be representative of the population that the cases are selected from; they should be chosen independent of exposure or non-exposure to reflect the proportion of exposed and unexposed individuals in the source population. The best controls are generally similar to cases except for the fact that they have not developed the disease (**52**). It is difficult to sample such a control group, but a group close to that ideal sample can be obtained by, for example, incidence density sampling which consists of drawing controls from a group of individuals, in a cohort, who are at risk at the moment where a case occurs (**53**). Another option is to use an unmatched control group and adjust for possible confounders (**54**). The sampling of controls can also be carried out by choosing for each case a control(s) that is similar for some variables such as age, sex, ethnicity (a process called pair-matching) or by sampling a control group that is overall similar to the case group for these variables (group matching). Matching can make an association less obvious or even mask the

association if cases and controls are matched for a variable that is correlated with the possible cause of the disease but is not itself a cause of the disease (overmatching) (**55**).

The results of an unmatched case-control study with one exposure of interest are often reported in a table similar to Table 3 and the odds ratio is calculated from the figures in the table. The odds ratio (OR) is used to express the ratio between exposed and non-exposed. The odds of an event is the probability of an event (for example a disease) occurring over the probability of it not occurring so the OR is the ratio of the odds in the exposed group over the odds in the unexposed group. The OR described here uses prospective likelihood which is based on the probability of disease given the exposure status (**56**). In the case of a rare disease, when the prevalence of the disease is less than 10% (**57**), the number of new occurring cases (incident cases) is very small compared to the non-diseased in both the exposed and unexposed group and the odds ratio is then a good approximation of the relative risk (RR) a measure often used for rare diseases and which represents the risk of developing a disease.

	Cases	**Controls**
Exposed	exposed cases (a)	exposed controls (b)
Unexposed	unexposed cases (c)	unexposed controls (d)
Total	a + c	b + d

Table 3: Summary of case-control study results.

$$OR = \frac{a/b}{c/d} \quad \textit{Equation (1)} \qquad RR = \frac{a/(a+b)}{c/(c+d)} \quad \textit{Equation (2)}$$

In rare disease (a+b) ≈ b and (c+d) ≈ d, hence $RR \approx \frac{a/b}{c/d} \approx OR$.

The advantage of case-control design is that there is no need to wait for incident cases; so the length of the study is relatively short and that contributes to cost effectiveness. This study design is suitable for rare conditions and more than one exposure can be

investigated in one study. The disadvantage of this design is that it can be difficult to find an appropriate control group and exposure is assessed retrospectively which is not the most reliable way of collecting data. Furthermore it can be difficult to determine if the exposure preceded the disease. The incidence of the disease in the population at risk is unknown unless the sampling fraction (ratio of sample size over population size) is known both for cases and for controls. Critically the sampling fraction may be different for cases and controls – one often aims to sample a high proportion of cases (to enhance power) while using a low proportion of potential controls (to avoid recruiting too many controls which is statistically inefficient). ESPRESSO-forte implements an unmatched case-control design with, by default, four controls for one case. If for example the number of cases is limited, power can be gained by increasing the number of controls but generally there is little gain of power over a ratio of four cases for one control (**52**).

2.4. REGRESSION ANALYSIS AND GENERALIZED LINEAR MODELS

In regression analysis mathematical models are used to describe the relationship between one or more input variables (also termed independent or explanatory variables or covariates) and an output (also known as dependent or outcome) variable. The aim of regression analysis is to determine the values of the parameters of the model that – based on an appropriate optimization criterion such as likelihood or least squares – best describe the relationship between the explanatory variables and the outcome variable. Regression analysis can enable us to understand and estimate relationships and predict future values of the outcome variable given observed values of the explanatories.

If an outcome variable y is quantitative and a given change in an explanatory variable x is associated with a fixed change in y, x and y are said to have a linear relationship and

the model is called linear regression. Such a model can be expressed mathematically as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad \textit{Equation (3)}$$

The above linear regression model has one explanatory variable or covariate (x) and two parameters (β_0 and β_1).The intercept, β_0, is the value of y when x is zero; it is therefore the y-coordinate of the point where the regression line cuts the y-axis. $\beta_1$reflects the change in magnitude of y for a unit change in x. ε is the error term, it represents random variability and may also incorporate errors in measurement, changes in the conditions under which measurements were carried out or the impact of other determinants of y which have not been measured. A conventional linear regression model assumes: a quantitative and normally distributed outcome and a normally distributed error term which has a constant variance across the data (homoscedasticity) and is independent of the covariate.

Other types of regression model can be considered if the relationship between the covariate and the outcome is not linear. For example, if the outcome variable is binary, a logistic regression model can be used to carry out regression analysis. Logistic regression models the changes in the natural logarithm of the odds of the outcome variable given some covariates; the model with one covariate (x) can be expressed mathematically as follows:

$$\ln\left(\frac{\mathrm{Y}}{1-\mathrm{Y}}\right) = \beta_0 + \beta_1 x \quad \textit{Equation (4)}$$

Where Y is the *probability* of success (the probability of the occurrence of an event y such as a disease); $\ln\left(\frac{\mathrm{Y}}{1-\mathrm{Y}}\right)$ or logit(y) is the logistic transform or logit transform; β_0 is the value of logit(y) when x is equal to zero, and so estimates the log of the odds

corresponding to the probability of response when all covariates are zero (here, x=0); $\beta_1$is the estimated effect of a one unit increase in x on the log odds of y. This implies that the relationship between x and logit(y) is modelled as being linear. Logistic regression models are very useful for association study designs with quantitative covariates where the results are expressed using odds ratios because the *exponential* transform conveniently relates the covariates to odds ratios (ORs). Thus, if we consider the simple model: $logit(y) = \beta_0 + \beta_1 x$, where x is a quantitative variable, β_1 estimates the increase in the log(odds) of a positive response given a one unit change in the explanatory variable x. This implies that the exponential of the coefficient (*i.e.* e^{β_1}), estimates the *multiplicative increase in the odds* associated with a one unit change in x. At the same time, $\beta_0$estimates the log-odds of a positive response when x=0 and, applying the *expit*, or *inverse logit* transform, this therefore means that, if x=0, the estimated probability of a positive response is $\frac{e^{\beta_0}}{1 + e^{\beta_1}}$.

Generalized linear models (GLMs) represent a broad class of regression models that allow for two important extensions of the conventional multiple linear regression model (**36**): (1) errors in estimation of the fitted values of the outcome variable in a GLM do not necessarily follow a normal (Gaussian) distribution; and (2) the functional relationship between the outcome variable and the covariates can be a non-linear function. The two regression models (linear and logistic) described in the two paragraphs above are therefore both GLMs.

In a GLM the functional relationship between the outcome variable and the linear predictor (a linear combination of covariates and coefficients) is known as the *link function*. By allowing functions other than the *identity* link that is found in a conventional linear model (where a linear combination of covariates and coefficients

[$\beta_0 + \beta_1 x_1 + \beta_2 x_2 ....$] directly predicts the expected value of the outcome variable) GLMs are greatly more flexible than standard linear regression models.

The relationship between outcome and linear predictor of a GLM, here with two covariates, is typically written as follows:

$$g(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad \textit{Equation (5)}$$

Where $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ is the linear predictor (often denoted eta [η]) and g(.) is the link function which links linear predictor (LP) and Y. The *inverse link function* is often denoted h(.) = g^{-1}(.). Each observed value of the outcome variable, y, is assumed to be distributed stochastically around its modelled expectation Y (often denoted μ) following one of the many distributions from *the exponential family* (**36**). The link function, g(.), is monotonic and therefore preserves the ordering of the relationship so that the expectation of y increases, decreases or remains constant if the LP respectively increases, decreases or remains constant. The parameters of the model are usually optimised via maximisation of the likelihood (**56**). The likelihood associated with a particular parameter value β_1 given y is computed by working out how probable y would be if the parameter was set to the specific value β_1 (**56**). In maximum likelihood estimation, the particular parameter values to explore may be chosen in several different ways. (1) In closed form (when a solution can be found in one step as in a linear model); (2) by systematic searching (*e.g.* based on the Iterative Reweighted Least Squares Algorithm [IRLS] which is closely related to Newton Raphson (**58**); or rarely, (3) randomly or by trial and error when closed form solutions and/or systematic search regimes are unavailable. However, the evaluation points are chosen, the likelihood of the data given the parameter values at each point is then calculated and likelihood

surface is derived. The peak value of likelihood across this surface jointly defines the maximum likelihood estimate (MLE) for every parameter being considered.

The precise form of the maximum likelihood calculations are determined by the particular error distribution that is assumed to apply. Thus the full formulation of a GLM is usually given as follows:

$$g(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$y \sim exponential.family(Y, .)$$

It is the second element that specifies that the observed values, y, are distributed with expectation Y (as predicted by the linear predictor and link function) and the error around those predictions follows an exponential family distribution which may include parameters relating to the variance function (.).

Table 4 gives a summary of the link and variance functions for the two GLMs used in ESPRESSO-forte to analyse simulated quantitative (normally distributed) and binary outcome data. Note that the variance function for the linear model is unrelated to Y. This implies that the errors are homoscedatic. But, the variance function does include an additional argument σ^2 which relates to the variance of the residuals. On the other hand, the variance function for logistic regression is *fully defined* by Y, and so the error distribution is heteroscedatic (with a specified functional form) but there is no need to estimate any additional parameters equivalent to σ^2 in the setting of the linear model.

Model	**Distribution**	**g(.)**	**Variance Function**
Linear	Normal	y	σ^2
Logistic	Binomial	logit(y)	Y(1-Y)

Table 4: GLM's link and variance functions analysed by ESPRESSO-forte.

For the sake of simplicity the mathematical expressions of the models presented in this section have included only one or at most two covariates; there could be however more than one covariate of any type. Under a main effect model ESPRESSO-forte allows for the fitting of GLM models with four covariates (two genetic and two environmental determinants) and three covariates (two determinants and the interaction term) under an interaction model. The next section explores the concepts of main effects and interactions.

2.5. MAIN AND INTERACTION EFFECT

An outcome variable may depend on one or more covariates and each covariate may have an effect not related to the effects of the other covariates. This is termed the *main effect*; it is the effect of each covariate fitted in the absence of interaction between covariates. Although the definition of main effect can be used for models with only one covariate it is mainly used when there is more than one covariate acting on an outcome variable. In the presence of more than one covariate if the combined observed effect is not equal to the sum of the single effects this could indicate an *interaction* between covariates.

If for example an outcome y is determined by two binary exposures X_A and X_B, there is no interaction if the effect of X_A is independent from the effect of X_B. Though, this assumes an ideal situation where the outcome is only affected by X_A and X_B, and there is no bias or all biases have been controlled. The example can be illustrated as in Table 5 which reports the levels of the outcome for the two values of X_A and the two values of X_B. The difference between the levels of the outcome under the two levels of the exposure X_A (in blue in Table 5) is the same regardless of what the status of the second exposure (X_B) is. It follows that the difference between the levels of the outcome under

the two levels of the exposure X_B (in red in Table 5) is also the same regardless of what the status of the other exposure (X_A) is.

	$X_{A.0}$	$X_{A.1}$	$X_{A.1} - X_{A.0}$
$X_{B.0}$	300	400	100
$X_{B.1}$	500	600	100
$X_{B.1} - X_{B.0}$	200	200	

Table 5: Example of two independent binary explanatory variables
The table reports the levels of the two independent binary explanatory variables X_A and X_B. The effect of X_A on the outcome is not influenced by the effect of X_B and the effect of X_B is also independent from that of X_A.

Returning to the above example to illustrate an interaction, if the difference between the levels of the outcome under the two levels of the exposure X_A is not the same for the two levels of the second exposure X_B, this indicates an interaction between the two determinants of the outcome i.e. the effect of one determinant is not independent from the effect of the other determinant as illustrated in Table 6.

	$X_{A.0}$	$X_{A.1}$	$X_{A.1} - X_{A.0}$
$X_{B.0}$	100	400	300
$X_{B.1}$	500	600	100
$X_{B.1} - X_{B.0}$	400	200	

Table 6: Example of two dependent binary explanatory variables.
The table reports the levels of two dependent binary explanatory variables X_A and X_B. The effect of X_A on the outcome is dependent upon the effect of X_B and vice versa.

So, in the absence of relevant biases, a departure from additivity in this setting indicates interaction. However biologically there could be an interaction even when the sum of the single effects is equal to the observed combined effects of the covariates; this could be the case if for example there are two antagonist types of interaction of the same magnitude that cancel out each other (**59**). Thus the absence of statistical interaction does not necessary imply an absence of biological interaction.

If a real interaction effect is not taken into account in the estimation of a joint effect this will lead to bias; the accuracy of the estimated relationship between the covariates and the outcome will be reduced. Typically, the predicted value Y will be overestimated for some combination of the covariate values and underestimated at others. The accuracy will depend, among other things, on the magnitude of the omitted interaction term relative to the main effect estimates; the larger that relative magnitude, the less accurate inferences based on the main effects alone will be. Because measurement errors in the values of covariates involved in an interaction may be amplified in multiplying each covariate by the other covariate in creating the product term, and by the combination of errors from each covariate, measurement error often has a larger impact on the precision of estimation (and hence statistical power) under interaction compared to main effect models.

Because interactions relate to the impact of one covariate at different levels of another covariate, the amount of information that is available to estimate them is generally smaller than for main effects (**7**). This means that it is generally more difficult to identify and prove the existence of an interaction effect than it is for main effects. Consequently, studies are often underpowered to investigate interaction (**60**) because most of them are designed to investigate main effect models (**61**).

2.6. DETAILS OF THE ESPRESSO-FORTE ALGORITHM

An ESPRESSO-forte power calculation involves repeatedly simulating a dataset with a number of key characteristics and seeing in what proportion of the simulations the effect of interest can be detected by an appropriate test of statistical inference, such as a p.value at a given level of statistical significance. The flowchart in Figure 13 shows the

three main blocks of the algorithm (parameters settings, data simulation and data analysis) and the different steps within each block.

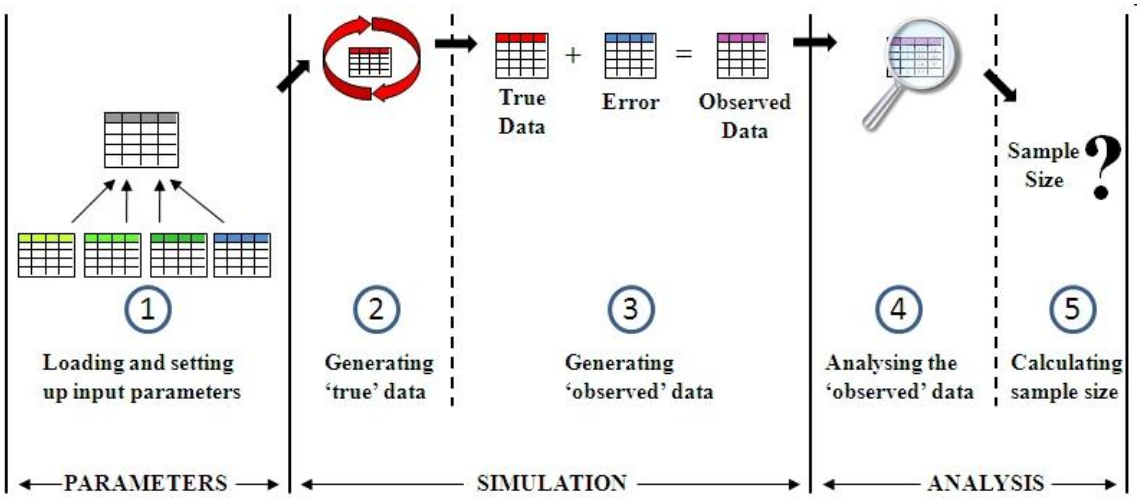


Figure 13: Flowchart of the main steps in ESPRESSO-forte simulation.

The details of each of the three main parts of ESPRESSO-forte are covered in this section. In the first part (step 1 in Figure 13) a series of input parameters required to run ESPRESSO-forte (general, genetic exposure, environmental exposure and other parameters) are loaded and merged into one input table. In the second part (steps 2 and 3 in Figure 13), an error free dataset which contains the true outcome and determinant values for each simulated individual is generated. Crucially, the word 'true' here refers not to the 'true' value of some real individual in the real world, but rather the true values (without error) of each simulated variable in each simulated subject within ESPRESSO-forte. Then a simulated error is generated for the outcome and for each of the covariates and added to the true data to produce the 'observed' data. The level of error depends on the input parameters reflecting the sensitivity and specificity of the assessments of binary variables or on the reliability of the assessment of quantitative variables. In the third and last part of the algorithm (steps 4 and 5 in Figure 13) the observed data generated in the simulation stage is analysed by GLM, the sample size

required to achieve the specified power is calculated and the empirical and modelled power achieved with the input sample size are estimated.

In sections 2.6.1, 2.6.2, and 2.6.3 each of the three main parts of the algorithm, respectively the parameters setting, the simulation of outcome and exposure data and the analysis of the simulated data, are explained in details.

2.6.1. INPUT PARAMETERS

The parameters are subdivided into four input tables to make the maintenance of the web-based version easier and more flexible. The first input table named *general.params* (see page 258 of the R package manual under Appendix 1) contains parameters that relate to all simulations such as the number of simulations and other parameters that relate to the outcome variable and to effects other than the genetic and environmental ones. The second table (table of genetic parameters, see page 256 of the R package manual under Appendix 1) contains parameters that relate to the genetic exposures (two bi-allelic SNPs, G_1 and G_2). The third input table (table of environment parameters, see page 255 of the R package manual under Appendix 1) contains parameters that relate to the environmental exposures (two environmental/life style exposures, $E_1$and E_2). The fourth input table which is not listed here but can be found in page 288 of the R package manual under Appendix 1) is optional and is only used if the user chooses to estimate the unknown sensitivity and specificity of the assessment of the alleles of the genetic variant or the sensitivity and specificity of the assessment of the binary environmental exposure.

Some of the outcome and covariates parameters may not be known to the user; in this case it is reasonable to use known values from pre-existing data. If there is no known

pre-existing data from which to generate reasonable parameter values, the user should consider running a sensitivity analysis to find out how sensitive the final estimated sample size is to the unknown parameters; if the estimated sample is relatively robust to the choice of parameters then a rough approximation is acceptable.

In the next three sub-sections the input parameters and their meaning are listed in Table 7, Table 8, Table 9 and Table 10which are followed by detailed explanations of the parameters and some considerations that would apply depending on what the parameters are set at.

2.6.1.1. General and other parameters

In this section, the input parameters in the table *general.params* are presented in two distinct tables for the sake of clarity but these two tables (Table 7 and Table 8) represent actually one table when running the software.

General parameters

Parameter	Description
scenario ID	An integer code for each simulation.
number of simulations	An integer that indicates the number of runs.
random number seed	An integer between 1 and 9999999.
number of cases	An integer indicating the number of cases for a binary outcome model.
number of controls	An integer indicating the number of controls for a binary outcome model.
number of subjects	An integer indicating the sample size for a quantitative outcome.
p.value	A value defining statistical significance.
power	A value between 0.1 and 1defining the desired level of statistical power
main effect/interaction	A discrete indicator that takes the value 0 for a main effect model, 1 for an interaction between genetic and environmental exposures, 2 for an interaction between the two genetic exposures and 3 for an interaction between the two environmental exposures.
display LD	A binary indicator set to 1 if the two SNPs that represent the genetic exposures are modelled as being in linkage disequilibrium and if a summary of the simulated LD values should be displayed on screen. The parameter is set to 0 if no summary should be displayed.
calculate sensitivity and specificity	A binary indicator variable set to 1 if the sensitivity and the specificity of the binary exposures should be estimated (e.g. if they are unknown) or 0 otherwise.

Table 7: Outline of the general parameters required by the algorithm.

The parameter ***scenario ID*** enables each simulated scenario to be identified in the output; there is no requirement that the IDs be sequential.

The parameter '***random number seed***' - technically, a *pseudo*-random number seed - is an integer between 1 and 9999999 that allows for the user to repeat precisely the same analysis on as many occasions as needed, to for example look at the results in more detail or to reproduce the results if an error is suspected. The parameter is passed on as an argument to an R pseudorandom number generator function which creates a sequence of numbers that behaves as if it were a random sequence even though it is actually deterministic. But, because it is actually deterministic, once a specific seed value has been specified, it is possible to generate precisely the same sequence of numbers by using the same seed.

The parameter '***number of simulations***' sets the number of simulation to run. The more simulations that are run, the better reflection the set of simulations will provide of the particular scenario that has been specified. However, the number of runs has a computational cost; the larger the number of simulations, the longer the runtime will be. It is probably better to run a small number of simulations (*e.g.* 10-100) for a first exploration and run rather more simulations (*e.g.* at least 500) for the final results.

The parameters '***number of cases***' and '***number of controls***' set the starting sample size for a binary outcome model. Here, 'starting' refers to the number of individuals in each simulation – this will in general be different to 'final' sample size, which is the sample size calculated in step 5 that is needed to generate the statistical power required. It is important to note that: (1) the ratio of the number of cases over the number of controls (cases/controls) is important because the sample size that is calculated at the end of the simulation (in step 5) is the required size for a case-control study with cases/controls

times as many controls as cases to produce the desired power to detect the simulated effect. In other words, the estimated power relates specifically to a case-control study with the chosen ratio of cases to controls and conclusions would in general be different if an alternative ratio was to be used; (2) the smaller the absolute number of cases and controls, the faster the program will run, but the more variable the results will be and therefore more runs will be required. If an inadequate number of cases and controls is specified, rare combinations may not occur at all in a dataset and this might generate misleading results or cause the model to fail. In the other hand very large number of cases and controls will cause the programme to run very slowly. The number of cases and controls that is specified must necessarily be higher if the model of interest includes, for example, an interaction that will be present in very few individuals in each simulation.

The parameter '***number of subjects***' sets the starting sample size for a quantitative outcome. Here there is no distinction into cases and controls; there is simply a number of individuals each of whom have a measured quantitative outcome. However, as for the number of case and controls; the smaller the absolute number of subjects, the faster the program will run, but the more variable the results will be and therefore more runs will be required.

The parameter '***p.value***' defines statistical significance (see paragraph on type 1 error in section 1.3.2.2). Instead of setting the parameter totally arbitrarily, it is probably more advisable to take into account the biological plausibility of the association between determinant and trait of interest and other elements that may obscure an existing association. However the choice of a relaxed p.value threshold can be legitimated by the aim of the investigator; if the aim is to not miss the true signal then one can afford to set

a non-stringent threshold just as a screening test whose main goal is to detect all affected individuals, for such test a high rate of false positive can be allowed.

The parameter '***power***' sets the statistical power to achieve; its value is up to the user but it should reflect the power required for detecting a bio-clinically meaningful main effect of an exposure or joint effect of several exposures. In ESPRESSO-forte the power is by default set arbitrarily to 80%.

The parameter '***main effect/interaction***' determines if there is an interaction between the covariates. This parameter indicates a main effect model, an interaction between genetic and environmental determinants, an interaction between the two genetic determinants and an interaction between two environmental determinants if set respectively to 0, 1, 2 and 3. Under a gene-environment interaction model, it is the first genetic determinant (G_1) and the first environmental determinant (E_1) that are used to form the interaction term; the results of the analysis should then be ignored for the other two covariates.

The parameter '***display LD***' is only relevant if the user chooses to model two SNPs (genetic exposures) as being in linkage disequilibrium (LD). In that case if the parameter is set to 1 a summary which consists of the target and simulated frequency of the major haplotype and the target and simulated level of LD between the two SNPs (the LD is measured as both a Pearson correlation coefficient and Lewontin's D) is displayed on screen after each simulation. This means that if the number of runs is 500, a summary will be displayed 500 times, thus this parameter is meant to be used only for a short number of simulations to verify that the target LD and the simulated LD do match.

The parameter '***calculate sensitivity and specificity***' indicates if the sensitivity and specificity of the exposures should be estimated empirically or not. It might be necessary to estimate these values if for example they are unknown to the user and where not available from the literature. For the genetic exposure, the function *sim.geno.sesp*, (see page 279of the manual under Appendix 1) that I wrote as part of the ESPRESSO-forte R package allows for the user to estimate the sensitivity and specificity required to generate the squared correlation between the 'true' alleles and the 'observed' alleles used to construct the genotypes, given the minor allele frequency of the SNP. For the binary environmental exposure, the sensitivity and specificity of the assessment of the exposure can be calculated using the function *sim.env.sesp* that I also wrote as part of the ESPRESSO-forte R package (see page 277 of the manual under Appendix 1). This function generates the appropriate sensitivities and specificities corresponding to the reliability that is required (given the prevalence of the 'at-risk' environmental determinant).

In the functions *sim.geno.sesp* and *sim.env.sesp,* the measurement error is represented as an incomplete correlation between the vector of "true" measurements and the vector of "observed" measurements. The calculated sensitivity and specificity values will overwrite the values specified in the genetic and environment parameters input tables. If the user chooses to estimate the sensitivity and specificity, the arguments to be passed on to the functions *sim.geno.sesp* and *sim.env.sesp* should be specified in the input table *sim.sesp.params* (see page 288 of the package manual under Appendix 1).

Parameters for the outcome and the other risk effects

Parameter	Description
outcome model	A binary indicator variable set to 0 if the outcome is to be modelled as binary and 1 if it is to be modelled as normally distributed
disease prevalence	The frequency of the binary outcome in the general population
outcome sensitivity	The sensitivity of the assessment of the binary outcome
outcome specificity	The specificity of the assessment of the binary outcome
outcome reliability	The reliability of the measurement of the quantitative outcome
baseline OR	The effect associated with the heterogeneity in baseline disease risk for a binary outcome
interactive OR	The effect associated with the interaction between two exposures for a binary outcome
interactive effect	The effect associated with the interaction between two exposures for a quantitative outcome

Table 8: Parameters related to the outcome, the baseline risk and the interaction term.

The parameter '***outcome model***' determines the type of the trait of interest which can be a binary trait or a quantitative trait.

The parameter '***disease prevalence***' represents the frequency of the binary outcome in the general population on which the study is to be based. If the true prevalence is very low, a very large number of subjects will have to be simulated before enough cases are generated and this will make the simulation process very slow. It might not even be possible to generate enough cases for an extremely low prevalence.

The parameters '***outcome sensitivity***' and '***outcome specificity***' represent the sensitivity and specificity of the assessment of disease. One of the main strengths of the ESPRESSO-forte algorithm is that it allows taking into account measurement error in both outcome and explanatory variables. The level of assessment error depends on the accuracy of the test used to ascertain the presence or absence of a binary trait (outcome variable) or a binary exposure (explanatory variable). The accuracy of a test (or more correctly, here, categorisation) may be defined by its sensitivity and specificity (**35**). Sensitivity is the proportion of all those with the disease who are correctly identified as diseased (categorised as having the disease). Specificity is the proportion of all those

without the disease who are correctly identified as not diseased (categorised as not having the disease) (**62**). An ideal test is characterized by a very high sensitivity and a very high specificity (**63**). However one of these two characteristics might be more important than the other, depending on the purpose of the categorisation.

The parameter '***outcome reliability***' represents the reliability of the assessment of a quantitative outcome. Reliability (test-retest reliability) is a characteristic of a measure which reflects the consistency of the observed measurement of a quantitative variable across several repeats. An estimate of the reliability is given by the ratio of the variance of the true measurement ($\sigma^2_{M.T}$) to the variance of the observed measurement ($\sigma^2_{M.OBS}$):

$$Reliability = \frac{\sigma^2_{M.T}}{\sigma^2_{M.OBS}} \qquad \textit{Equation (6)}$$

In Equation (6) the denominator can be replaced by $\sigma^2_{M.T} + \sigma^2_{M.E}$ (variance of the true measurement + variance of the error) because an observed measurement M_{OBS} is equal to the sum of a true measurement M_T and an error E_M ($M_{OBS} = M_T + E_M$). So the equation can be re-written as:

$$Reliability = \frac{\sigma^2_{M.T}}{\sigma^2_{M.T} + \sigma^2_{M.E}} \qquad \textit{Equation (7)}$$

In ESPRESSO-forte, Equation (7) is exploited to generate observed normal data with a specified reliability.

The parameter '***baseline OR***' represents the heterogeneity in disease risk arising from determinants not measured or not included in the model. The variance in baseline risk is assumed to follow a normal distribution on the logistic scale – *i.e.* a normally distributed error term that is added to the linear predictor. Set for example to10, it means that a person at high risk (an individual at the top 95% percentile of population

risk) is, all else being equal, at 10 times the odds of developing the disease compared to a person at low risk (bottom 5% percentile of population risk).

The parameter '***interactive OR***' represents the ratio of the odds ratios of two interacting exposures, for a binary outcome; it is the odds ratio associated with the interaction term. If, for example, a disease is determined by the interaction between a SNP and an environmental factor, '***interactive OR***' is the odds ratio associated with having an 'at risk' genotype rather than a 'not at risk' genotype in subjects exposed to the 'at risk' environmental determinant compared to the same odds ratio in subjects not exposed to the 'at risk' environmental determinant.

The parameter '***interactive effect***' represents the effect of the interaction term on a normally distributed outcome. It is the expected change in the magnitude of the outcome phenotype that results from a one unit change in the magnitude of the interaction term.

2.6.1.2. Parameters for the genetic exposure

Parameter	Description
genetic model	A binary indicator set to 1 if a genetic exposure (SNP) is to be modelled as additive and to 0 if it is to be modelled as binary.
MAF	The prevalence of the rarer (minor) allele.
genetic OR	The effect associated with the minor allele for a binary outcome model
genetic effect	The effect associated with the minor allele for a quantitative outcome model.
genetic sensitivity	The sensitivity of the assessment of the alleles that form the genotype
genetic specificity	The specificity of the assessment of the alleles that form the genotype
LD	A binary indicator set to 1 if the two SNPs are to be modelled as being in LD and to 0 if they are to be modelled as independent.
Target LD	The level of LD between the two SNPs, if they are to be modelled as in LD.

Table 9: Parameters for the genetic exposures.
The genetic exposures are two SNPs which can be modelled as being in linkage disequilibrium.

The parameter '***genetic model***' sets the parameterisation model for a genetic exposure (SNP). In ESPRESSO-forte the genetic model can be either binary (dominant or recessive) or additive. To explain these models in detail let us consider a binary trait (disease) determined by a biallelic SNP; the two alleles are *d* and *D*. There will then be three possible genotypes: *dd*, *dD* and *DD*. A genetic model describes how a phenotype (trait) is related - logically or quantitatively - to the alleles composing the underlying genotype. If a single copy of one of the alleles suffices to determine that an individual is 'at high risk' of expressing a binary phenotype and a second copy has no additional impact then that allele is said to be *dominant*. This is also true if a single copy of an allele has the same quantitative impact on a quantitative trait as two copies. In either case, the alternative allele only determines the level of phenotypic response if it is present in two copies and is then said to be *recessive*. In a co-dominant genetic model neither of the alleles determines the phenotype alone in heterozygous individuals, both alleles contribute to the phenotype but they do not necessarily have the same contribution.

- Dominant model

If *D* is the 'risk allele' (the allele that increases the risk of disease relative to the other allele); then any individual carrying that allele (heterozygous *dD* and homozygous *DD*) is at risk. Under this setting, if the genetic model is binary, one copy of the risk allele D is sufficient to reach the maximum risk. However if the dominance of *D* is incomplete i.e. the phenotype of the heterozygous *dD* is not identical to the phenotype of the homozygous *DD*, the risk is typically larger for the homozygous *DD*. Incomplete dominance can be regarded as one form of co-dominance.

- Recessive model

If *D* is the dominant allele it means *d* is the recessive allele. Even if *d* increases the risk of disease then it will only be homozygous *dd* individuals that express this increased genetic risk. One copy of *d* is not sufficient to put the individual at risk.

- Additive model

The additive genetic model is a special case of the co-dominant model under which the effect of the heterozygote genotype (1 copy of the allele) is precisely midway between the effects of the two homozygote genotypes. If *D* confers increased risk then under a logistic model if the *Dd* genotype confers k times the odds of disease compared to the dd genotype, then the *DD* genotype confers k^2 time the odds of the *dd* genotype and k times the odds of the *Dd* genotype. Here "additivity" is technically observed on the scale of log-odds (the fundamental scale of analysis of a logistic model), and the model can equally well be viewed as 'multiplicative' on the scale of odds.

Similarly, for a quantitative trait such as obesity measured by BMI, under the additive genetic model, each copy of the 'risk' allele increases the expected BMI by a certain magnitude (effect size). This means that the expected BMI associated with the heterozygote genotype is exactly midway between the expected value associated with the two homozygote genotypes. This reflects 'additivity' on the natural scale on which BMI is measured. If the expected BMI associated with the *Dd* genotype is h kg/m^2 higher than that associated with the *dd* genotype, the expected BMI associated with *DD* will be h kg/m^2 higher than *Dd* and 2h kg/m^2 higher than *dd.*

This parameter '***MAF***' represents the frequency of the rarer allele in the population. If this frequency is low the simulation runs for longer. It may not be possible to generate the specified number of individuals if the risk allele is extremely rare.

The parameter '***genetic OR***'is the odds ratio associated with one unit increase in the genetic covariate (the extent to which the odds of disease is multiplied by being 'at risk' rather than 'not at risk' in a binary genetic model).

The parameter '***genetic effect***' is the additive effect of each additional minor allele (the 'risk' allele) of a SNP, for a quantitative outcome model. If the SNP is binary, a second copy of the risk allele does not increase the risk further and if the SNP is additive each additional copy of the risk allele increases the risk. This explanation refers to a dominant model (complete dominance of the risk allele).

The parameters'***genetic sensitivity***' and '***genetic specificity***' reflects the accuracy of the measurements of the alleles. This accuracy depends on the genotyping platform, the technology used for the assays and the algorithm used to call the genotypes (**64**).

In ESPRESSO-forte the two SNPs that represent the genetic exposures can be modelled as being in linkage disequilibrium. The parameter '***LD***'is a binary indicator that takes the value 1 if the two SNPs are to be modelled as in LD and 0 if there are to be modelled as independent. The correlated SNPs are generated using a multivariate normal distribution function and the method developed in the R package *HapSim* (**65**). *HapSim* models a haplotype (defined in section 1.2.3.3) as a multivariate random variable with known marginal distributions and pairwise correlation coefficients. The package allows for the simulation of a SNP haplotype of several biallelic loci. In my implementation of the method for ESPRESSO-forte I limited the number of loci to two because I am generating only two SNPs in LD. My implementation of the method consisted in two main steps: (1) I compute the covariance matrix required to generate two correlated binary vectors of length *n* (each vector represents one SNP and *n* is the number of observations i.e. individuals); this is done by the function *make.cov.mat* (see

page 264 of the package manual under Appendix 1). (2) I use the covariance computed in (1) to generate a matrix of data that follow a multivariate normal distribution; this second step is carried out by the function *sim.LDsnps* (see page 282of the package manual under Appendix 1). For two loci, there are four possible haplotypes; the sum of the frequencies of the four possible haplotypes across the *n* simulated individual is 1 and the Lewontin D and r correlation values calculated from the single frequencies of the four haplotypes is equal to the target level of correlation (desired level of LD) specified in the first step. If the number of individuals to simulate is large, the programme runs more slowly. Because this setting could be very time consuming, it is preferable to set the sample size and the number of simulations to low values for an initial explorative analysis.

The parameter '***target LD***' represents the level of linkage disequilibrium between the two SNPs if they are to be modelled as in linkage disequilibrium (LD). The user should consider the minor allele frequencies of the two SNPs when setting the desired level of LD; the minor allele frequencies of the SNPs should not be markedly different. It is for example not possible to simulate an LD of 1 if one SNP has a MAF of 0.05 and the other a MAF of 0.4. As shown in the below equations where: the random variables X and Y represent two distinct SNPs; Corr(X,Y), the correlation between X and Y; Cov(X,Y), the covariance between X and Y; V(X) and V(Y), the variances of X and Y and E(X) and E(Y), the expectations of X and Y. As the covariance between X and Y tends toward zero (i.e. X and Y are independent) the correlation also tends toward zero.

$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$ and $Cov(X,Y) = E(X,Y) - E(X)E(Y)$

So $Corr(X,Y) = \frac{E(X,Y) - E(X)E(Y)}{\sqrt{V(X)V(Y)}}$.

The expectation of a binary variable X is given by $E(X) = np$ where p is the probability of success (here it is the MAF) and n the number of observations. It follows that correlation is dependent upon the difference in MAF between the two binary variables.

2.6.1.3. Parameters for the environmental/life style exposure

Parameter	Description
environmental exposure model	A discrete indicator variable set to 0 if the exposure is binary, 1 if the exposure follows the normal distribution and 2 if it follows the uniform distribution.
environment prevalence	The frequency of the 'at risk' environmental/lifestyle exposure in the study population
environment OR	The effect associated with the 'at risk' environmental/lifestyle exposure for a binary outcome model
environment effect	The effect associated with the 'at risk' environmental/lifestyle exposure for a quantitative outcome model
environment sensitivity	The sensitivity of the assessment of the binary 'at risk' environmental/lifestyle exposure
environment specificity	The specificity of the assessment of the binary 'at risk' environmental/lifestyle exposure
environment reliability	The reliability of the measurement of the quantitative 'at risk' environmental/lifestyle exposure
mean/lower limit	The mean value if the exposure is normally distributed or the lower limit of the exposure if it is uniformly distributed
sd/upper limit	The standard deviation if the exposure is normally distributed or the upper limit of the exposure if it is uniformly distributed
skewness	The magnitude of the skewness of the normally distributed exposure

Table 10: Parameters for the environmental/life style exposure.
This exposure can be modelled as a binary, normal or uniform variable.

In ESPRESSO-forte, the parameter '***environmental exposure model***'is set to 0 if the environmental exposure is binomially distributed, 1 if it is normally distributed and 2 if it is uniformly distributed.

The parameter '***environment prevalence***'represents the frequency of the 'at risk' environmental/lifestyle exposure in the general population on which the study is to be based.

The parameter '***environment OR***' is the odds ratio reflecting the ratio of the risk of developing the disease in subjects exposed to the 'at risk' level of the environmental exposure compared to those that are not exposed to this 'at risk' level.

The parameter '***environment effect***' represents the effect size of the 'at risk' environmental determinant; it reflects the expected change in the magnitude of the outcome that is related to one unit change in the 'at risk' environmental exposure.

The parameters '***environment sensitivity***' and '***environment specificity***' represent the sensitivity and specificity of the assessment of the binary environmental exposure. The measurement error in an environmental exposure is determined by considering the reliability of a latent quantitative variable that is assumed to underlie the binary variable under consideration. This is an approach that is used widely in genetic epidemiology and is sometimes called the latent threshold model (**66-69**). Basically, one assumes that there is a standardized normally distributed variable underlying the binary variable in question and if the value of this Gaussian variable exceeds a threshold T, the subject is "at risk" (binary variable = 1) and if it is less than T then the subject is "not at risk" (binary variable = 0). The value T is fixed at the value that corresponds to the correct prevalence of being "at risk" in the population under study. Assessment error may then be viewed as being quantified by the hypothetical reliability of that latent variable. In the setting of ESPRESSO-forte, therefore, it is effectively assumed that the observed values of a binary exposures are realisations of a latent threshold model that generates the desired exposure prevalence, and that measurement error is modelled by imperfect reliability of the normally distributed latent variable (*i.e.* by adding an appropriate level of normally distributed error to that latent variable).

The parameter '***environment reliability***' is the reliability of the assessment of a quantitative environmental exposure. Reliability (test-retest reliability) is a characteristic of a measure which reflects the consistency of the observed measurement of a quantitative variable across several repeats.

The parameters '***mean/lower limit***', '***sd/upper limit***' and '***skewness***' are sub-arguments used when the environmental exposure is modelled as quantitative. The parameter '***mean/lower limit***'represents the mean under a quantitative normal exposure and the lower limit under a quantitative uniform exposure; '***sd/upper limit***'represents the standard deviation under quantitative normal exposure and the upper limit under a quantitative uniform exposure. The parameter '***skewness***' sets the asymmetry of the probability distribution of the generated normal environmental exposure data; it takes positive values (> 0) for a right skewed distribution, negative values (< 0) for a left skewed distribution and 0 for a non-skewed distribution.

2.6.2. DATA SIMULATION

In this part of the algorithm, the task is to first generate the true (error free) exposure and outcome data (step 2 in Figure 13) and then generate errors which are then added to the true data to obtained the 'observed' data (step 3 in Figure 13).

In step 2, subjects are sampled, in batches of twenty thousand subjects, from the population on which the study is to be based (by default the maximum size of the population is set to 20 million individuals) until the specified number of cases and controls (under binary outcome) or the specified number of subjects (under quantitative outcome) are achieved. For each individual, error free genetic, environmental and subject effect (effect related to the heterogeneity in baseline disease risk) data are

generated; this effect data is then used to generate the 'true' outcome data. The effect and outcome data are then stored in a matrix where the rows represent individuals.

In step 3, the observed outcome and exposures data are generated using the same strategy: an error is simulated and added to the true data generated earlier in step 2. If the outcome or exposure is binary, the error is a binomially distributed vector generated with a probability (probability of disease for the outcome or probability of being at risk for the exposure) given by the misclassification rate. For a binary variable that takes values 0 and 1, where 0 is a negative test (non-diseased or control) and 1 a positive test (diseased or case); the misclassification rate from 0 to 1 (true control classified as case) is given by *1-specificity* and the misclassification rate from 1 to 0 (true case classified as control) is given by *1-sensitivity*. If the outcome or exposure is normal, the error is a normally distributed vector with a mean of zero and a variance $\sigma^2_{M.E}$ derived from *Equation (7)* in page 66:

$$\sigma^2_{M.E} = \left(\frac{\sigma^2_{M.T}}{Reliability}\right) - \sigma^2_{M.T} \qquad \textit{Equation (8)}$$

For the uniformly distributed environmental exposure, the error is a normally distributed vector with a mean of zero and a variance obtained in the same way as in *Equation (8)*. Although the uniformly distributed exposure is not encountered as often as the normally distributed exposure, it is useful to have this option in ESPRESSO-forte to model ranked data when it is more appropriate to use ordered quantitative measurement. If, for example, the aim is to compare two groups and if the variable of interest was measured differently in each group, so that a direct comparison is not a sensible approach to use, the data can be ranked to hence obtain uniformly distributed data, which can then be directly compared.

The subsections 2.6.2.1 to 2.6.2.7 explain in details how the true and observed exposure and outcome data are generated in ESPRESSO-forte. Under each subsection a mathematical formula was included, where necessary, to explain the method and some pseudo code was included to show how the method was implemented in R.

2.6.2.1. True genetic exposure data

The genotypes are formed from the two alleles (alleles *A* and *B*) of a biallelic SNP. Each allele is a binomially distributed vector where the MAF represents the probability of having the risk allele (i.e. probability of 'success'); the number of observations (i.e. number of trials) is *n*:

$$Allele \sim B(n, MAF)$$

The common allele is denoted by 0 and the rare allele (risk allele) is denoted by 1.

```
allele.A = rbinom (number.of.individuals, 1, MAF)
allele.B = rbinom (number.of.individuals, 1, MAF)
```

Under an additive genetic model the genotype of an individual is the sum of two alleles. Under a binary genetic model, the genotype is formed by summing up the two alleles and then assigning 1 (individual at risk) to any individual that has a sum greater than 0.

$$G_{ADDITIVE} = allele.A + allele.B$$
$$G_{BINARY} = G_{ADDITIVE} > 0$$

2.6.2.2. True environmental exposure data

Under a binary environmental exposure model, the exposure is a binomially distributed vector where the prevalence of the 'at risk' environmental exposure is the probability of being exposed to the 'at risk' environment (i.e. probability of 'success'); the number of observations (i.e. number of trials) is *n*:

$E_{BINARY} \sim B(n, prevalence)$

```
E_BINARY = rbinom (number.of.individuals, 1, prevalence)
```

Under a quantitative normal environmental exposure model, the exposure is a normally distributed vector with a specified mean and a standardised standard deviation of 1. Under a quantitative uniform exposure, the exposure is a uniformly distributed vector with a specified lower (minimum) and upper limit (maximum).

$E_{NORMAL} \sim N(mean, 1)$

$E_{UNIFORM} \sim U(minimum, maximum)$

```
E_NORMAL = skew.rnorm (number.of.individuals,mean,1,skewness)
E_UNIFORM = runif (number.of.individuals,minimum,maximum)
```

2.6.2.3. Effect related to the heterogeneity in baseline risk of disease

This component models the impact of the heterogeneity in baseline disease risk. The baseline odds ratio (for individual on 95th percentile vs. 5th population percentile) has to be converted into the corresponding variance $\sigma^2_{subject.effect}$ for a normally distributed effect because the variance in baseline risk $\sigma^2_{baseline.OR}$is assumed to follow a normal distribution on the logistic scale.

$$\sigma^2_{subject.effect} = \left[\frac{\log(\sigma^2_{baseline.OR})}{2 \times qnorm(0.95)}\right]^2 \quad \textit{Equation(9)}$$

The subject effect data is, then, a normally distributed vector with a mean of zero and a variance $\sigma^2_{subject.effect}$.

$S.effect \sim N(0, \sigma^2_{subject.effect})$

```
S.effect = rnorm (number.of.individuals,0, sqrt(σ^2 subject.effect))
```

2.6.2.4. True outcome data

Binary outcome

Under a main effect model, the linear predictor (LP) used to determine the statuses of the simulated individuals is constructed using the true exposure data already simulated (two genetic determinants, two environmental determinants and the subject effect).

$$LP_{MAIN.EFECT} = \beta_0 + \beta_{X1}X_1 + \beta_{X2}X_2 + \beta_{X3}X_3 + \beta_{X4}X_4 + s.effect$$

Under an interaction model only two covariates are used to construct the LP; the interaction term is the product of two interacting covariates.

$$LP_{X1.X2} = \beta_0 + \beta_{X1}X_1 + \beta_{X2}X_2 + \beta_{X1.X2(}X_1*X_2) + s.effect$$

As indicated earlier (in section 2.4) the beta values are the *log(OR)* of the covariates and $\beta_0 = log\left(\frac{disease\ prevalence}{1-disease\ prevalence}\right)$, here '*disease prevalence*' refers to the prevalence of the disease in the study population. The true outcome data is a binomially distributed vector where the probability of disease is given by mu, the expit transformation of the LP: $mu = \frac{e^{LP}}{1+e^{LP}}$.

$$OUT_{BINARY} = rbinom\ (number.of.individuals,\ 1,\ mu)$$

Quantitative outcome

The below R code lines give the LP under respectively the main effect and the interaction model.

$$LP_{MAIN.EFECT} = \beta_0 + \beta_{X1}X_1 + \beta_{X2}X_2 + \beta_{X3}X_3 + \beta_{X4}X_4 + s.effect$$
$$LP_{X1.X2} = \beta_0 + \beta_{X1}X_1 + \beta_{X2}X_2 + \beta_{X1.X2}\times(X_1*X_2) + s.effect$$

The beta values represent the effect sizes of the covariates and β_0 is the trait mean. The true outcome data is a normally distributed vector that has a mean equal to the LP and a standard deviation of 1.

$$OUT_{NORMAL} = rnorm\ (number.of.individuals, LP, 1)$$

2.6.2.5. Observed genetic exposure data

The observed genetic exposure data is generated by introducing some misclassification in the vectors of true alleles (alleles *A* and *B*) generated previously using the function *misclassify* which I wrote as part of the ESPRESSO-forte R package (see page 267 of the manual under Appendix 1). The function *misclassify* turns some risk alleles into non-risk alleles (1→0) and some non-risk alleles into risk alleles (0→1). The level of misclassification is determined by the misclassification rates: $error1_{\rightarrow 0}$ = 1 – sensitivity and $error_{0\rightarrow 1}$ = 1 - specificity. The newly generated alleles are then combined to construct the observed genotypes.

$$allele.A.obs = misclassify\ (allele.A,\ error_{1\rightarrow 0},\ error_{0\rightarrow 1})$$
$$allele.B.obs = misclassify\ (allele.B,\ error_{1\rightarrow 0},\ error_{0\rightarrow 1})$$
$$G_{ADDITIVE.OBS} = allele.A.obs + allele.B.obs$$
$$G_{BINARY.OBS} = G_{ADDITIVE.OBS} > 0$$

2.6.2.6. Observed environmental exposure data

Under a binary environmental exposure model, the error is equivalent to a case-control misclassification. The misclassifications rates are determined by the sensitivity and specificity of the assessment of the exposure:$error_{1\rightarrow 0}$ = 1 – sensitivity and $error_{0\rightarrow 1}$ = 1 - specificity. The observed data are then obtained by applying these misclassification rates to simulated true data using the function *misclassify*.

$$E_{BINARY.OBS} = misclassify\ (E_{BINARY},\ error_{1\rightarrow 0},\ error_{0\rightarrow 1})$$

Under a quantitative normal exposure, the reliability of the assessment of the exposure is used to compute the variance of the assessment error. This variance is used to generate a normally distributed error with a mean of zero which is then added to the vector of true exposure data to obtain the observed data.

$$Reliability = \frac{\sigma^2_{\text{TRUE.SCORE}}}{\sigma^2_{\text{TRUE.SCORE}} + \sigma^2_{\text{ERROR}}}$$

It follows that:

$$\sigma^2_{\text{TRUE.SCORE}} + \sigma^2_{\text{ERROR}} = \frac{\sigma^2_{\text{TRUE.SCORE}}}{Reliability}$$

Then:

$$\sigma^2_{\text{ERROR}} = \left(\frac{\sigma^2_{\text{TRUE.SCORE}}}{Reliability}\right) - \sigma^2_{\text{TRUE.SCORE}}$$

$$Error \sim N(0, \sigma^2_{\text{ERROR}})$$

$\texttt{Error = rnorm (number.of.individuals, 0, } \sigma_{\text{ERROR}}\texttt{)}$

$\texttt{E}_{\text{NORMAL.OBS}} \texttt{ = E}_{\text{NORMAL}} \texttt{ + Error}$

Under a quantitative uniform exposure, the error is a normally distributed vector that has a mean of 0 and a variance obtained using the same logic as for the quantitative normal error because the error variance, here, is a close approximation of the variance of a normally distributed error. The error is added to the true uniform exposure to obtain the observed exposure.

$$Error \sim N(0, \sigma^2_{\text{TRUE}})$$

$\texttt{Error = rnorm (number.of.individuals, 0, } \sigma_{\text{TRUE}}\texttt{)}$

$\texttt{E}_{\text{UNIFORM.OBS}} \texttt{ = E}_{\text{UNIFORM}} \texttt{ + Error}$

2.6.2.7. Observed outcome data

The same method as for the binary and quantitative exposure is used to generate the observed outcome data. The true outcome data was constructed using the simulated true exposure data. To obtain the observed outcome data, an error determined by the sensitivity and specificity of the assessment of disease (for binary outcome) or the reliability of the outcome measurement (for quantitative outcome) is generated and added to the true outcome data.

Binary outcome

The error is a case-control misclassification. The misclassification rates are used to introduce the appropriate level of error in the true outcome data to obtain the observed data.

$$OUT_{BINARY.OBS} = misclassify\ (OUT_{BINARY}, error_{1\rightarrow0}, error_{0\rightarrow1})$$

Quantitative outcome

The reliability of assessment of outcome is used to compute the variance of the assessment error. This variance is then used to generate a normally distributed error with a mean of zero which is then added to the vector of true outcome data to obtain the observed outcome data.

$$\sigma^2_{ERROR} = \left(\frac{\sigma^2_{TRUE.SCORE}}{Reliability}\right) - \sigma^2_{TRUE.SCORE}$$

$$Error \sim N(0, \sigma^2_{ERROR})$$

$$Error = rnorm\ (number.of.individuals,\ 0,\ \sigma_{ERROR})$$

$$OUT_{NORMAL.OBS} = OUT_{NORMAL} + Error$$

2.6.3. DATA ANALYSIS

In this part of the algorithm, the observed data generated in the simulation block of the algorithm are analysed by regression analysis (steps 4 in Figure 13) and the estimates obtained from the regression analysis are used to calculate the sample size required to achieve the power specified in the input parameters and to estimate the power achieved under the specified settings (steps 5 in Figure 13).

2.6.3.1. GLM analysis

Steps 2, 3 and 4, in Figure 13, are repeated for a number of times equal to the specified number of runs; after each run a dataset D is generated as illustrated in Figure 14. The dataset is a matrix that contains the observed exposure, the observed outcome, and some other information not relevant for the analysis section; each row of the matrix holds the records of one individual. After each run the data is analysed using a generalised linear model (GLM). The linear predictor of the GLM is constructed using the observed exposure effect data and is different from the linear predictor used to determine outcome statuses in the simulation phase. The estimates (beta, standard error and z-score), obtained from the GLM analysis, are stored in three distinct vectors.

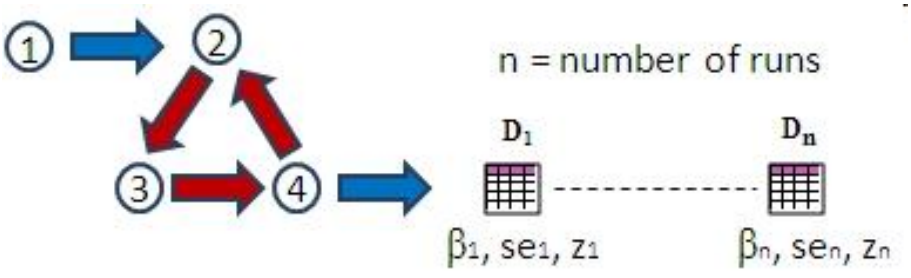


Figure 14: Graphical view of the GLM analysis in ESPRESSO-forte.
After each simulation a dataset is generated analysed and the estimates (beta, standard error and z-statistic) of the covariates stored.

The subsections below describe the GLMs fitted for binary and quantitative outcomes under main effect and interaction models.

Binary outcome

$$Logit(y) = \beta_0 + \beta_{X1}X_1 + \beta_{X2}X_2 + \beta_{X3}X_3 + \beta_{X4}X_4$$

$$Logit(y) = \beta_0 + \beta_{X1}X_1 + \beta_{X2}X_2 + \beta_{X1.X2} \times (X_1 * X_2)$$

Quantitative outcome

$$y = \beta_0 + \beta_{X1}X_1 + \beta_{X2}X_2 + \beta_{X3}X_3 + \beta_{X4}X_4$$

$$y = \beta_0 + \beta_{X1}X_1 + \beta_{X2}X_2 + \beta_{X1.X2} \times (X_1 * X_2)$$

2.6.3.2. Sample size calculation

To calculate the sample size required to achieve the desired power, I first estimate by how much the input study sample size needs to be inflated or shrunk (the relative change in standard error required) to reach the desired level of power. The relative change in standard error required is given by the ratio of the z-statistic required for the desired power (*z.power.required*) to the mean z-statistic obtained from the fitted GLM. The below formula shows how the sample size required (for difference in means between two groups) relates to the z-statistic required for the desired power, in a two-tailed test:

$$N = \frac{2\sigma^2(Z_\beta + Z_{\alpha/2})}{(\Delta\mu)^2}$$

For the sake of simplicity the above formula assumes two equal groups and the same variance, σ^2, for both groups. The sample size required N depends on the z-statistic required for the desired power which is the sum of the z-statistic of the desired power Z_β and the z-statistic of the desired level of statistical significance $Z_{\alpha/2}$. The effect size is represented by the difference in means $\Delta\mu$.

After each run the estimates of the GLM including the z-statistic are stored; the mean z-statistic is simply the average of the stored z-statistics. The sample size required to achieve the desired power is the product of the level of inflation or shrinkage required by the input sample size.

In ESPRESSO-forte, I obtain Z_β and $Z_{\alpha/2}$ by computing respectively the quantile values that correspond to the desired level of power (*desired.power*) and the desired level of statistical significance (*pvalue*); the z-statistic required for the desired power, denoted by `z.power.required` in ESPRESSO-forte, is then the sum of these two z-statistics as shown in the below formula where Q is the quantile function:

$$z.power.required = Q(desired.power) + Q(pvalue) = Z_\beta + Z_{\alpha/2}$$

In R the quantile function (with by default a mean of 0 and a standard deviation of 1) is '*qnorm*' so the z-statistic required for the desired power is obtained as follows:

```
z.power.required = qnorm (desired.power) + z.pvalue
```

where $z.pvalue = qnorm\left(1 - \frac{pvalue}{2}\right)$

The lines of code below show how the inflation/shrinkage required is computed for respectively main effect and interaction. The relative change in standard error required corresponds to a relative change on scale of square root of sample size; therefore the ratios in the below lines of code are squared. For a binary outcome the number of cases required and the number of controls required are calculated separately but the ratio of the number of cases to number of controls is preserved.

```
inflate.shrink.X = (z.power.required/mean.z.X)^2
inflate.shrink.X1X2 = (z.power.required/mean.z.X1X2)^2

sample.size.required.X = input.sample * inflate.shrink.X
```

`sample.size.required.`X_1X_2`= input.sample *inflate.shrink.`X_1X_2

2.6.3.3. Power estimation

The power is estimated both empirically (empirical power) and theoretically (modelled power). However it is the modelled power that is usually considered because under certain circumstances, explained below, the empirical power is not informative.

The empirical power is the proportion of simulations in which the z-statistic for the parameter of interest exceeds the z-statistic for the desired level of statistical significance (*z.pvalue*). Under respectively genetic effect and interaction, the empirical power is given by:

```
empirical.power.X = mean (z.X> z.pvalue)
empirical.power.X1X2 = mean (z. X1X2> z.pvalue)
```

The empirical power is not informative for extreme values of the standard error of the log odds ratio (i.e. when some of the counts (a, b, c or d) in Table 11 are extremely small) because the confidence interval becomes too wide as explained below (for a 95% confidence level):

	Cases	**Controls**
Exposed	a	b
Unexposed	c	d

Table 11: Contingency table reporting the results of a case-control study.

The approximate value of the standard error of the log odds ratio is given by:

$$SE = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{b}}$$

The limits of the confidence interval are given by:$Limit = \bar{x} \pm (SE \times 1.96)$,where $\bar{x}$represents the sample mean. This formula shows that as SE becomes larger the confidence interval becomes wider.

The modelled power should be considered if the empirical power is not informative. The modelled power is based on the ratio of the mean *beta* over the mean *standard error*, $\frac{E(\beta)}{E(SE)}$, and on the z-statistic for the desired level of statistical significance $Z_{\alpha/2}$. It is the probability that the z-statistic obtained from the GLM fit takes any value less than or equal to $\left(\frac{E(\beta)}{E(SE)} - Z_{\alpha/2}\right)$. So the modelled power is basically the cumulative distribution function (cdf) associated with the z-statistic, here a random variable, obtained from the GLM fit and this can be written mathematically as in the below formula where Z is the random z-statistic.

$$model.power = P\left(Z \leq \frac{E(\beta)}{E(SE)} - Z_{\alpha/2}\right)$$

In ESPRESSO-forte the ratio of the mean beta over the mean standard error is denoted *mean.z* and the R function *pnorm* is used to calculate the cdf:

```
mean.z = mean.beta / mean.se
model.power = pnorm (mean.z - z.pval)
```

2.7. HOW TO USE ESPRESSO-FORTE?

Currently the algorithm can be used as an R package or interactively from the Public Population Project in Genomics (P^3G) website:

http://www.p3gobservatory.org/powercalculator.htm

It is the ESPRESSO version developed in this thesis, ESPRESSO-forte, that forms the basis of the power calculator under the above link.

The R environment is required to use ESPRESSO-forte as an R package. The R development environment can be downloaded from the Comprehensive R Archive Network (CRAN) website (http://cran.r-project.org/). After installing R, the ESPRESSO-forte package can be installed by running, in R, the command **'install.packages ("ESPRESSO")'**which requires a connection to internet. Under UNIX operating system the package can downloaded from the CRAN repository of contributed packages (http://cran.r-project.org/web/packages/ESPRESSO/index.html) and installed by running, on the terminal, the command **'R CMD INSTALL ESPRESSO_1.1.tar.gz'**after setting correctly the path to the downloaded file. Under a WINDOWS operating system the R interface offers the possibility to download and install the package. Examples on how to use each of the 25 functions in the ESPRESSO-forte R package can be found in the package manual under Appendix 1. Users proficient in the R programming language can amend any of the functions to answer scientific questions that cannot be investigated directly using the downloadable version of the algorithm.

In the interactive version of the program that can be run directly from theP^3Gwebsite, the parameters are specified by typing numbers into appropriate boxes. The program is then run by clicking the 'Run calculation' button at the bottom of the screen and a summary of the answers appears in an output window.

If the algorithm is run from R, a summary of the simulation output is printed on the terminal screen and a more detailed output is stored as a '.csv' file (semicolon delimited file) in the working directory. In the interactive version a detailed output is also produced and can be downloaded as a '.csv' file. In addition to the main input parameters and the most relevant outputs (the sample size required to achieve the desired level of power and the modelled and empirical power), the detailed output also

contains the estimated effect sizes for each of the covariates. These estimates allow for the user to evaluate the level of shrinkage (shrinkage towards the null) of for example the ORs, in the presence of a non-differential error (an error which occurs with the same distribution in cases and controls) that is generated by the random error terms. The number of output values that are returned by the algorithm is restricted to the most relevant estimates to avoid unhelpful information overload; it is however possible to obtain the empirical estimates of almost all the input parameters by making minor amendments to some of the functions in order to store more estimates after each run.

In the first chapter I explained theoretically how statistical power is influenced by effect size, type I error and sample size. In the next two sections I apply ESPRESSO-forte to explore the influence on power (1) if an association is inferred using an observed genetic variant (here a SNP) in linkage disequilibrium with the unobserved causal variant and (2) if the genetic model of a SNP is not correctly specified.

2.8. EXPLORING THE IMPACT OF THE LEVEL OF LINKAGE DISEQUILIBRIUM BETWEEN CAUSAL AND OBSERVED GENETIC VARIANT ON POWER

2.8.1. INTRODUCTION

Alleles in linkage disequilibrium (LD) are more likely to segregate together (to be inherited together) than alleles that are not in LD (see section 1.2.3.3 for definition and calculation of LD). The stronger the LD is, the higher the probability is that the alleles will be co-inherited. This principle is used in genetic association studies to infer an association between an unobserved causal genetic variant and a particular trait, based

not on the causal variant itself, but on an observed genetic variant in close LD with the causal marker as shown graphically in Figure 15.

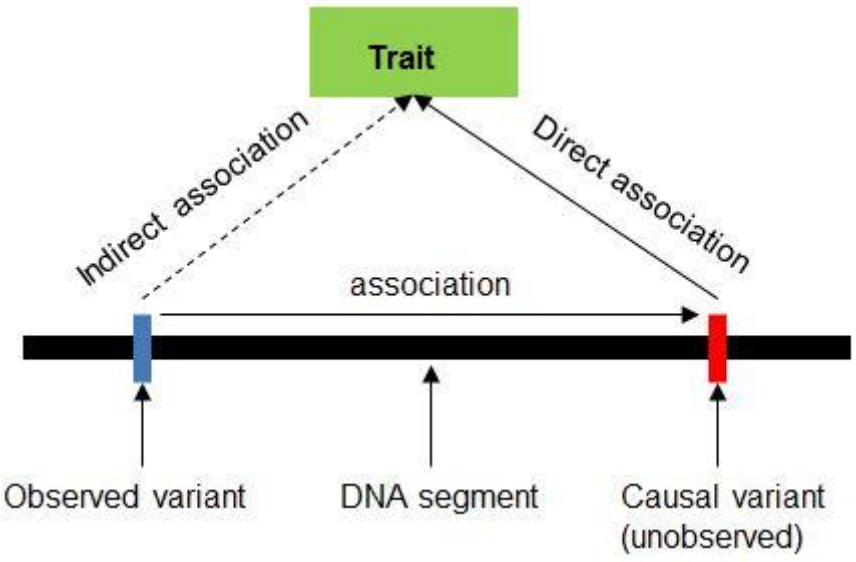


Figure 15: Inferring an association between an unobserved variant and a trait. The LD between an observed genetic variant and an unobserved causal genetic variant can be used to infer an association between the unobserved variant and a trait.

In an association study, when an observed variant is in perfect LD ($r^2 = 1$) with the unobserved causal variant, if the observed variant is genotyped it is as if the causal variant had been typed, provided that there are no genotyping errors. But if the level of LD between the two variants is less than 1, it cannot be assumed that the two variants are equivalent i.e. the observed marker does not carry all of the information that would have been available had the causal variant been genotyped. In this case when the observed variant is typed it is as if the causal variant was typed with a certain level of error; the magnitude of the error depends on the level of LD between the two variants, the weaker the LD is, the larger the error will be. The aim of the analysis in this section is to evaluate the influence of this error which is consequent solely on incomplete LD on the power of a hypothetical SNP genetic association study. Obviously the power is also influenced by the minor allele frequency of the genetic variant but the influence of

allele frequencies is not investigated here; the settings of the simulation ensure that the error influencing power depends solely on the incomplete LD between the observed and the causal variant.

2.8.2. METHODS

In ESPRESSO-forte, one modulates the sensitivity and specificity of the genotype assessment in order to generate an appropriate level of error in the observed genotypes (see section 2.6.2.5). In this analysis the first task was to calculate the sensitivity and specificity values that correspond to the specified levels of incomplete LD between observed variant and unobserved causal variant. The sensitivity and specificity levels were computed using the ESPRESSO-forte function *sim.geno.sesp.* The R documentation of the function is in page 279 of the package manual under Appendix 1. However, since the function is central to this analysis, the below section details the method implemented in the function.

DETAILS OF THE FUNCTION *'sim.geno.sesp'*

The function uses an approach called latent threshold model (explained in page 73). This approach assumes a standardized normally distributed variable underlying the binary variable of interest. In the context of the function *'sim.geno.sesp'* the underlying variable is a SNP allele where the allele that carries the risk ("at risk" allele) is given the value 1 and the other allele ("not at risk" allele) is 0. If the underlying variable, **X**, exceeds a threshold **T**, the subject has the "at risk" allele (allele = 1) and if it is less than **T** then the subject has the "not at risk" allele (allele = 0). The value **T** is fixed at the value that corresponds to the correct prevalence of being "at risk" in the population

under study. The assessment error is viewed here as being quantified by the hypothetical reliability of the underlying Gaussian variable. For the binary variable the assessment can be described as the level of correlation between the true and the observed measurements of the "at risk" allele.

Description of the function in 7 steps:

(1) Setting the input parameters

σ_1 $reliability\ of\ the\ measurement\ of\ the\ underlying\ variable\ X$

$\sigma_E = 1 - \sigma_1$ $error\ attached\ to\ the\ measurement\ of\ the\ underlying\ variable\ X$

p $prevalence\ of\ the$ *'at risk' allele in the study population*

r^2_{target} $target\ level\ of\ correlation\ between\ true\ and\ observed\ allele\ data$

ε $convergence\ value\ to\ compare\ the\ target\ and\ computed\ level\ of\ correlation\ between\ true\ and\ observed\ allele\ data$

n $sample\ size, the\ number\ of\ observations$

(2) Generating the normally distributed true and observed underlying X variable

$X_1 \sim N(0, \sigma_1^2)$ $true\ underlying\ variable\ for\ the\ n\ subjects\ at\ the\ specific\ SNP\ locus$

$E \sim N(0, \sigma_E^2)$ $measurement\ error\ data\ for\ the\ n\ subjects$

$X_2 = X_2 + E$ $observed\ underlying\ variable\ for\ the\ n\ subjects$

(3) Calculating the thresholds T_1 and T_2 for respectively the true and observed allele data (Q_p = the sample quantiles that correspond to the given prevalence p)

$T_1 = Q_p(X_1)$

$T_2 = Q_p(X_2)$

(4) Generating the binary vector,X_{1B} ,of true allele data by assigning 1 to all observations of X_1 that are less than the threshold T_1 and 0 to all observations of X_1 that are equal to or greater than the T_1. X_{1B} is hence equivalent to a binary variable with parameters n and p.

For the sake of clarity the subset of X_{1B} observations assigned 1 is named X_{1B1} and those assigned 0 represent X_{1B0}.In the below equations the lower x's denote the actual realisations of the variables.

$$X_{1B} \sim B(n,p)$$
$$x_{1B1} = 1 \ \ if \ \ x_1 < T_1$$
$$x_{1B0} = 0 \ \ if \ \ x_1 \geq T_1$$

(5) Generating the binary vector,X_{2B} ,of observed allele data by assigning 2 to all observations of X_2 that are less than the threshold T_2 and 1 to all observations of X_2 that are equal to or greater than the T_2. X_{2B} is hence equivalent to a binary variable with parameters n and p.

For the sake of clarity the subset of X_{2B} observations assigned 2 is named X_{2B2} and those assigned 1 represent X_{2B1}. In the below equations the lower x's denote the actual realisations of the variables.

$$X_{2B} \sim B(n,p)$$
$$x_{2B2} = 2 \ if \ \ x_2 < T_2$$
$$x_{2B1} = 1 \ if \ \ x_2 \geq T_2$$

(6) Tabulating the binary vectors X_{1B} vs. X_{2B} (each has two levels) and deriving the sensitivity and specificity of the assessment of the binary variable (i.e. the allele)

	X_{2B1} Observed non-risk	**X_{2B2}** Observed risk allele	

	allele		
X_{1B0} True non-risk allele	a	b	a + b Total number of individuals **not affected** by the risk allele
X_{1B1} True risk allele	c	d	c + d Total number of individuals **affected** by the risk allele

Table 12: Summary of the tabulation of X_{1B} versus X_{2B}.

$$sensitivity = \frac{d}{c+d} \qquad specificity = \frac{a}{a+b}$$

(7) Calculating the correlation between the true and observed binary allele data and comparing it to the target level of correlation

$$r^2_{empirical} = [Corr(X_{1B}, X_{2B})]^2$$

If $\left|r^2_{empirical} - r^2_{target}\right| \leq \varepsilon$:

The computed sensitivity and specificity values represent the sensitivity and specificity that correspond to the target correlation between the true and observed alleles, i.e. the allele is measured with the sensitivity and specificity values computed at step 6.

If $\left|r^2_{empirical} - r^2_{target}\right| > \varepsilon$:

The same process (steps 1 to 7) is re-run iteratively, with different sample size *n* and reliability σ_1, in a manner that minimises the value $\left|r^2_{empirical} - r^2_{target}\right|$, until convergence is reached.

For the purpose of this analysis ten levels of LD were investigated: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1. The MAF frequency of the genetic variant (a SNP) was set

arbitrarily to 0.1. For each of the levels of LD one sensitivity value and one specificity value were obtained (Table 15).

The influence of incomplete LD on power was investigated for a binary outcome and for a quantitative normal outcome. For both the binary and the quantitative normal outcome, there is only one covariate: the genetic exposure. The fitted GLM can be written as follows:

$$g(Y) = \beta_0 + \beta_1 G_1$$

In the above model Y is the outcome β_0 is the intercept value, β_1 is the effect size of the covariate G_1 which represents a genetic exposure (here a SNP). The terms in this model are explained in more details under the GLM section at page 52-53.

Under each of these two outcome models, ESPRESSO-forte was used to determine (1) the sample size required to achieve 80% power and (2) the power achieved with an input sample size of 2000 cases and 8000 controls (under binary outcome) and 5000 individuals (under quantitative outcome). Two genetic exposure models were analysed under each of the above settings: a binary and an additive SNP. The general and genetic parameters required by the algorithm to carry out the simulations are respectively under Table 13 and Table 14.

Parameter	**Value**
runs	1000
cases	2000
controls	8000
subjects	5000
outcome model	0/1
disease prevalence	0.1
baseline or	10
p.value	0.0001
power	0.8
sensitivity	1
specificity	1
reliability	1

Table 13: General and outcome parameters used in the analysis.
These input parameters were used for each of the ten levels of LD investigated. The outcomes analysed were binary (0) and normal (1).

Parameter	Value
genetic model	0 /1
MAF	0.1
OR	1.5
effect	0.25
sensitivity	0.39
specificity	0.93

Table 14: Parameters of the genetic exposure.
These input parameters were used for each of the ten levels of LD investigated. The SNP was simulated as binary (0) and additive (1).

2.8.3. RESULTS

The results are presented as comparative plots and tables that show how the magnitude of the error (indicated by the level of LD) affects sample size and power. Table 15 contains the levels of LD that were investigated and the corresponding sensitivity and specificity values calculated using the function *sim.geno.sesp* (see page 279 of the package manual under Appendix 1). As the LD decreases, the sensitivity and specificity also decrease but the specificity is better preserved i.e. for the same levels of LD, the sensitivity is lower than the specificity. However as power is less affected by low sensitivity it is unsurprising that even moderate levels of LD preserve power (see Table 16 and Table 17).

For the binary outcome, the power achieved with the additive SNP was higher than that achieved with the binary SNP, across the 10 levels of LD investigated (Table 16). For the quantitative outcome also the power achieved was larger with the additive SNP (Table 17). These observations were confirmed by the sample sizes required to achieve 80% power; under both outcome models, the sample size required was smaller for the additive SNP. The results are shown graphically in Figure 16 and Figure 17.

LD (r^2)	Sensitivity	Specificity
0.1	0.39	0.93
0.2	0.50	0.94
0.3	0.59	0.95
0.4	0.67	0.96
0.5	0.74	0.97
0.6	0.80	0.98
0.7	0.85	0.98
0.8	0.91	0.99
0.9	0.95	0.99
1.0	1.00	1.00

Table 15: Sensitivity and specificity values used in the analysis.
The above sensitivity and specificity figures correspond to the r^2values specified for a SNP that has a MAF of 0.1.

LD	Power achieved (2000 cases and 8000 controls)		Sample size (cases and controls) required for 80% power	
	Additive SNP	Binary SNP	Additive SNP	Binary SNP
0.1	5%	3%	46125	59655
0.2	20%	11%	24320	31490
0.3	44%	29%	16035	19965
0.4	66%	46%	12120	15560
0.5	83%	68%	9550	11720
0.6	92%	80%	8085	10025
0.7	97%	89%	6815	8515
0.8	99%	95%	5900	7405
0.9	100%	97%	5330	6640
1	100%	99%	4815	5680

Table 16: Power achieved with 2000 cases and 8000 controls.
The table also reports the sample sizes required to achieve 80% power under a binary outcome model and under different levels of LD. The ratio of cases to controls is 1:4.

LD	Power achieved (5000 subjects)		Sample size required for 80% power	
	Additive SNP	Binary SNP	Additive SNP	Binary SNP
0.1	7%	4%	19728	23851
0.2	28%	19%	10319	12214
0.3	56%	43%	6896	8082
0.4	76%	64%	5319	6199
0.5	91%	83%	4132	4738
0.6	96%	92%	3475	3978
0.7	99%	97%	2818	3420
0.8	100%	99%	2430	2918
0.9	100%	100%	2189	2628
1	100%	100%	1967	2359

Table 17: Power achieved with a sample size of 5000 subjects.
The table reports also the sample sizes required to achieve 80% power under a quantitative outcome model and under different levels of LD.

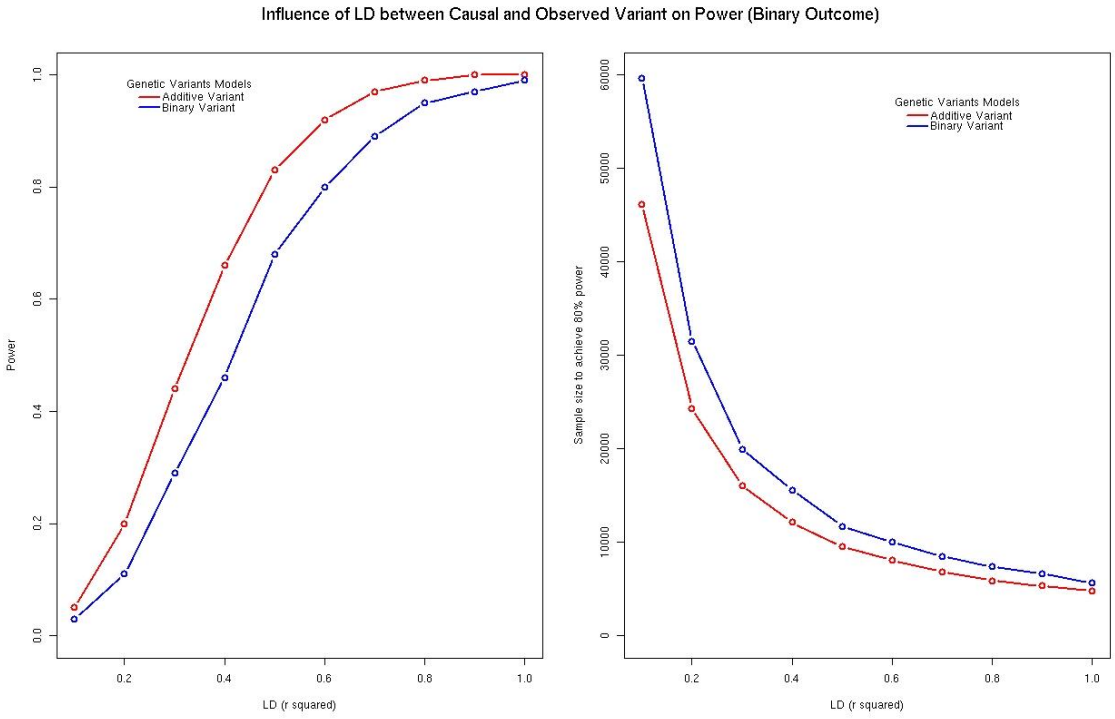


Figure 16: Binary outcome: impact of incomplete LD on power and sample size
The outcome is determined by an additive or a binary SNP.

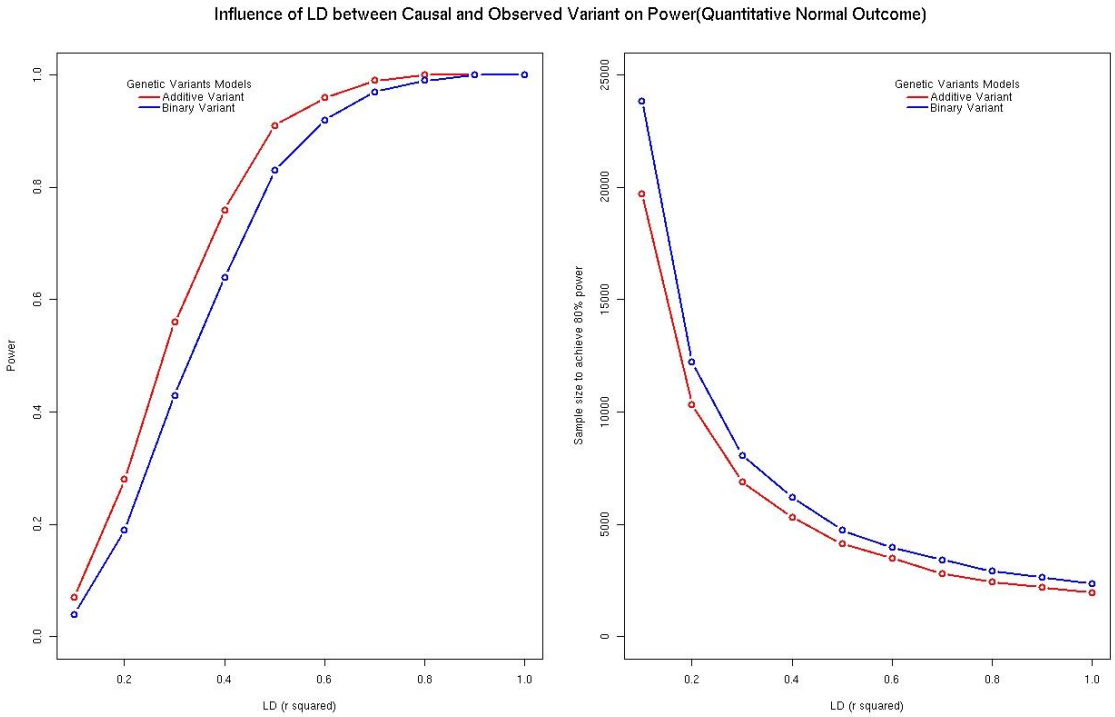


Figure 17: Quantitative outcome: impact of incomplete LD on power and sample size The outcome is determined by an additive or a binary SNP.

2.8.4. DISCUSSION

An association between an unobserved causal SNP and a trait can be inferred through LD between the causal SNP and an observed non-causal SNP (**18**). If the LD between the two variants is incomplete ($r^2 < 1$), the error in the estimated association has a component that depends on the level of LD. By measuring a variant not in perfect LD with the causal variant, it is as if we were measuring the causal SNP with an error equivalent to the incomplete LD.

The results of the analysis show that the sensitivity decreases markedly with a decreasing level of LD. The magnitude of the loss of power that is due to incomplete LD between the observed and the causal variant is larger for a binary SNP than for an additive SNP under both binary and quantitative outcome models. The relationship between the loss of power and the level of LD was not linear. As the LD decreases the

power is better preserved when the outcome is quantitative than when it is binary i.e. the lowest power setting was encountered when the outcome was binary. This suggests that the binary outcome model is more sensitive to incomplete LD. It is hence advisable, in a candidate gene study, to consider only markers that are very close to the supposed position of the causal variant, when inferring an indirect association between a SNP and an outcome; because markers that are physically close are in higher LD.

The results of this analysis are in the line with the argument that power can be gained by choosing a model based on additive allelic effects instead of a binary genetic model (**7**). However, modelling a binary SNP as additive represents in fact a misspecification of the genetic model. In the next section, I investigate the consequences of such genetic model misspecification on power, under different settings.

2.9. EXPLORING THE IMPACT OF GENETIC MODEL MISSPECIFICATION ON POWER

2.9.1. INTRODUCTION

In order to explore the impact of the misspecification of a genetic model using simulation-based power calculation, it is necessary to simulate the "true" data under one genetic model (the "true" model) and then to analyse those data as if they had been generated under an alternative model (the "misspecified" model). This was impossible using the original version of ESPRESSO (**7**), because a single model was specified as an argument to the function and this model was used both for simulation and analysis. But an exploration of the impact of misspecifying the genetic model is important, and this section describes the development and use of an extended version of ESPRESSO-forte that enables such an analysis to be undertaken.

Figure 18 provides a pictorial representation of one form of genetic model misspecification: a binary variant is erroneously analysed as additive and *vice-versa*.

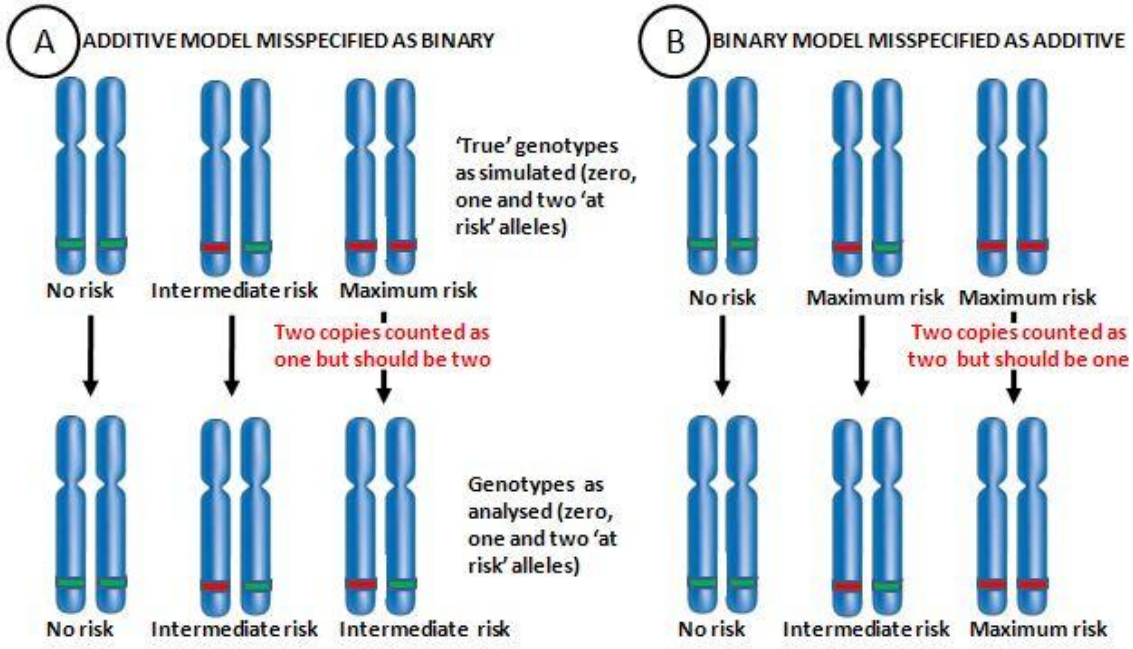


Figure 18: Graphical illustration of genetic model misspecification.
This illustration assumes a biallelic genetic variant. If an additive genetic model is analysed as binary (A) the modelled risk is underestimated for individuals homozygous for the risk allele (in red); two copies of the risk allele are treated as being equivalent to one copy. If a binary model is analysed as additive (B) the risk is overestimated for individuals homozygous for the risk allele in the fitted model it is as if the second copy increases further the modelled risk whilst this should not be the case in a binary model.

In this analysis ESPRESSO-forte was modified to investigate the impact of the error caused by genetic model misspecification on the statistical power of a SNP association study. The aim is to understand the implications of misspecifying the underlying genetic model in estimating statistical power for a binary or quantitative trait of interest. This analysis is important because the genetic model of each of the genetic determinants of a given trait are generally not known with certainty *a priori*; the results of this investigation could tell how detrimental genetic model misspecification is to power for binary and quantitative traits. If the impact on power is substantial then methods for a better ascertainment of the genetic models of traits determinant should be sought.

2.9.2. METHODS

In this analysis also only one genetic exposure (a SNP) is fitted as covariate. The fitted GLM model can be written as follows:

$$g(Y) = \beta_0 + \beta_1 G_1$$

In the above model Y is the outcome β_0 is the intercept value, β_1 is the effect size of the covariate G_1, the genetic exposure (here a SNP). The terms in this model are explained in more details under the GLM section at page 52-53.

For each of the two types of misspecification shown in Figure 18, the analysis was done in two steps: (1) simulation of data (2) analysis of the data.

(1) The alleles used to construct the genotypes are generated as explained in section 2.6.2. Under both genetic models (binary or additive), the genotype of an individual with no copy of the 'at risk' allele is 0. Under a binary model, the genotype of an individual with either one or two copies is 1. Under an additive model the genotype of an individual with one copy of the 'at risk' allele is 1 and that of an individual with two copies is 2. A linear predictor is used, as explained in details in section 2.6.2, to produce the outcome values.

(2) A logistic regression model is fitted and the sample size required to achieve a power of 80% is calculated (see section 2.6.3 for details of the analysis and sample size calculation in ESPRESSO-forte). The sample size is calculated twice: once assuming no misspecification of the genetic model and once assuming that the genetic model was misspecified. If a genetic model (binary or additive) is used in the simulation step to construct the genotypes and the same model is used for the genotypes fitted in the regression analysis then there is no misspecification. But if for example a binary genetic model is used in the

simulation step whilst an additive one is fitted in the logistic regression then a misspecification occurs.

The difference between the sample size value estimated under the "true" model and the one estimated under the "misspecified" model represents the sample size increase required to compensate for the loss of power caused by the misspecification. The parameters used for the analysis are under Table 18 and Table 19. The analysis was carried for ten different levels of minor allele frequency (0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45) to investigate whether the effect of the genetic model misspecification error on power is influenced by the MAF of the SNP; this, because the proportion of individuals, in the study sample size, whose genotypes were incorrectly assessed (individuals homozygous for the risk allele) depends on the MAF (here, the frequency of the 'at risk' allele). It is reasonable to hypothesize that the magnitude of the error resulting from the misspecification depends on the proportion of individuals whose genotypes were incorrectly assessed and hence on the MAF.

Parameter	**Value**
runs	1000
cases	1000
controls	4000
subjects	2500
outcome model	0/1
disease prevalence	0.1
baseline or	10
p.value	0.0001
power	0.8
sensitivity	1
specificity	1
reliability	1

Table 18: General and outcome parameters used in the analysis.
These input parameters were used for each of the ten levels of MAF investigated. The outcomes analysed were binary (0) and normal (1).

Parameter	Value
genetic model	0 /1
OR	1.5
effect	0.25
sensitivity	1
specificity	1

Table 19: Parameters of the genetic exposure.
These input parameters were used for each of the ten levels of MAF investigated. The genetic exposure was a SNP simulated as binary (0) and additive (1).

2.9.3. RESULTS

2.9.3.1. Power of case-control studies

The results in Table 20 report the percentage increase in sample size required to achieve a power of 80%when the genetic model of a binary or additive SNP was correctly or incorrectly specified under case-control settings (binary outcome). When an additive SNP was analysed as binary (Table 20), the misspecification causes a loss of power which increases with the increasing MAF of the SNP; hence the sample size required to compensate for the loss of power resulting from the error (Table 20) was larger for more frequent SNPs. When a binary SNP was erroneously analysed as additive (Table 20), the misspecification of the genetic model causes also a loss of power which increases with increasing MAF but the magnitude of the loss was slightly less than what was observed when an additive SNP was misspecified as binary.

MAF	Additive SNP			Binary SNP		
	Analysed as additive	Analysed as binary	% sample size increase	Analysed as binary	Analysed as additive	% sample size increase
0.01	9083	9139	1%	9238	9232	0%
0.05	1794	1866	4%	2012	2049	2%
0.1	968	1049	8%	1168	1240	6%
0.15	689	780	13%	924	1005	9%
0.2	571	669	17%	849	969	14%
0.25	499	605	21%	805	921	14%
0.3	464	587	27%	836	1018	22%
0.35	433	583	35%	877	1101	26%
0.4	413	602	46%	948	1263	33%
0.45	419	660	58%	1077	1514	41%

Table 20: Case-control study: number of cases required to achieve 80% power. The figures reported in table were the number of cases required to achieve 80% power when the genetic model of a binary or additive SNP was misspecified in a case-control study. The number of controls was 4 fold the number of cases.

2.9.3.2. Power for quantitative traits

The results in Table 21 report the percentage increase in sample size required to achieve a power of 80% when the genetic model of a binary or additive SNP was correctly or incorrectly specified whilst the outcome is a quantitative trait. When an additive SNP was analysed as binary (Table 21), the misspecification causes a loss of power which increases with the increasing MAF of the SNP; the more frequent the SNP is the larger the sample size required to compensate for the loss of power resulting from the error (Table 21). When a binary SNP was erroneously analysed as additive (Table 21), the misspecification of the genetic model causes also a loss of power which increases with increasing MAF. The magnitude of the loss was similar to what was observed when an additive SNP was misspecified except for MAFs of 0.2 and 0.25. The sample size increase required to compensate for the loss of power, when an additive SNP with a MAF of 0.2 was misspecified as binary, was nearly twice that required when a binary

SNP with the same MAF was fitted as additive. The sample size increase required, when a binary SNP with a MAF of 0.25 was fitted as additive, was nearly twice that required when an additive SNP with the same MAF was fitted as binary.

MAF	Additive SNP			Binary SNP		
	Analysed as additive	Analysed as binary	% sample size increase	Analysed as binary	Analysed as additive	% sample size increase
0.01	19028	18985	0%	19138	19340	1%
0.05	3725	3859	4%	4044	4107	2%
0.1	1963	2089	6%	2311	2421	5%
0.15	1422	1532	8%	1785	1956	10%
0.2	1106	1296	17%	1606	1732	8%
0.25	975	1095	12%	1433	1725	20%
0.3	861	1043	21%	1445	1757	22%
0.35	806	1025	27%	1506	1920	27%
0.4	759	1028	35%	1606	2121	32%
0.45	708	1045	48%	1748	2546	46%

Table 21: Quantitative outcome: number of cases required to achieve 80% power. The figures reported in table were the number of cases required to achieve 80% power when the genetic model of a binary or additive SNP was misspecified whilst the outcome was quantitative. The number of controls was 4 fold the number of cases.

2.9.4. DISCUSSION

The results showed that, in a SNP association study, power is lost if the genetic model of the SNP is not correctly specified; the amount of power lost is relatively small for very rare and rare SNPs (MAF 0.01 and 0.05) and larger for moderately common (MAF 0.1 to 0.25) and fairly common SNPs (MAF 0.3 to 0.45). If the SNP is binary the loss of power is more important than when it is an additive SNP that is incorrectly specified. The adverse effect of the genetic model misspecification on power was more pronounced for a binary outcome than for an additive outcome if the SNP is additive. If it is a binary SNP that is misspecified the impact on power is similar whether the outcome is quantitative or binary.

If the genetic variant that determines the outcome is additive and if complete dominance is assumed, each additional copy of the allele that carries the risk (risk allele) increases the risk so that individuals with two copies of the risk allele have a higher risk than those with one copy. If a *true* additive model is incorrectly specified as binary, the risk is underestimated for individuals homozygous for the risk allele because under a binary model the second risk allele does not increase the risk. This underestimation of the risk of some homozygous individuals that results from the genetic model misspecification represents an error that affects adversely the power. This explains why the misspecification in Figure 18A (additive SNP analysed as binary), causes a loss of power. The power loss increases with the increasing frequency of the risk allele because the proportion of individuals whose risk is underestimated becomes larger which causes a larger error.

Under a *true* binary genetic model an individual with two copies of the risk allele has the same risk as an individual with only one copy because an additional allele does not increase further the risk as can be seen in Figure 18B. Thus, if a binary model is incorrectly specified as additive, the risk for individuals homozygous for the risk allele is overestimated. The overestimation of the risk for homozygous subjects represents an error that decreases the power of a study. The proportion of homozygous individuals increases with increasing MAF and hence the proportion of individuals whose risk is overestimated becomes larger and that causes an increase in the magnitude of the error resulting from the model misspecification.

Genetic model misspecification has little effect on power for very rare SNPs (MAF < 0.05) because there are nearly no individuals carrying two copies of the risk allele when a SNP is very rare. In other word, there is virtually no individual whose risk is under or overestimated because the error arises only when there are subjects with two copies of

the ‘at risk’ allele. For moderately common and common SNPs, the error is responsible for a substantial loss of power, particularly for common SNPs (MAF ≥ 0.3).

In the next chapter, I use ESPRESSO-forte, developed as described in chapter 2, to explore the statistical power profile of a large Canadian cohort study given a range of possible sizes which that cohort could be planned to achieve at the end of the recruitment process. The analysis in chapter 3 constitutes an illustration of an important real-world application of the ESPRESSO-forte platform.

CHAPTER 3

3. ANALYSIS OF THE POWER OF THE CANADIAN PARTNERSHIP FOR TOMORROW COHORT PROJECT TO STUDY QUANTITATIVE TRAITS

This analysis was requested by The Canadian Partnership for Tomorrow (CPT) to inform the primary (and immediate) strategic decision to be made by the Canadian Partnership Against Cancer (CPAC) on whether to continue recruitment at a rate that is likely to produce a total of approximately 110000 participants by the end of March 2012, or alternatively to prioritise and step up recruitment with the aim of recruiting as many as 180000.

The request of CPT including two main tasks: (1) the analysis of the power profile of the cohort for the investigation of quantitative traits and (2) the analysis of the power of profile of the cohort for the investigation of binary traits. My task was to undertake the first analysis (investigation of the power profile of CPT for quantitative traits) and it is this work that is reported in this chapter. The investigation of the power profile of CPT for binary traits was completed by Professor Paul Burton. Both analyses (power profiles for quantitative and binary traits) were reported in one document sent to the CPT board. The text of this chapter is closely based on the text of the report sent to the CPT board, with the permission of Professor Paul Burton who jointly wrote that report with me.

3.1. INTRODUCTION

The Canadian Partnership for Tomorrow (CPT) is a pan-Canadian initiative funded by the Canadian Partnership Against Cancer (CPAC). It aims to create a national biobank/bio-repository to provide a platform for future research on common chronic disease including cancer and cardiovascular disease (**70**). CPT is to be based on the integration of five large provincial cohorts each recruiting several tens of thousands of middle-aged participants. The planned (target), current and projected final recruitment numbers for the CPT project (at the time of this thesis) are outlined in Table 22 (**70**). The estimated range of the probable final sample size is 110000-180000.

Name	Age-range at recruitment	Target sample size	Current number of recruits (approximate)	Likely sample size by 31/3/2012 (approximate)
Atlantic Cohort	40-69 years	30,000	11,000	15,000-25,000
British Columbia Cohort	40-69 years	40,000	11,000	15,000-25,000
CARTaGENE (Quebec)	35-69 years	20,000	20,000	20,000
Ontario Health Survey (Ontario)	35-69 years	150000	11,000	35,000-70,000
The Tomorrow Project (Alberta)	40-69 years	50,000	14,000	25,000-40,000
CPT Project overall	Predominantly 35-69 years	250,000	67,000	110000 – 180,000

Table 22: Planned, current and probable final recruitment configuration of the CPT.

Contemporary bioscience often demands huge sample sizes. This is because most diseases of public health importance are multi-factorial and effect sizes are typically small. In consequence, many of the scientific questions that are now being asked simply cannot be answered using data from one study or one biobank alone (**7, 71, 72**) and large collaborative consortia have been responsible for much of the recent progress in

human population genomics (**73-82**). The aim of this analysis is to assess the statistical power of CPT as a platform for research projects exploring quantitative traits as outcomes given its likely definitive sample size.

When designing a large infrastructural platform for future biomedical research that costs millions of dollars to set up, even a small misjudgement of the required sample size can have major financial and scientific implications. It is hence crucial, for the power and sample size calculations of a large platform such as CPT, to use an approach that takes realistic account of unavoidable complexities in biomedical datasets such as assessment error in outcome and explanatory data, and the likely impact of many causal factors of any complex disease that will not have been measured by even the most thorough of studies. The ESPRESSO-forte algorithm was used to carry out the calculations because unlike standard approaches it takes account of the complex elements mentioned above.

The detailed assumptions that were made to underpin the power calculations are outlined under section 3.2.4. These assumptions reflect a rational attempt to take appropriate account of the realistic uncertainty that exists in the analysis of any complex trait. Although this uncertainty is often ignored in conventional power calculations (**7**) it consistently increases sample size requirements and it is the ability to take realistic factors, such as these, into account that renders power calculation by simulation the preferred approach when undertaking power analyses for large cohorts and biobanks which demand vast investment of time and effort.

3.2. METHODS

The exploration of the power profiles of CPT for quantitative outcomes was based on the estimates of participant-to-participant variation (standard deviation) of an extensive

range of critical disease-related traits. These traits were originally drawn up for the power calculations of the CARTaGENE project (**83**) carried out by Paul Burton and Catherine Boileau

The outcome variables investigated in the current analysis (*i.e.* included in this thesis) are either physical measures or biochemical and haematological parameters. The scientific rationales that justified the inclusion of these variables in this study are reported under sections 3.2.1 and 3.2.2; they are the same as those that justified the inclusion of the variables in the CARTaGENE project. The set of outcome variables to analyse and their assumed distribution (mean and standard deviation) are under Table 23 to Table 38. These distributions were derived from a literature review (as referenced), and/or from empirical data derived from the optimization phase of the CARTaGENE project. The biomedical scenarios for which CPT power profiles were investigated are detailed in Table 39 under section 3.2.3.

3.2.1. SCIENTIFIC RATIONALES AND STATISTICAL DISTRIBUTIONS OF THE PHYSICAL VARIABLES ANALYSED AS OUTCOMES

3.2.1.1. Arterial stiffness and central blood pressure

Unrelated studies have reported that central arterial stiffness is increased in elderly (**84**) and in patients with coronary artery disease (**85**), myocardial infarction (**86**), heart failure(**87**), hypertension (**88**), stroke (**89**), diabetes (**90**), end stage renal disease (**91, 92**) and hypercholesterolemia (**90**). The Anglo-Scandinavian Cardiac Outcomes Trial (ASCOT) study has suggested that central aortic blood pressure (BP) is a better predictor of cardiovascular mortality and morbidity than peripheral. A recent study found that more than 70% of individuals with high-normal brachial pressure and those

with stage 1 hypertension have similar aortic pressure; this suggests that central pressure cannot be reliably inferred from peripheral pressure (**93**).The measures required (Table 23) were obtained using a SphygmoCor, a non-invasive device that derives the central aortic pressure waveform from the pressure pulse measured at peripheral sites.

Outcome	Mean	Standard Deviation
Aortic Systolic BP	118 mm/Hg	18.3
Aortic Diastolic BP	82 mm/Hg	11.1
Aortic Pulse Pressure	36 mm/Hg	10.4
Aortic augmentation index	28.6 mm/Hg	11.4

Table 23: Variables that relate to arterial stiffness and central blood pressure. The table reports the distributions (in the CARTaGENE project) of the set of outcome variable.

3.2.1.2. Electrocardiogram (ECG)

ECG is a non-invasive tool for investigating cardiac arrhythmias and other cardio-electrical abnormalities in epidemiological studies. It is a key component of a cardiovascular work-up in the clinical setting. A full 12-lead ECG could not be used in all of the projects in CPT. This is in part because of the time and resources that would be required, in part because of the consequent need to act clinically in response to the potentially subtle abnormalities that may be found, and in part because the preparation of a full ECG requires the individual to partially undress and some subjects are uncomfortable with that. In consequence, a limited ECG (four lead ECG) was undertaken allowing derivation of the RR and QS intervals, respectively the interval between two consecutive heart beats and the interval between the Q and S waves of the QRS complex (a graphical depiction of the electrical energy generated by the heart, in an ECG). These are clinically relevant because 10 milliseconds (ms) increase in the QS interval is associated with a 15% higher risk for incident heart failure, a 13% higher risk for coronary heart disease, and a 17% higher risk for mortality in patients with chronic

kidney disease (**94**). And also, abnormalities in the variability and duration of the QT interval have been reported to be associated with life-threatening arrhythmias (**95**). The QT interval is the time elapsed from the beginning of a QRS complex to the T wave; in the below table it is the corrected QT interval, QTc - ratio of the QT interval by the square root of the preceding RR-interval (**96**)-, that is reported

Outcome	Mean	Standard Deviation
RR-interval	796 ms	107
QS-interval	100 ms	19
QTc-interval	426 ms	61

Table 24: Variables that relate to cardiac function.
The table reports the distributions (in the CARTaGENE project) of the set of outcome variable.

3.2.1.3. Peripheral Blood pressure

High peripheral blood pressure is a well-known risk factor for coronary heart disease, stroke and several other vascular, cerebrovascular and renal diseases. Approximately 50% of cardiovascular disease can be attributed to supraoptimal blood pressure due to its strong causal relationship. The prevalence of hypertension and the prescription of antihypertensive drugs have increased in Canada (**97**). Diabetes and obesity are strongly associated with hypertension and cardiovascular disease and 51.9% of Canadian diabetics are hypertensive (**98**). The measurements in Table 25 represent the mean of 3 measurements carried out by an automated device (Colin Press-Mate Prodigy II 2200) that uses the oscillometric method for assessing blood pressure.

Outcome	Mean	Standard Deviation
Systolic BP (mean of 3 measurements)	126 mm/Hg	18.2
Diastolic BP (mean of 3 measurements)	81 mm/Hg	10.7

Table 25: Variables that relate to blood pressure.
The table reports the distributions (in the CARTaGENE project) of the set of outcome variable.

3.2.1.4. Lung function

Spirometry is the gold standard for the diagnosis and assessment of chronic obstructive pulmonary disease (COPD) because it is the most reproducible, standardized and objective way of measuring airflow limitation (**99**). Spirometry is a long-term predictor for overall survival rates in both genders (**100**) and it has been reported as a predictor for cardiovascular and cerebrovascular disease and for death from all causes (**101**), as well as for lung cancer and chronic lung disease (**102**). The values for the forced expiratory volume in 1 second (FEV_1, the volume of air that can been forcibly expelled in 1 second) and the forced vital capacity (FVC, the volume of air that can be expelled after full inspiration) reported in Table 26 were obtained using a portable USB spirometer (MiniSpir). The values reported in the table below represent the proportion of predicted values (similar to percent predicted) and not the raw measures which are usually measured in litres per second (FEV_1) and litres (FVC).

Outcome	Mean	Standard Deviation
FEV_1 proportion of predicted	0.96	0.202
FVC proportion of predicted	0.94	0.203
FEV_1/FVC	1.03	0.113

Table 26: Variables that relate to lung function.
The table reports the distributions (in the CARTaGENE project) of the set of outcome variable.

3.2.1.5. Bone density

Bone mineral density (BMD) is a reliable predictor of cardiovascular (in white men) and other causes combined (in whites and blacks) as well as death from all causes in white men and blacks (**103**). Femoral dual x-ray absorptiometry (DXA)and ultrasound have a similar capacity to predict the risk of hip fracture (**104**). Quantitative calcaneal ultrasound appears to have the same ability as BMD to predict osteoporotic fracture and may predict fracture independent of bone mineral density (**105**). The values in Table 27

were obtained using the Achilles QUS Systems. This device a bone ultrasonometer that evaluates the bone status in the heel by measuring the speed at which sound travels through the bone in m/s and the amount of sound absorbed by the bone (broadband ultrasound attenuation – BUA) in db/MHz (**106-108**).

Outcome	Mean	Standard Deviation
Os calcis BUA index	33.44 dB/MHz	0.522
Os calcis bone density t-score	-0.07	1.1
Os calcis percent normal bone density	99.00%	17.56

Table 27: Variables that relate to bone density.
The table reports the distributions (in the CARTaGENE project) of the set of outcome variable.

3.2.1.6. Grip strength

Hand grip strength is a predictor of functional limitations and disability in old age as well as all cause and cardiovascular mortality (**109, 110**). Hand grip strength was reported as highly predictive of functional limitations and disability 25 years later, among healthy 45 to 68 years old men (**111**). For women, low hand grip strength was associated with an increased risk of developing incident vertebral fracture. Low hand grip strength was also associated with lower bone mass. The grip strength values in Table 28 represent the mean of two measurements obtained using a Digital Hydraulic Hand Dynamometer.

Outcome	Mean	Standard Deviation
Grip strength left hand (mean of two measures)	32.8 Kg	12.1
Grip strength right hand (mean of two measures)	33.8 Kg	12.4

Table 28: Grip strength for left and right hand.
The table reports the distributions (in the CARTaGENE project) of these two variables.

3.2.1.7. Bioimpedance

Bioimpedance analysis is non-invasive tool used to determine body composition. It allows for the level of many body fluids and tissues to be estimated. These measures when combined appropriately with data on age, sex, weight and height can be a good indicator of the absolute and relative amounts of adipose and lean tissue. It is a clinical method and a potential field for evaluating skeletal muscle mass. Body composition modulates common carotid artery remodelling independently of metabolic and atherosclerotic factors. It is also a good indicator for the assessment of nutritional status in individuals with bowel disorders (**112**). The measurements in Table 29 were obtained using a device (TANITA, TBF -10) that measures simultaneously leg-to-leg impedance and body weight.

Outcome	Mean	Standard Deviation
Fat mass by bioimpedance	22.38 Kg	10
Lean mass by bioimpedance	51.61 Kg	10.14
Percent body fat by bioimpedance	30.56 Kg	10.26

Table 29: Variables that relate to bioimpedance.
The table reports the distributions (in the CARTaGENE project) of these two variables.

3.2.1.8. Weight and height

Weight and height are measured in many large-scale epidemiologic studies because they are key predictors and reflectors of health, and changing health, across many systems including the cardiovascular, gastrointestinal, metabolic, musculoskeletal, and respiratory systems. Weight is an informative indicator of body fatness when height is adequately taken into account. Information on height is important as it refines the interpretation of a number of other key measures such as weight, body fat content, and lung function. The weight measures in Table 30 were obtained using the same device as

for the bioimpedance; the height measures represent an average of two measurements obtained using a portable stadiometer.

Outcome	Mean	Standard Deviation
Body weight	66.82Kg (female) 80.96Kg (male)	13.4
Height	160.0 cm (female) 173.0 cm (male)	6.4

Table 30: Body weight and height.
The table reports the distributions (in the CARTaGENE project) of these two variables.

3.2.1.9. Body Mass Index (BMI)

BMI is derived directly from weight and height, it is the ratio of weight to height. There is now clear evidence that a BMI above ≈25 kg/m^2 increases the risk of developing ischaemic heart disease(**113**), ischaemic stroke (**114**), type 2 diabetes, osteoarthritis and various types of cancer (**115**).

Outcome	Mean	Standard Deviation
BMI	26.12 Kg/m^2 (female) 27.05 Kg/m^2 (male)	4.7

Table 31: Body Mass Index.
The table reports the distributions (in the CARTaGENE project) of BMI.

3.2.1.10. Waist-Hip circumference

Excessive presence of fat in the intra-abdominal cavity may be harmful. Intra-abdominal fat mass can be inferred reasonably well from waist circumference (WC). It has been reported in some clinical studies that, within each sex, waist circumference is highly correlated with intra-abdominal fat mass estimated by ultrasonography and MRI (**116, 117**). The ratio of waist circumference to hip circumference (WC/HP) has been shown to be a good clinical tool for assessing the risk of cardiovascular diseases (CVD)(**118**). This ratio has also been found to be more relevant and closely related to

CVD risk factors than BMI(**119**). WC/HP is positively associated with both arterial stiffness and early atherosclerosis markers such as common carotid arteries-intima-media thickness (**120**). WC and WC/HC are better indicators of risk of central adiposity in postmenopausal women(**121**). WC is also a good predictor of pulmonary function(**122**). Both waist and hip circumference were measured as an average of two measurements carried out using a circumference measuring tape.

Outcome	Mean	Standard Deviation
Waist circumference(WC)	85.2 cm (female) 96.3 cm (male)	11.9
Hip circumference(HC)	103.3 cm (female) 101.6 cm (male)	10.3
WC/HC ratio	0.83 cm (female) 0.95 cm (male)	0.073

Table 32: Waist and hip circumferences and the ratio of the two.
The table reports the distributions (in the CARTaGENE project) of these variables.

3.2.1.11. Cognitive function

It is well recognised that only a limited number of population-based biobanks concentrate rigorously on assessing cognitive function and psychosocial determinants. In the UK, for example, this has led to joint action by the Medical Research Council (MRC), Wellcome Trust and the Economic and Social Research Council (ESRC) in an attempt to enhance this element of the underlying science. It is important, not only because many of the traits that fall in these domains are of direct interest as disease-related phenotypes in their own right (*e.g.* depression, anxiety, cognitive function), but also because they reflect key intermediates in causal pathways that might lead to novel insights into disease aetiology or to the development of enhanced treatment mechanisms.

The CARTaGENE project used widely spread tests and validated paradigms to measure cognitive function. These tests are self- administered using a touch screen computer

platform. Because the parameters to be assessed under this domain are all algorithmic (generally linear) scales, their means and standard deviations will be sensitive to their precise distribution in the Quebec population and yet because they were not included in the CARTaGENE optimization phase, these distributions are unknown. Consequently, rather than estimating the minimum absolute effect sizes that will be detectable for each scale individually, a generic table (Table 40), indicating the fraction of a standard deviation that will be detectable for each scenario, was constructed.

3.2.1.12. Anxiety and depression

Clinical associations between psychiatric illness and chronic medical conditions are supported by a substantial literature. Most research focused on depression found that depression can adversely affect selfcare and increase the risk of incident medical illness, complications and mortality (**123, 124**). The burden of disease attributable to depression is alarming: depression is becoming the major source of disability, second only to cardiovascular diseases; it is the main source of disability in the Canadian workplace (**125**). Depression has repeatedly been associated with morbidity and mortality from several diseases. There is for instance growing evidence indicating that depression is an important primary and secondary risk factor for coronary heart disease (CHD)(**126-128**). The prevalence of depression is three times greater in people with type 2 diabetes than in the general population(**129**).

For anxiety and depression too, a generic table (Table 40), indicating the fraction of a standard deviation that will be detectable for each scenario, was constructed.

3.2.2. SCIENTIFIC RATIONALES AND STATISTICAL DISTRIBUTIONS OF THE BIOCHEMICAL AND HAEMATOLOGICAL VARIABLES ANALYSED AS OUTCOMES

An analysis of biochemical and haematological parameters on fresh blood is useful because: (1) it provides a series of valuable quantitative traits which are meaningful in their own right as complex traits that are worthy of aetiological study; (2) it provides a series of quantitative traits that reflect intermediate traits that lie on the causal pathways leading to a number of complex binary traits that are of scientific interest; (3) it includes a number of "health screening" parameters that are of interest to potential recruits and therefore provide a tangible "return" for agreeing to participate.

Although most of the parameters returned by an automated biochemical or haematological analysis are of direct value clinically (*e.g.* white blood cell count) many have little or no obvious role in the epidemiological setting. Even though these parameters were analysed, there is no attempt to justify them here from a scientific perspective.

3.2.2.1. Insulin, glucose and glycosylated haemoglobin (HbA1C)

These measures reflect activity in causal pathways involving the metabolic syndrome and diabetes which both have a substantial importance to the public health. All the non-diabetic subjects with the morning appointments were asked to fast. This is because when fasting, the plasma glucose levels increases in the body; in subjects without diabetes, insulin is produce to re-balance the level of plasma glucose whilst in subjects with diabetes the level of plasma glucose remains high either because not enough insulin is produced or because the insulin is not used effectively. Consequently when

glucose levels are tested, subjects with diabetes have much higher glucose levels than those without diabetes.

Outcome	Mean	Standard Deviation
Insulin (in non-diabetics)	107.2 uU/ml	107.2
Glucose (in non-diabetics)	4.84 mmol/L	0.925
HbA1C (in non-diabetics)	5.54 %	0.494

Table 33: Insulin, Glucose and HbA1C.
The table reports the distributions (in the CARTaGENE project) of these variables.

3.2.2.2. Cholesterol components and triglycerides

These measures are linked positively or negatively with cardiovascular disease and they also reflect causal pathways associated with the metabolic syndrome.

Outcome	Mean	Standard Deviation
Total cholesterol	5.27 mmol/L	0.986
LDL cholesterol	2.85 mmol/L	0.844
HDL cholesterol	1.56 mmol/L	0.487
LDL:HDL ratio	1.84	0.832
Triglycerides	1.93	1.14

Table 34: Cholesterol components included in the analysis and triglycerides.
The table reports the distributions (in the CARTaGENE project) of these variables.

3.2.2.3. Uric acid

Hyperuricaemia is associated with renal disease, hypertension and a number of haematological conditions. It is the primary pathophysiology underlying gout. Hyperuricaemia is highly prevalent among individuals suffering from metabolic syndrome (**130, 131**)and some studies have found an association between plasma uric acid and the incidence of type 2 diabetes (**106, 107**).

Outcome	Mean	Standard Deviation
Uric acid	296.0 umol/L	74.98

Table 35: Uric acid.
The table reports the distributions (in the CARTaGENE project) of the variable.

3.2.2.4. Thyroid hormones

Hypothyroidism and hyperthyroidism are amongst the most frequent endocrine conditions. To understand both pituitary and thyroid dysfunction and properly interpret thyroid function, it is necessary to measure thyroid stimulating hormone (TSH) and free thyroxine (free T4).

Outcome	Mean	Standard Deviation
Free T4	15.06 pmol/L	1.95
TSH	2.17 mIU/L	1.99

Table 36: Thyroid hormones.
The table reports the distributions (in the CARTaGENE project) of these variables.

3.2.2.5. Creatinine

Creatinine is a marker of renal function and provides a convenient and easily derived reflection of the glomerular filtration rate. The reciprocal of the creatinine concentration often declines in a relatively linear fashion and longitudinal monitoring of creatinine therefore provides a particularly straightforward way to investigate determinants of the rate of decline of renal function over time (for example in diabetes). Creatinine was not measured in the optimization phase of CARTaGENE and so an approximate mean and standard deviation were inferred from the literature (**132**).

Outcome	Mean	Standard Deviation
Creatinine	97.2 umol /L	44.2

Table 37: Creatinine.
The table reports the distributions of the variable inferred from the literature.

3.2.2.6. Haemoglobin and mean red blood cell volume

Because many estimates generated by standard haematology screen are important in clinical practice but of limited value in the epidemiological setting, the study focuses on just two measures which are commonly abnormal because they are associated with

anaemia (and its particular causes) and are determined by a number of key factors including nutritional status, chronic ill health and cancer.

Outcome	Mean	Standard Deviation
Haemoglobin	143.7 g/L	13.5
Mean cell volume	91.1 fL	5.06

Table 38: Haemoglobin and red blood cell volume.
The table reports the distributions (in the CARTaGENE project) of these variables.

3.2.3. THE BIOMEDICAL SCENARIOS INVESTIGATED

The power profiles of CPT were analysed for the six scenarios summarised in Table 39. For each outcome and under each scenario, an iterative approach was used in ESPRESSO-forte to determine the minimum estimated effect that can be detected with a power of 80%± 2% using the probable final samples sizes of CPT (110000 and 180000). The iterative approach consisted of looping through a range of effect sizes until reaching the smallest effect that ensures a power of 80%; these minimum effects were referred to as *minimum detectable effect sizes* (MDES). Section 3.3.1 explains how an MDES should be interpreted.

Scenario	Minor Allele Frequency (MAF)	Prevalence of 'at risk ' environmental determinant	Mathematical model
1 - Common determinants	0.30	0.50	Main effects only
2 - Moderately common determinants	0.10	0.20	Main effects only
3 - Uncommon determinants	0.05	0.10	Main effects only
4 - Common determinants	0.30	0.50	Main effects + interaction
5 - Moderately common determinants	0.10	0.20	Main effects + interaction
6 - Uncommon determinants	0.05	0.10	Main effects + interaction

Table 39: The six scenarios that were explored in constructing each power profile.

3.2.4. ANALYTIC ASSUMPTIONS ABOUT THE OUTCOME AND THE GENETIC AND ENVIRONMENTAL DETERMINANTS

For each of the scenarios 1, 2 and 3 in Table 39, the GLM model fitted in ESPRESSO-forte consist of one outcome (the quantitative trait being analysed) and one covariate; the 3 scenarios were analysed twice, once with a SNP as covariate and once with an environmental factor as covariate. The GLM model for these main effect scenarios can be written as follows:

$$g(Y) = \beta_0 + \beta_1 x_1$$

Where Y is the outcome β_0 is the intercept value, β_1 is the effect size of the covariate x_1. The terms in this model are explained in more details under the GLM section at page 52-53.

For each of the scenarios 4, 5 and 6 in Table 39, the GLM model fitted in ESPRESSO-forte consists of one outcome (the quantitative trait being analysed) and two interacting covariates (a SNP and an environmental factor). The GLM model for these interaction scenarios can be written as follows:

$$g(Y) = \beta_0 + \beta_1 G_1 + \beta_2 E_1 + \beta_3 G_1 E_1$$

Where Y is the outcome, β_0 is the intercept value, β_1 is the effect size of the SNP, β_2 is the effect size of the environmental factor and β_3 is the effect size of the interaction term G_1E_1.

The genetic determinants were modelled as SNPs using an additive genetic model, as is now most commonly used (**24**); the genotyping error was taken as being equivalent to the error that arises when the genotype at a locus of interest is inferred from the

genotype of an observed marker (with the same allelic distribution) that is in linkage disequilibrium with the unobserved causal variant at the locus of interest at r^2=0.8. This corresponds to the weakest LD with HapMap 2 markers on the Affymetrix 500K chip (**133**). The environmental determinants were modelled as binary, and measurement error was introduced by assuming an underlying latent variable with a reliability of 0.7. This reflects a moderate level of measurement error corresponding, for example, to blood pressure measurement in the Intersalt Study (**134**). Gene-environment interactions were modelled using product terms again assuming an additive genetic model. Significance tests for genetic main effects and interactions were based on p.value <0.0001 (i.e. assuming vague candidate genes) or p.value $<10^{-7}$ (genome wide association studies), while non-genetic effects were tested at p.value <0.01. Unless otherwise specified, power estimation was based on the standard deviation and on the measurement reliability of the trait being considered as obtained from the analysis of the CARTaGENE optimization phase. When no firm evidence to the contrary was available to determine the likely measurement reliability of the quantitative trait being considered, it was taken to be 0.7.

3.3. RESULTS

3.3.1. INTERPRETATION OF THE MINIMUM DETECTABLE EFFECT SIZE

It is important to understand how an MDES should be interpreted. To illustrate the interpretation let us use Table 48 which provides the power profile for systolic blood pressure (mean of 3 measurements) measured conventionally in a clinic setting.

Conventional (peripheral) blood pressure is measured as the mean of 3 measurements. The device chosen (Colin Prodigy II Vital Signs Monitor OM-2200**)** is an automated

device that uses the oscillometric method for assessing blood pressure. The measurement process is both quick (2-3 min) and simple.

The population distribution of the variable reported in Table 48 (systolic blood pressure) is: mean = 126 mm/Hg and standard deviation = 18.2. In the body of the table, the MDES for the environmental main effect for the moderately common exposure was reported as 0.9323mm/Hg. This scenario (see Table 2) invokes a binary environmental exposure with a prevalence of 0.2 (20%). The reported results therefore imply that if the final sample size of CPT was 110000 participants, if conventional clinic blood pressure was measured using the standard operating protocol (SOP) outlined in the second paragraph above, and if scientific interest focused on the impact of a binary environmental exposure which had realistic characteristics corresponding to those outlined in section 3.2.4, the power calculations would indicate that there was an 80% chance of detecting, at p.value < 0.01, a real effect that corresponded to that environmental determinant increasing (or decreasing) systolic blood pressure (SBP) by 0.9323 mm/Hg. Similarly, if interest focused on a rare SNP in a genome wide association study (GWAS) there would be an 80% chance of detecting the effect of a SNP with a minor allele frequency of 0.05 (5%) at p.value < 10^{-7} (for genome-wide inference) if that SNP really increased or decreased SBP by at least 1.4806 mm/Hg.

Alternatively, if it was known that the SOP that was actually used for measuring SBP across the CPT project produces an SBP distribution with a rather higher standard deviation than the SOP for the CARTaGENE project (*e.g.* 21.1 mm/Hg not 18.2), then the corresponding entries (0.0512 and 0.0813) could be read from Table 40 and then multiplied by 21.1. This would generate estimated MDESs of 0.0512 × 21.1 ≈ 1.08 mm/Hg (instead of 0.93) and 0.0813 × 21.1 = 1.72 mm/Hg (instead of 1.48) implying,

as would be expected, that if SBP is fundamentally more variable, the effect size that can reliably be detected must inevitably be larger. Reassuringly, if the value 18.2 is fed into Table 40 the values 0.93 and 1.48 are obtained, in other words the same answers that were obtained directly from Table 48.

3.3.2. TABULATED POWER PROFILES

Each of the tables below reports the MDES results for one outcome variable. The tables are subdivided into two parts; the top part (A.) summarises the expected power profile if CPT ultimately recruits a total of 110000 participants, while the bottom part (B.) overviews the corresponding profile given 180000 recruits. The variables cognitive function, anxiety and depression were treated as standardized variables (see section 3.3.2.1); the results for these variables are therefore derived from Table 40.

3.3.2.1. Generic standardized variable

If there is a specific reason to believe that the population distribution of a particular quantitative trait is markedly different in Quebec compared to elsewhere in Canada and that the standard deviation that has been reported from CARTaGENE may therefore be misleading for CPT as a whole; the corresponding power profile for that particular trait could be obtained by treating it as a standardized variable(mean = 0, SD = 1) and multiplying the tabulated MDES values in Table 40 by the known standard deviation of the trait.

The same method can be used to obtain the power profile of a trait that was measured using a different approach (SOP) or equipment: the tabulated MDES values in Table 40 are multiplied by the known standard deviation of the trait under the approach that is to be used.

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0291sd	0.0379sd	0.0372sd	0.0902sd	0.1176sd
Moderately common determinants	0.0440sd	0.0573sd	0.0512sd	0.1895sd	0.2469sd
Uncommon determinants	0.0624sd	0.0813sd	0.0745sd	0.3700sd	0.4821sd
B. 180000 recruits					
Common	0.0227sd	0.0296sd	0.0291sd	0.0705sd	0.0919sd
Moderately common determinants	0.0344sd	0.0448sd	0.0400sd	0.1481sd	0.1930sd
Uncommon determinants	0.0488sd	0.0636sd	0.0582sd	0.2892sd	0.3768sd

Table 40: Minimal detectable effect sizes for variables that have mean=0 and SD=1.

3.3.2.2. Aortic systolic blood pressure

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.5323mm/Hg	0.6936mm/Hg	0.6802mm/Hg	1.6513mm/Hg	2.1516mm/Hg
Moderately common determinants	0.8051mm/Hg	1.0491mm/Hg	0.9374mm/Hg	3.4677mm/Hg	4.5184mm/Hg
Uncommon determinants	1.1425mm/Hg	1.4887mm/Hg	1.3635mm/Hg	6.7703mm/Hg	8.8217mm/Hg
B. 180000 recruits					
Common	0.4161mm/Hg	0.5422mm/Hg	0.5317mm/Hg	1.2909mm/Hg	1.6820mm/Hg
Moderately common determinants	0.6294mm/Hg	0.8201mm/Hg	0.7328mm/Hg	2.7108mm/Hg	3.5322mm/Hg
Uncommon determinants	0.8931mm/Hg	1.1638mm/Hg	1.0659mm/Hg	5.2926mm/Hg	6.8962mm/Hg

Table 41: Minimal detectable effect sizes for aortic systolic blood pressure.

3.3.2.3. Aortic diastolic blood pressure

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.3229mm/Hg	0.4207mm/Hg	0.4126mm/Hg	1.0016mm/Hg	1.3051mm/Hg
Moderately common determinants	0.4883mm/Hg	0.6363mm/Hg	0.5686mm/Hg	2.1034mm/Hg	2.7407mm/Hg
Uncommon determinants	0.6930mm/Hg	0.9030mm/Hg	0.8270mm/Hg	4.1066mm/Hg	5.3509mm/Hg
B. 180000 recruits					
Common	0.2524mm/Hg	0.3289mm/Hg	0.3225mm/Hg	0.7830mm/Hg	1.0202mm/Hg
Moderately common determinants	0.3818mm/Hg	0.4974mm/Hg	0.4445mm/Hg	1.6443mm/Hg	2.1425mm/Hg
Uncommon determinants	0.5417mm/Hg	0.7059mm/Hg	0.6465mm/Hg	3.2103mm/Hg	4.1830mm/Hg

Table 42: Minimal detectable effect sizes for aortic diastolic blood pressure.

3.3.2.4. Aortic pulse pressure

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.3025mm/Hg	0.3942mm/Hg	0.3866mm/Hg	0.9384mm/Hg	1.2228mm/Hg
Moderately common determinants	0.4576mm/Hg	0.5962mm/Hg	0.5327mm/Hg	1.9707mm/Hg	2.5678mm/Hg
Uncommon determinants	0.6493mm/Hg	0.8460mm/Hg	0.7749mm/Hg	3.8476mm/Hg	5.0134mm/Hg
B. 180000 recruits					
Common	0.2365mm/Hg	0.3081mm/Hg	0.3022mm/Hg	0.7336mm/Hg	0.9559mm/Hg
Moderately common determinants	0.3577mm/Hg	0.4661mm/Hg	0.4164mm/Hg	1.5406mm/Hg	2.0074mm/Hg
Uncommon determinants	0.5076mm/Hg	0.6614mm/Hg	0.6057mm/Hg	3.0078mm/Hg	3.9192mm/Hg

Table 43: Minimal detectable effect sizes for aortic pulse pressure.

3.3.2.5. Aortic augmentation index

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.3316mm/Hg	0.4321mm/Hg	0.4237mm/Hg	1.0287mm/Hg	1.3404mm/Hg
Moderately common determinants	0.5015mm/Hg	0.6535mm/Hg	0.5839mm/Hg	2.1602mm/Hg	2.8148mm/Hg
Uncommon determinants	0.7117mm/Hg	0.9274mm/Hg	0.8494mm/Hg	4.2176mm/Hg	5.4955mm/Hg
B. 180000 recruits					
Common	0.2592mm/Hg	0.3378mm/Hg	0.3312mm/Hg	0.8041mm/Hg	1.0478mm/Hg
Moderately common determinants	0.3921mm/Hg	0.5109mm/Hg	0.4565mm/Hg	1.6887mm/Hg	2.2004mm/Hg
Uncommon determinants	0.5564mm/Hg	0.7250mm/Hg	0.6640mm/Hg	3.2970mm/Hg	4.2960mm/Hg

Table 44: Minimal detectable effect sizes for aortic augmentation index.

3.3.2.6. RR-Interval

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	3.1124msecs	4.0555msecs	3.9771msecs	9.6551msecs	12.5806msecs
Moderately common determinants	4.7075msecs	6.1339msecs	5.4809msecs	20.2757msecs	26.4192msecs
Uncommon determinants	6.6803msecs	8.7044msecs	7.9722msecs	39.5858msecs	51.5803msecs
B. 180000 recruits					
Common	2.4331msecs	3.1703msecs	3.1090msecs	7.5477msecs	9.8347msecs
Moderately common determinants	3.6800msecs	4.7951msecs	4.2846msecs	15.8502msecs	20.6528msecs
Uncommon determinants	5.2222msecs	6.8045msecs	6.2322msecs	30.9457msecs	40.3222msecs

Table 45: Minimal detectable effect sizes for RR-Interval.

3.3.2.7. QS-Interval

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	1.0181msecs	1.3266msecs	1.3009msecs	3.1582msecs	4.1151msecs
Moderately common determinants	1.5398msecs	2.0064msecs	1.7928msecs	6.6322msecs	8.6418msecs
Uncommon determinants	2.1851msecs	2.8472msecs	2.6077msecs	12.9486msecs	16.8721msecs
B. 180000 recruits					
Common	0.7959msecs	1.0370msecs	1.0170msecs	2.4689msecs	3.2170msecs
Moderately common determinants	1.2037msecs	1.5685msecs	1.4015msecs	5.1846msecs	6.7556msecs
Uncommon determinants	1.7082msecs	2.2258msecs	2.0386msecs	10.1224msecs	13.1895msecs

Table 46: Minimal detectable effect sizes for QS-Interval.

3.3.2.8. QTc-Interval

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	1.7744msecs	2.3120msecs	2.2673msecs	5.5043msecs	7.1721msecs
Moderately common determinants	2.6837msecs	3.4969msecs	3.1246msecs	11.5590msecs	15.0614msecs
Uncommon determinants	3.8084msecs	4.9623msecs	4.5449msecs	22.5676msecs	29.4056msecs
B. 180000 recruits					
Common	1.3871msecs	1.8074msecs	1.7724msecs	4.3029msecs	5.6067msecs
Moderately common determinants	2.0980msecs	2.7336msecs	2.4426msecs	9.0361msecs	11.7740msecs
Uncommon determinants	2.9771msecs	3.8792msecs	3.5529msecs	17.6419msecs	22.9874msecs

Table 47: Minimal detectable effect sizes for QTc-Interval.

3.3.2.9. Systolic blood pressure

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.5294mm/Hg	0.6898mm/Hg	0.6765mm/Hg	1.6423mm/Hg	2.1399mm/Hg
Moderately common determinants	0.8007mm/Hg	1.0433mm/Hg	0.9323mm/Hg	3.4488mm/Hg	4.4937mm/Hg
Uncommon determinants	1.1363mm/Hg	1.4806mm/Hg	1.3560mm/Hg	6.7333mm/Hg	8.7735mm/Hg
B. 180000 recruits					
Common	0.4138mm/Hg	0.5392mm/Hg	0.5288mm/Hg	1.2838mm/Hg	1.6728mm/Hg
Moderately common determinants	0.6259mm/Hg	0.8156mm/Hg	0.7288mm/Hg	2.6960mm/Hg	3.5129mm/Hg
Uncommon determinants	0.8883mm/Hg	1.1574mm/Hg	1.0601mm/Hg	5.2637mm/Hg	6.8585mm/Hg

Table 48: Minimal detectable effect sizes for systolic blood pressure.

3.3.2.10. Diastolic blood pressure

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.3112mm/Hg	0.4055mm/Hg	0.3977mm/Hg	0.9655mm/Hg	1.2581mm/Hg
Moderately common determinants	0.4708mm/Hg	0.6134mm/Hg	0.5481mm/Hg	2.0276mm/Hg	2.6419mm/Hg
Uncommon determinants	0.6680mm/Hg	0.8704mm/Hg	0.7972mm/Hg	3.9586mm/Hg	5.1580mm/Hg
B. 180000 recruits					
Common	0.2433mm/Hg	0.3170mm/Hg	0.3109mm/Hg	0.7548mm/Hg	0.9835mm/Hg
Moderately common determinants	0.3680mm/Hg	0.4795mm/Hg	0.4285mm/Hg	1.5850mm/Hg	2.0653mm/Hg
Uncommon determinants	0.5222mm/Hg	0.6805mm/Hg	0.6232mm/Hg	3.0946mm/Hg	4.0322mm/Hg

Table 49: Minimal detectable effect sizes for diastolic blood pressure.

3.3.2.11. Forced expiratory volume in 1second (FEV1)

(FEV_1 was reported as the proportion of predicted values)

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0059	0.0077	0.0075	0.0182	0.0238
Moderately common determinants	0.0089	0.0116	0.0103	0.0383	0.0499
Uncommon determinants	0.0126	0.0164	0.0151	0.0747	0.0974
B. 180000 recruits					
Common	0.0046	0.0060	0.0059	0.0142	0.0186
Moderately common determinants	0.0069	0.0091	0.0081	0.0299	0.0390
Uncommon determinants	0.0099	0.0128	0.0118	0.0584	0.0761

Table 50: Minimal detectable effect sizes for FEV_1.
As indicated at the top of the table FEV1 was measured as the proportion of predicted.

3.3.2.12. Forced vital capacity (FVC)

(FVC was reported as the proportion of predicted values)

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0059	0.0077	0.0075	0.0183	0.0239
Moderately common determinants	0.0089	0.0116	0.0104	0.0385	0.0501
Uncommon determinants	0.0127	0.0165	0.0151	0.0751	0.0979
B. 180000 recruits					
Common	0.0046	0.0060	0.0059	0.0143	0.0187
Moderately common determinants	0.0070	0.0091	0.0081	0.0301	0.0392
Uncommon determinants	0.0099	0.0129	0.0118	0.0587	0.0765

Table 51: Minimal detectable effect sizes for FVC
As indicated at the top of the table FEV1 was measured as the proportion of predicted.

3.3.2.13. FEV1/FVC ratio

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0033	0.0043	0.0042	0.0102	0.0133
Moderately common determinants	0.0050	0.0065	0.0058	0.0214	0.0279
Uncommon determinants	0.0071	0.0092	0.0084	0.0418	0.0545
B. 180000 recruits					
Common	0.0026	0.0033	0.0033	0.0080	0.0104
Moderately common determinants	0.0039	0.0051	0.0045	0.0167	0.0218
Uncommon determinants	0.0055	0.0072	0.0066	0.0327	0.0426

Table 52: Minimal detectable effect sizes for FEV_1/FVC ratio.

3.3.2.14. Os calcis BUA index

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0152dB/MHz	0.0198dB/MHz	0.0194dB/MHz	0.0471dB/MHz	0.0614dB/MHz
Moderately common determinants	0.0230dB/MHz	0.0299dB/MHz	0.0267dB/MHz	0.0989dB/MHz	0.1289dB/MHz
Uncommon determinants	0.0326dB/MHz	0.0425dB/MHz	0.0389dB/MHz	0.1931dB/MHz	0.2516dB/MHz
B. 180000 recruits					
Common	0.0119dB/MHz	0.0155dB/MHz	0.0152dB/MHz	0.0368dB/MHz	0.0480dB/MHz
Moderately common determinants	0.0180dB/MHz	0.0234dB/MHz	0.0209dB/MHz	0.0773dB/MHz	0.1008dB/MHz
Uncommon determinants	0.0255dB/MHz	0.0332dB/MHz	0.0304dB/MHz	0.1510dB/MHz	0.1967dB/MHz

Table 53: Minimal detectable effect sizes for os calcis BUA index.

3.3.2.15. Os calcis bone density t-score

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0320	0.0417	0.0409	0.0993	0.1293
Moderately common determinants	0.0484	0.0631	0.0563	0.2084	0.2716
Uncommon determinants	0.0687	0.0895	0.0820	0.4070	0.5303
B. 180000 recruits					
Common	0.0250	0.0326	0.0320	0.0776	0.1011
Moderately common determinants	0.0378	0.0493	0.0440	0.1629	0.2123
Uncommon determinants	0.0537	0.0700	0.0641	0.3181	0.4145

Table 54: Minimal detectable effect sizes for os calcis bone density t-score.

3.3.2.16. Os calcis percent normal bone density

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.5108	0.6655	0.6527	1.5845	2.0646
Moderately common determinants	0.7726	1.0066	0.8995	3.3275	4.3357
Uncommon determinants	1.0963	1.4285	1.3083	6.4965	8.4650
B. 180000 recruits					
Common	0.3993	0.5203	0.5102	1.2387	1.6140
Moderately common determinants	0.6039	0.7869	0.7032	2.6012	3.3894
Uncommon determinants	0.8570	1.1167	1.0228	5.0786	6.6174

Table 55: Minimal detectable effect sizes for os calcis percent normal bone density.

3.3.2.17. Grip strength left hand

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.3520Kg	0.4586Kg	0.4497Kg	1.0918Kg	1.4227Kg
Moderately common determinants	0.5323Kg	0.6936Kg	0.6198Kg	2.2929Kg	2.9876Kg
Uncommon determinants	0.7554Kg	0.9843Kg	0.9015Kg	4.4765Kg	5.8329Kg
B. 180000 recruits					
Common	0.2751Kg	0.3585Kg	0.3516Kg	0.8535Kg	1.1121Kg
Moderately common determinants	0.4162Kg	0.5422Kg	0.4845Kg	1.7924Kg	2.3355Kg
Uncommon determinants	0.5905Kg	0.7695Kg	0.7048Kg	3.4995Kg	4.5598Kg

Table 56: Minimal detectable effect sizes for left hand grip strength.

3.3.2.18. Grip strength right hand

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.3607Kg	0.4700Kg	0.4609Kg	1.1189Kg	1.4579Kg
Moderately common determinants	0.5455Kg	0.7108Kg	0.6352Kg	2.3497Kg	3.0617Kg
Uncommon determinants	0.7742Kg	1.0087Kg	0.9239Kg	4.5875Kg	5.9775Kg
B. 180000 recruits					
Common	0.2820Kg	0.3674Kg	0.3603Kg	0.8747Kg	1.1397Kg
Moderately common determinants	0.4265Kg	0.5557Kg	0.4965Kg	1.8368Kg	2.3934Kg
Uncommon determinants	0.6052Kg	0.7886Kg	0.7222Kg	3.5862Kg	4.6729Kg

Table 57: Minimal detectable effect sizes for right hand grip strength.

3.3.2.19. Fat mass by bioimpedance

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.2909Kg	0.3790Kg	0.3717Kg	0.9023Kg	1.1758Kg
Moderately common determinants	0.4400Kg	0.5733Kg	0.5122Kg	1.8949Kg	2.4691Kg
Uncommon determinants	0.6243Kg	0.8135Kg	0.7451Kg	3.6996Kg	4.8206Kg
B. 180000 recruits					
Common	0.2274Kg	0.2963Kg	0.2906Kg	0.7054Kg	0.9191Kg
Moderately common determinants	0.3439Kg	0.4481Kg	0.4004Kg	1.4813Kg	1.9302Kg
Uncommon determinants	0.4881Kg	0.6359Kg	0.5824Kg	2.8921Kg	3.7684Kg

Table 58: Minimal detectable effect sizes for fat mass by bioimpedence.

3.3.2.20. Lean mass by bioimpedance

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.2950Kg	0.3843Kg	0.3769Kg	0.9150Kg	1.1922Kg
Moderately common determinants	0.4461Kg	0.5813Kg	0.5194Kg	1.9215Kg	2.5037Kg
Uncommon determinants	0.6331Kg	0.8249Kg	0.7555Kg	3.7514Kg	4.8881Kg
B. 180000 recruits					
Common	0.2306Kg	0.3004Kg	0.2946Kg	0.7153Kg	0.9320Kg
Moderately common determinants	0.3487Kg	0.4544Kg	0.4060Kg	1.5021Kg	1.9572Kg
Uncommon determinants	0.4949Kg	0.6448Kg	0.5906Kg	2.9326Kg	3.8212Kg

Table 59: Minimal detectable effect sizes for lean mass bioimpedence

3.3.2.21. Percent body fat by bioimpedance

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.2984%	0.3889%	0.3814%	0.9258%	1.2063%
Moderately common determinants	0.4514%	0.5882%	0.5256%	1.9442%	2.5333%
Uncommon determinants	0.6406%	0.8346%	0.7644%	3.7958%	4.9459%
B. 180000 recruits					
Common	0.2333%	0.3040%	0.2981%	0.7237%	0.9430%
Moderately common determinants	0.3529%	0.4598%	0.4108%	1.5198%	1.9804%
Uncommon determinants	0.5007%	0.6525%	0.5976%	2.9673%	3.8664%

Table 60: Minimal detectable effect sizes for percent body fat by bioimpedence

3.3.2.22. Body weight

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.3898Kg	0.5079Kg	0.4981Kg	1.2091Kg	1.5755Kg
Moderately common determinants	0.5895Kg	0.7682Kg	0.6864Kg	2.5392Kg	3.3086Kg
Uncommon determinants	0.8366Kg	1.0901Kg	0.9984Kg	4.9575Kg	6.4596Kg
B. 180000 recruits					
Common	0.3047Kg	0.3970Kg	0.3894Kg	0.9452Kg	1.2316Kg
Moderately common determinants	0.4609Kg	0.6005Kg	0.5366Kg	1.9850Kg	2.5864Kg
Uncommon determinants	0.6540Kg	0.8522Kg	0.7805Kg	3.8754Kg	5.0497Kg

Table 61: Minimal detectable effect sizes for body weight

3.3.2.23. Height

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.19cm	0.24cm	0.24cm	0.58cm	0.75cm
Moderately common determinants	0.28cm	0.37cm	0.33cm	1.21cm	1.58cm
Uncommon determinants	0.40cm	0.52cm	0.48cm	2.37cm	3.09cm
B. 180000 recruits					
Common	0.15cm	0.19cm	0.19cm	0.45cm	0.59cm
Moderately common determinants	0.22cm	0.29cm	0.26cm	0.95cm	1.24cm
Uncommon determinants	0.31cm	0.41cm	0.37cm	1.85cm	2.41cm

Table 62: Minimal detectable effect sizes for height.

3.3.2.24. Body Mass Index

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.1367Kg/m2	0.1781Kg/m2	0.1747Kg/m2	0.4241Kg/m2	0.5526Kg/m2
Moderately common determinants	0.2068Kg/m2	0.2694Kg/m2	0.2407Kg/m2	0.8906Kg/m2	1.1605Kg/m2
Uncommon determinants	0.2934Kg/m2	0.3823Kg/m2	0.3502Kg/m2	1.7388Kg/m2	2.2657Kg/m2
B. 180000 recruits					
Common	0.1069Kg/m2	0.1393Kg/m2	0.1366Kg/m2	0.3315Kg/m2	0.4320Kg/m2
Moderately common determinants	0.1616Kg/m2	0.2106Kg/m2	0.1882Kg/m2	0.6962Kg/m2	0.9072Kg/m2
Uncommon determinants	0.2294Kg/m2	0.2989Kg/m2	0.2737Kg/m2	1.3593Kg/m2	1.7712Kg/m2

Table 63: Minimal detectable effect sizes for BMI

3.3.2.25. Waist circumference (WC)

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.3461cm	0.4510cm	0.4423cm	1.0738cm	1.3991cm
Moderately common determinants	0.5235cm	0.6822cm	0.6096cm	2.2550cm	2.9382cm
Uncommon determinants	0.7429cm	0.9681cm	0.8866cm	4.4025cm	5.7365cm
B. 180000 recruits					
Common	0.2706cm	0.3526cm	0.3458cm	0.8394cm	1.0938cm
Moderately common determinants	0.4093cm	0.5333cm	0.4765cm	1.7628cm	2.2969cm
Uncommon determinants	0.5808cm	0.7568cm	0.6931cm	3.4416cm	4.4844cm

Table 64: Minimal detectable effect sizes for waist circumference (WC).

3.3.2.26. Hip circumference (HC)

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.2996cm	0.3904cm	0.3828cm	0.9294cm	1.2110cm
Moderately common determinants	0.4532cm	0.5905cm	0.5276cm	1.9518cm	2.5432cm
Uncommon determinants	0.6431cm	0.8379cm	0.7674cm	3.8106cm	4.9652cm
B. 180000 recruits					
Common	0.2342cm	0.3052cm	0.2993cm	0.7266cm	0.9467cm
Moderately common determinants	0.3542cm	0.4616cm	0.4124cm	1.5258cm	1.9881cm
Uncommon determinants	0.5027cm	0.6550cm	0.5999cm	2.9789cm	3.8815cm

Table 65: Minimal detectable effect sizes for hip circumference (HC).

3.3.2.27. WC/HC ratio

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0021	0.0028	0.0027	0.0066	0.0086
Moderately common determinants	0.0032	0.0042	0.0037	0.0138	0.0180
Uncommon determinants	0.0046	0.0059	0.0054	0.0270	0.0352
B. 180000 recruits					
Common	0.0017	0.0022	0.0021	0.0051	0.0067
Moderately common determinants	0.0025	0.0033	0.0029	0.0108	0.0141
Uncommon determinants	0.0036	0.0046	0.0043	0.0211	0.0275

Table 66: Minimal detectable effect sizes for WC/HC ratio.

3.3.2.28. Insulin (in non-diabetics)

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	3.1182uU/ml	4.0630uU/ml	3.9845uU/ml	9.6731uU/ml	12.6041uU/ml
Moderately common determinants	4.7163uU/ml	6.1453uU/ml	5.4911uU/ml	20.3136uU/ml	26.4686uU/ml
Uncommon determinants	6.6928uU/ml	8.7207uU/ml	7.9871uU/ml	39.6598uU/ml	51.6767uU/ml
B. 180000 recruits					
Common	2.4376uU/ml	3.1762uU/ml	3.1148uU/ml	7.5618uU/ml	9.8531uU/ml
Moderately common determinants	3.6869uU/ml	4.8040uU/ml	4.2926uU/ml	15.8798uU/ml	20.6914uU/ml
Uncommon determinants	5.2320uU/ml	6.8173uU/ml	6.2438uU/ml	31.0035uU/ml	40.3976uU/ml

Table 67: Minimal detectable effect sizes for insulin (in non-diabetics).

3.3.2.29. Glucose (in non-diabetics)

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0269mmol/l	0.0351mmol/l	0.0344mmol/l	0.0835mmol/l	0.1088mmol/l
Moderately common determinants	0.0407mmol/l	0.0530mmol/l	0.0474mmol/l	0.1753mmol/l	0.2284mmol/l
Uncommon determinants	0.0578mmol/l	0.0752mmol/l	0.0689mmol/l	0.3422mmol/l	0.4459mmol/l
B. 180000 recruits					
Common	0.0210mmol/l	0.0274mmol/l	0.0269mmol/l	0.0652mmol/l	0.0850mmol/l
Moderately common determinants	0.0318mmol/l	0.0415mmol/l	0.0370mmol/l	0.1370mmol/l	0.1785mmol/l
Uncommon determinants	0.0451mmol/l	0.0588mmol/l	0.0539mmol/l	0.2675mmol/l	0.3486mmol/l

Table 68: Minimal detectable effect sizes for glucose (in non-diabetics).

3.3.2.30. HbA1C (in non-diabetics)

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0144%	0.0187%	0.0184%	0.0446%	0.0581%
Moderately common determinants	0.0217%	0.0283%	0.0253%	0.0936%	0.1220%
Uncommon determinants	0.0308%	0.0402%	0.0368%	0.1828%	0.2381%
B. 180000 recruits					
Common	0.0112%	0.0146%	0.0144%	0.0348%	0.0454%
Moderately common determinants	0.0170%	0.0221%	0.0198%	0.0732%	0.0954%
Uncommon determinants	0.0241%	0.0314%	0.0288%	0.1429%	0.1862%

Table 69: Minimal detectable effect sizes for HbA1C (in non-diabetics).

3.3.2.31. Cholesterol

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0287mmol/l	0.0374mmol/l	0.0366mmol/l	0.0890mmol/l	0.1159mmol/l
Moderately common determinants	0.0434mmol/l	0.0565mmol/l	0.0505mmol/l	0.1868mmol/l	0.2435mmol/l
Uncommon determinants	0.0616mmol/l	0.0802mmol/l	0.0735mmol/l	0.3648mmol/l	0.4753mmol/l
B. 180000 recruits					
Common	0.0224mmol/l	0.0292mmol/l	0.0286mmol/l	0.0696mmol/l	0.0906mmol/l
Moderately common determinants	0.0339mmol/l	0.0442mmol/l	0.0395mmol/l	0.1461mmol/l	0.1903mmol/l
Uncommon determinants	0.0481mmol/l	0.0627mmol/l	0.0574mmol/l	0.2852mmol/l	0.3716mmol/l

Table 70: Minimal detectable effect sizes for cholesterol.

3.3.2.32. LDL cholesterol

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0246mmol/l	0.0320mmol/l	0.0314mmol/l	0.0762mmol/l	0.0992mmol/l
Moderately common determinants	0.0371mmol/l	0.0484mmol/l	0.0432mmol/l	0.1599mmol/l	0.2084mmol/l
Uncommon determinants	0.0527mmol/l	0.0687mmol/l	0.0629mmol/l	0.3122mmol/l	0.4069mmol/l
B. 180000 recruits					
Common	0.0192mmol/l	0.0250mmol/l	0.0245mmol/l	0.0595mmol/l	0.0776mmol/l
Moderately common determinants	0.0290mmol/l	0.0378mmol/l	0.0338mmol/l	0.1250mmol/l	0.1629mmol/l
Uncommon determinants	0.0412mmol/l	0.0537mmol/l	0.0492mmol/l	0.2441mmol/l	0.3181mmol/l

Table 71: Minimal detectable effect sizes for LDL cholesterol.

3.3.2.33. HDL cholesterol

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0142mmol/l	0.0185mmol/l	0.0181mmol/l	0.0439mmol/l	0.0573mmol/l
Moderately common determinants	0.0214mmol/l	0.0279mmol/l	0.0249mmol/l	0.0923mmol/l	0.1202mmol/l
Uncommon determinants	0.0304mmol/l	0.0396mmol/l	0.0363mmol/l	0.1802mmol/l	0.2348mmol/l
B. 180000 recruits					
Common	0.0111mmol/l	0.0144mmol/l	0.0142mmol/l	0.0344mmol/l	0.0448mmol/l
Moderately common determinants	0.0167mmol/l	0.0218mmol/l	0.0195mmol/l	0.0721mmol/l	0.0940mmol/l
Uncommon determinants	0.0238mmol/l	0.0310mmol/l	0.0284mmol/l	0.1408mmol/l	0.1835mmol/l

Table 72: Minimal detectable effect sizes for HDL cholesterol.

3.3.2.34. LDL/HDL ratio

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0242	0.0315	0.0309	0.0751	0.0978
Moderately common determinants	0.0366	0.0477	0.0426	0.1577	0.2054
Uncommon determinants	0.0519	0.0677	0.0620	0.3078	0.4011
B. 180000 recruits					
Common	0.0189	0.0247	0.0242	0.0587	0.0765
Moderately common determinants	0.0286	0.0373	0.0333	0.1232	0.1606
Uncommon determinants	0.0406	0.0529	0.0485	0.2406	0.3135

Table 73: Minimal detectable effect sizes for LDL/HDL ratio.

3.3.2.35. Triglycerides

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0332mmol/l	0.0432mmol/l	0.0424mmol/l	0.1029mmol/l	0.1340mmol/l
Moderately common determinants	0.0502mmol/l	0.0654mmol/l	0.0584mmol/l	0.2160mmol/l	0.2815mmol/l
Uncommon determinants	0.0712mmol/l	0.0927mmol/l	0.0849mmol/l	0.4218mmol/l	0.5495mmol/l
B. 180000 recruits					
Common	0.0259mmol/l	0.0338mmol/l	0.0331mmol/l	0.0804mmol/l	0.1048mmol/l
Moderately common determinants	0.0392mmol/l	0.0511mmol/l	0.0456mmol/l	0.1689mmol/l	0.2200mmol/l
Uncommon determinants	0.0556mmol/l	0.0725mmol/l	0.0664mmol/l	0.3297mmol/l	0.4296mmol/l

Table 74: Minimal detectable effect sizes for triglycerides.

3.3.2.36. Uric acid

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	2.1810umol/l	2.8419umol/l	2.7869umol/l	6.7658umol/l	8.8158umol/l
Moderately common determinants	3.2988umol/l	4.2983umol/l	3.8407umol/l	14.2081umol/l	18.5132umol/l
Uncommon determinants	4.6812umol/l	6.0996umol/l	5.5865umol/l	27.7397umol/l	36.1448umol/l
B. 180000 recruits					
Common	1.7050umol/l	2.2216umol/l	2.1786umol/l	5.2890umol/l	6.8916umol/l
Moderately common determinants	2.5788umol/l	3.3601umol/l	3.0024umol/l	11.1070umol/l	14.4724umol/l
Uncommon determinants	3.6594umol/l	4.7683umol/l	4.3672umol/l	21.6851umol/l	28.2557umol/l

Table 75: Minimal detectable effect sizes for uric acid.

3.3.2.37. Free Thyroxine (Free T4)

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0567pmol/l	0.0739pmol/l	0.0725pmol/l	0.1760pmol/l	0.2293pmol/l
Moderately common determinants	0.0858pmol/l	0.1118pmol/l	0.0999pmol/l	0.3695pmol/l	0.4815pmol/l
Uncommon determinants	0.1217pmol/l	0.1586pmol/l	0.1453pmol/l	0.7214pmol/l	0.9400pmol/l
B. 180000 recruits					
Common	0.0443pmol/l	0.0578pmol/l	0.0567pmol/l	0.1376pmol/l	0.1792pmol/l
Moderately common determinants	0.0671pmol/l	0.0874pmol/l	0.0781pmol/l	0.2889pmol/l	0.3764pmol/l
Uncommon determinants	0.0952pmol/l	0.1240pmol/l	0.1136pmol/l	0.5640pmol/l	0.7348pmol/l

Table 76: Minimal detectable effect sizes for free thyroxine.

3.3.2.38. Thyroid stimulating hormone (TSH)

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.0579mlU/l	0.0754mlU/l	0.0740mlU/l	0.1796mlU/l	0.2340mlU/l
Moderately common determinants	0.0876mlU/l	0.1141mlU/l	0.1019mlU/l	0.3771mlU/l	0.4913mlU/l
Uncommon determinants	0.1242mlU/l	0.1619mlU/l	0.1483mlU/l	0.7362mlU/l	0.9593mlU/l
B. 180000 recruits					
Common	0.0453mlU/l	0.0590mlU/l	0.0578mlU/l	0.1404mlU/l	0.1829mlU/l
Moderately common determinants	0.0684mlU/l	0.0892mlU/l	0.0797mlU/l	0.2948mlU/l	0.3841mlU/l
Uncommon determinants	0.0971mlU/l	0.1266mlU/l	0.1159mlU/l	0.5755mlU/l	0.7499mlU/l

Table 77: Minimal detectable effect size sfor thyroid stimulating hormone.

3.3.2.39. Creatinine

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	1.2857umol/l	1.6752umol/l	1.6429umol/l	3.9884umol/l	5.1968umol/l
Moderately common determinants	1.9446umol/l	2.5338umol/l	2.2641umol/l	8.3756umol/l	10.9133umol/l
Uncommon determinants	2.7595umol/l	3.5956umol/l	3.2932umol/l	16.3523umol/l	21.3070umol/l
B. 180000 recruits					
Common	1.0051umol/l	1.3096umol/l	1.2843umol/l	3.1178umol/l	4.0626umol/l
Moderately common determinants	1.5202umol/l	1.9808umol/l	1.7699umol/l	6.5475umol/l	8.5314umol/l
Uncommon determinants	2.1572umol/l	2.8108umol/l	2.5744umol/l	12.7832umol/l	16.6565umol/l

Table 78: Minimal detectable effect sizes for creatinine.

3.3.2.40. Haemoglobin

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.3927g/l	0.5117g/l	0.5018g/l	1.2182g/l	1.5873g/l
Moderately common determinants	0.5939g/l	0.7739g/l	0.6915g/l	2.5581g/l	3.3333g/l
Uncommon determinants	0.8428g/l	1.0982g/l	1.0058g/l	4.9945g/l	6.5078g/l
B. 180000 recruits					
Common	0.3070g/l	0.4000g/l	0.3923g/l	0.9523g/l	1.2408g/l
Moderately common determinants	0.4643g/l	0.6050g/l	0.5406g/l	1.9998g/l	2.6057g/l
Uncommon determinants	0.6589g/l	0.8585g/l	0.7863g/l	3.9044g/l	5.0874g/l

Table 79: Minimal detectable effect sizes for haemoglobin.

3.3.2.41. Mean red blood cell volume

	Genetic main effect	Genetic main effect	Environment main effect	G×E interaction	G×E interaction
P value	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**	**0.01**	**10^{-4} (candidate gene study)**	**10^{-7} (GWAS)**
A. 110000 recruits					
Common	0.1472fl	0.1918fl	0.1881fl	0.4566fl	0.5949fl
Moderately common determinants	0.2226fl	0.2901fl	0.2592fl	0.9588fl	1.2494fl
Uncommon determinants	0.3159fl	0.4116fl	0.3770fl	1.8720fl	2.4392fl
B. 180000 recruits					
Common	0.1151fl	0.1499fl	0.1470fl	0.3569fl	0.4651fl
Moderately common determinants	0.1740fl	0.2268fl	0.2026fl	0.7496fl	0.9767fl
Uncommon determinants	0.2470fl	0.3218fl	0.2947fl	1.4634fl	1.9068fl

Table 80: Minimal detectable effect sizes for mean red blood cell volume.

3.4. CONCLUSIONS

It has recently become clear that tens of thousands of subjects are often required to study quantitative disease-related phenotypes, because allelic effect sizes may be as small as one tenth of a standard deviation, or even less (**25, 77, 78**). To date, the majority of small effect sizes that have been found to be associated with quantitative traits pertain to genetic main effects (**25, 77, 78**). However, there is no reason to believe that relevant gene-environment interactions will be any larger. A gene-environment interaction simply represents the difference in the magnitude of a genetic effect between two different population subgroups defined by exposure to an environmental determinant. If the genetic effect in both groups is likely to be very small, it is also likely that the difference (the interaction effect) will be small. But, despite the extensive difficulties, the exploration of gene-environment interactions is fundamental to our

ultimate understanding of the causal architecture of complex diseases. This is because the processes that underpin human evolution are based primarily on selection on the basis of interactions between the genome and the contemporaneous environment to which individuals are exposed. In consequence, it is inevitable that many of the chronic diseases that affect contemporary society will have a primary basis in the way in which our genome interacts with the modern day environment.

Given the considerations outlined in the preceding paragraph, the potential contribution of CPT to study the aetiological architecture of quantitative traits is obvious. Scrutiny of the power profiles of the quantitative variables tabulated individually demonstrates that given a sample size of 110000 or 180000, genetic and environmental main effects associated with any quantitative variables that are collected across the whole CPT project will all be able to be studied with substantial power – effect sizes as small as 1/12th of a standard deviation will reliably be detectable even under the most challenging scenario (uncommon genotype [MAF = 0.05] with testing at p.value$< 10^{-7}$ under a GWAS). But, as would be anticipated (**7, 49**) the power to detect gene-environment interactions is considerably less strong. Given the central relevance of such interactions (see above), it is important to note that a sample size of 180,000 rather than 110000 would markedly enhance the capacity to study gene-environment interactions when the interacting determinants are both other than common. For example, when both determinants are moderately common (MAF = 0.10, prevalence of environmental determinant = 0.2) and testing is at p.value $< 10^{-4}$, a sample size of 110000 will enable detection of an effect equivalent to 0.19 of a standard deviation. A sample size of 180000 will support detection of 0.15 of a standard deviation. Similarly, when both determinants are uncommon (MAF = 0.05, prevalence of environmental determinant = 0.1) the detectable effect sizes will be 0.37 and 0.29 respectively. The

larger sample size will also allow additional scope for data sub-setting. For example, the data from Table 40 indicate that if a separate analysis is required, for example, in women aged > 65, and if this subgroup represents 10% of the overall data set, the minimal detectable effect size for a main effect associated with an uncommon environmental determinant (prevalence = 10%)would be 0.24 (nearly one quarter of a standard deviation) with a sample size of 110000 compared to 0.18 (less than one fifth of a standard deviation) with a sample size of 180000.

These differences are not trivial. It is clear that if the scientific focus in relation to quantitative traits - that are collected across CPT as whole - is directed solely at main effects of genetic and environmental determinants across the data set as a whole, CPT will provide a research platform that will be highly competitive at an international level and will provide more than adequate power given a sample size of 110000. But, if there is any potential interest in gene-environment interactions (which there logically should be) and any interest in being able to study key subsets of the overall data set (for example, men alone) then the power to detect effect sizes of plausible magnitude will be limiting, and there would then be no doubt that a larger sample size such as 180,000 would potentially be more useful.

This analysis shows the importance of achieving large sample sizes for the investigation of the genetic and environmental/life style factors underlying complex traits. It is however equally important to ensure good quality data by attempting to limit errors and control random variation during data collection and processing. Chapter 4 investigates the pattern of measurement errors that can arise from delays in the definitive processing and storage of biosamples as new participants are enrolled into a growing biobank and

explores their potential impact on the statistical power of association studies based on biosample measures generated by that study.

CHAPTER 4

4. UNDERSTANDING THE EFFECTS OF VARIATION IN SAMPLE COLLECTION AND HANDLING ON THE POWER OF GENETIC ASSOCIATION STUDIES

The analysis I carried out in this chapter follows a question put through to Professor Paul Burton by Dr Tim Peakman, Executive Director of UK Biobank. Tim wanted to know the impact of pre-analytical variations, introduced through samples processing and storage, on studies that uses the biobank data. It was not possible to answer this question with the initial version of ESPRESSO developed by Professor Paul Burton because that version did not allow for the analysis of quantitative environmental variables. After I completed the development of ESPRESSO-forte Paul asked me to investigate Tim's question. A paper, that is now awaiting submission, was subsequently written about this analysis with me as first author.

4.1. INTRODUCTION

A biobank may be defined as *"an organized collection of human biological material(e.g., blood, urine, or extracted DNA) and associated information stored for one or more research purposes"*(**72, 135**). Most contemporary biobanks are large by design because the aetiological determinants (genes, environment and interactions) of complex diseases are typically weak (*e.g.* relative risks between 1.1 and 1.3) and their resolution therefore demands many subjects (**7**) with data that are both accurate and

precise. Crucially, errors introduced through poor assessment or physical measurement or because of inconsistent/or inappropriate standard operating procedures for collecting, processing, storing or analysing biosamples can seriously impair data quality. This can dramatically reduce the statistical power of a study, particularly if one is studying gene-environment interactions (**7, 49**). Given the vast cost and effort that is needed to establish and maintain a contemporary biobank, even a small loss of power can impact substantially on the balance of costs and benefits of developing adequately powered study resources. The quality and future utility of biological samples can be affected by factors arising during the collection, transport, processing and storage of biosamples (**136**). It is therefore crucial to use carefully selected and validated protocols that minimise any changes in the quantity or nature of the constituents (biological analytes) of each biosample and allow for the further re-use of the samples (**137-140**). It is for this reason that certain procedures (**138, 141**) that ensure minimal pre-analytical variability between samples are published - as best practice guidelines for biological resource centres - by organisations involved in the conceptualization, design and conduct of samples collection, processing and analysis. This includes the Public Population Project in Genomics (P^3G) (**138, 140, 141**) and the International Society for Biological and Environmental Repositories (ISBER) (**138, 140-144**).

A critical issue is the impact of any delay between sample collection and the processing step that definitively stabilizes that sample (typically, the needle-to-freezer time). This is because some biological analytes are not stable under certain conditions and their concentration changes over time. For example, the concentration of aspartame transaminase (AST), a biochemical analyte present in red blood cells, changes over one day (24 hours) by 15.2% and 1.5% under respectively21°C and 4°C (**145**). If there is any tendency for a particular analyte to degrade (or accumulate) over time ahead of

stabilization, any delay in definitive processing will introduce measurement error. If the rate of degradation is very similar in all samples then a standard operating procedure that requires a fixed (though non-zero) delay till processing (*e.g.* 24 hours) will ensure that all samples to be analysed will be similarly affected and biostatistical/epidemiological analyses may therefore be unbiased. But, if case and control samples are processed under protocols that incur a different processing time – or distribution of processing times - serious systematic bias may arise. Additional problems arise if the rate of degradation varies markedly from subject to subject. Then *any* delay in processing will introduce random error that will usually reduce statistical power; even if every sample is subject to the *same* delay. Furthermore, the magnitude of the consequent bias will become steadily more serious as the duration increases.

Unfortunately, although the bioscience might generally favour a common protocol with a constant minimal delay, this may prove to be extremely expensive. For example, in a multi-centre study, it may require that every collection site has a local capacity for state-of-the-art processing and storage rather than restricting such facilities to a single central facility. In a nationwide study this may well be unaffordable. There are three possible solutions to prevent or minimize pre-analytical variability: (1) use a common standard operating procedure (SOP) involving local processing for all studies; (2) set up a large study with rapid sample transportation and central processing such that any delays are acceptable; or (3) carefully assess the impact of biosample deterioration so that, where possible, its impact can be taken into account in the analysis.

Because the optimal standard operating procedure (taking account of both scientific rigour and cost) may vary from analyte to analyte – and possibly from study to study - it is clear that a sound quantitative understanding is required of the manner in which individual analytes degrade or accumulate in unprocessed samples. As a minimum, we

need to know the time course of the degradation, and the variability of that time course between subjects. Given these data, it may then be possible to select a standard operating procedure that provides an acceptable balance between science and cost for the set of analytes that are most crucial for the proposed project. This chapter describes an analysis that explores these issues in this way.

The analysis we describe is a joint venture between UK Biobank and the University of Leicester and represents part of the international biobank harmonization programs of P^3G and BioSHARE-EU (**146**). It uses a set of samples that were originally collected by UK Biobank, before definitive data collection began, with the express purpose of exploring the stability of analytes in the period prior to definitive storage. Although these data have previously been analysed with a similar intention in mind, the analysis that was used (**147**) invoked a very particular "cost of errors" model which does not necessarily capture the impact of certain potentially important classes of degradation or accumulation. To be specific, the previous analysis focused on estimating the probability that a given analyte would markedly degrade (or accumulate) over a given period, but less attention was paid to the impact of differing rates of degradation (or accumulation) in different subjects.

UK Biobank is a large biobank of 500,000 participants aimed at investigating the role of genetic factors, environmental exposures and lifestyle in the causes of major diseases of late and middle age (**148**). Although the data used in this project are from a UK Biobank pre-pilot study, the conclusions of the analysis are potentially generalisable to other biobanks and large scale biosample collections. Furthermore, although UK Biobank has completed its primary sample collection and so its standard operating protocols for that collection are immutable, the analysis we describe will be of value to UK Biobank in future collection sweeps. Finally, due to the large sample sizes required to investigate

the causes of complex disease, it is often necessary to combine data from more than one platform (**71, 149, 150**). The results of my analysis can potentially inform an interpretation of the comparison of analytes between platforms that have used SOPs with unavoidably different needle to freezer delays and of meta-analyses synthesising data across multiple biobanks.

4.2. METHODS

The work described in this chapter entails three complementary analyses of the concentration over time (before definitive long-term storage) of 47 different biological analytes. In the first analysis, stability is investigated by estimating the proportionate change in concentration attributable to delay in processing. The second analysis explores heterogeneity in the rate of change of concentration between samples from different subjects. In the third analysis the impact of delay in sample processing on the power of genetic association studies is estimated through computer simulations using the ESPRESSO-forte power calculator.

The data used in this analysis are from a pre-pilot study set up during the design phase of UK Biobank, to devise Standard Operating Procedures (SOPs) for all analytes. They consist of the measured values of 47 blood and urine analytes collected from 40 subjects at a variety of times between 0 and 36 hours after initial blood sampling. All samples were kept at 4°C until entering definitive long-term frozen storage. The 40 subjects were healthy volunteers and are not among the 500,000 participants ultimately recruited into UK Biobank.

The structure of the data is hierarchical with three levels (subject, time point and replicate measure). Measurements were taken at four time points (0, 12, 24 and 36

hours) for 19 analytes and at two time points (0 and 24 hours) for the other 28 analytes. These time points represent the time elapsed since the collection of the sample; 0 hour means the assay was carried out immediately after sample collection (Figure 19). Two replicate measurements were taken at each nominal time point except at 24 hours in those analytes studied at four time points – here, four measurements were taken but two were used to study the effect of freeze/thaw and, as these two measurements were not true replicates, they were not included in my analysis. A total of 320 measurement values were therefore analysed for the 19 analytes with measurement at four distinct time points and 160 values in the other 28 analytes. The analyte *C Reactive Protein* was excluded from the analysis because its data were censored; all measurement values < 0.2 were reported as 0.2, this distorted the correlation structure both within and between subjects. The remaining 46 analytes were analysed one at a time.

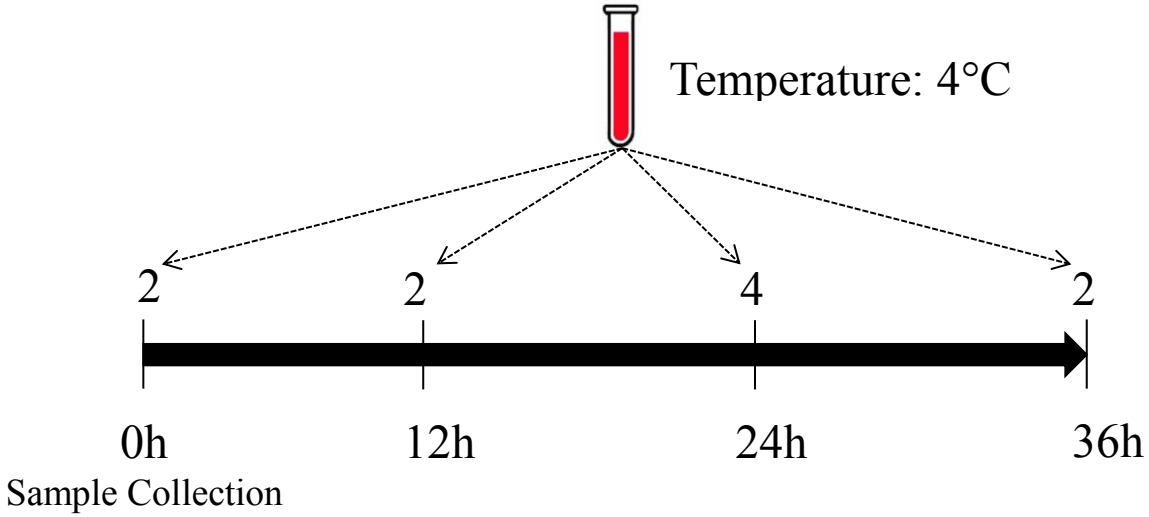


Figure 19: Time points of the repeated measurement.
Measurements were taken over 24 or 36 hours for each analyte and each subject.

4.2.1. FIRST ANALYSIS: ESTIMATING THE PROPORTION OF THE VARIANCE IN ANALYTE CONCENTRATION THAT MAY BE ATTRIBUTED TO DELAY IN PROCESSING

The observed variability between samples from different participants combines the real biological heterogeneity between subjects with random measurement error and with the pre-analytical variability caused by processing delays. Although the variability due to delay in processing is in fact a real biological effect – *i.e.* the change in concentration (usually degradation) of a biological analyte over time is *real* - it leads to a loss of information from the sample, and is in that sense as undesirable as random measurement errors. To estimate the proportion of the observed variability between subjects that can be attributed to delay in processing; a three-level variance component model was fitted in MLwiN 2.1, a software tool that fits multilevel models to complex hierarchical data (**151**).

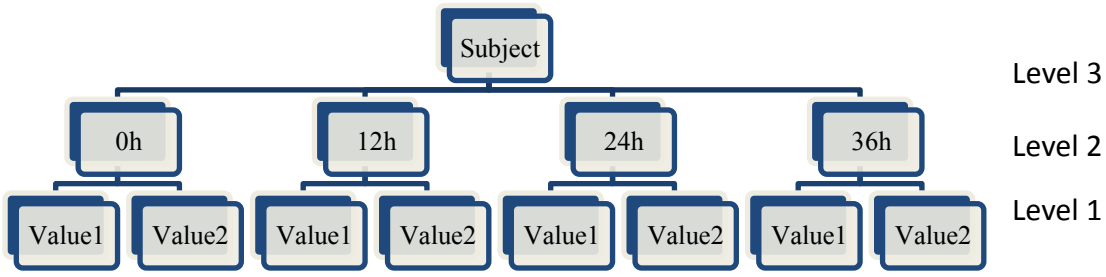


Figure 20: Graphical view of the three-level model fitted in MLwiN.

Algebraically, this 'random intercept' model may be expressed as follows:

$$\boldsymbol{y_{ijk} = \beta_{0ijk} x_0} \qquad \textbf{(MODEL 1)}$$

Where: i, j and k represents respectively the lowest (replicate measurement level), the middle (processing time level) and the highest level (subject level).

$$\beta_{0ijk} = \beta_0 + v_{0k} + w_{0jk} + e_{0ijk}$$

$x_0 = 1$

$v_{0k} \sim N(0, \sigma^2_{v0})$

$w_{0jk} \sim N(0, \sigma^2_{w0})$

$e_{0ijk} \sim N(0, \sigma^2_{e0})$

In Figure 20: Level 1 is the level of replicated measurements; the variance at this level, σ^2_{e0}, reflects random measurement error. Level 2 reflects measurement time-point; variance at this level, σ^2_{w0}, incorporates variability arising from delay in processing. Level 3 is the subject level; variance at this level, σ^2_{v0}, captures the real biological difference between subjects, and it is this - informative - variability that might be expected to provide the basis for useful scientific enquiry. The distribution of the residuals at levels 1, 2 and 3, respectively e_{0ijk}, w_{0jk} and v_{0k} was explored and verified as Gaussian using normal probability plots. The model was fitted without explanatory variables; i.e. with x_0 taking the constant value of 1.00, all subjects are modelled as sharing a common mean intercept of β_0. Specifically, this means that in analysing the data with four time points, the nominal times were treated as a non-ordered categorical variable. As an alternative an ordinal parameterization would have been possible and would have been more powerful for detecting a weak but consistent decline (or increase) in concentration over time. But it was considered preferable to minimise any assumptions and to treat each time point as an independent entity rather than assuming a natural order. Given that consistent changes over time were easily detected for many of the analytes with four time points anyway - despite this small loss of statistical power - it would appear that this decision was not unreasonable. For analytes measured at only two time points, the two models are equivalent.

On the basis of this modelling, the proportion of variability that may be attributed to delay in measurement is estimated as the ratio of the variance at the second level of the model (level 2) over the sum of the variances at all 3 levels: $\frac{\sigma^2_{w0}}{(\sigma^2_{v0}+\sigma^2_{w0}+\sigma^2_{e0})}$

4.2.2. SECOND ANALYSIS: ESTIMATING THE PROPORTION OF HETEROGENEITY BETWEEN SUBJECTS THAT IS DUE TO VARIABILITY IN THE RATE OF DEGRADATION OF ANALYTES

In considering the stability of a biological analyte over time, there are three fundamental scenarios (Figure 21): (1) Samples are stable and there is no change in the concentration of the analyte over time; (2) samples are unstable, but the rate of change in concentration is the same in all samples from all individuals; (3) samples are unstable and the rate of change in concentration varies from sample to sample.

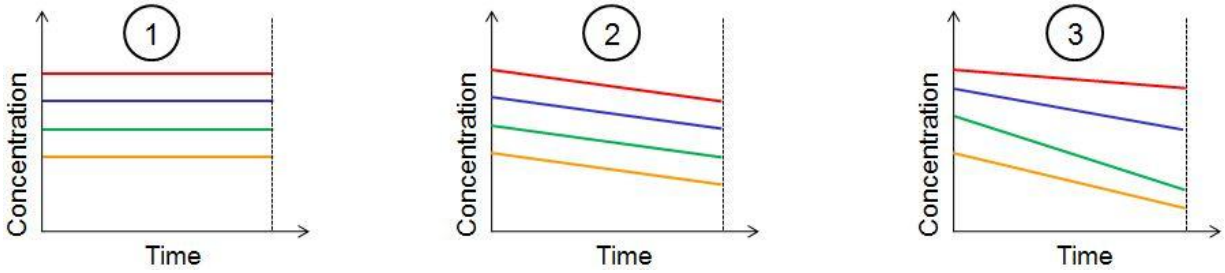


Figure 21: Variation in biological analyte concentration over time.
The above graphs illustrate the comparison of 4 samples under the 3 scenarios mentioned in the paragraph above.

These three scenarios may be modelled using a series of nested two-level multilevel models (Figure 22).

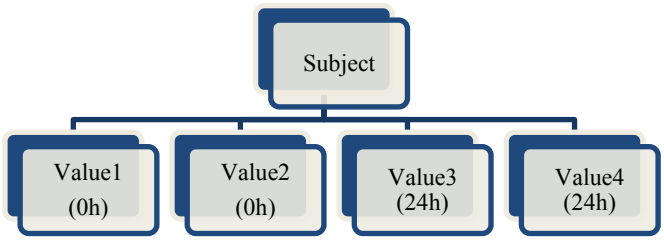


Figure 22: Graphical view of the two-level model.
This graph illustrates the model for an analyte measured at two time points (0 and 24 hours).

Scenario 3 is the general case and is captured by the model:

$$\boldsymbol{y_{ik} = \beta_{0ik} x_0 + \beta_{1k} x_{1_{ik}}} \qquad \textbf{(MODEL 2)}$$

Where i and k represent respectively the lower (replicate measurement level) and the upper level (subject level) and the nominal processing time is modelled as a fixed covariate (rather than as an extra level in the random effects hierarchy, as in Model 1),

$\beta_{0ik} = \beta_0 + u_{0k} + e_{0ik}$

$\beta_{1k} = \beta_1 + u_{1k}$

$x_0 = 1$

$x_{1ik} = a\ dummy\ variable\ coding\ nominal\ time\ (0 = 0\ h, 1 = 24\ h)$ for the i^{th} sample of the k^{th} subject.

Under this model:

$\beta_0 = mean\ concentration\ of\ analyte\ at\ nominal\ time\ 0\ h$

$\beta_1 = meandifferencebetweenconcentrationofanalyteat\ 0\ and\ 24h$

$u_{0k} \sim N(0, \sigma^2_{u0})$ = *subject level random effect reflecting the variance between subjects in the expected concentration of analyte at 0h.*

$u_{1k} \sim N(0, \sigma^2_{u1})$ = *subject level random effect reflecting the variance between subjects in the expected rate of change in concentration between 0 and 24 h.*

$e_{0ik} \sim N(0, \sigma^2_{e0}) = \textit{residual error at level 1.}$

Under the general model (scenario 3) the expected change of concentration between 0 and 24 hours may be non-zero ($\beta_1 \neq 0$) and the rate of change may vary from subject to subject ($\sigma^2_{u1} > 1$). Scenario 2 is a special case of scenario 3 in which the slope may be non-zero ($\beta_1 \neq 0$) but the rate of change is the same in samples from all subjects ($\sigma^2_{u1} = 0$). Scenario 1 is a special case of scenario 2 in which the slope is identically zero in all subjects ($\beta_1 = 0$ and $\sigma^2_{u1} = 0$). The coefficients of these models may be used to estimate the overall variance at 0 *h* and the overall variance at 24 *h* (see Figure 23).

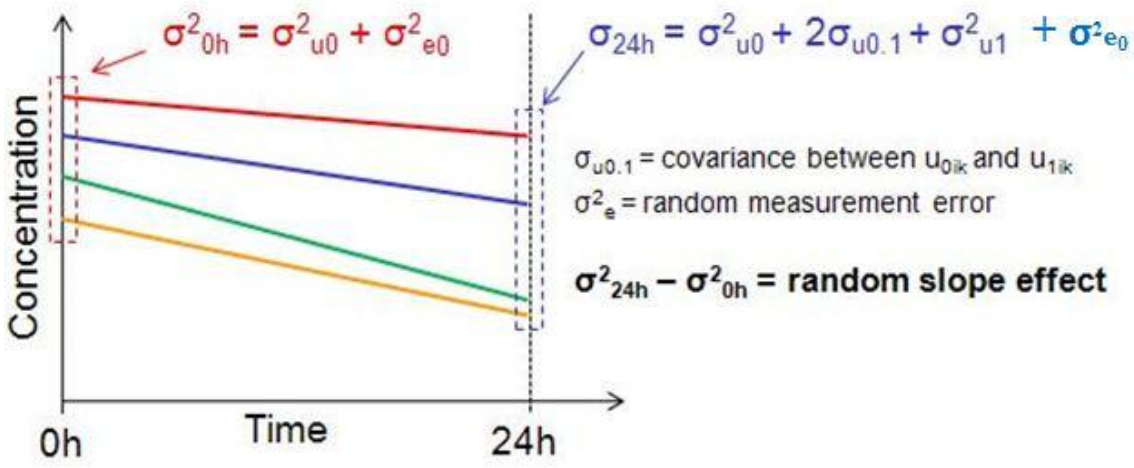


Figure 23: Using variance function to determine overall variance at 0 and 24h.

The sign and magnitude of the covariance term $\sigma_{u0.1}$ in comparison to the subject variance in slope (σ^2_{u1}) determine the relative magnitudes of the overall variance at 24 *h* (σ^2_{24h}), and that at base-line (σ^2_{0h}). Specifically, if the covariance term is sufficiently negative, the variance at 24*h* may be lower than at 0*h* and, at first glance, this might lead one to believe that a *gain* in power could potentially be achieved by measuring all samples at 24 *h* rather than 0*h*. But, this is misleading. Here, variability reflects a

combination of the intercept variance, reflecting the real biological difference between subjects - which holds the scientific information that is generally of interest to an analysis - and random error and slope variance that both degrade the information that may be extracted. There can be no *gain* in power at 24 *h*, the between subject variance in slope can only lead to loss of information – it just happens that if there is a negative covariance between intercept and slope, subjects with a high base-line value will tend to exhibit a greater decline in concentration than those with low base-line values. The lines depicting concentration (Figure 24) will then tend to converge from left to right and, in the extreme a zero variance at 24 *h* (all lines meet at a point) will indicate a total loss of the information carried by the between subject variance in intercept at 0 *h*.

In order to model the loss of power consequent upon between subject variance in slope if all samples are measured at 24 *h* rather than 0 *h*, it should instead be noted that (measurement error aside) the observed analyte concentrations at base-line (0 h) are 'real', while those at 24 *h* are contaminated not only by measurement error but also by the between subject variability in the slope. In consequence, the overall assessment error (around the true biological effects) increases from $\sigma^2_{e0}$to σ^2_{e0}+ $\sigma^2_{u1}$and it is this increase in error variance that should be incorporated into the power calculation.

The term $\sigma_{u0.1}$ that was central to the characteristics of the 'intuitive' approach plays no explicit role in the interpretation outlined. However, the term was included in MODEL 2 to reflect the covariance between u_{0k} and u_{1k}, and this is essential if consistent estimates of all model terms (including σ^2_e and σ^2_{u1}) are to be obtained; the term was also included in the calculation when estimating the average variance of errors around the true biological effects. The model contained no covariance term between u_{1k} and e_{ik} (these random effects are assumed conditionally independent), and no covariance

term is therefore required in calculating their combined impact ($u_{1k} + e_{ik}$) on the analyte measure at 24 h.

4.2.3. THIRD ANALYSIS: ESTIMATING THE IMPACT OF DELAYS IN SAMPLES PROCESSING ON THE POWER OF GENETIC EPIDEMIOLOGY CASE-CONTROL STUDIES

This aim of this analysis is to determine the amount of power lost due to variance of slope between individuals (σ^2_{u1}). This was done by estimating, through simulations, the increase in sample size required to compensate for the loss of power in a hypothetical association study that investigates a binary outcome determined by an interaction between a genetic variant and an environmental factor; the analytes measurements were considered as the biological responses to environmental factors. The simulations were carried out using the ESPRESSO-forte algorithm.

The input parameters of the outcome and the genetic, environment determinants are respectively under Table 81, Table 82 and Table 83. Sections 2.6.1, 2.6.2and 2.6.3 explain in details how data are generated and analysed by ESPRESSO-forte. The same analysis was run for seven different interactions odds-ratios (1.05, 1.1, 1.2, 1.33, 1.5, 1.75, and 2.0) using the same input parameters reported on the 3 input tables (Table 81, Table 82 and Table 83). So only the difference between these scenarios is the interactions odd-ratios.

For each of the seven interactions odds-ratios mentioned above, the GLM model fitted in ESPRESSO-forte consists of one outcome (a binary trait) and two interacting covariates (a SNP and an environmental factor). The GLM model for these interaction scenarios can be written as follows:

$$g(Y) = \beta_0 + \beta_1 G_1 + \beta_2 E_1 + \beta_3 G_1 E_1$$

Where Y is the outcome, β_0 is the intercept value, β_1 is the effect size of the SNP, β_2 is the effect size of the environmental factor and β_3 is the effect size of the interaction term G_1E_1

Parameter	Value
number of runs	500
number of cases	10000
number of controls	40000
interaction	yes
outcome model	binary
disease prevalence	0.2
subject effect	12.36
p.value	0.0001
power required	0.8
sensitivity outcome	1
specificity outcome	1

Table 81: General and outcome parameters used in the analysis.
Seven scenarios were analysed - one scenario for each of the seven interaction OR.

Parameter	Value
genetic model	binary
MAF	0.1
or	1.0
sensitivity	0.95
specificity	0.95

Table 82: Parameters for the genetic determinant
These input parameters were used for each of the seven interactions OR investigated. The sensitivity and specificity values are the actual sensitivity and specificity of the assessment of the individual alleles that form the genotypes.

Parameter	Value
exposure model	quantitative
prevalence "at risk" exposure	0.1
reliability	1
or	1.0
sensitivity	0.99
specificity	0.99

Table 83: Parameters for the environmental exposure.
These input parameters were used for each of the seven interactions OR investigated.

The sample size required to reach a power of 80% was calculated under the two below scenarios A and B. The difference between the sample size calculated under scenario A and the one calculated under scenario B represents the sample size increase required to compensate for the loss of power caused by the slope variance(σ^2_{u1}) if measurements are delayed by 24 hours. For each of the seven interaction ORs and for each analytes, the difference of sample size between scenario A and scenario B was calculated and the results were reported as the ratio of the mean sample size for scenario B over the mean sample size for scenario A i.e. the multiplicative increase.

4.2.3.1. Scenario A

All measurements are taken at 0h; there is no variability arising due to slope variance because there was no delay in the processing. Therefore only the random measurement error affects the power - all other variability at baseline relates to true (biologically meaningful) variation between individuals (i.e. σ^2_{u0}). The true outcome D_1 (error-free data), is therefore generated with a mean β_0 (mean intercept at time 0h) and a variance σ^2_{u0} . The observed outcome $D_2$is obtained by adding the random measurement error E_R to D_1.

$D_1 \sim N(\beta_0, \sigma^2_{u0})$

$E_R \sim N(0, \sigma^2_{e0})$

$D_2 = D_1 + E_R$

4.2.3.2. Scenario B

Measurements are taken at 24h; there is a delay in the processing. Therefore both the random measurement error and the error arising from the variability of slope between subjectsaffect the power. The true outcome D_1 is generated with the same variance σ^2_{u0}

as in scenario 1 (because the between subject variance remains the same) and a mean $\beta_0+\beta_1$ ($\beta_1$is the average slope). The observed outcome $D_2$is obtained by adding the random measurement error E_R and the error due to slope variance S_R toD_1.

$D_1 \sim N(\beta_0 + \beta_1, \sigma^2_{u0})$

$E_R \sim N(0, \sigma^2_{e0})$

$S_R \sim N(0, \sigma^2_{u1})$

$D_2 = D_1 + E_R + S_R$

4.3. RESULTS

4.3.1. FIRST ANALYSIS

Using the three level model with *no* heterogeneity in slope between subjects (MODEL 1) the proportion of the observed variability that can be attributed to processing time (σ^2_{u0}) is ≥ 10% for 16 out of 46 analytes (see Table 84). Eight analytes have between 5% and 10% of the observed variability attributable to delay in processing. For the remaining 22 analytes the contribution of the variance resulting from delay in processing represents less than 5% of the observed variance. Bicarbonate seems particularly sensitive to delayed processing with 61% of the observed variability coming from σ^2_{u0}. For MCHC, basophils, total protein, albumin and potassium, σ^2_{u0} accounts for between 28% and 38% of the total variability.

Analyte	Mean	Between subject variance (σ^2_{v0})	Variance due to delay in processing (σ^2_{w0})	Residual (σ^2_{e0})	$\sigma^2_{w0} / (\sigma^2_{v0} + \sigma^2_{w0} + \sigma^2_{e0})$
Bicarbonate	26.0938 (mmol/L)	0.6787	1.2875	0.1563	61%
Albumin	41.6880 (g/L)	4.5735	2.8563	0.1687	38%
Total Protein	68.7312 (g/L)	14.4590	9.1812	2.4062	35%
MCHC	33.0156 (g/dL)	0.1440	0.1119	0.0582	36%
Potassium	4.0597 (mmol/L)	0.0754	0.0326	0.0015	30%
Basophils	0.0445($\times 10^9$/L)	0.0003	0.0001	0.00008	28%
Eosinophils	0.1377 ($\times 10^9$/L)	0.0063	0.0013	0.0005	16%
Calcium	2.2345 (mmol/L)	0.0070	0.0013	0.0001	15%
Packed Cell Volume	0.4212 (-)	0.0020	0.0004	0.00002	15%
Haemoglobin	13.8962 (g/dL)	2.2303	0.3641	0.004	14%
Haemoglobin A1C	3.3047 (%)	0.0507	0.0110	0.0173	14%
Sodium	137.6812 (mmol/L)	2.4703	0.4719	0.6063	13%
Red Blood Cell	4.5452 ($\times 10^{12}$/L)	0.2359	0.0330	0.0028	12%
Chloride	106.3250 (mmol/L)	3.8819	0.5500	0.3875	11%
Platelet Count	240.3508 ($\times 10^9$/L)	2873.8896	346.1551	46.6603	11%
Haemoglobin A1CX	0.4842(%)	0.0035	0.0004	0.0003	10%
Glucose	5.3384 (mmol/L)	0.2277	0.0194	0.0023	8%
Bilirubin	14.7856 (µmol/L)	16.9216	1.4912	2.2736	7%
Glucose (F. Oxalate)	5.6100 (mmol/L)	0.2478	0.0176	0.0019	7%
Magnesium	0.8854 (mmol/L)	0.0027	0.0002	0.00005	6%
Lymphocytes	1.9247 ($\times 10^9$/L)	0.3463	0.0195	0.0068	5%
Insulin	7.2404 (mIU/L)	17.5079	0.9419	0.0494	5%
Cholesterol	5.2481 (mmol/L)	1.0490	0.0539	0.0024	5%
Monocytes	0.4131 ($\times 10^9$/L)	0.0275	0.0015	0.0019	5%
Fibrinogen	3.0472 (g/L)	0.4428	0.0206	0.0074	4%
High Density Lipid	1.5668 (mmol/L)	0.1888	0.0068	0.0018	3%
Amylase	71.5250 (IU/L)	559.6944	15.5124	2.0375	3%
White Cell Count	6.4375 ($\times 10^9$/L)	6.1661	0.1674	0.0501	3%
Alkaline Phosphatase	61.3954 (IU/L)	307.2807	7.4825	2.4989	2%
Creatinine	89.8375 (µmol/L)	160.5050	3.8938	14.9	2%
Mean Cell Volume	92.7881(fL)	17.5896	0.3760	0.0489	2%
AST	22.9688 (IU/L)	29.1178	0.6063	0.7437	2%
Neutrophils	3.9184 ($\times 10^9$/L)	4.5051	0.0680	0.0273	1%
MCH	30.6269 (pg)	1.5992	0.0229	0.0458	1%
CK MB Fraction	4.7951 (IU/L)	21.8518	0.2348	0.1457	1%
Urinary Urea	246.2291 (mmol/L)	11998.9326	127.6297	27.7411	1%
Triglyceride	1.2653 (mmol/L)	0.5557	0.0056	0.001	1%
Creatinine Kinase	106.5875 (IU/L)	2655.4072	20.7750	2.0125	1%
Gamma GT	30.5960 (IU/L)	354.5936	2.1418	3.0115	1%
ALT	23.5438 (IU/L)	133.4745	0.7000	0.5687	1%
Inorganic Phosphorus	1.0374 (mmol/L)	0.0205	0.0001	0.0016	0%
Uric Acid	301.9563 (µmol/L)	3336.3438	12.2125	3.6188	0%
Blood Urinary Nitrogen	2.6881 (mmol/L)	0.3189	0.0009	0.0016	0%
Urinary Calcium	2.2970 (mmol/L)	1.9764	0.0050	0.0272	0%
Urinary Sodium	96.1281 (mmol/L)	2605.4973	0.0833	9.0594	0%
Urinary Potassium	49.6527 (mmol/L)	651.4268	0.6566	6.8331	0%

Table 84: Proportion of the observed variability attributable to processing time. The analytes are ranked by decreasing contribution of delay in processing to observed variability between subjects- reported in the last column. The first column contains the names of the biochemical analytes. The second column contains the mean of the measurements across the 40 subjects for the fitted 3-level model. The third column holds the variance at the subject level. The fourth column contains the variance arising from the delay in processing – processing time level. The fifth column contains the 'the variance that represents the random measurement error.

4.3.2. SECOND ANALYSIS

Prior to fitting the full two-level model with a random intercept and random slope (MODEL 2) for each analyte, the nested model with a constant but non-zero slope ($\beta_1 \neq 0$, $\sigma^2_{u1} = 0$) was fitted. On removing the constraint on the variance ($\sigma^2_{u1} > 0$) the likelihood ratio test showed a highly significant improvement in fit (Table 85) for almost all analytes, suggesting that there was substantive evidence of between subject heterogeneity in slope for most analytes. This suggests that the impact of delayed processing should be explored carefully for most analytes.

Analyte	Mean (β_0)	Slope (β_1)	Expected % change per 12hours	Between subjects variance (σ^2_{u0})	Between slopes variance (σ^2_{u1})	Residual (σ^2_{e0})	$\chi 2$	P.Value
Bicarbonate	26.688 (mmol/L)	-1.188	-2.22	1.506	1.165	0.156	64.687	8.78E-16
Potassium	4.107 (g/L)	-0.032	-0.77	0.093	0.007	0.018	188.206	7.83E-43
Albumin	42 (g/L)	-0.763	-0.91	6.766	5.131	0.169	151.673	7.47E-35
Total Protein	69.488 (g/dL)	-1.513	-1.09	20.840	16.075	2.406	59.583	1.17E-14
MCHC	32.9 (mmol/L)	0.231	0.35	0.166	0.170	0.058	32.018	1.53E-08
Basophils	0.046 (×109/L)	-0.002	-2.47	0.000	0.000	0.000	31.491	2.00E-08
Calcium	2.26 (×109/L)	-0.017	-0.74	0.008	0.000	0.000	689.672	5.27E-152
Eosinophils	0.157 (mmol/L)	-0.039	-12.40	0.009	0.001	0.001	27.142	1.89E-07
Packed Cell Volume	0.428 (-)	-0.014	-1.68	0.002	0.001	0.000	142.267	8.50E-33
Haemoglobin A1C	3.261 (g/dL)	0.029	0.89	0.057	0.002	0.023	113.637	1.56E-26
Haemoglobin	14.078 (%)	-0.363	-1.29	2.482	0.597	0.004	270.890	7.26E-61
Red Blood Cell	4.604 (mmol/L)	-0.118	-1.29	0.259	0.052	0.003	118.800	1.16E-27
Glucose (F. Oxalate)	5.651 (×1012/L)	-0.027	-0.48	0.253	0.004	0.011	224.603	8.96E-51
Chloride	106.15 (mmol/L)	0.350	0.16	4.159	0.978	0.388	23.268	1.41E-06
Platelet Count	245.738 (×109/L)	-11.408	-2.32	3269.157	548.204	46.660	86.048	1.76E-20
Sodium	138.131 (%)	-0.300	-0.22	2.364	0.084	0.762	0.000	1.00E+00
Insulin	6.837 (mmol/L)	0.268	3.92	17.593	0.239	0.417	0.000	1.00E+00
Bilirubin	15.225 (µmol/L)	-0.879	-2.89	18.188	2.210	2.274	6.519	1.07E-02
Fibrinogen	3.146 (mmol/L)	-0.066	-2.09	0.470	0.002	0.016	183.268	9.37E-42
Glucose	5.378 (mmol/L)	-0.027	-0.49	0.229	0.004	0.013	163.010	2.49E-37
Magnesium	0.888 (×109/L)	-0.006	-0.32	0.003	0.004	0.000	59.021	1.56E-14
Lymphocytes	1.966 (mIU/L)	-0.082	-2.09	0.389	0.032	0.007	45.793	1.31E-11
Cholesterol	5.403 (mmol/L)	-0.103	-1.91	1.133	0.004	0.027	92.411	7.04E-22
Haemoglobin A1CX	0.486 (×109/L)	-0.001	-0.23	0.004	0.000	0.001	704.404	3.30E-155
Monocytes	0.424 (g/L)	-0.022	-2.59	0.029	0.002	0.002	9.628	1.92E-03
High Density Lipid	1.638 (mmol/L)	-0.045	-2.76	0.226	0.001	0.004	364.579	2.83E-81
Amylase	73.051 (IU/L)	-3.050	-2.09	599.198	21.722	2.038	85.765	2.03E-20
White Cell Count	6.515 (×109/L)	-0.155	-1.19	6.061	0.311	0.050	56.796	4.83E-14
Alkaline Phosphatase	62.471 (IU/L)	-2.146	-1.72	318.718	10.379	2.497	38.308	6.04E-10
Mean Cell Volume	93.049 (µmol/L)	-0.521	-0.28	17.309	0.480	0.049	80.493	2.92E-19
AST	23.05 (fL)	-0.163	-0.35	29.938	1.186	0.744	13.050	3.03E-04
Creatinine	91.023 (IU/L)	-0.790	-0.87	173.816	0.403	16.770	0.000	1.00E+00
Neutrophils	3.923 (×109/L)	-0.009	-0.11	4.313	0.136	0.027	49.600	1.89E-12
MCH	30.603 (pg)	0.049	0.08	1.621	0.044	0.046	6.294	1.21E-02
CK MB Fraction	4.905 (IU/L)	-0.218	-2.22	24.768	0.417	0.146	60.985	5.75E-15
Triglyceride	1.308 (mmol/L)	-0.028	-2.15	0.584	0.001	0.004	414.335	4.17E-92
Creatinine Kinase	108.225 (mmol/L)	-3.275	-1.51	2673.261	30.825	2.013	105.437	9.80E-25
ALT	23.788 (IU/L)	-0.488	-1.02	135.898	1.162	0.569	18.373	1.82E-05
Gamma GT	31.324 (IU/L)	-0.488	-1.56	389.502	0.540	3.738	0.000	1.00E+00
Inorganic Phosphorus	1.042 (IU/L)	-0.010	-0.46	0.021	0.000	0.002	0.275	6.00E-01
Uric Acid	300.875 (mmol/L)	2.163	0.36	3328.925	19.749	3.619	49.927	1.60E-12
Blood Urinary Nitrogen	2.69 (µmol/L)	-0.0037	-0.07	0.3174	0.0019	0.0016	8.9756	2.74E-03
Urinary Urea	246.5983 (mmol/L)	-0.2461	-0.10	11609.8408	24.6165	101.8849	0.0000	1.00E+00
Urinary Sodium	96.3476 (mmol/L)	-0.1463	-0.15	2629.6777	0.5562	8.3056	0.0000	1.00E+00
Urinary Calcium	2.305 (mmol/L)	-0.0056	-0.24	1.9686	0.0044	0.025	0.0000	1.00E+00
Urinary Potassium	50.0015 (mmol/L)	-0.2349	-0.47	686.5211	0.6045	6.4321	0.0000	1.00E+00

Table 85: Heterogeneity in the rate of change of analyte concentration in 24 hours. The first column contains the names of the biochemical analytes. The second column contains the mean of the measurements across the 40 subjects. The third column holds the rate of change in the analyte's concentration over 24 hours. The fourth column contains the percentage change in analyte concentration per 12 hours. The fifth column contains the 'real' biological variance between subjects (i.e. the biological difference of interest for the study). The sixth column holds the variance that represents the heterogeneity in the rate of change of the analyte's concentration. The random measurement error in recorded in the sixth column. The seventh column holds the chi-squared values for the comparison between the two fitted models (the model with a random intercept only vs. the model with the random slope and random intercept). The significance of the difference of fit between the two models is reported as p-value in the last column.

4.3.3. THIRD ANALYSIS

The sample size increase compensating for the loss of power resulting from delayed processing as estimated in the second analysis - and the consequent impact of slope heterogeneity - is expressed as a multiplicative factor in Table 86. Five analytes demanded an increase > 2, that is, more than 100% increase in sample size; fourteen analytes required a sample size increase of between 10% and 81%; twenty two analytes required a sample size increase of between 1% and 9%. The loss of power, and hence the required increase in sample size, is largest when σ^2_{u1} is large.

Analyte	Mean (β_0)	Slope (β_1)	Between subjects variance (σ^2_{u0})	Between slopes variance (σ^2_{u1})	Residual (σ^2_{e0})	Multiplicative increase required to compensate for power loss
Potassium	4.107 (g/L)	-0.032	0.093	0.007	0.018	3.20
Basophils	0.046 (×109/L)	-0.002	0	0	0	2.31
Bicarbonate	26.688 (mmol/L)	-1.188	1.506	1.165	0.156	2.28
Albumin	42 (g/L)	-0.763	6.766	5.131	0.169	2.28
Total Protein	69.488 (g/dL)	-1.513	20.84	16.075	2.406	2.22
MCHC	32.9 (mmol/L)	0.231	0.166	0.17	0.058	1.81
Calcium	2.26 (×109/L)	-0.017	0.008	0	0	1.41
Packed Cell Volume	0.428 (-)	-0.014	0.002	0.001	0	1.26
Haemoglobin	14.078 (%)	-0.363	2.482	0.597	0.004	1.25
Chloride	106.15 (mmol/L)	0.35	4.159	0.978	0.388	1.25
Red Blood Cell	4.604 (mmol/L)	-0.118	0.259	0.052	0.003	1.18
Glucose (F. Oxalate)	5.651 (×1012/L)	-0.027	0.253	0.004	0.011	1.18
Eosinophils	0.157 (mmol/L)	-0.039	0.009	0.001	0.001	1.17
Haemoglobin A1CX	0.486 (×109/L)	-0.001	0.004	0	0.001	1.17
Platelet Count	245.738 (×109/L)	-11.408	3269.157	548.204	46.66	1.16
Glucose	5.378 (mmol/L)	-0.027	0.229	0.004	0.013	1.16
Magnesium	0.888 (×109/L)	-0.006	0.003	0.004	0	1.13
Insulin	6.837 (mmol/L)	0.268	17.593	0.239	0.417	1.12
Bilirubin	15.225 (µmol/L)	-0.879	18.188	2.21	2.274	1.11
Fibrinogen	3.146 (mmol/L)	-0.066	0.47	0.002	0.016	1.09
Monocytes	0.424 (g/L)	-0.022	0.029	0.002	0.002	1.09
Cholesterol	5.403 (mmol/L)	-0.103	1.133	0.004	0.027	1.08
Lymphocytes	1.966 (mIU/L)	-0.082	0.389	0.032	0.007	1.07
High Density Lipid	1.638 (mmol/L)	-0.045	0.226	0.001	0.004	1.07
Creatinine	91.023 (IU/L)	-0.79	173.816	0.403	16.77	1.07
Haemoglobin A1C	3.261 (g/dL)	0.029	0.057	0.002	0.023	1.06
White Cell Count	6.515 (×109/L)	-0.155	6.061	0.311	0.05	1.06
AST	23.05 (fL)	-0.163	29.938	1.186	0.744	1.03
Neutrophils	3.923 (×109/L)	-0.009	4.313	0.136	0.027	1.03
Blood Urinary Nitrogen	2.69 (µmol/L)	-0.0037	0.3174	0.0019	0.0016	1.03
Amylase	73.051 (IU/L)	-3.05	599.198	21.722	2.038	1.02
Alkaline Phosphatase	62.471 (IU/L)	-2.146	318.718	10.379	2.497	1.02
ALT	23.788 (IU/L)	-0.488	135.898	1.162	0.569	1.02
Uric Acid	300.875 (mmol/L)	2.163	3328.925	19.749	3.619	1.02
Urinary Urea	246.5983 (mmol/L)	-0.2461	11609.8408	24.6165	101.8849	1.02
Sodium	138.131 (%)	-0.3	2.364	0.084	0.762	1.01
Mean Cell Volume	93.049 (µmol/L)	-0.521	17.309	0.48	0.049	1.01
Creatinine Kinase	108.225 (mmol/L)	-3.275	2673.261	30.825	2.013	1.01
Inorganic Phosphorus	1.042 (IU/L)	-0.01	0.021	0	0.002	1.01
Urinary Sodium	96.3476 (mmol/L)	-0.1463	2629.6777	0.5562	8.3056	1.01
Urinary Calcium	2.305 (mmol/L)	-0.0056	1.9686	0.0044	0.025	1.01
Urinary Potassium	50.0015 (mmol/L)	-0.2349	686.5211	0.6045	6.4321	1.00
MCH	30.603 (pg)	0.049	1.621	0.044	0.046	0.99
CK MB Fraction	4.905 (IU/L)	-0.218	24.768	0.417	0.146	0.99
Triglyceride	1.308 (mmol/L)	-0.028	0.584	0.001	0.004	0.98
Gamma GT	31.324 (IU/L)	-0.488	389.502	0.54	3.738	-

Table 86: Sample size increase required to compensate for the power loss.
The columns 1 to 6 are as defined in Table 85. The values in the last column represent the sample size multiplicative increase required to compensate for the power loss caused by the bias arising from slope heterogeneity reported in Table 85. The sample size multiplicative increase values reported for the 3 last analytes (0.99, 0.99 and 0.98) seem to indicate a gain of power but this is due to the stochastic nature of the simulations; these figure are the average multiplicative increase across the 7 interactions OR investigated; for some ORs the increase was slightly greater than 1.0 and slightly less than 1.0 for others. There was no multiplicative increase calculated for the analyte Gamma GT because there was not any heterogeneity in slope for this analyte ($\sigma^2_{u1} = 0$).

4.4. DISCUSSION

The work reported in this chapter builds directly on the earlier paper of Jackson et al.(**147**). The relevant component of their analysis (see their table 1 and figure 1) focussed on four statistical measures of the impact of needle-freezer time delay: (1) the expected proportionate change (generally a decline) in the concentration of each analyte as a percentage of the level of that same analyte at baseline, with a 95% prediction interval for the change in a given individual; (2) the predicted probability of a consistently negative (or positive) change in the level in a given future individual; (3) a likelihood ratio test for "significant" evidence of heterogeneity of slope (decline or increase) between individuals; (4) the proportion of the total variance at a given time point that can reasonably be explained by this heterogeneity of slope.

In brief summary they demonstrated that the percentage change in analyte concentration per 12 hours of delay was generally small. It was less than 3% in absolute magnitude for all but two analytes: insulin at +3.9%; and eosinophils count at -12%. Only four analytes (calcium, fibrinogen, cholesterol, HDL cholesterol) exhibited a high (>90%) probability of a consistently negative or positive change in an arbitrary future individual– none was so consistent that the posterior probability exceeded 97.5%. No analytes exhibited a >90% probability of a consistently positive change. They concluded that their results: *"suggest that any instability in assay results up to 36 h is likely to be small in comparison with between individual differences and assay error, and that a single assay measurement at any time between 0 and 36 h should give a representative value of the analyte concentration at time zero for that individual."*

I started by fitting a similar range of models, though using a different modelling environment, MLwiN 2.1 (**151**), and broadly confirm the equivalent estimates published by Jackson *et al* (**147**). But I chose to interpret these parameters from a rather different perspective and this led me on to undertake additional analyses that I feel shed additional light on the need to consider and adjust for processing delay in designing and using large biobanks. The need for this additional perspective is illustrated by consideration of why biobanks exist at all, and why consistent standard operating procedures are so important. To be specific, contemporary bioscience has a focus on dissecting the weak aetiological signals that collectively represent the architecture of the common complex diseases. Even in the most ideal of worlds, weak signals can demand vast sample sizes to generate adequate statistical power (**7**). But the real world is not ideal – the actual biological signals that might in theory be detected are almost always degraded by a variety of factors such as measurement error, and including the impact of delay in final processing. Rational biobanking demands an understanding of which of these factors can be addressed, with what impact and at what cost to time and resources. But, the true balance between costs and benefits can often only be determined with empirical evidence.

One of the key decisions that must be undertaken in designing a new biobank is whether to invest a substantive amount of resources in enabling rapid local processing and definitive storage to prevent biosamples degradation, or to invest those same resources recruiting more participants under a more flexible regime that means that biosamples will be transported centrally before final processing and that some will therefore be subject to significant delay. It is the type of empirical study described in the Jackson *et al* paper and now re-analysed in the current chapter that enables such a decision to be taken rationally. But, I would argue that the mathematics underpinning that decision can

only be fully informed if one directly assesses the loss of power that is likely to emanate from degradation of the biological signal. Jackson *et al*, clearly demonstrate that any instability is likely to be small in comparison to the between individual differences and to assay error. They also demonstrate that the probability of a *consistently* negative or positive trend in analyte levels is less than 90% for all but four analytes. My findings are entirely consistent with these basic conclusions, but I would argue that what typically matters in terms of the impact of delayed processing on statistical power, is the quantitative effect of these factors on the power to detect relevant changes in the true biological signal. In other words, even if the expected rate of change with time for a given analyte is zero (no consistent change), if there is nevertheless marked heterogeneity in the rate of change between subjects (as exhibited by many analytes, both in our analysis and in that of Jackson *et al*), then this can lead to substantial degradation of the true biological signal, and a meaningful loss of statistical power for many of the classes of analysis that are likely, in the real world, to be undertaken. Furthermore, as outlined earlier (see Figure 23), this cannot be evaluated by simply comparing the variance at baseline with that at the time that processing was finalised. Rather, one must work directly with the parameters reflecting the biological signal and its degradation via slope variance – the power must necessarily fall with delay time, even if, as is possible, the total variance appears to decrease with time.

If we are to design biobanks using rational standard operating procedures, and are to analyse data from biobanks in a manner that enables us to take proper account of the likely loss of power, the important analyses reported by Jackson *et al* and confirmed in the first part of this chapter, need to be supplemented by the analyses addressed in the second phase of our work. These are detailed in Table 86 and are reflected in the

additional interpretation fundamental to the current chapter. This interpretation must necessarily be analyte specific and must be quantitative.

Thus, for example, Table 84 demonstrates that given the distribution of processing times fundamental to this empirical study, 30% of the variance in the level of potassium can reasonably be attributed to factors related to a delay in processing; either a consistent decrease /increase in concentration or random variation in slope between subjects. Table 85 indicates that, in this case, the latter is more important. In fact, the consistent change over time represents an estimated -0.77% decrease in concentration over 12 hours (in agreement with the findings of Jackson *et al* who reported a -0.77% decrease and 64% probability of a consistent negative trend. In contrast, there is strong - highly significant - evidence of heterogeneity in slope between participants (Table 85), and the additional variance arising directly from this variability of slope if processing is delayed by 24 hour is of a similar order of magnitude to the between subject variance at baseline. This implies that there could be a substantial degradation of biological signal if processing is delayed by 24 hour. To quantify the impact on statistical power of this degradation in signal, Table 86 demonstrates that if one does want to analyse potassium levels assayed in biosamples that are processed 24 hour after collection, then the required sample size should be increased by a factor by a factor of more than three – compared to what would be anticipated if all samples had been processed immediately after collection.

As more and more analyses involve sharing and meta-analysing data over several biobanks (**7**), the need to reflect carefully on the impact of a range of different SOPs becomes ever more important. As emphasised in this analysis, this requires careful consideration of the difference between the impact of a consistent rate of

decline/increase and of a rate that varies markedly between subjects. The former may be circumvented if all biosamples are subject to the same delay, but is seriously problematic (though potentially correctable – at least in theory) if, for example, cases and controls are derived from different biobanks with different delays. The latter implies that processing delay should, where possible, be constricted in all subjects, but if it is unavoidable then analytic adjustment must be made – for example by modifying expectations in terms of realistic statistical power.

The empirical study initially carried out by Jackson *et al*, and now subject to additional analysis and interpretation in the present chapter illustrates an important way forward. Like UK Biobank (**148**), all new biobanks need to determine clear and rational standard operating procedures for sample pre-processing (**147, 148, 152**) that ensure consistency for things that can reasonably be kept constant, and ensures faithful recording of parameters that may then vary between subjects, in order that they can later be adjusted for in analysis. Adjustment may involve tentative correction of mean bias when delay time is known and an analyte is subject to a known and consistent rate of decline/increase in concentration over time. This may take the form of a correction of the primary analysis (where the rate of decline is known with near certainty) or, more often, a sensitivity analysis that can only produce a clear answer if both the primary (unadjusted) analysis and the adjusted analysis generate equivalent conclusions. Where the problem is due to variation in rate of decline/increase in concentration over time between subjects, the key issue is to ensure that statistical power calculations take proper account of the distribution of delay times (particularly, differences between different studies in a meta-analysis). In addition, some adjustment may sometimes be required in meta-analyses to down-weight the influence of studies with greater variability - because processing delay is longer - relative to those with shorter delay.

In the analyses carried out so far, in this thesis, I have investigated the impact on power of errors associated with outcome and explanatory variables in studies involving single nucleotide polymorphisms (SNPs) as the genetic determinants of interest. Some studies are now investigating the potential role of copy number variants (CNVs), another type of polymorphism, in causing disease and influencing complex traits. Some CNVs have already been found to be associated with complex traits. In the next chapter, CNV measurement errors are explored to estimate how well previously discovered CNVs can be measured using different SNP platforms and how this might impact upon the statistical power of association studies involving CNVs. The aim of this work is to help inform the design and analysis of future CNV association studies.

CHAPTER 5

5. ESTIMATING THE ACCURACY OF CNV MEASUREMENTS USINGINTENSITY DATA FROM SNP GENOTYPING PLATFORMS

5.1. INTRODUCTION

It is important to measure accurately the copy number level of an individual at a given copy number variant (see section 1.2.3.2) locus in order to investigate this type of genetic association without bias. CNV genotypes are subject to random measurement errors; if ignored, such errors will result in an underestimation of the true effects of the genotypes on the trait of interest. To understand the true relationship between CNVs and disease it is, hence, crucial to determine accurately the number of copies within individuals.

If a CNV association study aims to use CNV measurements from SNP genotyping platform data (see the paragraph about SNP genotyping under section 1.2.3.1), it can be useful to know how accurately CNVs can be measured from the platforms in order to make a decision about the CNVs to call from that data. If some CNV and platform characteristics are good predictors of the accuracy of CNV measurements, this information could be used by an investigator to choose the SNP platform from which the CNVs investigated in the study can be most accurately measured. If, as is often the case, the platform has already been decided for other reasons, then this knowledge might help to guide whether to proceed with CNV association studies and on whether pooling information across studies might be needed to achieve adequate power for

association testing. This project investigates some CNV and SNP platform characteristics to find out how they relate to the accuracy of CNV calls.

In this analysis, the approach used to estimate the accuracy of CNV measurements from SNP genotyping platform data was to compare copy numbers called from SNP genotyping intensity data with copy numbers called from array-comparative genomic hybridization (aCGH) platform data, using the same set of samples. The aCGH intensity data was considered as the gold standard because aCGH is a reliable technology to detect CNVs and it is currently the most reliable data available in the sample set used in this project for calling CNVs.The SNP intensity data in this study was from the Illumina Infinium 1.2M array, one of the most recent SNP arrays for which probe intensity data was available for the selected set of samples. The results of the comparison between copy numbers from the SNP platform data and from the gold standard, expressed as a correlation between the two, indicates the level of accuracy of the measurements obtained from the SNP platform data.

The correlation was checked against some CNV characteristics to investigate whether those characteristics can help to predict the level of accuracy. The characteristics that are informative about the accuracy of CNV measurements can be used to decide whether or not the intensity data of a specific platform should be used to call certain CNVs. The accuracy of CNV measurements from SNP platforms with different probe densities was also estimated; the results can help to decide whether or not certain CNVs should be called from platforms with comparatively low probe densities. If for example the probe density of the platform does not allow for an adequate level of accuracy of CNV calls, the investigator may decide to use data from a different platform, run new assays to specifically detect the relevant CNV(s) or exclude from his study the CNV(s) that is not well called from the platform of concern. Some SNP platforms contain CNVI

probes, CNV-targeted markers designed using the same strategy as SNP probes(see paragraph on CNV detection and genotyping under section 1.2.3.2), in addition to the SNP probes; the accuracy of CNV calls was estimated using SNP and CNVI probes separately and together. This analysis can also inform the decision about the choice of platform (with or without CNVI probes) that provides a better accuracy of CNV calls. A summary of the accuracy of measurement of a specific CNV using a particular genotyping platform can be used as a key parameter for the ESPRESSO-forte power calculator (chapter2) for a simulation based analysis to determine the impact of copy number accuracy on the statistical power of a study that investigates the association between the CNV and a trait.

This project aims to:

- Estimate how accurately known CNVs can be measured from the Illumina Infinium 1.2M array and determine the proportion of CNV genotype information carried by SNP probes and by CNVI probes.
- Investigate whether key CNV characteristics, including CNV frequency, level of linkage disequilibrium between the relevant CNV and a HapMap SNP, number of CNV classes, type of polymorphism and CNV length, are associated with how well a CNV can be measured.
- Estimate the level of accuracy when copy numbers are measured using data from SNP genotyping platforms older than the 1.2M array.
- Assess the impact of the accuracy of copy number measurement on the power of a CNV association study that uses CNV genotype information derived from a SNP platform analysed in this study.

The results of the analysis can provide useful information for the choice of CNVs to include in a study and inform recommendations for investigators planning to use CNV measurements, derived from SNP platform intensity, in candidate CNV or genome-wide CNV studies.

5.2. METHODS

5.2.1. DATA DESCRIPTION

The two datasets used for the analysis consist of normalised probe intensities from an aCGH platform and normalised probe intensities from an Illumina 1.2M platform. A set of 1284 samples from the 1958 British Birth Cohort (1958BC) (**153**) for which data available on both the aCGH and the 1.2M data, was extracted for the analysis. The set of CNVs selected for the analysis was chosen from the WTCCC CNV study (WTCCC+) (**154**) that was informed by the study of Conrad et al (**17**) which used a custom designed array to detect CNVs (see the paragraph on recent CNV maps under section 1.2.3.2 for a summary of the Conrad study).

5.2.1.1. aCGH data

The aCGH data for 1284 individuals from the 1958BC are from the WTCCC+ study. The WTCCC+ study designed an aCGH experiment to measure copy number variation at 3432 polymorphic CNV loci and study association with eight common diseases. The aCGH dataset contains normalised CNV probe intensities with an average of 10 probes per CNV. The probe intensities were normalized using different methods and the optimum normalization method for each CNV was available from the WTCCC+ study.

For the purpose of this analysis an initial set of 3215 independent CNVs which showed an appropriate clustering (the clusters reflecting different CNV classes) in the WTCCC+ study were selected. The optimum normalization methods for the probe intensities of the selected CNVs were $log_2\left(\frac{R}{G}\right)$, the $\log_2$ ratio of the red channel (test DNA) and green channel (reference DNA) or $log_2\left(\frac{QNorm(R)}{QNorm(G)}\right)$, the transformed $\log_2$ ratio of the quantile normalized red and green channel.

5.2.1.2. Illumina 1.2 data

For the same 1284 individuals, the Illumina 1.2M dataset consists of the x and y intensities, one intensity channel for the A-allele and one for the B-allele, from which genotypes were called as part of the a recent WTCCC genome-wide study of common diseases (WTCCC2), an extension of the WTCCC1 study (**24**). The platform measured a total of 1238733 probes (SNPs and CNVIs) for each sample. Of the initial 3215 CNVs initially selected, 2700 had SNP and/or CNVI probes within their boundaries in the 1.2M platform and were therefore included in further analysis.

5.2.2. CNVTOOLS ALGORITHM

For both aCGH and Illumina 1.2M, the copy number status of each of the individuals in the sample set was determined using CNVtools (**155**), an algorithm designed to analyse CNV array data. CNVtools takes as input a one-dimensional signal, for each sample, obtained by summarizing the intensities across all the probes within the CNV, and fits a mixture model to the one-dimensional data.

5.2.2.1. Summary methods

The first paragraph of this section explains how summary methods were used in the WTCCC+ study to summarise probe intensities and the second paragraph details how the summary was carried out in this analysis.

Summary methods in WTCCC+

Several different approaches to summarising information across probes within a CNV were compared. The intensities of the probes within each CNV were summarised using first the mean summary method in which the single value for each sample was the statistical mean of the probe intensities. Alternatively, the intensities were summarised using principal component analysis (PCA) which transforms the data and attenuates the effect of outliers (the mean method is more sensitive to outliers that cause the average value to be inflated or deflated). The summary method that allowed for the best clustering of the signal was chosen. After summarising the data by mean or PCA, the clustering was refined by linear discriminant analysis to find a linear transformation called linear discriminant function (LDF) that transforms the predictors (the one-dimensional signal values) of the copy numbers into values that allow for a better segregation between CNV classes.

Summary methods used for this analysis

The optimal summary method (mean, mean +LDF, PCA, PCA+LDF) for each of the 2700 CNVs was known for the aCGH data from the WTCCC+ study. I applied the same summary method for each CNV to the 1.2M probe intensities. Since any signal that is a proxy for the number of copies can be used to summarize probe intensities (**155**), I combined the x and y intensities of the CNV and SNP probes in the Illumina 1.2M dataset by the sum of x and y to obtain one value for each probe and used the predefined summary method (mean or PCA) from the WTCCC+ study to summarise the

information across probes within the CNV boundaries. I systematically applied LDF to the output of the predefined summary method to optimize the clustering for both the aCGH and the 1.2M data.

To fit the model appropriately, the algorithm requires the different classes of the CNV to be distinct i.e. to form separated clusters, with no or limited overlap, on the histogram of the one-dimensional signal (see histogram in Figure 24); if the clustering is not efficient the mixed model will not fit correctly.

5.2.2.2. CNVtools parameters

To fit a mixture of gaussians to the summarised data CNVtools requires (1) a vector of numeric values (*start.mean*) to set the starting values for the means in the likelihood process, (2) a formula (*model.mean*) that describes the linear model for the location of the mean signal intensity (a linear model with free means or a linear model with means proportional to the number of copies), (3) a formula (*model.var*) similar to *model.mean* but used to model the variances, (4) an integer for the number of CNV classes and (5) an integer for the number of iterations to maximise the likelihood.

In the WTCCC+ study, some CNVs could not be appropriately fitted at the first attempt using a random *start.mean* value; for those CNVs the WTCCC+ study determined the optimal *start.mean*. For the aCGH data, I used the *start.mean* values available from the WTCCC+ study to set the starting values for the CNVs, where required (i.e. where the model could not be successfully fitted with random *start.mean* values). For the 1.2M data, I allowed CNVtools to pick the starting points randomly. The number of copy number classes I fitted in CNVtools, for each CNV, was available from the WTCCC+ study. For the aCGH data, I used the optimal *mean.model* values where available from the WTCCC+ paper. When information about the *mean.model* parameter was not

available from WTCCC+, I used an iterative approach to determine the values of *model.mean* and *model.var*, for aCGH, that gave the highest correlation between aCGH and 1.2M copy numbers. To estimate copy numbers from the 1.2M intensity data, I used the same *model.mean* and *model.var* values that I used to estimate copy numbers from the aCGH intensity data, hence assuming that the optimal parameters are the same for the aCGH and the 1.2M data.

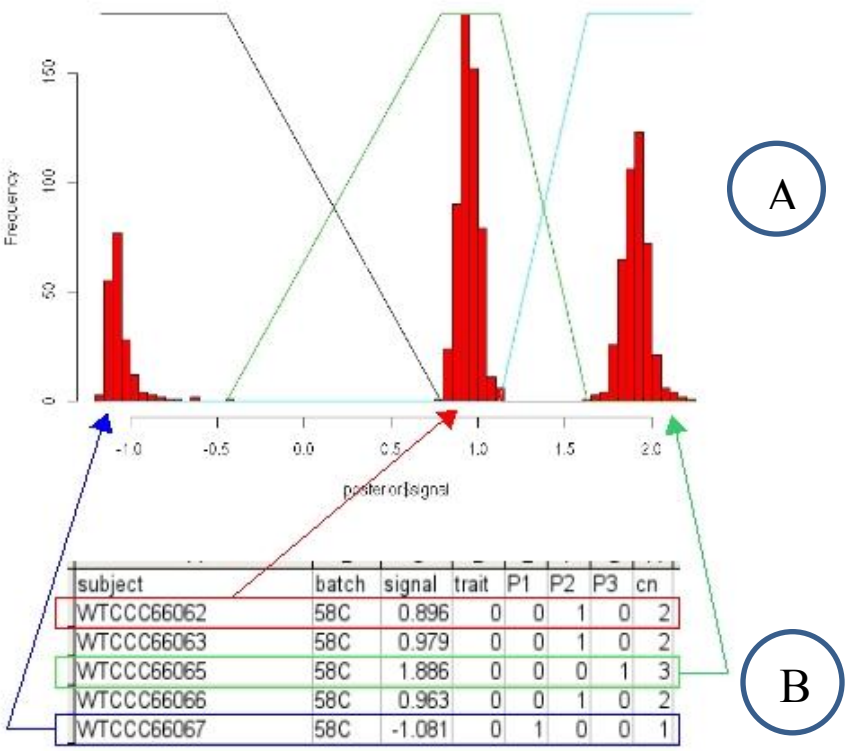
5.2.2.3. Convergence of the model

CNVtools uses an expectation-maximisation (EM) approach to estimate the maximum likelihood (**155**). After fitting a mixture model it is possible to verify the status of the fit to find out if the model converged. For each CNV, and for both aCGH and 1.2M, I specified 5 EM iterations to maximize the likelihood. For aCGH, for each CNV I checked the status of the fitted model and only considered CNV calls where the model converged because failure to reach convergence could indicate underestimated or overestimated number of clusters, constrained/relaxed values for *model.mean* and *model.var*, or presence of one or more outliers that cause the posterior probability to rise again after falling to zero (**155**). Since there was a risk that the model converges to a local maximum which did not represent the optimal solution, convergence was also confirmed by visual examination to ensure there was no discrepancy of the posterior probabilities of the CNV classes.

5.2.2.4. Obtaining copy numbers

The copy numbers were determined separately for the samples using the aCGH and the Illumina 1.2M data. The copy number status of an individual was determined by calculating the posterior probability of belonging to each CNV class; the highest

posterior probability indicated the most likely copy number of an individual. Figure 24 illustrates an analysis where the clustering was unambiguous; the CNV was fitted in CNVtools with 3 classes.



subject	batch	signal	trait	P1	P2	P3	cn
WTCCC66062	58C	0.896	0	0	1	0	2
WTCCC66063	58C	0.979	0	0	1	0	2
WTCCC66065	58C	1.886	0	0	0	1	3
WTCCC66066	58C	0.963	0	0	1	0	2
WTCCC66067	58C	-1.081	0	1	0	0	1

Figure 24: Example of an unambiguous clustering generated by CNVtools. Three individuals (in blue, red and green) were, respectively, assigned to copy number classes 1, 2 and 3 represented by the three clusters on the histogram (A). The coloured lines on the histogram plot represent the posterior probability for each of the three copy number classes. Columns 'P1', 'P2' and 'P3' in the table (B), hold the posterior probabilities of belonging to copy number class 1, 2 or 3. The column 'signal' represents the one-dimensional signal where each value represents a summary of the probe intensities for one individual (see section 5.2.2.1 on how probe intensities are summarized). The column 'cn' holds the copy numbers.

5.2.3. EVALUATING THE ACCURACY OF COPY NUMBER CALLS FROM ILLUMINA 1.2M PLATFORM

The accuracy of copy numbers called from the 1.2M data was assessed by comparing copy numbers called from the 1.2M data with those called from the gold standard

(aCGH) data; the correlation between the copy number calls from the aCGH and the 1.2M platforms was used to represent the level of accuracy of copy numbers called from 1.2M data. For each CNV, the correlation between two vectors (one for each platform) of n copy number values (where is the n is the number of individuals) was computed.

I carried out sample quality control (QC) checks prior to running the CNVtools analysis. Samples were excluded if they met one or more of the following four exclusion criteria, any sample excluded from one dataset (aCGH or 1.2M) was also excluded from the other dataset:

1) Derivative log ratio spread (DLRS) greater than 0.35 (DLRS is a measure of the variability of $\log_2$[R/G] across all probes), irrespective of which normalisation was chosen. This criterion was applied to the aCGH dataset.
2) Low signal intensity in the aCGH data (signal intensity < 100 for either the red or the green channel).This criterion was applied to the aCGH dataset.
3) Standard deviation of the x and y intensities, across all probes located in autosomal chromosomes, greater than 0.65. For each sample in the 1.2M dataset, the standard deviations of the x and y intensities, across all probes located in autosomal chromosomes, were calculated; a large standard deviation indicates poor genotyping quality. This criterion was applied to the 1.2M dataset.
4) Presence in the WTCCC2 study samples exclusion list (samples in that list were excluded from WTCCC2 for meeting QC criteria including missing heterozygosity, related individuals, ethnic outliers, unknown identity, outlying mean x or y intensity); since the 1.2M intensity data used in this analysis are from the WTCCC2 study it is reasonable to exclude samples that were excluded from WTCCC2. This criterion was applied to the 1.2M dataset.

After the CNVtools analysis and before calculating the correlation between 1.2M and aCGH copy numbers, CNV QC checks were carried out on the 2700 selected set of CNVs and CNVs were excluded if they met one or more of the following three exclusion criteria:

1) The fitted model did not converge.
2) The clustering failed (cluster plots were visually inspected).
3) More than 5% of the individuals had missing intensities at one or more probes within the CNV boundaries.

The first two QC filters (non-converged model and failed clustering) were applied only to CNV measurements from the aCGH data because a failure to converge or to cluster, on the 1.2M data, may just indicate that the CNV is not "callable" from the 1.2M data. The CNVs that passed the two first QC filters are those for which there was a gold standard set of copy number measurements i.e. it is assumed that the CNVs were measured correctly from the aCGH data.

The correlation between 1.2M and aCGH copy numbers was calculated using SNP probes only, CNVI probes only and finally both SNP and CNVI probes combined, from the 1.2M data to assess whether one type of probe (SNP or CNVI) carries more copy number genotype information, in the 1.2M data. The correlation coefficients were calculated before and after applying LDF to the optimal summary. Where applying LDF did not improve the correlation, the correlation coefficient obtained prior to LDF was reported.

5.2.4. EVALUATING THE EFFECT OF CNV CHARACTERISTICS ON THE ACCURACY OF COPY NUMBER CALLS FROM ILLUMINA 1.2M PLATFORM

The influence of some CNV characteristics, including minor allele frequency (MAF), level of linkage disequilibrium with a HapMap SNP (**11**), number of copy number classes of the CNV, type of variation (deletion, duplication, deletion and duplication) and CNV size, on the accuracy of copy numbers called from the Illumina 1.2M data, was investigated.

Consistent with the WTCCC+ study (**154**), I used the copy numbers called from the aCGH intensity data to calculate the minor allele frequency of CNVs called with 2 or 3 classes using the formula $\text{MAF} = \frac{2n_0+n_1}{2n}$ where n_0 and n_1 are the genotype counts for the rare homozygote and heterozygote respectively and n the total genotype count. The linkage disequilibrium values I used in the analysis were available from the WTCCC+ study which calculated linkage disequilibrium as the Pearson r^2 value between CNV genotypes and SNP genotypes using samples from the WTCCC+ study itself and samples from previous WTCCC studies (**154**). For each CNV, the number of classes I used in this analysis was from the WTCCC+ which determined CNV classes after a visual inspection of the histogram of probe intensities. The information about CNV type that I used was available from WTCCC+; the type of variations were reported by Conrad et al. (**17**) who classified CNVs with a diploid copy number of 0, 1 or 2 as deletions, CNVs with a diploid copy number of 2, 3 or 4 as duplications and CNVs with more than 3 possible diploid copy numbers as multiallelic (CNVs involving duplications and deletions).

5.2.5. EVALUATING THE EFFECT OF EARLIER VERSUS NEWER GENERATION SNP GENOTYPING PLATFORM ON THE ACCURACY OF COPY NUMBER CALLS

The probe intensities of the 1958 British Birth Cohort (58C) samples used in the first comparison (aCGH vs. 1.2M) were not available for Illumina 660 (ILM660K) and 610 (ILM610K), two platforms that contain SNP and CNVI probes, and for Illumina 300 (ILM300K), a platform without CNVI probes. The 58C intensity data from the 1.2M platform was used to generate datasets that closely resembled intensity data from ILM660K, ILM610K and ILM300K by selecting subsets of probes from the 1.2M which were known to feature on the ILM660K, ILM610K and ILM300K. For example, for ILM660K I identified 1.2M probes not present in the ILM660K platform and excluded them. A similar approach was taken to generate datasets that mimicked ILM610K and ILM300K platforms. This strategy assumes that the three platforms differ only on probe density (the number of probes on a genotyping array); the implication of this assumption will be discussed later.

The accuracy of copy numbers measured from ILM660K, ILM610K, and ILM300K was evaluated using the same strategy as for 1.2M. The correlation between copy numbers measured from the gold standard (aCGH) and those measured from ILM660K, ILM610K and ILM300K reflects the level of accuracy of copy number calls obtained using probe intensities from the three SNP genotyping platforms. The accuracy was evaluated using SNP and CNVI probes combined and using SNP probes only where CNVI probes were not present. The levels of accuracy of copy number called from 1.2M, ILM660K, ILM610K and ILM300K were compared.

5.2.6. IMPACT OF COPY NUMBER CALLINACCURACY ON THE POWER OF AN ASSOCIATION STUDY

This part of the analysis was an illustration of how the estimated level of accuracy of CNV calls can be used in ESPRESSO-forte (chapter2), developed as part of this thesis, to evaluate the effect of copy number measurement errors on the statistical power of a candidate CNV association study that investigates the association between a CNV measured on the 1.2M and a hypothetical binary outcome.

The accuracy of CNV calls was assessed as the level of correlation between copy numbers called from 1.2M intensity data and copy numbers called from the aCGH intensity data (gold standard). The maximum accuracy, which corresponds to a correlation r = 1, denotes the absence of measurement error in copy numbers called from the 1.2M data; the largest measurement error was obtained for the lowest accuracy (r = 0). The ESPRESSO-forte algorithm uses assessment error in the outcome and in the genetic determinant (genotype) to calculate (1) the power that can be achieved under the specified sample size and (2) the sample size required to achieve the specified desired level of power. In this analysis, the estimated measurement error of one of the CNVs analysed in this project (CNVR5101.1) was transferred to ESPRESSO-forte to evaluate its effect on power.

In ESPRESSO-forte the genetic determinant is a SNP; CNVR5101.1 is a biallelic CNV and its analysis is hence similar to the analysis of a biallelic SNP under a binary genetic model, so the analysis did not require major changes in the algorithm. CNVR5101.1 consisted of a single deletion; the genotypes were coded as 0 and 1, where 1 is the 'at risk' genotype (one deletion) and 0 the common homozygous (no deletion). The user function '*sim.geno.sesp'* that I wrote as part of the ESPRESSO-forte R package

(documented in page 279 of the algorithm's manual under Appendix 1) was used to calculate the sensitivity and specificity that corresponds to the accuracy (correlation r between 1.2M and aCGH genotypes) that was estimated for CNVR5101.1. This function calculates the sensitivity and specificity required to generate the squared correlation between two simulated binary vectors where one vector represents the 'true' (error free) measurements and the other vector represents the 'observed' ('true' measurement + error) measurements. In this analysis, the sensitivity and specificity values reflect the incomplete correlation between copy numbers called from 1.2M and those called from the gold standard. These sensitivity and specificity values were used by ESPRESSO-forte to calculate the misclassification rates in the two binary vectors that represent the alleles used to construct the observed genotypes (as explained in section 2.6.2.5).

The power achieved for a hypothetical binary outcome determined by CNVR5101.1 and the sample size required to achieve a power of 80% were calculated under three scenarios:

(1) In the first scenario, it was assumed that CNVR5101.1 was measured without error (r = 1) and the outcome was also assumed to be measured without error (sensitivity=1 and sensitivity=1).
(2) In the second scenario, CNVR5101.1 was fitted with its actual measurement error as evaluated in this chapter (the estimated level of correlation between 1.2M and gold standard copy numbers for this CNV) whilst the outcome was still assumed to be measured without error.
(3) In the third scenario, both the outcome and the genetic determinant (CNVR5101.1) were assumed to be measured with an error. The error on CNVR5101.1 was the same as the one mentioned in the second scenario and

the sensitivity and specificity of the outcome were set arbitrarily to 0.72 and 0.99 respectively because these values are reasonable.

The difference between the power values achieved under scenario 1 and scenario 2 indicated the amount of power loss caused by the measurement error on CNVR5101.1 (since the outcome was assumed to be assessed without error, the power loss observed was solely due to measurement error on CNVR5101.1). The difference between the power values achieved under scenario 1 and scenario 3 indicated the amount of power loss if both the outcome and the genetic determinant are measured with some error.

Likewise the sample size required to achieve a power of 80% under the first scenario was compared to those required to achieve the same level of power under the second and third scenario. The difference indicated the increase in sample size required to compensate for the power loss caused by the measurement error on CNVR5101.1 genotype (scenario 1 vs. scenario2) and the increase in sample size required to compensate for the power loss if both the outcome and the determinant are measured with some error (scenario 1 vs. scenario 3).

5.3. RESULTS

5.3.1. SAMPLE QUALITY CONTROL

This paragraph is a summary of samples excluded from the analysis for meeting one or more of the four QC criteria in the second paragraph of section 5.2.3. Any sample excluded from one of the two datasets (aCGH or 1.2M) was also excluded from the other.

1) Seven samples with a DLRS > 0.35 were excluded from the aCGH dataset and not replaced by their replicate because the replicate did not exist or because the DLRS of the replicate was not greater than 0.30.

2) Two samples were excluded from the aCGH dataset for low signal intensity.

3) Ten samples were excluded from the 1.2M dataset because the standard deviation of the x and/or y intensity was greater than 0.65 (Figure 25).

4) A total of 80 samples excluded from the WTCCC2 study were excluded from the 1.2M dataset.

Of the initial 1284 samples, 99 were excluded; the analysis was carried out with the remaining 1185 samples.

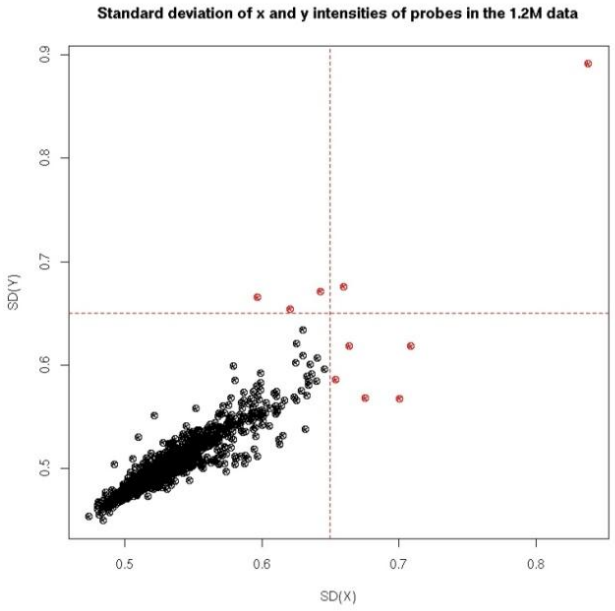


Figure 25: Standard deviations of x and y intensities from Illumina 1.2M data. In red, the 10 samples excluded for x or y intensity greater than 0.65.

5.3.2. CNV QUALITY CONTROL

Figure 26 is a summary of CNV exclusions, from the aCGH dataset, according to the three CNV quality control criteria listed in the third paragraph of section 5.2.3. For 605 CNVs called from the aCGH data, the fitted model did not converge and the CNVs were excluded (step 1 in Figure 26); CNVs with 2 copy number classes were overrepresented in this exclusion list, they represented approximately 56% of the CNVs that did not converge whilst only 36% of all CNVs had 2 classes. All the CNVs with 2 classes excluded for no convergence were rare (MAF < 0.05). A total of 485 CNVs were excluded for failing to cluster (step 2 in Figure 26). These CNVs were excluded after visual inspection of the posterior plots which for the majority showed that all the samples were assigned to one copy number class or that the posterior lines did not follow the contours of the clusters. The 1610 CNVs which passed the two QC filters mentioned above are the ones for which there was a gold standard (highlighted in red on Figure 26).

For a total of113 CNVs, the number of samples with missing intensity at one or more probes represented more than 5% of all 1185 samples (step 3 in Figure 26); this was observed only in the 1.2M data, none of the probes in the aCGH data presented missing intensities. These 113 CNVs were assigned a correlation coefficient of zero which means that they were considered as not “callable” from the 1.2M data.

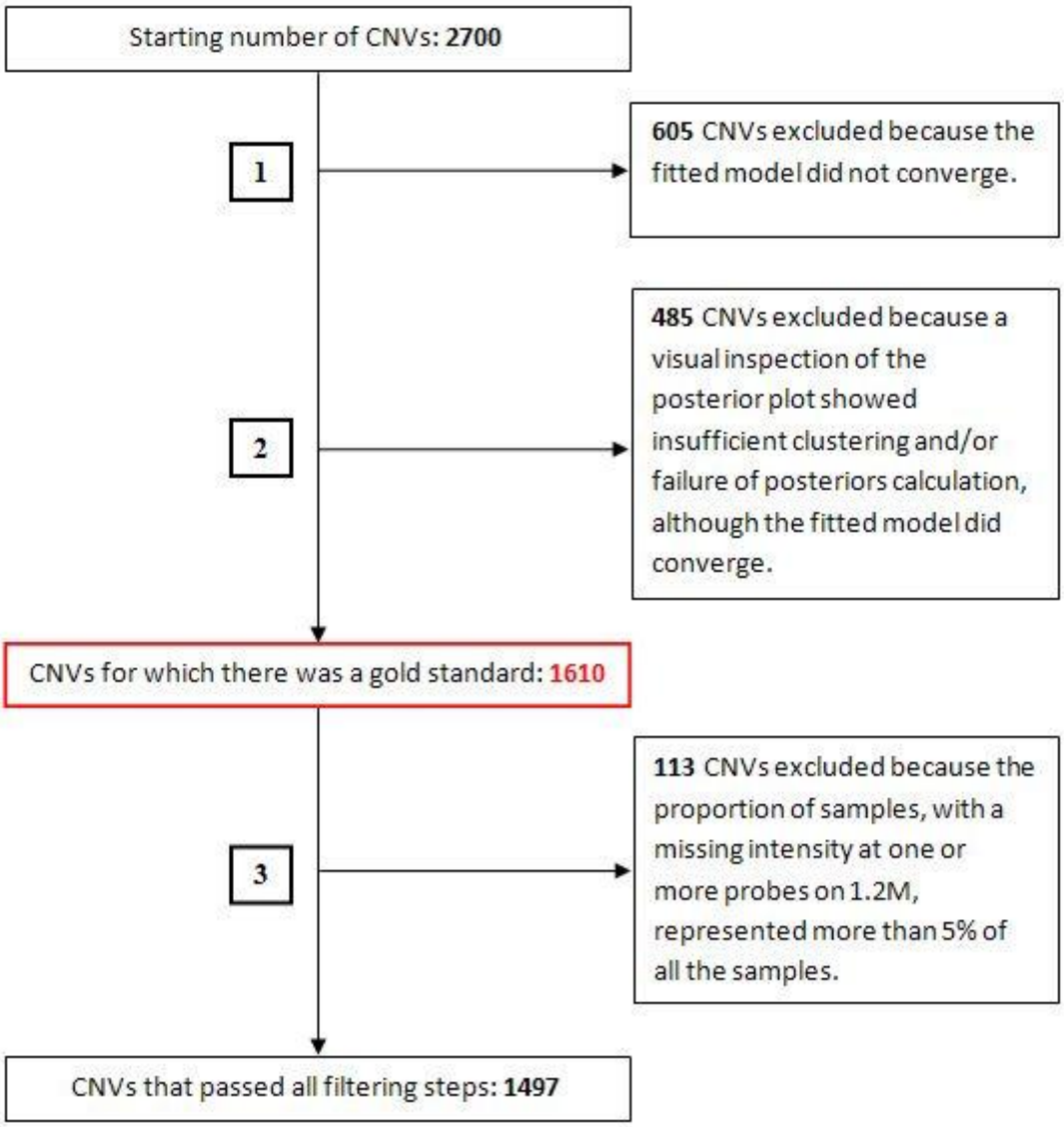


Figure 26: Summary of CNV exclusions.
This chart shows the number of CNVs for which clustering and copy number assignment went well for aCGH, highlighted in red, and the number of CNVs that passed all CNV QC checks.

5.3.3. ACCURACY OF CNV CALLS FROM THE ILLUMINA 1.2M DATA

Table 87 reports the number of CNVs by levels of correlation between copy numbers measured from the aCGH and 1.2M intensity data; these correlations represent the levels of accuracy of copy number calls from the 1.2M data. A proportion of 49% of

the 1610 CNVs tested had a correlation equal to or greater than 0.8. Figure 27 shows the distribution of correlation coefficients.

Correlation (r)	**≥ 0.8**	**0.5 ≤ r < 0.8**	**< 0.5**
Number of CNVs	791 (49%)	231	588

Table 87: CNV counts by level of accuracy.
The accuracy of copy number calls was assessed as the correlation between copy numbers from aCGH and 1.2M. The percentage represents the corresponding proportion in the 1610 CNVs tested.

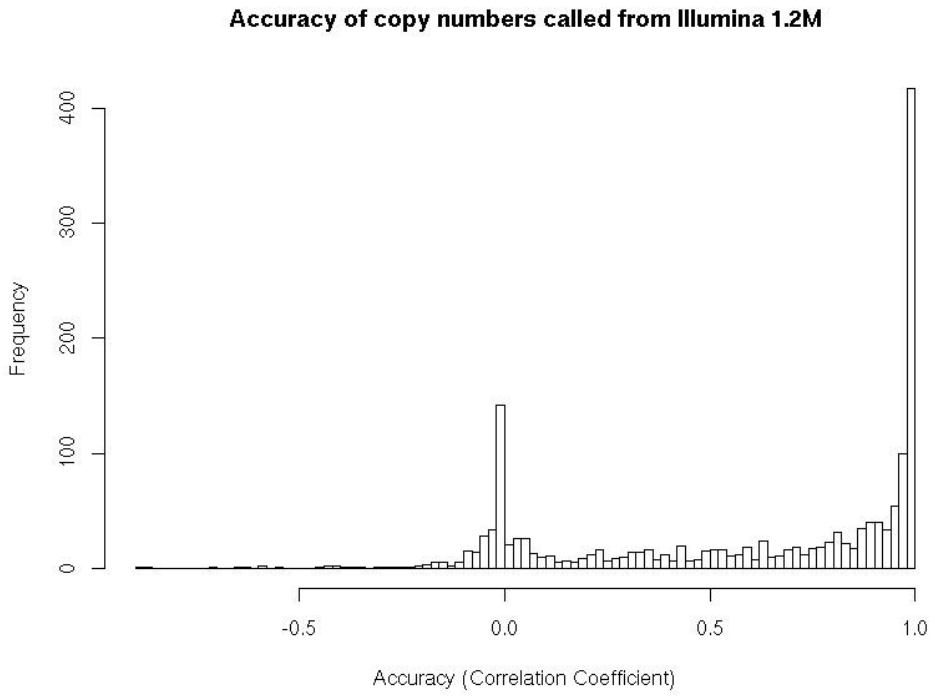


Figure 27: Distribution of levels of accuracy when SNPs and CNVI were combined.
The number of CNVs called with no accuracy (r=0) was inflated because of the CNVs that were assigned a correlation of 0 when the number of samples with missing intensity at one or more probe represented more than 5% of all 1185 samples.

Most of the correlations reported in Table 87 were positive; however, there were 46 CNVs with a negative correlation of less than -0.1. It is reasonable to expect some correlations of up to -0.3 to occur by chance but some CNVs had very high negative correlations; one of such CNVs is CNVR551.3 ($r = -0.9$). For this CNV and the other 4 that showed a high negative correlation, a large number of individuals assigned a copy

number of 1 in the aCGH data were assigned a copy number of 3 in the 1.2M data and vice-versa. To further investigate this apparent flip between copy numbers of 3 and 1, the CNV calling was repeated for CNVR551.3 using each probe separately (rather than by summarising across multiple probes).

CNVR551.3 had 9 probes (5 SNPs and 4 CNVIs) within its boundaries on the 1.2M data and 10 probes on the aCGH data. First a CNVtools analysis was carried out using the probes in the 1.2M dataset one by one and then calculating the correlation with aCGH calls (which were generated using all the probes within the boundaries of the CNV). Then another analysis was carried using the aCGH probes one by one whilst using all the probes on the 1.2M to generate the 1.2M calls. The results in Table 88 indicate that depending on the probe included some individuals were assigned to the same class on both aCGH and 1.2M and a positive correlation was calculated or they were assigned to different copy number classes (to class 1 in aCGH and class 3 in 1.2M or the other way round) and a negative correlation was then calculated. Since negative correlations were observed even when the probe intensities were not summarised, this shows that the reversing of the order of copy number classes (negative correlation) was not attributable to the summary methods.

1.2M		**aCGH**	
Probes	**Correlation (r)**	**Probes**	**Correlation (r)**
cnvi0068801	-0.89	7278	0.12
cnvi0104294	-0.78	7279	0.51
rs11807244	-0.89	7280	0.1
rs11807542	0.74	7281	-0.16
rs12758021	-0.38	7282	-0.2
rs12759359	0.67	7283	0.02
rs16833833	-0.8	7284	-0.88
rs28444730	-0.45	7285	-0.87
rs28704525	-0.9	7286	0.77
		7287	-0.08

Table 88: Level of accuracy of copy number calls achieved using each probe alone.

5.3.3.1. Influence of CNV size on the accuracy of copy number calls

Forty eight percent of the 1610 CNVs tested were shorter than 3kb; 41% were between 3 and 22kb and 11% were greater than 22kb in length. The proportion of CNVs with a correlation ≥ 0.8between aCGH and 1.2M CNV measures in each of the three size categories, in Table 89, decreases with increasing CNV size. The results suggest that the copy number status of shorter CNVs is measured with a higher accuracy and that the accuracy decreases with increasing CNV length as shown by the gradient of the line of best fit on Figure 28. However, for any given CNV size, there was considerable variation in the accuracy of calling and 16% of CNVs of size ≥22kb could still be called accurately (r≥ 0.8).

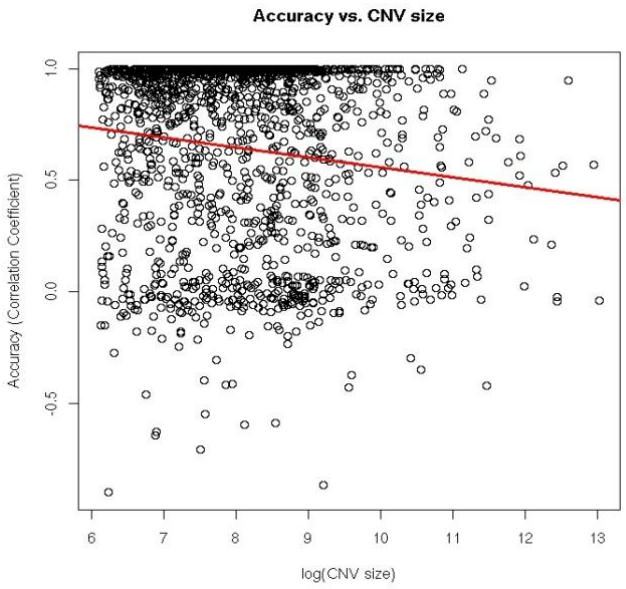


Figure 28: Plot of accuracy versus CNV size.
The line of best fit is in red.

Correlation (r)	size <3kb	3kb ≤ Size < 22kb	Size ≥ 22kb
≥ 0.8	444 (58%)	312 (47%)	35 (16%)
0.5 ≤ r < 0.8	113	90	28
< 0.5	212	260	116
Total	769	662	179

Table 89: CNV count by CNV size and level of accuracy.
The accuracy of copy number calls was assessed as the correlation between copy numbers from aCGH and 1.2M. The percentages represent the corresponding proportions in the number of CNVs within the size category.

5.3.3.2. Influence of the number and type of probes on the accuracy of copy number calls

Number of probes versus accuracy of copy number calls

Figure 29 shows the relationship between the number of probes within CNV boundaries in the 1.2M data and the levels of accuracy for the three CNV length categories in Table 89. The plots in Figure 29 were stratified by the three CNV size categories to take into account the fact that the number of probes may be partly a function of the size of the CNV. The graphs indicated that: There was no strong correlation between number of probes and accuracy for any CNV length category although the accuracy of calls increased very slightly for CNVs shorter than 3kb, and decreased very slightly for CNVs of length greater than 22kb. Similar trends were observed when finer intervals were chosen to stratify the CNV size characteristic (see Figure 38 under Appendix 2).

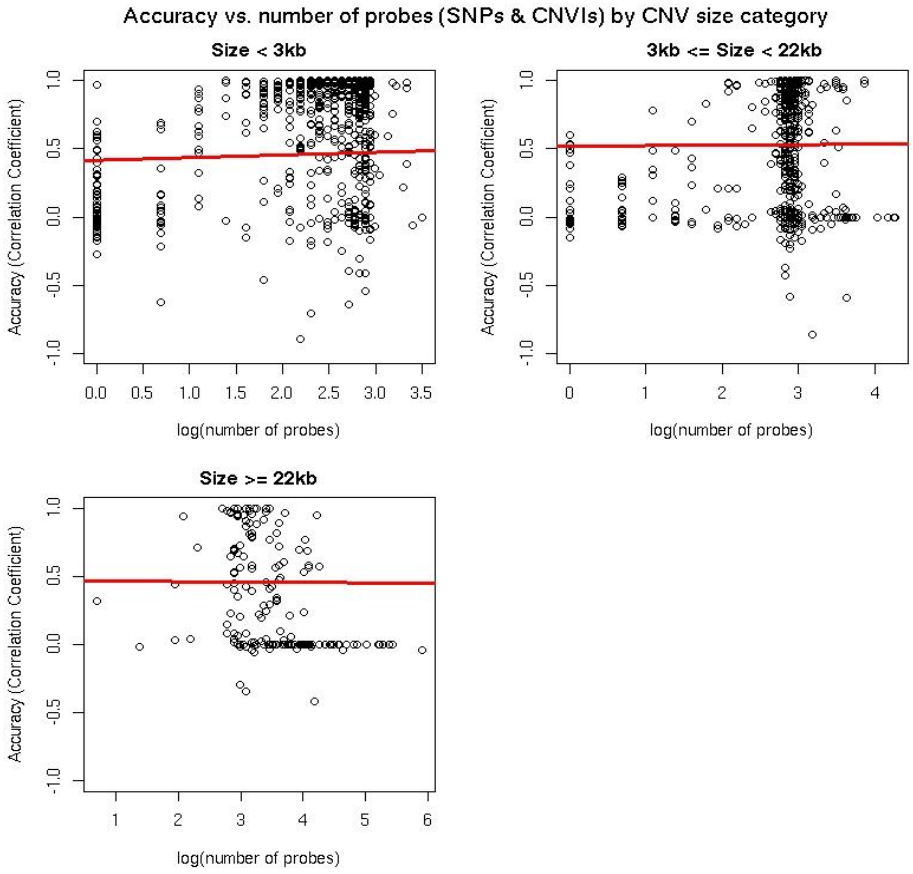


Figure 29: Plots of number of probes against accuracy.
The accuracy of copy number calls was assessed as the correlation between copy numbers from aCGH and 1.2M for four CNV size categories. The red line across each plot represents the line of best fit.

Type of probes (SNP or CNVI) versus accuracy of copy number calls

A total of 1276 CNVs had both SNP and CNVI probes within their boundaries on the Illumina 1.2M platform. The histograms in Figure 30 show similar distributions of accuracy whether SNP probes only or CNVI probes only were analysed. The proportion of CNVs with a correlation ≥ 0.8 was 49% for CNVIs probes only and 42% for SNP probes only(Table 90). This seems to indicate that copy numbers were measured slightly more accurately with CNVI probes alone than with SNP probes alone on the 1.2M platform. When both types of probes were combined the accuracy was further improved (52% of the CNVs tested had an r ≥ 0.8).

Correlation (r)	SNPs only	CNVIs only	SNPs & CNVIs
≥ 0.8	536 (42%)	619 (49%)	660 (52%)
0.5 ≤ r < 0.8	228	174	190
< 0.5	512	483	426
Total	1276	1276	1276

Table 90: CNV count by type of probes and level of accuracy.
The accuracy of copy number calls was assessed as the correlation between copy numbers from aCGH and 1.2M. The percentages represent the corresponding proportions in the 1276 CNVs tested.

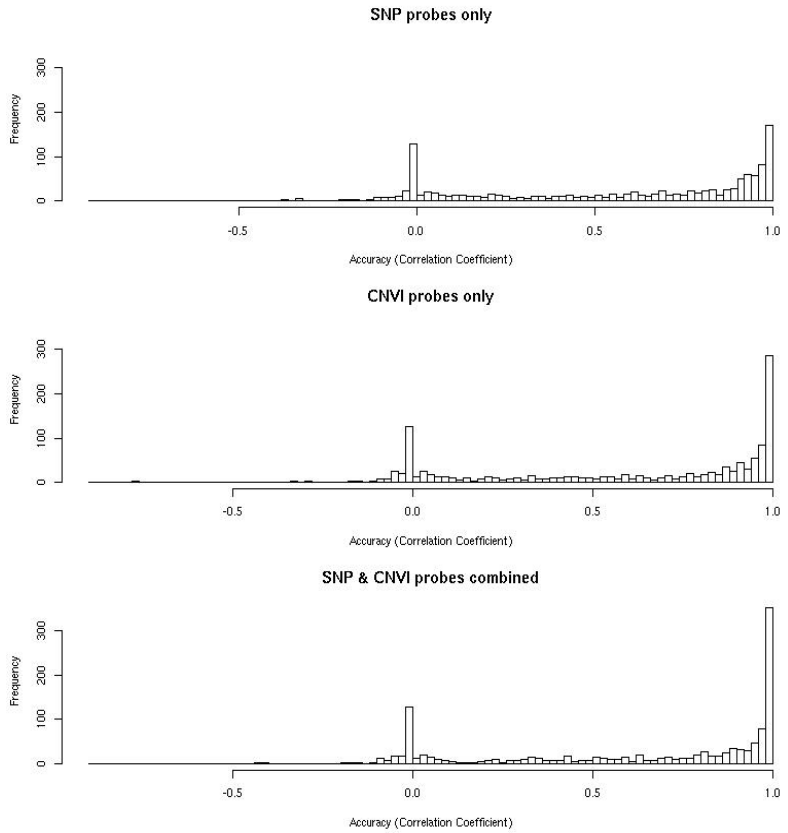


Figure 30: Distribution of accuracy by type of probe.
The above histograms show the distribution of CNV calls accuracy when 1.2M SNP and CNVI probes were analysed separately or together for the 1276 CNVs which had both SNP and CNVI probes within their boundaries on the 1.2M platform.

5.3.3.3. Influence of MAF on the accuracy of copy number calls

Table 91 reports the number of CNVs by level of accuracy of copy number calls and minor allele frequency (calculated as described in 5.2.4). Twenty seven percent of the 1543 CNVs which had 2 or 3 copy number classes, were rare (MAF < 0.05) and 73%

were common CNVs (MAF ≥ 0.05). The difference between the accuracy of CNV measurement for common and rare CNVs is shown in Figure 31, where the mean and median accuracy (r) of the common CNVs were above 0.8 whilst both the mean and the median were below 0.5 for the rare CNVs. Furthermore 58% of the common CNVs had a correlation ≥ 0.8 compared to only 28% for the rare CNVs. Similar observations to those above were made when the common CNVs (MAF ≥ 0.05) were divided into 3 intervals (see Figure 39 and Table 100 under Appendix 3): the results indicated a better accuracy for the more frequent CNVs.

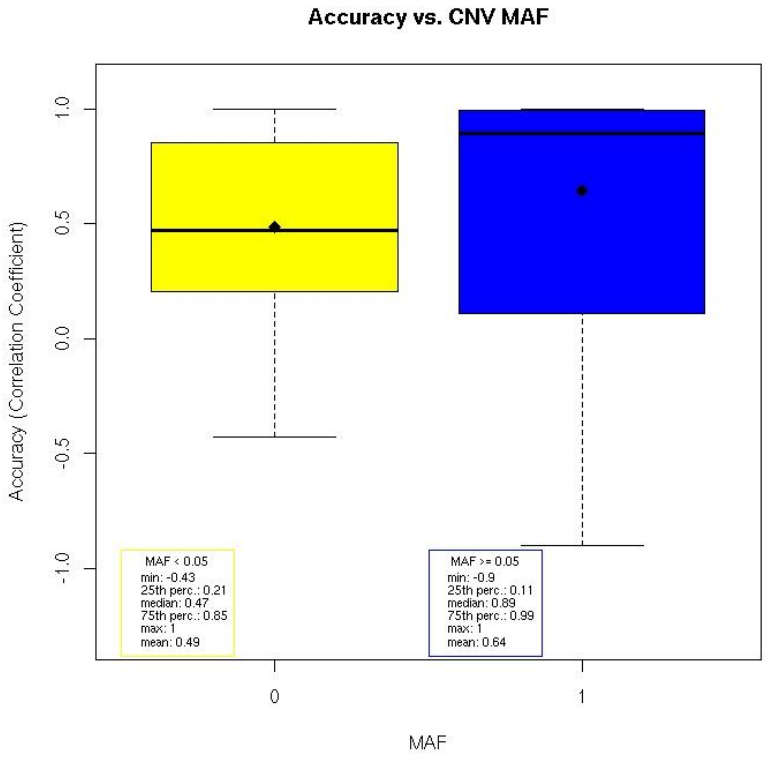


Figure 31: Plot of accuracy of copy number calls by CNV MAF.
The black lines on the boxplots are the medians and the black diamonds represent the statistical means. The extremities of the whiskers represent the minimum and maximum values. The corresponding summary statistics are shown.

Correlation (r)	MAF < 0.05	MAF ≥ 0.05
≥ 0.8	115 (28%)	657 (58%)
0.5 ≤ r < 0.8	83	125
< 0.5	215	348
Total	413	1130

Table 91: CNV count by MAF and level of accuracy.
The accuracy of copy number calls was assessed as the correlation between copy numbers from aCGH and 1.2M. The percentages represent the corresponding proportions in the MAF category.

5.3.3.4. Influence of CNV type on the accuracy of copy number calls

Fifty nine percent of the 1508 CNVs, for which information about the type of variation was available, were deletions, 28% were duplications and 13% consisted of variations involving duplication and deletion. Table 92 reports the accuracy of CNV calls by CNV type: 72% of the CNVs involving duplication and deletion, 48% of the duplications and 49% of the deletions had an accuracy ≥ 0.8; this observation was shown graphically in Figure 32 which shows that the mean and median correlations were greater than 0.5 for all 3 types of CNV tested. The overall level of accuracy was roughly the same for duplications and deletions and higher for duplication/deletion type CNVs. However when the relationship between accuracy of CNVs calls and CNV type was stratified by the number of CNV classes (Table 93), the results showed that CNVs involving duplications and deletion were less accurately measured than simple duplications and deletions when the number of CNV classes was greater than 3.

Correlation (r)	Duplication	Duplication/Deletion	Deletion
≥ 0.8	200 (48%)	144 (72%)	433 (49%)
0.5 ≤ r < 0.8	53	20	141
< 0.5	168	37	312
Total	421	201	886

Table 92: CNV count by CNV type and level of accuracy.
The accuracy of copy number calls was assessed as the correlation between copy numbers from aCGH and 1.2M. The percentages represent the corresponding proportions in each CNV type.

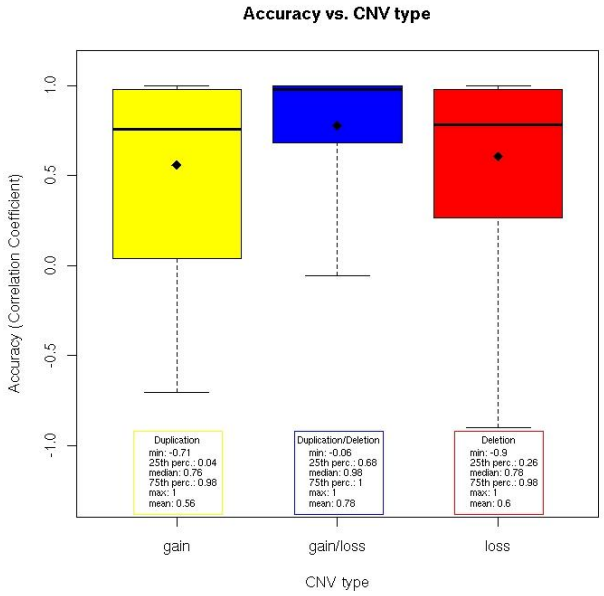


Figure 32: Plot of accuracy of copy number calls by CNV type.
The black lines on the boxplots are the medians and the black diamonds represent the statistical means. The extremities of the whiskers represent the minimum and maximum values. The corresponding summary statistics are shown.

Copy number classes	CNV Type	r ≥ 0.8	r < 0.8	Total count
2	Duplication	36 (32 %)	76 (68 %)	112
	Duplication/Deletion	6 (38 %)	10 (62 %)	16
	Deletion	89 (25 %)	269 (75 %)	358
				486
3	Duplication	159 (56 %)	127 (44 %)	286
	Duplication/Deletion	136 (80 %)	34 (20 %)	170
	Deletion	332 (67 %)	167 (33 %)	499
				955
> 3	Duplication	5 (22 %)	18 (78 %)	23
	Duplication/Deletion	2 (13 %)	13 (87 %)	15
	Deletion	12 (41 %)	17 (59 %)	29
				67

Table 93: CNV counts by accuracy and CNV type stratified by number of CNV copy number classes.
*The percentages show the proportion of CNVs called with high accuracy(**r ≥ 0.8**) and low accuracy (r<0.8) for each CNV type within each copy number class category.*

5.3.3.5. Influence of number of CNV copy number classes on the accuracy of copy number calls

Thirty six percent of the 1610 CNVs analysed had two copy number classes, 60% had three classes and 4% had more than 3 classes. Sixty-five percent of the CNVs with3copy number classes had a high accuracy of calling on the 1.2M platform (r≥ 0.8); 25% of CNVs with two copy number classes had a high accuracy and 28% of CNVs with a number of copy number classes greater than 3 had a high accuracy. Figure 33 shows the difference in level of accuracy between CNVs by number of classes. The level of accuracy in 3-class CNVs (mean and median > 0.75) was markedly higher than that of CNVs with more than 3 classes; 2-class CNVs were measured with the lowest accuracy (69% of the 2-class CNVs were rare). The mean correlation was0.51 for CNVs with a number of classes greater than 3, but these CNVs represent only 4% of the 1610 CNVs tested (Table 94).

Correlation (r)	2 Classes	3 Classes	> 3 Classes
≥ 0.8	143 (25%)	629 (65%)	19 (28%)
0.5 ≤ r < 0.8	98	110	23
< 0.5	337	226	25
Total	578	965	67

Table 94: CNV count by number of copy number classes and level of accuracy. The accuracy of copy number calls was assessed as the correlation between copy numbers called from aCGH and 1.2M. The percentages represent the corresponding proportions in each category of CNV copy number class.

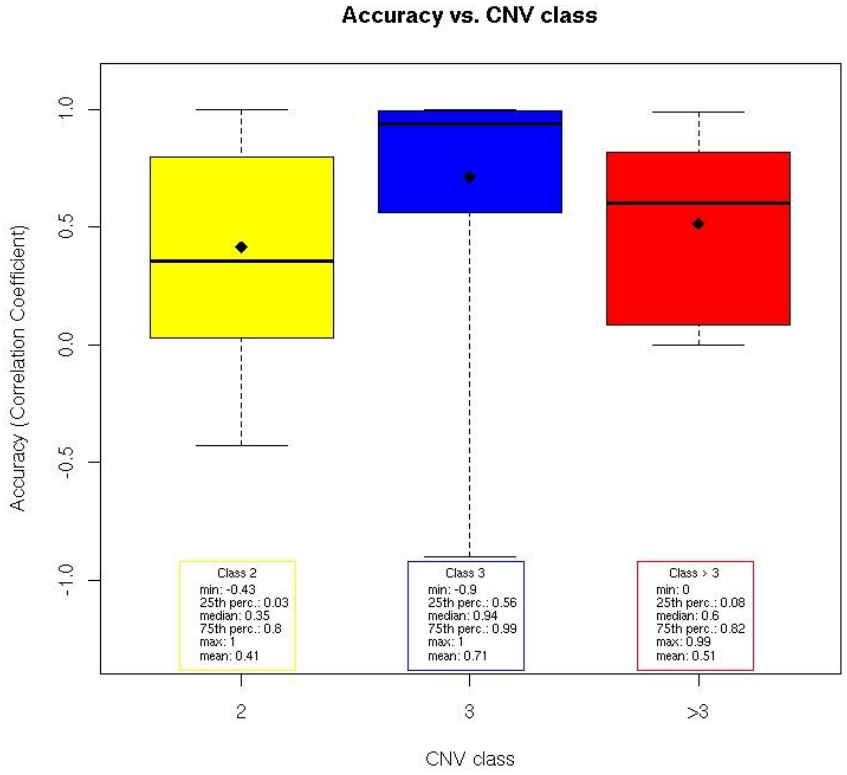


Figure 33: Plot of accuracy of copy number calls by copy number classes. The black lines on the boxplots are the medians and the black diamonds represent the statistical means. The extremities of the whiskers represent the minimum and maximum values. The corresponding summary statistics are shown.

5.3.3.6. Influence of linkage disequilibrium with a HapMap SNP on the accuracy of copy number calls

A measure of linkage disequilibrium with a HapMap SNP was available for 1581 CNVs. Of these, 83% were in high linkage disequilibrium (linkage disequilibrium ≥ 0.75) with a HapMap SNP. More than half (52%) of the 636 CNVs in weak linkage

disequilibrium (linkage disequilibrium< 0.5) with a HapMap SNP had a correlation ≥ 0.5 between aCGH and 1.2M genotypes; this suggests that CNVs that are not well tagged by a HapMap SNP could be measured with relatively good accuracy on the 1.2M platform. Figure 34 shows that the mean and median correlations were ≥ 0.5 for the four levels of linkage disequilibrium (LD) considered and both the mean and the median increased with increasing LD. Table 95 shows an increasing proportion of CNVs with a correlation ≥ 0.8 as LD with a HapMap SNP increases.

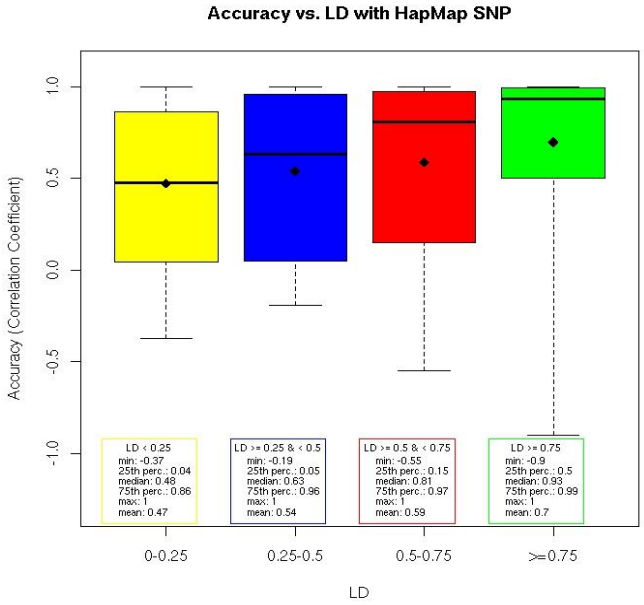


Figure 34: Plot of accuracy of copy number calls by level of LD with HapMap SNP. The black lines on the boxplots are the medians and the black diamonds represent the statistical means. The extremities of the whiskers represent the minimum and maximum values. The corresponding summary statistics are shown.

Correlation (r)	**LD < 0.25**	**0.25 ≤ LD < 0.5**	**0.5 ≤ LD < 0.75**	**LD ≥ 0.75**
≥ 0.8	124 (29%)	93 (44%)	72 (51%)	502 (62%)
0.5 ≤ r < 0.8	81	30	18	100
< 0.5	220	88	51	202
Total	425	211	141	804

Table 95: CNV count by level of LD with HapMap SNP and level of accuracy. The accuracy of copy number calls was assessed as the correlation between copy numbers from aCGH and 1.2M. The percentages represent the corresponding proportions in the LD category.

5.3.4. ACCURACY OF COPY NUMBER CALLS FROM OLDER SNP GENOTYPING PLATFORMS

5.3.4.1. Overlapping probes between Illumina 1.2M, 660, 610 and 300 platforms

Figure 35 shows the overlap of probes between 1.2M, ILM660K and ILM610K platforms. Each of the platforms contains SNP and CNVI probes but the name of some probes changed from one platform to another. The probe counts were obtained after mapping the positions and using one probe ID (the name of the probe on the 1.2M platform) for each SNP and CNVI probe. More than 99% of the probes (SNPs and CNVIs) on ILM660K were also on 1.2M and 90.6% of the probes (SNPs and CNVIs) on ILM610K were present on 1.2M.

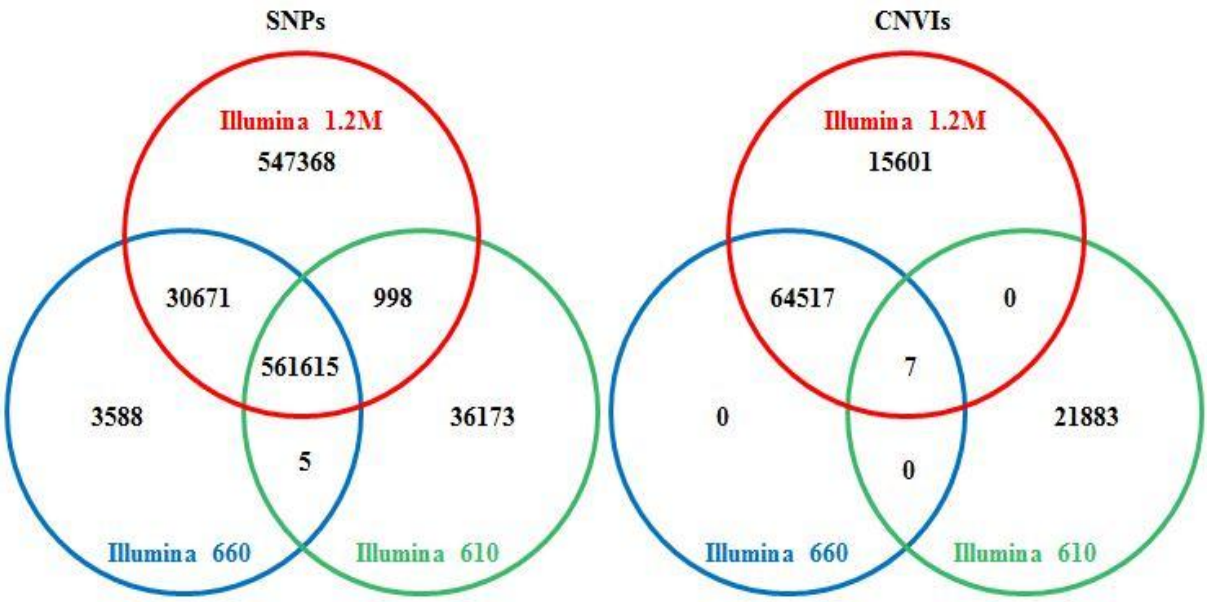


Figure 35: SNP and CNVI probes overlap in Illumina 1.2M, 660 and 610 platforms.

Given the noticeable difference in the number of non-overlapping CNVI probes between ILM660K and ILM610K, I decided to analyse the accuracy of copy number measurements from both platforms data.

Figure 36 shows the SNP probe overlap between 1.2M and ILM300K platforms; 99% of probes on ILM300K were also on the 1.2M platform. ILM300K does not contain CNVI probes.

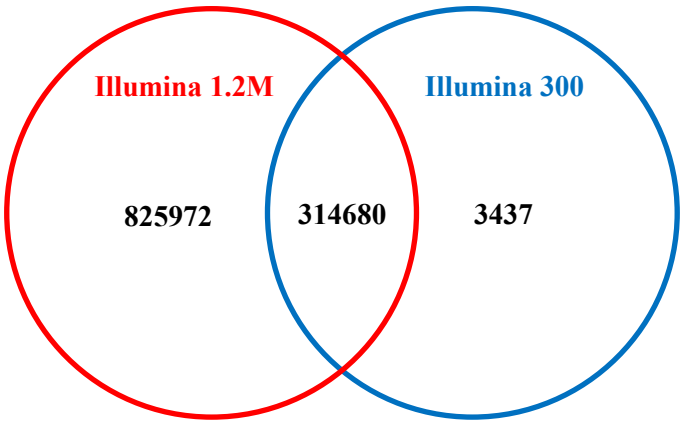


Figure 36: SNP probes overlap in Illumina 1.2M and 300 platforms.

The 1610 CNVs for which there was a gold standard were analysed to evaluate the accuracy of calls on ILM660K, ILM610K and ILM300K. Table 96 summarises the number of CNVs with and without probes within their boundaries on each of the three platforms. The older the platform, the larger the proportion of CNVs without probes within their boundaries; this observation is consistent with the lower probe density of older platforms.

Platform	CNVs with probes (SNPs and/or CNVIs) within their boundaries	CNVs without probes within their boundaries
1.2M	1610	0
ILM660K	1553	57
ILM610K	1351	259
ILM300K	1110	500

Table 96: Number of CNVs with and without probes within their boundaries.
The figures in this table represent the number of CNVs among the 1610 CNVs for which there was a gold standard.

The CNVs which did not have probes within their boundaries cannot be called from the SNP genotyping platforms; those CNVs were assigned a correlation coefficient of zero

to reflect the inability to accurately measure the CNV on the relevant SNP genotyping platform.

5.3.4.2. Accuracy of copy number calls from Illumina 660, 610 and 300

For each platform, CNVs were assigned a correlation coefficient of zero when 5% or more of all the samples presented missing intensities at one or more probes within the CNV boundaries (third QC criteria under section 5.3.2); the number of CNVs that met this criterion are reported in Table 97. Some CNVs that clustered when called from the 1.2M data did not cluster when called using data from ILM660K, ILM610K and ILM300K; these CNVs were also assigned a correlation of 0 i.e. it was not possible to call them from those platforms; the counts for these CNVs are reported under Table 97.

Platform	CNVs where more than 5% of all samples have missing intensities at one or more probes	CNVs that clustered with 1.2M data but not with this platform data
1.2M	113	-
ILM660K	97	1
ILM610K	40	13
ILM300K	11	13

Table 97: Counts of CNVs excluded for meeting the above two QC criteria. All these CNVs were assigned a correlation coefficient of zero.

The results of the evaluation of the accuracy of calls from each platform are reported in Table 98. For ILM660K, 49% of the 1610 CNVs tested had a correlation ≥ 0.8. Of the 1553 CNVs which had SNP and/or CNVI probes within their boundaries in the ILM660K data, 1236 contained both SNP and CNVI probes. Thirty-eight percent of these CNVs had a correlation ≥ 0.8 when SNP probes alone were included in the analysis, 44% had a correlation ≥ 0.8 when CNVI probes alone were considered, and 50% when SNP and CNVI probes were combined. As was observed for the 1.2M data, the accuracy on ILM660K was improved when SNP and CNVI probes were combined.

For ILM610K, there were no CNVI probes within the boundaries of the 1610 CNVs tested. Forty-six percent of the 1610CNVs tested had a correlation ≥ 0.5 and 31% had a correlation ≥ 0.8. This indicates that, overall (across the 1610 for which there was a gold standard), copy numbers were measured with a lower accuracy using probe intensity data from the ILM610K compared to using probe intensity data from 1.2M and ILM660K.

For ILM300Kthere were no CNVI probes within the boundaries of the 1610 CNVs tested. Thirty-six percent of the 1610CNVs tested had a correlation ≥ 0.5 and 23% had a correlation ≥ 0.8. This suggests an overall lower accuracy of copy number calls when probe intensity data from ILM300K was used to measure CNVs, in comparison with 1.2M, ILM660K and ILM610K.

Correlation (r)	1.2M	ILM660K	ILM610K	ILM300K
≥ 0.8	791 (49%)	781 (49%)	499 (31%)	364 (23%
0.5 ≤ r < 0.8	231	219	247	217
< 0.5	588	610	864	1029
Total count	1610	1610	1610	1610

Table 98: CNV counts by level of accuracy and SNP platform.
The accuracy of copy number calls was assessed as the correlation between copy numbers from aCGH and Illumina 1.2M, 660, 610 and 300 platforms based on all the probes and on the 1610 CNVs for which there was a gold standard. The percentages represent the corresponding proportions in the 1610 CNV tested.

5.3.4.3. Comparing accuracy of copy number calls from Illumina 1.2M, 660, 610 and 300 SNP platforms

Figure 37 compares the four SNP genotyping platforms, investigated in this analysis, to assess the extent to which existing data from older platforms can be used to study CNVs. The CNVs that did not cluster in any of the platforms were assigned a correlation level of 0. Figure 37 shows a decreasing accuracy of copy number measurements from the latest platform included in this analysis (1.2M) to the oldest

(ILM300K). The accuracy was roughly the same for copy number calls from 1.2M and ILM660K data but markedly lower for ILM610K and ILM300K.

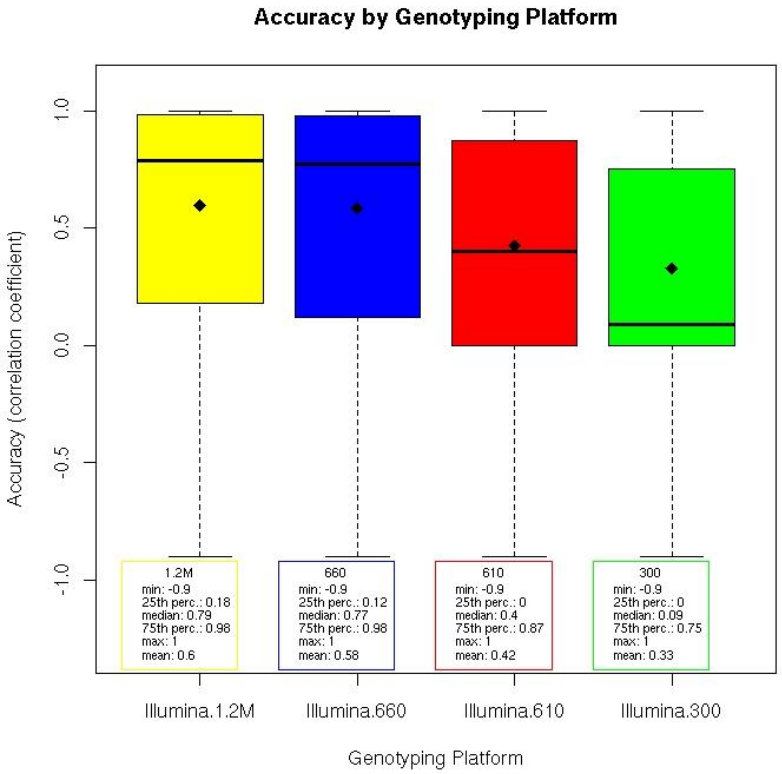


Figure 37: Plot of accuracy by SNP genotyping platform.
The plots are based on the 1610 CNVs for which there was a gold standard. The black lines on the boxplots are the medians and the black diamonds represent the statistical means. The extremities of the whiskers represent the minimum and maximum values. The corresponding summary statistics are shown.

5.3.5. ESTIMATING THE IMPACT OF COPY NUMBER CALLS INACCURACY ON POWER

Table 99 contains the relevant parameters used for the power and sample size calculations in ESPRESSO-forte. The measurement error of CNVR5101.1, r = 0.71, corresponded to a sensitivity of 0.72 and a specificity of 0.99. All the parameters, except the sensitivity and specificity of the measurement of CNVR5101.1 and the outcome, were the same in all three scenarios detailed on page 193. So the differences between the power and sample size values calculated under the first scenario and those

calculated under the second and third scenarios were due to the measurement error on CNVR5101.1 (in scenario 2) and to measurement errors on the outcome and on CNVR5101.1 (in scenario 3).

The power achieved was 0.86 under the first scenario (no measurement error i.e. r = 1) and 0.36 under the second scenario (r = 0.71). The measurement error on the determinant was responsible of the loss of more than half of the power that could be achieved in the absence of measurement error. The OR (estimated empirically) shrank from 1.47, under the first scenario, to 1.32 under the second scenario. The sample size required to achieve a power of 80% was 1820 cases and 7279 controls, under the first scenario (r =1) and 3565 cases and 14261 controls under the second scenario (r = 0.71). This means that the sample size required to achieve a power of 80% has to be increased by 96% to compensate for the loss of power caused by the measurement error on the genetic determinant.

In the third scenario, in addition to the measurement error on the genotype of CNVR5101.1, there was also a measurement error on the outcome which was arbitrarily assumed to be measured with a sensitivity of 0.72 and a specificity of 0.99. The power decreased from 86% to 0%, the OR shrank from 1.47 to 1.1 and the initial sample size has to be multiplied by 19.2 to compensate for the loss of power resulting from the two measurements errors.

PARAMETERS	SCENARIO 1	SCENARIO 2	SCENARIO 3
Simulation and outcome parameters			
number of runs	1000	1000	1000
number of cases	2000	2000	2000
number of controls	8000	8000	8000
outcome model	binary	binary	binary
disease prevalence	0.1	0.1	0.1
p.value	1.00E-04	1.00E-04	1.00E-04
power required	0.8	0.8	0.8
sensitivity of the assessment of outcome	1	1	0.72
specificity of the assessment of outcome	1	1	0.99
Genotype parameters			
genetic model	binary	binary	binary
MAF	0.05	0.05	0.05
OR	1.5	1.5	1.5
sensitivity	1	0.72	0.72
specificity	1	0.99	0.99
Results			
Power achieved with 2000 cases and 8000 controls	0.86	0.36	0
Sample size required to achieve a power of 80% (cases/controls)	1820/7279	3565/14261	34919/139675
Estimated OR	1.47	1.32	1.1

Table 99: Parameters and results of the analysis
The OR of CNVR5101.1 was arbitrarily set to 1.5. The results show power and sample size figures under the first scenario (no measurement error on both the outcome and determinant), under the second scenario (no measurement error on outcome but determinant measured with an error) and under third scenario (measurement errors on both outcome and determinant.

5.4. DISCUSSION

5.4.1. OVERVIEW, STRENGHTS AND WEAKNESESSES OF THE STUDY

The power achieved by a genetic association study depends partially on the quality of the measurements of the genetic determinants of the trait of interest. If a CNV is used as genetic determinant in a genetic association study that uses CNV measurements from SNP genotyping platform data, it might be important to know how accurately the CNV

was measured from that data in order to take into account the level of measurement error and interpret the findings correctly.

In this project, the accuracy of CNV measurements from four SNP genotyping platforms was evaluated by comparing copy numbers called from these platforms with copy numbers called, for the same individuals, from an array-CGH platform which can be considered as a gold standard in copy number genotyping. The results of the comparison, expressed as a level of correlation between copy number calls from the SNP platforms and the aCGH platform, indicate how well the CNVs were measured from the SNP platforms.

CNVtools was used to determine the copy number statuses of the individuals. This algorithm allows for the tuning of some parameters to optimize the clustering of each CNV. However given the large number of CNVs to investigate in this project and the fact that CNVtools allows only for the analysis of one CNV at a time, it was necessary to automate the process. A script was written in R to use CNVtools in 'batch mode'. This allowed for a considerable gain in efficiency but it was not possible to try all possible combinations of parameters to cluster appropriately each CNV. The copy number statuses of the individuals were determined based on the most likely copy number which was indicated by the highest posterior probability; so all the comparisons between copy numbers from SNP platforms and from aCGH were based on the most likely CNV genotype (copy numbers). To verify that this approach was reasonable another strategy was used to assign copy numbers to individuals; in this strategy, the copy number statuses of individuals were determined based on all the posterior probabilities and not only on the highest probability. Reassuringly the results (see Appendix 4) of the comparisons between aCGH and 1.2M copy numbers carried out using CNV genotypes determined through this second strategy were similar to those

obtained when only the highest posterior probability was used to assign individuals to copy number classes.

Information about the optimal CNVtools parameters for the set of CNVs that were studied was obtained from the WTCCC+ paper where available. When this information was not available the optimal parameters were chosen, through an iterative approach, as the combination of parameters that yields the highest accuracy (largest proportion of CNVs with high correlation) across all the CNVs. So the parameters used might not be the optimal ones for each CNV; some CNVs may have been better clustered using different parameters. Each of the 2600 clustering plots (one for each CNV), generated through the automated process and using the aCGH data, was visually inspected and the 1610 CNVs that clustered appropriately and had correct posterior probabilities were considered to be those for which a gold standard measure was available.

The set of individuals from the 1958 British Birth Cohort, from whom the aCGH and 1.2M data was available for this analysis had not been genotyped on the Illumina 660, 610 and 300 platforms. I generated datasets mimicking each of these platforms by thinning the available 1.2M platform data aiming to obtain a probe content identical to the older platforms. So, the variability in the level of accuracy between the SNP genotyping platforms investigated in this chapter reflects the difference in probe density between the platforms. There have been developments in the chemistry and other aspects of the genotyping technology between these platforms (including the ability to run “multiplex” assays across multiple samples simultaneously) that may not be fully reflected in these mimicked datasets.

The measurement error on the genotype of a CNV, called from the Illumina 1.2M platform, was used in the ESPRESSO-forte algorithm to assess the impact of CNV call

inaccuracy on the statistical power of a hypothetical association study that investigates the association between the selected CNV and a binary outcome. This analysis illustrates how a CNV call accuracy estimated in this analysis can be used to analyse the effect on power, using the ESPRESSO version that was developed as part of this thesis.

5.4.2. KEY FINDINGS

5.4.2.1. Summary of the results

The accuracy of copy number calls was measured as the level of correlation between copy numbers called from SNP genotyping platform data and from aCGH platform data. In the paragraphs below high accuracy refers to a correlation ≥ 0.8.

The analysis of the accuracy of calls from the 1.2M platform dataset was carried out for the 1610 CNVs for which there was a gold standard set of copy number calls (section 5.3.2). The results of the analysis showed that 49% of the 1610 CNVs tested were measured with a high accuracy on the Illumina 1.2M platform. The accuracy was optimal when SNP and CNVI probes were combined compared to when these two types of probes were analysed separately.

The accuracy of copy number callson the 1.2M platform decreased with increasing CNV length; 59% of the CNVs shorter than 3kb were measured with a high accuracy whilst this proportion was 50% for CNVs of between 3 and 22kb in length and 29% for CNVs longer than 22kb. However, a cross-tabulation of CNV frequency against CNV size (under Appendix 5) indicated that the accuracy of calls was not driven by CNV length alone; the proportion of CNVs measured with an accuracy r ≥ 0.8 under each CNV length category were not similar for the two frequency categories. For example, for CNVs shorter than 3kb, the majority (61%) of the common CNVs (MAF ≥ 0.05)

were measured with an accuracy r ≥0.8 whilst only 33% of rare CNVs (MAF < 0.05) were measured with the same level of accuracy; if the accuracy was driven by CNV length alone, the proportions observed under these two MAF categories would have been similar.

Common CNVs (MAF > 0.05) were measured with a greater accuracy on the 1.2M platform than rare CNVs (MAF < 0.05); 58% of the common CNVs were measured with a high accuracy whilst only 28% of the rare CNVs were measured with the same level of accuracy. CNVs with 3 copy number classes were better called than CNVs with 2 classes with respectively 65% and 25% of highly accurate calls for 3-class and 2-class CNVs. CNVs classified as duplications/deletions were genotyped more accurately on the 1.2M array than simple deletions or duplications; the respective proportions of CNVs called with a high accuracy in each of these types of variation were 72%, 49%, and 48%. CNVs reported by the WTCCC+ study as well tagged by HapMap SNPs were better called from the 1.2M platform; 62% of CNVs in high linkage disequilibrium ($r^2 \geq 0.75$) with a HapMap SNP were measured with a high accuracy on the 1.2M platform. The accuracy increased with increasing level of linkage disequilibrium (LD) but even CNVs in low LD with HapMap SNPs were relatively well measured on the 1.2M platform: 44% of the CNVs with a level of LD between 0.25 and 0.5 ($0.25 \leq LD < 0.5$) and 29% of the CNVs with a level of LD between 0.02 and 0.25 ($0.02 \leq LD < 0.25$) were called with a high accuracy on the 1.2M platform

The results of cross-tabulations between the CNV characteristics (Table 93 in section 5.3.3.4 and Table 101 to Table 109 in Appendix 5) showed that the accuracy of calls was not driven by one CNV characteristic alone. I carried out chi-squared tests of independence to assess the correlation between the CNV characteristics considered in this project (CNV size, minor allele frequency, CNV type, number of classes and LD

with HapMap SNP). The results of the tests (reported in Table 110under Appendix 5) indicate that all the characteristics were significantly correlated with each other except CNV type and LD where the p.value was slightly greater than 0.05.

The comparison of the accuracy of calls between Illumina 1.2M and three older SNP genotyping platforms (Illumina 660, 610 and 300) showed that 49% of the 1610 CNVs tested were measured with a high accuracy on the Illumina 1.2M whilst the proportions of CNVs measured with the same level of accuracy were respectively 49%, 31% and 23% on the Illumina 660, 610and 300 platforms. These figures indicate that accuracy of copy number calls on Illumina 660 was roughly similar to that on 1.2M and that CNVs were less well measured on the two oldest platforms (Illumina 610 and 300) investigated in this analysis. On the 1.2M and 660 platforms, the accuracy of calls was higher when information from CNVI probes and SNP probes were combined.

5.4.2.2. About the negative correlations between aCGH and 1.2M copy numbers

A low correlation of between 0 and -0.3 can be expected to occur by chance; however 17 (1%) of the 1610 CNVs analysed had a correlation lower than -0.3. In section 5.3.3, the correlation was calculated using one probe at time for CNVR551.3 which had the highest negative correlation; the results of this investigation showed that the correlation was not negative for all the probes. It was not possible to know, with the available data, if the eventual error that causes this occurs on the aCGH or the 1.2M data or whether if it is a characteristic of the CNV itself. So a negative correlation should be interpreted as an impossibility to determine the copy number statuses of individuals for the tested CNV.

5.4.2.3. About accuracy versus CNV characteristics on the 1.2M platform

The observed trends between high accuracy and CNV characteristics should be carefully interpreted as it is not clear what characteristics are the best predictors of accuracy. As an illustration, when the accuracy was checked against CNV length (Table 89) the results showed that shorter CNVs were better called than longer CNVs; but when these results were checked against CNV frequency (Table 109 under Appendix 5) it appeared that shorter CNVs were more accurately measured but that was only true for common CNVs; shorter CNVs were poorly measured when they were rare.

Similar remarks can be made about the influence of the number of CNV classes on accuracy; CNVs with 3 copy number classes seemed more accurately measured than CNVs with 2 copy number classes, but here again the checks against the frequency indicate completely opposite observations for rare and common CNVs within that class category: among the 3-class CNVs, the common ones were well measured whilst the rare ones were poorly measured; this suggests that rare CNVs are measured with lower accuracy. Since 69% of the 2-class CNVs that were analysed were rare this can explain why the 2-class CNVs were less well measured. Furthermore the 2-class CNVs might be in fact 3-class CNVs where one class is missing: if the minor allele is rare, individuals homozygous for the minor allele will be even rarer and it will be as if only two classes exist (**7, 154**).

Conrad et al found that duplications and CNVs involving both duplications and deletions were "*more difficult*" to genotype than deletions (**17**). In this analysis, when the accuracy of calls was checked against the types of variations it was observed that CNVs involving duplications and deletions were better called than simple deletions (Table 92). When the figures reported in Table 92 were stratified by the number of

CNV classes (Table 93), the conclusions did not change for simple deletions and duplications with deletions being slightly more accurately measured than duplications. But it appeared that actually CNVs involving duplications and deletions were less accurately measured than deletions when the number of classes was greater than 3 and this is in accordance with the observation of Conrad et al. Only 15 (7%) of the 201 CNVs involving duplications and deletions had more than 3 classes; the lower proportion of complex CNVs in the subset that was analysed in this study could explain the higher accuracy reported in Table 92, for this CNV type.

5.4.2.4. About the influence of probes on the accuracy of calls

The accuracy of CNV calls from the Illumina 1.2M platform does not increase as the number of probes within the CNV increases (even when taking into account the overall CNV length); this suggests that some probes carry more genotype information than others. If the probes that carry most of the information are present then the copy numbers are measured with a relatively high accuracy even if there are only a limited number of probes. To investigate this, the level of accuracy was evaluated iteratively by removing one probe at a time and re-calculating correlation for one CNV (CNVR1063.1). The accuracy was markedly reduced when some probes were removed or remained unchanged when some other probes were excluded. For CNVR1063.1, the accuracy was r = 0.93 when all the probes were included, it decreased to r = 0.81 when the SNP probe rs10195936 was excluded but remained unchanged when SNP probe rs13003047 or CNVI probe cnvi0103164 were excluded. This suggests that rs10195936 carries more copy number information than rs13003047 and cnvi0103164. Similar results were observed for the probes on the aCGH platform, some probes were more informative then others; for CNVR1063.1, probe 13325 carries more information than

probe 13334. This can explain why the accuracy of calls does not necessarily increase with increased number of probes.

5.4.2.5. About the accuracy of calls by type of probe (SNP probe vs. CNVI probe)

The proportion of CNVs measured with a high accuracy using CNVI probes alone was higher than that of SNP probes. This may be because CNVI probes are specifically designed to map CNVs. A difference in GC-content (the proportion of guanine and cytosine nucleotides) between SNP and CNVI probes might also explain the difference in accuracy; if SNP probes have larger GC-content the signal-to-noise ratio might be lower for SNP probes and the intensity used to determine the number of copies will then be less accurately measured (**156**). This is because if the GC content is high (> 55%), the denaturation of the DNA (i.e. separation of the two DNA strands to obtain single DNA strands), may not be complete making it impossible for the allele specific oligonucleotide to hybridise with the target DNA (see paragraph on SNP genotyping under section 1.2.3.1). The level of accuracy achieved when SNPs and CNVIs were combined was higher than when the probes were used separately; this suggests that SNPs and CNVIs do not carry the same type of information.

5.4.2.6. About the comparison of the accuracy of calls on different SNP platforms

The comparison of copy number calls accuracy across the four SNP genotyping platforms included in this analysis (results in section5.3.4) showed that platforms with a higher density of probes allow capture of more CNV genotype information and hence ensure a higher accuracy of calls, however, as mentioned in section 5.4.2.4, because some probes may carry more genotype information whilst others may increase the level

of noise, a large number of probes does not always improve the accuracy of calls. This could probably explain why ILM610 had a level of accuracy similar to that of 1.2M which has a higher probe density. In this analysis, the probe intensity data from older platforms was generated by thinning the intensity data from Illumina 1.2M, so the implicit assumption was that the platforms differ only by probe density. The technology evolved from one platform to another. So the inaccuracy of calls, on the three older platforms evaluated here represents the level of error that could be expected due to lower probe density alone. The true error on the earlier genotyping platforms might be higher than what was estimated in this project because the signal-to-noise ratio is lower in earlier platforms; due to improved technology and laboratory experience over time the latest platform have a larger signal-to-noise ratio. Nevertheless, the results are informative about the minimum level of error that should be expected if intensity data from Illumina 660, 610 and 300 were used to determine copy number statuses.

5.4.3. RECOMMENDATIONS

If intensity data from the Illumina 1.2M platform are to be used in a CNV association study and if a decision is to be made about what CNVs to measure, based on how well they can be called from that data, it is not advisable to use one CNV characteristic alone as indicator of how accurately the CNV can be called from the 1.2M intensity data. This is because the CNV characteristics analysed in this project were related and it is not clear which characteristics drive the accuracy of CNV measurements.

The analysis in this chapter can inform about how accurately each of 1610 CNVs, for which there was a gold standard, was measured from each of the four SNP platforms analysed in this project. The results can be used by investigators as a guide to decide if a particular CNV can be measured accurately using intensity data from one of the four

Illumina platforms. The results can also be used as a guide to decide if it is worth calling a particular CNV from one of the four SNP platforms. However for a comprehensive study of all CNVs, new assays should be run for the CNVs that could not be called accurately from any of the four platforms.

For a genome-wide CNV association study it is preferable to not use intensity data from the Illumina 610 and 300 platforms. This is because 69% and 77%, respectively, of the genotypes called from the Illumina 610 and Illumina 300 data were not accurately measured. For a candidate CNV association, any of the four SNP platforms could be used provided that the investigator chooses CNVs that were called accurately from the region of interest (the region to test the CNV association). If for a specific platform, none of the CNVs located in the particular region to test for association was measured accurately then that platform should not be used for the candidate CNV study.

The findings of this analysis suggest that where a choice of platform is available for de novo CNV measurement it would be preferable to use a platform that has both SNP and CNVI probes and where existing datasets are being used then information from both SNP probes and CNVI probes should be used in CNV studies.

CHAPTER 6

6. GENERAL CONCLUSION

6.1. INTRODUCTION

This thesis explores the key factors that influence power in large scale genetic association studies. As scientists, our control of statistical power depends mainly on the number of participants we choose to enrol into an analysis (sample size) and on the quality of the data and samples we decide to collect (**7, 71**). The first element of the work described in this thesis was therefore to develop a tool that enables for more realistic calculation of sample size by taking into account errors in outcome and exposure measures that alter the quality of the data. The ESPRESSO-forte power calculator developed in this thesis builds on an earlier version of the software which could not allow for many of the classes of analysis carried out in this thesis. The newly developed tool was subsequently used to explore the statistical power profile of an existing large cohort i.e. to estimate the minimum effect that could be detected by a study using the entire data generated by a large multi-centre Canadian cohort to investigate quantitative traits. The tool was also used to find out how errors and biases that relate to biobank procedures, for the collection and processing of biosamples, influence the power of the association studies that subsequently use that biobank data. Although, to date, most genetic association studies have primarily investigated single nucleotide polymorphisms, genome-wide association data are widely available from which the association between copy number variants and disease can be tested(**15**). The accuracy of CNV genotype assessment, using different SNP platforms, was explored to inform future studies that may use existing SNP platforms data to measure CNVs. By

choosing a platform that allows for the most accurate measurements, the investigator can improve the potential statistical power of his/her study.

6.2. SUMMARY OF THE CHAPTERS

The first chapter introduces some key concepts in genetics and genetic epidemiology that are centrally relevant to my thesis. It was essential to have a clear understanding of those concepts in order to undertake the analyses carried out in the thesis. The chapter ended with a graphical explanation on how effect size, sample size and type I error affect statistical power.

The second chapter goes through the details of the development of ESPRESSO-forte. Most crucially, the algorithm allows for elements not taken into account in conventional calculators to be considered in the power and sample size calculations for stand-alone case-control and cohort studies and for case-control analyses nested in cohort studies. ESPRESSO-forte was implemented as an open source R package to allow for researchers proficient in the R programming language to use it in a flexible way and to access the code which they could eventually alter to answer scientific questions that require some modification to the downloadable version. Analyses can be run interactively using the web-based version of the software hosted within the P^3G website. The influence of minor allele frequency, level of linkage disequilibrium between the observed and the causal variant and genetic model misspecification on power were explored in Chapter 2 after describing the building and implementation of the ESPRESSO-forte algorithm. The development of ESPRESSO-forte enabled me to undertake the subsequent analyses carried out in chapter 3 and 4 and to explore the impact of CNV measurement accuracy on statistical power, in chapter 5.

The results of the analysis in chapter 3 answer a question raised by the Canadian Partnership for Tomorrow project (CPT); that is, what would be the power of this multi-provincial cohort to study quantitative traits, relevant for cancer studies, given its likely final size of either 110000 or 180000 participants? For each of the chosen traits, the minimum detectable effect size was calculated under several biomedical scenarios. These power analyses demonstrated that CPT can provide a world leading platform for studying the etiological architecture of quantitative traits and for conducting exposure-based studies with either of the two potential final sample sizes; but, all else being equal, 180000 participants would undoubtedly be more informative than 110000. However, consideration should only be given to increasing sample size provided it was first ensured that the phenotyping protocol (for outcomes and exposures) was fit-for-purpose and would produce high quality data that could effectively be pooled – between provinces - across all of the component studies. This is because the gain of power obtained through larger sample sizes would be negated if errors related to outcome and exposure measurements are not minimized by a carefully developed protocol that limits measurement errors and ensures adequate harmonization to enable effective data sharing and pooling.

The fourth chapter consists of an exploration of the UK Biobank protocols for the collection and storage of biosamples. The aim was to estimate the magnitude of the pre-analytical variation i.e. “non-biological” variation introduced if the processing of the samples was delayed for up to 24 or 36 hours after their collection, and to investigate the effect of such variation on the power of association studies that then use these biobank data. The results showed that (1) the majority of the analytes investigated in my analysis, were stable over a period of 36 hours, (2) for some analytes, the concentration declines or increases over the same period of time but there was no heterogeneity in the

rate of decline or increase across the individuals; and (3)for some analytes, the concentration declines or increases over the same period of time with a significant heterogeneity in the rate of decline or increase across all the individuals. For the analytes that are stable, i.e. there is no significant change in concentration over the period of time between the collection of the sample and the processing (quantification of the analyte), there is no bias resulting from the delay in processing and it is perfectly acceptable to quantify the analytes at any time point within the processing time period. For the analytes for which there is a decline or increase in concentration but no heterogeneity in the rate of change across individuals, there will be a systematic bias if the samples are not processed at the same time point; a delay in processing however is acceptable if this delay is the same for all samples or if the difference in change of concentration is to be calibrated thereafter (if the samples are not processed at the same time point). If an analyte exhibits a change of concentration over time and a significant heterogeneity in the rate of change of concentration across individuals, the solution to avoid a systematic bias is to process all samples immediately after the samples collection as there will be a systematic bias with any delay even if all samples are processed at the same time point because the magnitude of the change in concentration will not be the same across the individuals. If the pre-analytical variation occurring due to delays in sample processing is ignored it affects adversely the power of the studies that use the data which may then not be well powered to detect existing association.

Most genetic association studies undertaken to date have investigated the potential role of single nucleotide polymorphisms (SNPs) in causing disease, there are however an increasing number of studies that are now looking at the corresponding role of copy number variants (CNVs). The accuracy of CNV measurements has not yet reached that of SNPs. In Chapter 5, the accuracy of CNV genotypes measured from some SNP

genotyping platforms was assessed. The results showed that the accuracy of a CNV measurement depends on the characteristics of the particular CNV to be assessed - a set of characteristics that are all correlated. Amongst the four contemporary SNP platforms investigated, the Illumina Infinium 1.2Mappeared, overall, to be the best platform to measure the particular set of CNVs investigated in this analysis but the Illumina 660 platform also offers a good level of accuracy considering its lower probe density in comparison with Illumina 1.2M. These findings can inform future studies that plan to use SNP platforms data to call CNVs.

6.3. FURTHER DEVELOPMENT WORK

The documentation of the web-based ESPRESSO-forte has been updated to reflect the new elements of the algorithm. Likewise, the current graphical user interface (GUI), under the P^3G website, needs now to be rebuilt to reflect the new features of the software. In ESPRESSO-forte, it is possible to model two genetic variants that are in linkage disequilibrium; such analysis can however be very time consuming particularly if the number of subjects to simulate is large. The run-time could be improved by writing the required function in a programming language such as C which is computationally more efficient.

An extension of ESPRESSO-forte is currently being developed that would enable the design of a prospective analysis to be entered, the impact on expected statistical power to be generated, and appropriate modifications of meta-analysis weights to be derived. But none of this will be practicable unless, in keeping with the analysis carried out based on the UK Biobank biosamples, standard operating procedures are carefully thought through, well described and are freely available to biobank users on request.

With the current version of ESPRESSO-forte, it is not possible to carry out power calculations that precisely represent reality, when an analysis involves a large number of genetic variants. This is because ESPRESSO-forte allows for the modelling of up to two genetic exposures and two environment/life style exposures only. The reason that this is not *seriously* problematic is that it is often possible to undertake a perfectly valid analysis of key components of such an analysis (*e.g.* two SNPs in linkage disequilibrium) and it is only if the inferences based on those two SNPs in isolation would differ substantially from their equivalents if all SNPs were considered in totality that there would be problem. In many settings there is no substantive difference at all. For example, if a GWAS analysis deals sequentially with one million separate variant-disease associations, it is perfectly acceptable to model one of those associations in isolation. On the other hand, there are some settings – for example involving the derivation of haplotypes and generation of inferences based up them – where individual SNPs cannot be considered alone (or even on a two-by-two basis incorporating linkage disequilibrium) and so the current ESPRESSO-forte approach is potentially restrictive. It is therefore desirable to extend the software by implementing additional methods that *do* enable the joint consideration of a large number of genetic variants.

The current version of ESPRESSO-forte assumes independent observations and is hence unsuitable for directly estimating the power of analyses based on longitudinal or family data. Another useful extension would therefore be the incorporation of methods allowing for the correlation of observational units.

6.4. FINAL CONCLUSIONS

Many of the recent successes in genomic epidemiology – in particular the identification of replicable associations - have been achieved primarily through large sample sizes (**7,**

24, 154). This move to using much larger sample sizes has been critical, primarily because of the generally modest - or weak – effect of the genetic variants underlying these conditions. The work undertaken, in this thesis, confirms the vital importance of very large sample sizes in the investigation of the mechanisms of complex diseases. However, this work confirms that of others (**24, 49, 157**) in showing that a large sample size is not a panacea and that the gain of power obtained through increased number of participants can be jeopardised if the outcome and exposures data are of poor quality or if they cannot reasonably be pooled between different studies (**7, 71, 72**). This is because the biases and errors built into a dataset can cause a profound loss of power (**18**) and if data are simply too different to reasonably be pooled then the potential desirability of the consequent sample size increase becomes, in practice, irrelevant.

It is therefore crucial to take into full account the uncertainty of outcome and exposure measurements, in the power analysis at the design stage of a study, to ensure that the probability of finding an existing true association is not overestimated. By substantially overestimating the power of a study, one might ultimately conclude that no association exists between an outcome and an exposure of interest whilst in actuality that association does exist but the analysis that was used had almost no real chance of detecting it. This not only leads to a failure to detect real effects but, in addition, because any form of statistical inference leads inevitably to a predictable proportion of false positives (associations that are declared as real when in fact they are not), widespread low power can lead directly to a situation in which most positive associations that are reported are false positives. This is reflected in a persistent failure to replicate and can lead to a serious waste of scientific resources. This consideration is not merely an esoteric possibility that might theoretically occur. Rather, it can strongly be argued to have been precisely the situation that pertained in genetic and genomic

epidemiology before successful projects such as the Wellcome Trust Case Control Consortium led to a step change in the average sample size of genetic association studies. The blunt truth of the power calculations undertaken using ESPRESSO-forte software provides a valuable reality check and thereby supports the move to more realistic power analysis and sample size calculation; the analyses in later chapters provide practical examples of applications of the software to real world scientific problems (chapters 3 and 4) and to a theoretical problem that undoubtedly exists (chapter 5).

Finally, the large sample sizes required for the investigation of complex conditions are often achieved through pooling of data (meta-analysis) from different sources and platforms. Yet, these platforms (large cohorts and biobanks) have not always used the same standard operating procedures (SOPs) for collecting, transporting, storing and processing the data and samples from which key information are obtained. It is therefore important that studies develop protocols that minimize pre-analytical variability of both data and samples. In addition, we must develop and implement approaches to retrospective harmonization that can optimise the extraction of valid information across a series of legacy data sets to be pooled. In this regard, ESPRESSO-forte can serve as a tool to promote effective harmonization endeavours by clearly demonstrating how many studies need to be brought together in order to answer a particular scientific question of interest.

BIBLIOGRAPHY

7. BIBLIOGRAPHY

1. Who/Fao. Diet, Nutrition, and the Prevention of Chronic Diseases. WHO/FAO; 2009.

2. Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. *Lancet* 2005 Sep 10-16;**366**(9489): 941-51.

3. Gibson G. Decanalization and the origin of complex disease. *Nature reviewsGenetics* 2009 Feb;**10**(2): 134-40.

4. Davey Smith G, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR. Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* 2005 Oct 22-28;**366**(9495): 1484-98.

5. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009 Oct 8;**461**(7265): 747-53.

6. Palmer LJ, Cardon LR. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* 2005 Oct 1;**366**(9492): 1223-34.

7. Burton PR, Hansell AL, Fortier I, et al. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *International journal of epidemiology* 2009 Feb;**38**(1): 263-73.

8. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA : the journal of the American Medical Association* 2008 Mar 19;**299**(11): 1335-44.

9. Hey J. What's so hot about recombination hotspots? *PLoS biology* 2004 Jun;**2**(6): e190.

10. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. *PathoGenetics* 2008 Nov 3;**1**(1): 4.

11. International HapMap C, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007 Oct 18;**449**(7164): 851-61.

12. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature* 2006 Nov 23;**444**(7118): 444-54.

13. Poplawski T, Stoczynska E, Blasiak J. Non-homologous DNA end joining--new proteins, new functions, new mechanisms. *Postepy biochemii* 2009;**55**(1): 36-45.

14. Illumina Inc. *Genome-Wide DNA analysis BeadChips*. 2010 [cited 2011 08.09.2011]; Available from: http://www.illumina.com/Documents/products/datasheets/datasheet_infiniumhd.pdf

15. Wain LV, Armour JA, Tobin MD. Genomic copy number variation, human health, and disease. *Lancet* 2009 Jul 25;**374**(9686): 340-50.

16. McCarroll SA, Kuruvilla FG, Korn JM, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics* 2008 Oct;**40**(10): 1166-74.

17. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. *Nature* 2009 Oct 7.

18. Cordell HJ, Clayton DG. Genetic association studies. *Lancet* 2005 Sep 24-30;**366**(9491): 1121-31.

19. Lewontin RC. On measures of gametic disequilibrium. *Genetics* 1988 Nov;**120**(3): 849-52.

20. Lewontin RC. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* 1964 Jan;**49**(1): 49-67.

21. Gaut BS, Long AD. The lowdown on linkage disequilibrium. *The Plant Cell* 2003 Jul;**15**(7): 1502-6.

22. Ito T, Chiku S, Inoue E, et al. Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. *American Journal of Human Genetics* 2003 Feb;**72**(2): 384-98.

23. Altshuler D, Hirschhorn JN, Klannemark M, et al. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* 2000 Sep;**26**(1): 76-80.

24. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007 Jun 7;**447**(7145): 661-78.

25. Repapi E, Sayers I, Wain LV, et al. Genome-wide association study identifies five loci associated with lung function. *Nat Genet* 2010 Jan;**42**(1): 36-44.

26. McCarroll SA, Huett A, Kuballa P, et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nature genetics* 2008 Sep;**40**(9): 1107-12.

27. Nakajima T, Kaur G, Mehra N, Kimura A. HIV-1/AIDS susceptibility and copy number variation in CCL3L1, a gene encoding a natural ligand for HIV-1 co-receptor CCR5. *Cytogenetic and genome research* 2008;**123**(1-4): 156-60.

28. Ingelman-Sundberg M, Sim SC, Gomez A, Rodriguez-Antona C. Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacology & therapeutics* 2007 Dec;**116**(3): 496-526.

29. Morton NE. *Outline of Genetic Epidemiology*. New York: S. Karger; 1982.

30. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res* 2007;**16**(4): 309-30.

31. Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *International journal of epidemiology* 2004 Feb;**33**(1): 30-42.

32. Smith GD, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS medicine* 2007 Dec;**4**(12): e352.

33. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science (New York, NY)* 1996 Sep 13;**273**(5281): 1516-7.

34. Elston R, Olston J, Palmer L. *Biostatistical Genetics and Genetic Epidemiology*. UK: John wiley & Sons Ltd; 2003.

35. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research*. USA: Blackwell Science Ltd; 2002.

36. McCullagh P, Nelder JA. *Generalized Linear Models*. USA: CHAPMAN & HALL/CRC; 1989.

37. Barrett JH, Sheehan NA, Cox A, Worthington J, Cannings C, Teare MD. Family based studies and genetic epidemiology: theory and practice. *Human heredity* 2007;**64**(2): 146-8.

38. Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2002 Jun;**11**(6): 505-12.

39. Hattersley AT, McCarthy MI. What makes a good genetic association study? *Lancet* 2005 Oct 8;**366**(9493): 1315-23.

40. Burton PR. Helping Doctors to Draw Appropriate Inferences from the Analysis of Medical Studies. *Statistics in medicine* 1994 Sep 15;**13**(17): 1699-713.

41. Burton PR, Gurrin LC, Campbell MJ. Clinical significance not statistical significance: a simple Bayesian alternative to p values. *J Epidemiol Commun H* 1998 May;**52**(5): 318-23.

42. Swinscow TD, Campbell MJ. *Statistics at Square One*. London: BMJ Books; 2002.

43. Rothman KJ. *Epidemiology, An Introduction*. New York: Oxford university Press, Inc.; 2002.

44. Grunkemeier GL, Jin R. Power and sample size: how many patients do I need? *The Annals of Thoracic Surgery* 2007 Jun;**83**(6): 1934-9.

45. Lenth RV. Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician* 2001;**55**(N/A): 187-8-93.

46. Schultz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005;**365**(N/A): 1348-9;50;51;52;53.

47. Seo J, Gordish-Dressman H, Hoffman EP. An interactive power analysis tool for microarray hypothesis testing and generation. *Bioinformatics (Oxford, England)* 2006 Apr 1;**22**(7): 808-14.

48. Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS genetics* 2009 May;**5**(5): e1000477.

49. Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ. The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *International journal of epidemiology* 2003 Feb;**32**(1): 51-7.

50. Ihaka R. R: Past and future history. *Comp Sci Stat* 1998;**30**: 392-6.

51. Wunsch H, Linde-Zwirble WT, Angus DC. Methods to adjust for bias and confounding in critical care health services research involving observational data. *Journal of critical care* 2006 Mar;**21**(1): 1-7.

52. Grimes DA, Schulz KF. Compared to what? Finding controls for case-control studies. *Lancet* 2005 Apr 16-22;**365**(9468): 1429-33.

53. Richardson DB. An incidence density sampling program for nested case-control analyses. *Occup Environ Med* 2004 Dec;**61**(12): e59.

54. Bland GM, Altman DG. Statistics notes: Matching. *BMJ* 1994;**309**(NK): 1128-.

55. Sorensen ST, Gillman MW. Matching in case-control studies. *BMJ* 1994;**309**(NK): 1128-.

56. Clayton D, Hills M. *Statistical Models in Epidemiology*. New York: Oxford Science Publications; 1993.

57. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA : the journal of the American Medical Association* 1998 Nov 18;**280**(19): 1690-1.

58. Kuk AYC, Cheng YW. The Monte Carlo Newton-Raphson algorithm. *J Stat Comput Sim* 1997;**59**(3): 233-50.

59. Rothman KJ, Greenland S, Lash L. *Modern Epidemiology*. Philadelphia, PA 19106 USA: LIPPINCOTT WILLIAMS & WILKINS; 2008.

60. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Statistics in medicine* 1983 Apr-Jun;**2**(2): 243-51.

61. Marshall SW. Power for tests of interaction: effect of raising the Type I error rate. *Epidemiologic perspectives & innovations : EP+I* 2007 Jun 19;**4**: 4.

62. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. *BMJ (Clinical research ed)* 1994 Jun 11;**308**(6943): 1552.

63. Motulsky H. *Intuitive Biostatistics*. New York: Oxford university Press, Inc.; 1995.

64. Carvalho BS, Louis TA, Irizarry RA. Quantifying uncertainty in genotype calls. *Bioinformatics* 2010 January 15, 2010;**26**(2): 242-9.

65. Montana G. HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. *Bioinformatics* 2005 Dec 1;**21**(23): 4309-11.

66. Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics* 1965;**29**: 51-71.

67. Hopper JL. Variance components for statistical genetics: applications in medical research to characteristics related to human diseases and health. *Stat Methods Med Res* 1993;**2**(3): 199-223.

68. Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of Genetic Epidemiology*. New York: Oxford University Press, Inc.; 1993.

69. Todorov AA, Suarez BK. Liability of model. *Biostatistical Genetics and Genetic Epidemiology*. Chichester: John Wiley & Sons; 2002. p. 430-5.

70. Borugian MJ, Robson P, Fortier I, et al. The Canadian Partnership for Tomorrow Project: building a pan-Canadian research platform for disease prevention. *Can Med Assoc J* 2010 Aug 10;**182**(11): 1197-201.

71. Burton P, Fortier I, Deschenes M, Hansell A, Palmer L. Biobanks and Biobank Harmonization. In: Palmer l, Burton P, Davey Smith G, editors. *An Introduction to Genetic Epidemiology* Bristol: Policy Press; 2011.

72. Burton P, Fortier I, Knoppers B. The Global Emergence of Epidemiological Biobanks; Opportunities and Challenges. Building the evidence for using genetic information to improve health and prevent disease. In: Khoury M, Gwinn M, Bradley L, Little J, Higgins J, Ioannidis J, editors. *Human Genome Epidemiology (Second Edition).* New York: Oxford University Press; 2010.

73. Burton PR, Clayton DG, Cardon LR, et al. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat Genet* 2007 Nov;**39**(11): 1329-37.

74. Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 2007;**447**(17529967): 1087-93.

75. Frayling TM, Timpson NJ, Weedon MN, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007 May 11;**316**(5826): 889-94.

76. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* 2009 Jun 9;**106**(23): 9362-7.

77. Newton-Cheh C, Eijgelsheim M, Rice KM, et al. Common variants at ten loci influence QT interval duration in the QTGEN Study. *Nat Genet* 2009 Apr;**41**(4): 399-406.

78. Newton-Cheh C, Johnson T, Gateva V, et al. Genome-wide association study identifies eight loci associated with blood pressure. *Nature Genetics* 2009 Jun;**41**(6): 666-76.

79. Saxena R, Voight BF, Lyssenko V, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007;**316**(17463246): 1331-6.

80. Scott LJ, Mohlke KL, Bonnycastle LL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007;**316**(17463248): 1341-5.

81. Stacey SN, Manolescu A, Sulem P, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* 2007 Jul;**39**(7): 865-9.

82. Zeggini E, Weedon MN, Lindgren CM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007;**316**(17463249): 1336-41.

83. CARTaGENE. *Cartagene, the world within you*. 2008 [cited 2012 23.07]; Available from: http://www.cartagene.qc.ca/index.php?option=com_content&task=view&id=2&Itemid=60

84. Franklin SS, Gustin Wt, Wong ND, et al. Hemodynamic patterns of age-related changes in blood pressure. The Framingham Heart Study. *Circulation* 1997 Jul 1;**96**(1): 308-15.

85. Gatzka CD, Cameron JD, Kingwell BA, Dart AM. Relation between coronary artery disease, aortic stiffness, and left ventricular structure in a population sample. *Hypertension* 1998 Sep;**32**(3): 575-8.

86. Hirai T, Sasayama S, Kawasaki T, Yagi S. Stiffness of systemic arteries in patients with myocardial infarction. A noninvasive method to predict severity of coronary atherosclerosis. *Circulation* 1989 Jul;**80**(1): 78-86.

87. Nichols WW, Pepine CJ. Ventricular/vascular interaction in health and heart failure. *Comprehensive therapy* 1992 Jul;**18**(7): 12-9.

88. Benetos A, Adamopoulos C, Bureau JM, et al. Determinants of accelerated progression of arterial stiffness in normotensive subjects and in treated hypertensive subjects over a 6-year period. *Circulation* 2002 Mar 12;**105**(10): 1202-7.

89. Laurent S, Katsahian S, Fassot C, et al. Aortic stiffness is an independent predictor of fatal stroke in essential hypertension. *Stroke* 2003 May;**34**(5): 1203-6.

90. Wilkinson IB, Prasad K, Hall IR, et al. Increased central pulse pressure and augmentation index in subjects with hypercholesterolemia. *Journal of the American College of Cardiology* 2002;**39**(11897443): 1005-11.

91. Blacher J, Guerin AP, Pannier B, Marchais SJ, Safar ME, London GM. Impact of aortic stiffness on survival in end-stage renal disease. *Circulation* 1999;**99**(10318666): 2434-9.

92. Safar ME, London GM, Plante GE. Arterial stiffness and kidney function. *Hypertension* 2004 Feb;**43**(2): 163-8.

93. McEniery CM, Yasmin, McDonnell B, et al. Central pressure: variability and impact of cardiovascular risk factors: the Anglo-Cardiff Collaborative Trial II. *Hypertension* 2008 Jun;**51**(6): 1476-82.

94. Kestenbaum B, Rudser KD, Shlipak MG, et al. Kidney function, electrocardiographic findings, and cardiovascular events among older adults. *Clinical journal of the American Society of Nephrology : CJASN* 2007 May;**2**(3): 501-8.

95. Tavernier R, Jordaens L, Haerynck F, Derycke E, Clement DL. Changes in the QT interval and its adaptation to rate, assessed with continuous electrocardiographic recordings in patients with ventricular fibrillation, as compared to normal individuals without arrhythmias. *European heart journal* 1997 Jun;**18**(6): 994-9.

96. Barr CS, Naas A, Freeman M, Lang CC, Struthers AD. QT dispersion and sudden unexpected death in chronic heart failure. *Lancet* 1994 Feb 5;**343**(8893): 327-9.

97. Onysko J, Maxwell C, Eliasziw M, Zhang JX, Johansen H, Campbell NR. Large increases in hypertension diagnosis and treatment in Canada after a healthcare professional education program. *Hypertension* 2006 Nov;**48**(5): 853-60.

98. Mo F, Pogany LM, Li FC, Morrison H. Prevalence of diabetes and cardiovascular comorbidity in the Canadian Community Health Survey 2002-2003. *TheScientificWorldJournal* 2006;**6**: 96-105.

99. Pauwels RA, Buist AS, Calverley PM, Jenkins CR, Hurd SS. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary. *American journal of respiratory and critical care medicine* 2001 Apr;**163**(5): 1256-76.

100. Schunemann HJ, Dorn J, Grant BJ, Winkelstein W, Jr., Trevisan M. Pulmonary function is a long-term predictor of mortality in the general population: 29-year follow-up of the Buffalo Health Study. *Chest* 2000 Sep;**118**(3): 656-64.

101. Hole DJ, Watt GC, Davey-Smith G, Hart CL, Gillis CR, Hawthorne VM. Impaired lung function and mortality risk in men and women: findings from the Renfrew and Paisley prospective population study. *BMJ* 1996;**313**(8819439): 711-5.

102. Eberly LE, Ockene J, Sherwin R, Yang L, Kuller L. Pulmonary function as a predictor of lung cancer mortality in continuing cigarette smokers and in quitters. *Int J Epidemiol* 2003 Aug;**32**(4): 592-9.

103. Mussolino ME, Madans JH, Gillum RF. Bone mineral density and mortality in women and men: the NHANES I epidemiologic follow-up study. *Ann Epidemiol* 2003 Nov;**13**(10): 692-7.

104. Lochmuller EM, Zeller JB, Kaiser D, et al. Correlation of femoral and lumbar DXA and calcaneal ultrasound, measured in situ with intact soft tissues, with the in vitro failure loads of the proximal femur. *Osteoporos Int* 1998;**8**(10326066): 591-8.

105. Gregg EW, Kriska AM, Salamone LM, et al. The epidemiology of quantitative ultrasound: a review of the relationships with bone mass, osteoporosis and fracture risk. *Osteoporos Int* 1997;**7**(9166387): 89-99.

106. Diaz-Guerra GM, Gil-Fraguas L, Jodar E, et al. Quantitative ultrasound of the calcaneus in long-term liver or cardiac transplantation patients. *Journal of clinical densitometry : the official journal of the International Society for Clinical Densitometry* 2006 Oct-Dec;**9**(4): 469-74.

107. Lim YW, Chan L, Lam KS. Broadband ultrasound attention reference database for southeast Asian males and females. *Annals of the Academy of Medicine, Singapore* 2005 Oct;**34**(9): 545-7.

108. Funke M, Kopka L, Vosshenrich R, et al. Broadband ultrasound attenuation in the diagnosis of osteoporosis: correlation with osteodensitometry and fracture. *Radiology* 1995 Jan;**194**(1): 77-81.

109. Gale CR, Martyn CN, Cooper C, Sayer AA. Grip strength, body composition, and mortality. *Int J Epidemiol* 2007 Feb;**36**(1): 228-35.

110. Ruiz JR, Sui X, Lobelo F, et al. Association between muscular strength and mortality in men: prospective cohort study. *BMJ* 2008;**337**(18595904).

111. Rantanen T, Guralnik JM, Foley D, et al. Midlife hand grip strength as a predictor of old age disability. *JAMA* 1999;**281**(10022113): 558-60.

112. Valentini L, Schaper L, Buning C, et al. Malnutrition and impaired muscle strength in patients with Crohn's disease and ulcerative colitis in remission. *Nutrition* 2008 Jul-Aug;**24**(7-8): 694-702.

113. Allison MA, Michael Wright C. Body morphology differentially predicts coronary calcium. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 2004 Mar;**28**(3): 396-401.

114. Hu G, Tuomilehto J, Silventoinen K, Sarti C, Mannisto S, Jousilahti P. Body mass index, waist circumference, and waist-hip ratio on the risk of total and type-specific stroke. *Archives of internal medicine* 2007 Jul 9;**167**(13): 1420-7.

115. Office of Nutrition Policy and Promotion HC. Canadian Guidelines for Body Weight Classification in Adults. Ottawa: Health Canada Publications Centre; 2003.

116. Ribeiro-Filho FF, Faria AN, Azjen S, Zanella MT, Ferreira SR. Methods of estimation of visceral fat: advantages of ultrasonography. *Obesity research* 2003 Dec;**11**(12): 1488-94.

117. Schreiner PJ, Terry JG, Evans GW, Hinson WH, Crouse JR, Heiss G. Sex-specific associations of magnetic resonance imaging-derived intra-abdominal and subcutaneous fat areas with conventional anthropometric indices. The Atherosclerosis Risk in Communities Study. *Am J Epidemiol* 1996;**144**(8712190): 335-45.

118. Lean ME, Han TS, Seidell JC. Impairment of health and quality of life in people with large waist circumference. *Lancet* 1998;**351**(9525361): 853-6.

119. Zhu S, Wang Z, Heshka S, Heo M, Faith MS, Heymsfield SB. Waist circumference and obesity-associated risk factors among whites in the third National Health and Nutrition Examination Survey: clinical action thresholds. *Am J Clin Nutr* 2002 Oct;**76**(4): 743-9.

120. Czernichow S, Bertrais S, Oppert JM, et al. Body composition and fat repartition in relation to structure and function of large arteries in middle-aged adults (the SU.VI.MAX study). *Int J Obes (Lond)* 2005 Jul;**29**(7): 826-32.

121. Donato GB, Fuchs SC, Oppermann K, Bastos C, Spritzer PM. Association between menopause status and central adiposity measured at different cutoffs of waist circumference and waist-to-hip ratio. *Menopause* 2006 Mar-Apr;**13**(2): 280-5.

122. Chen Y, Rennie D, Cormier YF, Dosman J. Waist circumference is associated with pulmonary function in normal-weight, overweight, and obese subjects. *Am J Clin Nutr* 2007 Jan;**85**(1): 35-9.

123. Evans DL, Charney DS. Mood disorders and medical illness: a major public health problem. *Biological psychiatry* 2003 Aug 1;**54**(3): 177-80.

124. Kessler RC, Ormel J, Demler O, Stang PE. Comorbid mental disorders account for the role impairment of commonly occurring chronic physical disorders: results from the National Comorbidity Survey. *Journal of occupational and environmental medicine / American College of Occupational and Environmental Medicine* 2003 Dec;**45**(12): 1257-66.

125. Vasiliadis HM, Lesage A, Adair C, Wang PS, Kessler RC. Do Canada and the United States differ in prevalence of depression and utilization of services? *Psychiatr Serv* 2007 Jan;**58**(1): 63-71.

126. Blumenthal JA. Depression and coronary heart disease: association and implications for treatment. *Cleveland Clinic journal of medicine* 2008 Mar;**75 Suppl 2**: S48-53.

127. Carney RM, Freedland KE, Steinmeyer B, et al. Depression and five year survival following acute myocardial infarction: a prospective study. *Journal of affective disorders* 2008 Jul;**109**(1-2): 133-8.

128. Jiang W. Impacts of depression and emotional distress on cardiac disease. *Cleveland Clinic journal of medicine* 2008 Mar;**75 Suppl 2**: S20-5.

129. Potyralska MM, Krawczyk AK. [Depression in patients with type 2 diabetes mellitus--clinical and therapeutical implications]. *Wiad Lek* 2007;**60**(18350720): 449-53.

130. Sui X, Church TS, Meriwether RA, Lobelo F, Blair SN. Uric acid and the development of metabolic syndrome in women and men. *Metabolism: clinical and experimental* 2008 Jun;**57**(6): 845-52.

131. Nan H, Qiao Q, Soderberg S, et al. Serum uric acid and components of the metabolic syndrome in non-diabetic populations in Mauritian Indians and Creoles and in Chinese in Qingdao, China. *Metabolic syndrome and related disorders* 2008 Mar;**6**(1): 47-57.

132. Brown JR, Cochran RP, Dacey LJ, et al. Perioperative increases in serum creatinine are predictive of increased 90-day mortality after coronary artery bypass graft surgery. *Circulation* 2006 Jul 4;**114**(1 Suppl): I409-13.

133. Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. *Nat Genet* 2006;**38**(16715099): 659-62.

134. Dyer AR, Elliott P, Shipley M. Urinary electrolyte excretion in 24 hours and blood pressure in the INTERSALT Study. II. Estimates of electrolyte-blood pressure associations corrected for regression dilution bias. The INTERSALT Cooperative Research Group. *Am J Epidemiol* 1994;**139**(8166144): 940-51.

135. Genomics PPPi. *The Biobank Lexicon*. 2005 [cited 2011 29.12.2011]; Available from: http://p3g.org/

136. Holland NT, Smith MT, Eskenazi B, Bastaki M. Biological sample collection and processing for molecular epidemiological studies. *Mutation research* 2003 Jun;**543**(3): 217-34.

137. Bowen RA, Hortin GL, Csako G, Otanez OH, Remaley AT. Impact of blood collection devices on clinical chemistry assays. *Clin Biochem* 2010 Jan;**43**(1-2): 4-25.

138. Hallmans G, Vaught JB. Best practices for establishing a biobank. *Methods in molecular biology* 2011;**675**: 241-60.

139. Lehmann S, Roche S, Allory Y, et al. Preanalytical guidelines for clinical proteomics investigation of biological fluids. *Annales de Biologie Clinique* 2009 Nov-Dec;**67**(6): 629-39.

140. Vaught JB. Blood collection, shipment, processing, and storage. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2006 Sep;**15**(9): 1582-4.

141. Public Population Project in G. P3G Observatory - Lexicon. Public Population Project in Genomics; 2005.

142. Hallmans G, Vaught JB. Best practices for establishing a biobank. *Methods in molecular biology (Clifton, NJ)*;**675**: 241-60.

143. International Society for B, Environmental R. ISBER Best Practices For Repositories. ISBER; 2009.

144. Public Population Project in G. Sample Collection and Processing. Public Population Project in Genomics; 2009.

145. Clark S, Youngman LD, Palmer A, Parish S, Peto R, Collins R. Stability of plasma analytes after delayed separation of whole blood: implications for epidemiological studies. *International journal of epidemiology* 2003 Feb;**32**(1): 125-30.

146. BioSHARE-EU. Biobank Standardisation and Harmonisation for Research Excellence in the European Union (BIOSHARE-EU) In: Groningen AZ, editor. Groningen, Netherlands: CORDIS RTD-PROJECTS; 2010.

147. Jackson C, Best N, Elliott P. UK Biobank Pilot Study: stability of haematological and clinical chemistry analytes. *International journal of epidemiology* 2008 Apr;**37 Suppl 1**: i16-22.

148. Peakman TC, Elliott P. The UK Biobank sample handling and storage validation studies. *International journal of epidemiology* 2008 Apr;**37 Suppl 1**: i2-6.

149. Wolfson M, Wallace SE, Masca N, et al. DataSHIELD: resolving a conflict in contemporary bioscience--performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol* 2010 Oct;**39**(5): 1372-82.

150. Fortier I, Burton PR, Robson PJ, et al. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010;**39**(20813861): 1383-93.

151. Rasbash J, Charlton C, Browne WJ, Healy M, Cameron B. *MLwiN Version 2.1*. UK: Centre for Multilevel Modelling, University of Bristol; 2009.

152. Dunn WB, Broadhurst D, Ellis DI, et al. A GC-TOF-MS study of the stability of serum and urine metabolomes during the UK Biobank sample collection and preparation protocols. *International journal of epidemiology* 2008 Apr;**37 Suppl 1**: i23-30.

153. Power C. Cohort profile: 1958 British birth cohort (National Child Development Study). *International journal of epidemiology* 2006;**35**(1): 34-41.

154. Wellcome Trust Case Control C, Craddock N, Hurles ME, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 2010 Apr 1;**464**(7289): 713-20.

155. Barnes C, Plagnol V, Fitzgerald T, et al. A robust statistical method for case-control association testing with copy number variation. *Nature genetics* 2008 Oct;**40**(10): 1245-52.

156. Kitchen RR, Sabine VS, Simen AA, Dixon JM, Bartlett JM, Sims AH. Relative impact of key sources of systematic noise in Affymetrix and Illumina gene-expression microarray experiments. *Bmc Genomics* 2011;**12**: 589.

157. Schwartz D, Collins F. Medicine. Environmental biology and human disease. *Science* 2007;**316**(17478705): 695-6.

APPENDIX

8. APPENDIX

Appendix 1

This appendix contains the R-package manual of the ESPRESSO-forte algorithm developed in this thesis.
P.S: The original name 'ESPRESSO' was kept for the ESPRESSO-forte R package submitted to the Comprehensive R Archive Network (CRAN) because 'ESPRESSO-forte' was not a valid package name according to CRAN standards.

Package 'ESPRESSO'

May 20, 2012

Type Package

Title Power Analysis and Sample Size Calculation

Version 1.3

Date 2011-04-01

Author Amadou Gaye under the supervision of Prof Paul Burton

Maintainer Amadou Gaye <ag239@le.ac.uk>

Description The package allows for the Estimation of Sample-size and Power by Exploring Simulated Study Outcomes. It supports simulation-based power calculation for stand-alone case-control studies and for case-control analyses nested in cohort studies that take account of realistic assessment error.

Depends MASS

License GPL-2
LazyLoad yes

R topics documented:

ESPRESSO Package for power analysis and sample size calculation

Description

A package to estimate sample-size and power by exploring simulated study outcomes. It supports simulation-based power calculation for stand-alone case-control studies and for case-control analyses nested in cohort studies that take account of realistic assessment error.

Details

ESPRESSO(Estimating Sample-size and Power in R by Exploring Simulated Study Outcomes)allows for the calculation of the sample size required to achieve a desired statistical power in a case control study. It also allows one to calculate the power achieved with a specified sample size. The simulated dataset consists of a binary or a continuous outcome and two genetic and two environmental determinants. The functions *sim.CC.data* and *sim.QTL.data* simulate the outcome (phenotype) and the initial effects data considered as the true measures of the determinants. The function *make.obs.data* adds some error to the effect data generated by *sim.CC.data* or sim.QTL.data to obtain the observed measures of the determinants. The function *regr.analysis* carries out a regression analysis of the covariates (genetic variants, environmental exposures and interaction term) over the outcome. The function *samplsize.calc* calculates the sample sizes required to achieve the desired power under the specified effect model (main effect or interaction). The functions e*mpirical.power.calc* and *model.power.calc* calculate, respectively, the empirical power and the theoretical power achieved under the specified sample size.

Author(s)

Amadou Gaye under the supervision of Prof. Paul Burton

Maintainer

Amadou Gaye <ag239@le.ac.uk>

References

Burton, P.R., Hansell, A.L., Fortier, I., Manolio, T.A., Khoury, M.J., Little, J. & Elliott, P. 2009,Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology, International journal of epidemiology,vol. 38, no. 1, pp.263-273.

Examples

```
# This example illustrates how to use the main functions of the
# package.
# load input control files and make one table of parameters
data(general.params)
data(gen.params)
data(env.params)
s.temp  <-  merge(general.params, gen.params)
s.parameters  <-  merge(s.temp, env.params)
# create up to 20m subjects in blocks of 20k until required
number
# of cases and controls is achieved. in general the only problem
# in achieving the required number of cases will occur if the
# disease prevalence is very low.
```

```
allowed.sample.size <- 20000000
block.size <- 20000
# tracer to monitor iterations
#trace.interval <- 10
# total number of scenarios
numscenarios  <-  dim(s.parameters)[1]
# scenario to start with
start.at.scenario  <-  1
# number of scenarios to run
stop.at.scenario  <-  numscenarios
for(j in start.at.scenario : stop.at.scenario)
{
set.seed(s.parameters$seed.val[j])
# general parameters
scenario.id  <-  s.parameters$scenario.id[j]
seed.val  <-  s.parameters$seed.val[j]
numsims  <-  s.parameters$numsims[j]
numcases  <-  s.parameters$numcases[j]
numcontrols  <-  s.parameters$numcontrols[j]
num.subjects  <-  s.parameters$num.subjects[j]
is.interaction  <-  s.parameters$interaction[j]
pheno.model  <-  s.parameters$pheno.model[j]
disease.prev  <-  s.parameters$disease.prev[j]
or.int  <-  s.parameters$or.int[j]
int.efkt  <-  s.parameters$int.efkt[j]
if(is.interaction != 0) {or.int  <-  s.parameters$or.int[j]}
sigma.subject  <-  s.parameters$RR.5.95[j]
pval  <-  s.parameters$p.val[j]
power  <-  s.parameters$power[j]
pheno.error  <-  c(1-s.parameters$sensitivity.pheno[j],
1-s.parameters$specificity.pheno[j])
reliability.pheno  <-  s.parameters$reliability.pheno[j]
# genetic determinants parameters
is.add  <-  c(s.parameters$model.geno1[j],
s.parameters$model.geno2[j])
MAF  <-  c(s.parameters$MAF.geno1[j], s.parameters$MAF.geno2[j])
or.geno  <-
c(s.parameters$or.geno1[j],s.parameters$or.geno2[j])
geno.efkt  <-  c(s.parameters$geno1.efkt[j],
s.parameters$geno2.efkt[j])
LD  <-  s.parameters$LD[j]
R.target  <-  s.parameters$R.target[j]
display  <-  s.parameters$display[j]
geno.error  <-  c(1-s.parameters$sensitivity.geno[j],
1-s.parameters$specificity.geno[j])
# environmental determinants parameters
env.expo  <-  c(s.parameters$model.env1[j],
s.parameters$model.env2[j])
reliability.env  <-  c(s.parameters$reliability.env1[j],
s.parameters$reliability.env2[j])
env.prev  <-
c(s.parameters$env1.prev[j],s.parameters$env2.prev[j])
env.mean.lowlm  <-  c(s.parameters$env1.mean.lowlm,
s.parameters$env2.mean.lowlm)
env.stdev.uplm  <-  c(s.parameters$env1.stdev.uplm,
s.parameters$env2.stdev.uplm)
or.env  <-  c(s.parameters$or.env1[j],s.parameters$or.env2[j])
env.efkt  <-
c(s.parameters$env1.efkt[j],s.parameters$env2.efkt[j])
env.error  <-  c(1-s.parameters$sensitivity.env[j],
1-s.parameters$specificity.env[j])
```

```
skewness  <-  c(s.parameters$skewness1,s.parameters$skewness2)
# the covariance matrix required to generate 2 variants with
# the desired ld
cor.mat  <-  matrix(c(1,R.target,R.target,1),2,2) # cor. matrix
cov.mat.req  <-  make.cov.mat(cor.mat, c(1-MAF[1], 1-MAF[2]))
# if the required covariance matrix is not positive-definite get
# the nearest positive-definite matrix (tolerance = 1e-06)
if(!is.posdef(cov.mat.req, 0.000001)){
cov.mat.req  <-  make.posdef(cov.mat.req, 0.000001)
}
# empty vectors for results of the analyses of each simulation
# in the scenario
# genotype
beta.geno1.results <- rep(NA,numsims)
se.geno1.results <- rep(NA,numsims)
z.geno1.results <- rep(NA,numsims)
beta.geno2.results <- rep(NA,numsims)
se.geno2.results <- rep(NA,numsims)
z.geno2.results <- rep(NA,numsims)
# environment
beta.env1.results <- rep(NA,numsims)
se.env1.results <- rep(NA,numsims)
z.env1.results <- rep(NA,numsims)
beta.env2.results <- rep(NA,numsims)
se.env2.results <- rep(NA,numsims)
z.env2.results <- rep(NA,numsims)
# interaction
beta.int.results <- rep(NA,numsims)
se.int.results <- rep(NA,numsims)
z.int.results <- rep(NA,numsims)
# tracer to detect exceeding max allowable sample size
sample.size.excess  <-  0
# generate and analyse datasets one at a time
for(s in 1:numsims)
{
if(pheno.model == 0){ # under binary outcome
# generate cases and controls untill the required number
# of cases, controls and sample size is achieved
sim.matrix  <-  sim.CC.data(block.size, numcases, numcontrols,
allowed.sample.size, is.interaction,
disease.prev, MAF, is.add, R.target,
LD, cov.mat.req, display, or.geno, env.expo,
env.mean.lowlm, env.stdev.uplm, env.prev,
or.env, skewness, or.int, sigma.subject,
pheno.error)
}else{ # under quantitative outcome model
# generate the specified number of subjects
sim.matrix  <-  sim.QTL.data(num.subjects, is.interaction,
MAF, is.add, R.target, LD, cov.mat.req,
display, geno.efkt, env.expo, env.mean.lowlm,
env.stdev.uplm,env.prev, env.efkt, skewness,
int.efkt, reliability.pheno)
}
# add appropriate errors to produce observed genotypes
observed.data  <-  get.observed.data(is.interaction, sim.matrix,
geno.error, is.add, MAF, env.expo, env.prev,
env.error, reliability.env)
sim.df  <-  observed.data$sim.df
# data analysis
glm.estimates  <-  regr.analysis(is.interaction, pheno.model,
sim.df)
```

```
# genetic variants estimates
beta.geno1.results[s]  <-  glm.estimates[1]
se.geno1.results[s]  <-  glm.estimates[2]
z.geno1.results[s]  <-  glm.estimates[3]
beta.geno2.results[s]  <-  glm.estimates[4]
se.geno2.results[s]  <-  glm.estimates[5]
z.geno2.results[s]  <-  glm.estimates[6]
# environment estimates
beta.env1.results[s]  <-  glm.estimates[7]
se.env1.results[s]  <-  glm.estimates[8]
z.env1.results[s]  <-  glm.estimates[9]
beta.env2.results[s]  <-  glm.estimates[10]
se.env2.results[s]  <-  glm.estimates[11]
z.env2.results[s]  <-  glm.estimates[12]
# interaction estimates
beta.int.results[s]  <-  glm.estimates[13]
se.int.results[s]  <-  glm.estimates[14]
z.int.results[s]  <-  glm.estimates[15]
# print tracer after every nth dataset created
# if(s %% trace.interval ==0)cat("\n",s,"of",numsims,
# "completed in scenario",scenario.id)
}
cat("\n\n")
# summary of primary parameter estimates
# genetic variants
mean.beta.geno1  <-  round(mean(beta.geno1.results),3)
mean.se.geno1  <-  round(sqrt(mean(se.geno1.results^2)),3)
mean.model.z.geno1  <-  mean.beta.geno1/mean.se.geno1
mean.beta.geno2  <-  round(mean(beta.geno2.results),3)
mean.se.geno2  <-  round(sqrt(mean(se.geno2.results^2)),3)
mean.model.z.geno2  <-  mean.beta.geno2/mean.se.geno2
mean.model.z.geno  <-  c(mean.beta.geno1/mean.se.geno1,
mean.beta.geno2/mean.se.geno2)
# environments
mean.beta.env1  <-  round(mean(beta.env1.results),3)
mean.se.env1  <- round(sqrt(mean(se.env1.results^2)),3)
mean.model.z.env1  <-  mean.beta.env1/mean.se.env1
mean.beta.env2  <-  round(mean(beta.env2.results),3)
mean.se.env2  <- round(sqrt(mean(se.env2.results^2)),3)
mean.model.z.env2  <-  mean.beta.env2/mean.se.env2
mean.model.z.env  <-  c(mean.beta.env1/mean.se.env1,
mean.beta.env2/mean.se.env2)
# interaction
if(is.interaction == 0){
mean.beta.int  <-  NA
mean.se.int  <-  NA
mean.model.z.int  <-  NA
}else{
gen.params 11
mean.beta.int <- round(mean(beta.int.results),3)
mean.se.int <- round(sqrt(mean(se.int.results^2)),3)
mean.model.z.int  <-  mean.beta.int/mean.se.int
}
mean.betas  <-  c(mean.beta.geno1, mean.beta.geno2,
mean.beta.env1, mean.beta.env2, mean.beta.int)
# calculate the sample size required under each model
sample.sizes.required  <-  samplsize.calc(numcases,numcontrols,
num.subjects, pheno.model,
is.interaction, pval, power,
mean.model.z.geno,mean.model.z.env,
mean.model.z.int)
```

```
# calculate empirical power ie simply the proportion of
# simulations in which the z statistic for the parameter of
# interest exceeds the z statistic for the desired level of
# statistical significance
empirical.power  <-  empirical.power.calc(is.interaction,pval,
z.geno1.results,z.geno2.results,z.env1.results,
z.env2.results, z.int.results)
# calculate the power reached under the initial sample size
model.power  <-  model.power.calc(is.interaction, pval,
mean.model.z.geno, mean.model.z.env,
mean.model.z.int)
# return critical results and print a summary
res  <-  get.critical.results(j, is.interaction, pheno.model,
is.add, env.expo, sample.sizes.required, empirical.power,
model.power, mean.betas)
}
```

---

empirical.power.calc Calculates empirical power

Description

Determines the proportion of simulations in which the z-statistic for the parameter of interest exceeds the z-statistic for the desired level of statistical significance.

Usage

```
empirical.power.calc(is.interaction=0, pval,
z.geno1.results,z.geno2.results, z.env1.results, z.env2.results,
z.int.results)
```

Arguments

`is.interaction`
: Effect model: main effects=0, Gene-Environment interaction=1, Gene-Gene interaction=2 and Environment-Environment interaction=3

`pval`
: Cut-off p-value defining statistical significance

`z.geno1.results`
: Vector of z-statistics for the main effect of genetic variant 1

`z.geno2.results`
: Vector of z-statistics for the main effect of genetic variant 2

`z.env1.results`
: Vector of z-statistics for the main effect of environment 1

`z.env2.results`
: Vector of z-statistics for the main effect of environment 2

`z.int.results`
: Vector of z-statistics for interaction effect

Value
A list containing:

`empirical.power.geno1`
Empirical power under the main effect of genetic variant 1

`empirical.power.geno2`
Empirical power under the main effect of genetic variant 2

`empirical.power.env1`
Empirical power under the main effect environment 1

`empirical.power.env2`
Empirical power under the main effect environment 2

`empirical.power.int`
Empirical power under Interaction

Author(s)
Amadou Gaye

See Also
samplsize.calc, model.power.calc

Examples

```
# set outcome type to binary
pheno.model <- 0
# set the model
is.interaction  <- 0 # main effect model
# number of runs
numsims  <-  10
# empty vectors to store the results of each run
# gene
beta.geno1.results <- rep(NA,numsims)
se.geno1.results <- rep(NA,numsims)
z.geno1.results <- rep(NA,numsims)
beta.geno2.results <- rep(NA,numsims)
se.geno2.results <- rep(NA,numsims)
z.geno2.results <- rep(NA,numsims)
# environment
beta.env1.results <- rep(NA,numsims)
se.env1.results <- rep(NA,numsims)
z.env1.results <- rep(NA,numsims)
beta.env2.results <- rep(NA,numsims)
se.env2.results <- rep(NA,numsims)
z.env2.results <- rep(NA,numsims)
# interaction
beta.int.results <- rep(NA,numsims)
se.int.results <- rep(NA,numsims)
z.int.results <- rep(NA,numsims)
# tracer to detect exceeding max allowable sample size
sample.size.excess  <-  0
# generate and analyse datasets one at a time
for(s in 1:numsims)
{
if(pheno.model == 0){# under binary outcome
# generate cases and controls untill the required number
# of cases, controls and sample size is achieved
sim.matrix  <-  sim.CC.data(20000,2000,8000,20000000,0,0.1,
```

```
c(0.1,0.1),c(0,0),0.7,0,cov.mat.req=NULL, FALSE,c(1.5,1.5),
c(0,0),c(3.3,3.3),c(1,1),c(0.1,0.1),c(1.5,1.5),c(0,0),1.8,
12.36,c(0,0))
}else{# under quantitative outcome model
# generate the specified number of subjects
sim.matrix  <-  sim.QTL.data(1000,0,c(0.1,0.1),c(0,0),
0.7,0,cov.mat.req=NULL,FALSE,c(0.25,0.25),c(0,0),c(3.3,3.3),
c(1,1),c(0.1,0.1),c(0.25,0.25),c(0,0),0.5,0.9)
}
# add appropriate errors to produce observed genotypes
observed.data  <-  get.observed.data(0, sim.matrix,c(0.95,0.95),
c(0,0),c(0.1,0.1),c(0,0),c(0.1,0.1),c(0.85,0.85),c(0.8,0.8))
sim.df  <-  observed.data$sim.df
# data analysis
glm.estimates  <-  regr.analysis(is.interaction, pheno.model,
sim.df)
# genetic variants estimates
beta.geno1.results[s]  <-  glm.estimates[1]
se.geno1.results[s]  <-  glm.estimates[2]
z.geno1.results[s]  <-  glm.estimates[3]
beta.geno2.results[s]  <-  glm.estimates[4]
se.geno2.results[s]  <-  glm.estimates[5]
z.geno2.results[s]  <-  glm.estimates[6]
# environment estimates
beta.env1.results[s]  <-  glm.estimates[7]
se.env1.results[s]  <-  glm.estimates[8]
z.env1.results[s]  <-  glm.estimates[9]
beta.env2.results[s]  <-  glm.estimates[10]
se.env2.results[s]  <-  glm.estimates[11]
z.env2.results[s]  <-  glm.estimates[12]
# interaction estimates
beta.int.results[s]  <-  glm.estimates[13]
se.int.results[s]  <-  glm.estimates[14]
z.int.results[s]  <-  glm.estimates[15]
# print tracer after every nth dataset created
# if(s %% trace.interval ==0)cat("\n",s,"of",numsims,
# "completed in scenario",scenario.id)
}
cat("\n\n")
# summary of primary parameter estimates
env.params 5
# genetic variants
mean.beta.geno1  <-  round(mean(beta.geno1.results),3)
mean.se.geno1  <-  round(sqrt(mean(se.geno1.results^2)),3)
mean.model.z.geno1  <-  mean.beta.geno1/mean.se.geno1
mean.beta.geno2  <-  round(mean(beta.geno2.results),3)
mean.se.geno2  <-  round(sqrt(mean(se.geno2.results^2)),3)
mean.model.z.geno2  <-  mean.beta.geno2/mean.se.geno2
mean.model.z.geno  <-  c(mean.beta.geno1/mean.se.geno1,
mean.beta.geno2/mean.se.geno2)
# environments
mean.beta.env1  <-  round(mean(beta.env1.results),3)
mean.se.env1  <- round(sqrt(mean(se.env1.results^2)),3)
mean.model.z.env1  <-  mean.beta.env1/mean.se.env1
mean.beta.env2  <-  round(mean(beta.env2.results),3)
mean.se.env2  <- round(sqrt(mean(se.env2.results^2)),3)
mean.model.z.env2  <-  mean.beta.env2/mean.se.env2
mean.model.z.env  <-  c(mean.beta.env1/mean.se.env1,
mean.beta.env2/mean.se.env2)
# interaction
if(is.interaction == 0){
```

```
mean.beta.int  <-  NA
mean.se.int  <-  NA
mean.model.z.int  <-  NA
}else{
mean.beta.int <- round(mean(beta.int.results),3)
mean.se.int <- round(sqrt(mean(se.int.results^2)),3)
mean.model.z.int  <-  mean.beta.int/mean.se.int
}
# calculate empirical power ie simply the proportion of
# simulations in which the z statistic for the parameter of
# interest exceeds the z statistic for the desired level of
# statistical significance
empirical.power  <-  empirical.power.calc(is.interaction,1e-04,
z.geno1.results,z.geno2.results,z.env1.results,
z.env2.results, z.int.results)
```

---

env.params Parameters to simulate environmental exposures data

---

Description

A table of scenarios (rows) and parameters (columns).

Usage

```
data(env.params)
```

Format

A data frame with 24 observations for the following21 variables:

`scenario.id`
: Scenario number

`model.env1`
: Models of the first environmental exposure: 0 for binary, 1 for quantitative-normaland 2 for quantitative-uniform

`model.env2`
: Models of the first environmental exposure: 0 for binary, 1 for quantitative-normaland 2 for quantitative-uniform

`reliability.env2`
: Reliability of the assessment of quantitative exposure 1

`reliability.env2`
: Reliability of the assessment of quantitative exposure 2

`env1.prev`
: Prevalence of environment 1

`env2.prev`
: Prevalence of environment 2

`env1.mean.lowlm`

Mean measure for environment 1 under quantitative-normal model and lower limit under quantitative-uniform model

`env2.mean.lowlm`
Mean measure for environment 2 under quantitative-normal model and lower limit under quantitative-uniform model

`env1.stdev.uplm`
Standard deviation under quantitative-normal model and upper limit under quantitative-uniform model, environment 1

`env2.stdev.uplm`
Standard deviation under quantitative-normal model and upper limit under quantitative-uniform model, environment 2

`or.env1`
Odds ratio for environment 1

`env1.efkt`
Effect size for environment 1

`or.env2`
Odds ratio for environment 1

`env2.efkt`
Effect size for environment 2

`sensitivity.env`
Sensitivity of the assessment to environment 1

`specificity.env`
Specificity of the assessment to environment 2

`skewness1`
Determines skewness under quantitative-normal model for environment 1;rightskeweddistribution if set to a positive value and left-skewed when set to a negative value

`skewness2`
Determines skewness under quantitative-normal model for environment 2; right skewed distribution if set to a positive value and left-skewed when set to a negative value

Examples

```
data(env.params)
```

---

gen.params Parameters to simulate genetic data

Description

The table contains scenarios (rows) and parameters (columns).

Usage

```
data(gen.params)
```

Format

A data frame with 24 observations for the following14 variables.

`scenario.id`
: Scenario number; each row stores parameters for one scenario

`model.geno1`
: Genetic model of the first variant: 0 for binary and 1 for additive

`model.geno2`
: Genetic model of the second variant: 0 for binary and 1 for additive

`MAF.geno1`
: Minor allele frequency of genetic variant 1

`MAF.geno2`
: Minor allele frequency of genetic variant 2

`or.geno1`
: Odds-ratio of genetic variant 1

`geno1.efkt`
: Effect of genetic variant 2

`or.geno2`
: Odds-ratio of genetic variant 1

`geno2.efkt`
: Effect of genetic variant 2

`LD`
: Sets independence or LD between the two genetic variants: 0 for independence and 1 for LD

`R.target`
: Correlation coefficient required if the alleles of the two genetic variants are in LD

`display`
: If TRUE, a summary is printed on screen

`sensitivity.geno`
: Sensitivity of the assessment of genotype

`specificity.geno`
: Specificity of the assessment of genotype

Examples

```
data(gen.params)
```

general.params Main parameters of the simulations

Description

A table of scenarios (rows) and parameters (columns).

Usage

```
data(general.params)
```

Format

A data frame with 24 observations for the following18 variables.

`scenario.id`
: The id of the scenario (each row stores parameters for one scenario

`seed.val`
: Seed value

`numsims`
: Number of runs for each simulation

`numcases`
: Number of cases under binary outcome

`numcontrols`
: Number of controls under binary outcome

`num.subjects`
: Number of subjects under continuous outcome

`interaction`
: Effect model: main effects=0, Gene-Environment interaction=1, Gene-Gene interaction=2 and Environment-Enviroment interaction=3

`pheno.model`
: Type of the outcome; 0 for binary and 1 for continuous

`disease.prev`
: Prevalence of the binary outcome

`or.int`
: Odds ration of the interaction

`int.efkt`
: Interaction effect

`RR.5.95`
: The baseline odds ratio for subjects on 95 percent population centile versus 5 percentile. This parameter reflects the heterogeneity in disease risk arising from determinants that have not been measured or have not been included in the model. If this parameter is set to 10, the implication is that a high risk subject (someone at the upper 95 percent centile of population risk) is, all else being equal, at 10 times the odds of

developing disease compared to someone else who is at low risk (individual at the lower 5 percent centile of population risk).

`p.val`
Cut-off p-value defining statistical significance

`power`
Desired power

`sensitivity.pheno`
Sensitivity of the assessment of binary outcome

`specificity.pheno`
Specificity of the assessment of binary outcome

`reliability.pheno`
Reliability of the assessment of continuous outcome

`sim.sesp.geno.env`
Tells if sensitivity and specificity values should be simulated for genotypicand environmental exposures assessment; set to 1 to simulate.

Examples

```
data(general.params)
```

---

get.critical.results Summarizes the main results

Description

Gets the number of cases and controls or subjects and the empirical and theoretical power under each model and prints a summary on the screen.

Usage

```
get.critical.results(scenario, is.interaction = 0, pheno.model =
0, is.add = c(0, 0), env.expo = c(0, 0), sample.sizes.required,
empirical.power, model.power, mean.betas)
```

Arguments

`scenario`
Scenario number

`is.interaction`
Effect model: main effects=0, Gene-Environment interaction=1, Gene-Gene interaction=2 and Environment-Enviroment interaction=3

`pheno.model`
Type of the outcome; 0 for binary and 1 for continuous

`is.add`
Genetic models of the two variants: 0 for binary model and 1 for additive model.

`env.expo`
: Models of the environmental exposures

`sample.sizes.required`
: Number of cases and controls or number of subjects required to achieve thedesired power.

`empirical.power`
: Estimated empirical power

`model.power`
: Calculated theoretical power

`mean.betas`
: Mean beta value of each of the determinants

Value

A table containing the following variables:

`models`
: Model of each of the covariates

`numcases`
: Number of cases required to achieve the desired power under each model

`numcases`
: Number of controls required to achieve the desired power under each model.

`powers1`
: Estimated empirical power under each model

`powers2`
: Power achieved under each model with the specified sample size

`models`
: Model of each of the covariates

`numsubjects`
: Number of subjects required to achieve the desired power under each model

`est.ORs`
: Estimated odds-ratios - due to shrinkage toward the null resulting from misclassification

Author(s)

Amadou Gaye

Examples

```
# scenario number
j  <-  1
# main effect model
is.interaction  <-  1
# outcome
pheno.model  <-  1
```

```
# models of the genetic variants
is.add  <-  c(0,0)
# models of the environmental exposures
env.expo  <-  c(0,0)
# Estimated sample sizes required for a continuous outcome
sample.sizes.required  <-  c(1000, 1300, 2000, 2400, 7000)
# Estimated values for empirical power
empirical.power  <-  c(0.2,0.3,0.56,0.6,0.15)
# power values calculated from the set number of subjects
model.power  <-  c(0.18,0.27,0.58,0.59,0.17)
# mean beta values for each determinant
mean.betas  <-  c(0.18,0.18,0.10,0.10,0.15)
# return critical results and print a summary
res  <-  get.critical.results(j, is.interaction, pheno.model,
is.add, env.expo, sample.sizes.required, empirical.power,
model.power, mean.betas)
```

---

get.observed.data Generates exposure data with some error

Description

Uses functions make.obs.geno and make.obs.env to generate effect data with a set level oferror.

Usage

```
get.observed.data(is.interaction = 0, true.data, geno.error =
c(0.05, 0.05), is.add = c(0, 0), MAF = c(0.1, 0.1), env.expo =
c(0, 0), env.prev = c(0.1, 0.1), env.error = c(0.15, 0.15),
reliability.env=c(0.8,0.8))
```

Arguments

`is.interaction`
: Effect model: main effects=0, Gene-Environment interaction=1, Gene-Gene interaction=2 and Environment-Enviroment interaction=3

`true.data`
: Input table of simulated data considered as true data

`geno.error`
: Misclassification rates in the assessment of genotypes

`is.add`
: Genetic models of the two variants: 0 for binary model and 1 for additive model

`MAF`
: Minor Allele frequencies

`env.expo`
: Model of the exposure: binary=1, quantitative-normal=1 or quantitative-uniform=2

`env.prev`
: Prevalence of the two environmental exposures

`env.error`
: Misclassification rates in environmental exposures assessment: 1-sensitivityand 1-specificity

`reliability.env`
: Reliability of the assessment of quantitative exposures

Value

A matrix containing 11 variables

Author(s)

Amadou Gaye

See Also

make.obs.geno, make.obs.env

Examples

```
# load the 'true' data
data(init.data)
# adds some error to the 'true' exposure data and generate
# 'observed' data
obs.data  <-  get.observed.data(0, init.data, c(0.05, 0.05),
c(0, 0), c(0.1, 0.1), c(0, 0), c(0.1, 0.1),
c(0.15, 0.15), c(0.8, 0.8))
```

---

init.data Simulated genotypes, phenotypes and environmental exposure data

---

Description

A table of simulated true data. The number of rows represents the number of cases and controls (under binary outcome) or the number of subjects (under continuous outcome).

Usage

```
data(init.data)
```

Format

A data frame with 10000 observations for the following11 variables:

`id`
: Scenario id

`cc.U`
: Phenotypes

`geno1.U`
: Genotype for the first genetic variant

`geno2.U`
: Genotype for the second genetic variant

`allele.A1`
: Alelle A of genetic variant 1

`allele.B1`
: Alelle B of genetic variant 1

`allele.A2`
: Alelle A of genetic variant 2

`allele.B2`
: Alelle B of genetic variant 2

`env1.U`
: Exposure data for environment 1

`env2.U`
: Exposure data for environment 2

`int.U`
: Data for the interaction term

Examples

```
data(init.data)
```

---

is.posdef Tells if a matrix is positive definite

Description

Checks if any of the eigenvalues of the matrix is smaller than the set tolerance value.

Usage

```
is.posdef(matrix, tolerance = 1e-06)
```

Arguments

`matrix`
: Input matrix

`tolerance`
: A constant

Value

TRUE or FALSE

Author(s)

Amadou Gaye

See Also

make.posdef

Examples

```
# Example 1
# a positive definite matrix
mat1  <-  matrix(c(0.9999934,0.9999914,0.9999914,0.9999934),2,2)
# check if the matrix is positive definite
is.posdef(mat1, 0.000001)
```

```
# Example 2
# a non positive definite matrix
mat2  <-  matrix(c(0.9999924,0.9999924,0.9999924,0.9999924),2,2)
# check if the matrix is positive definite
is.posdef(mat2, 0.000001)
```

---

make.cov.mat Generates the covariance matrix required to achieved the desired LD

Description

Finds the covariance values required to achieve the specified frequency of major allele haplotype.

Usage

```
make.cov.mat(cor.mat, freqs)
```

Arguments

`cor.mat`

Correlation matrix

`freqs`

Major allele frequencies of the two snps

Value

A 2X2 covariance matrix

Author(s)

Amadou Gaye

References

Montana, G. 2005, HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients., Bioinformatics, vol. 21 (23), pp.4309-4311.

See Also

sim.LDsnps

Examples

```
# MAF of the first snp
maf1  <-  0.1
# MAF of the second snp
maf2  <-  0.1
# target LD
R.target  <-  0.8
# correlation matrix
cor.mat  <-  matrix(c(1,R.target,R.target,1),2,2)
# covariance matrix required to generate 2 variants with the
desired LD
cov.mat.req  <-  make.cov.mat(cor.mat, c(1-maf1, 1-maf2))
```

---

make.obs.env Adds some error to environmental exposure data

Description

Adds a set level of error to simulated binary or quantitative data (the true data) to obtain data with a larger variance (the observed data). The level of error is determined by the misclassification rates in binary data and by the set level of variance in the quantitative data.

Usage

```
make.obs.env(env.data, env.expo = 0, env.prev = 0.1, env.error =
c(0.15, 0.15), reliability.env = 0.8)
```

Arguments

`env.data`
: A vector of environmental measures that represents the true data

`env.expo`
: Model of the exposure: binary=1, quantitative-normal=1 or quantitative-uniform=2

`env.prev`
: Prevalence of the environmental determinant

`env.error`
: Misclassification rates: 1-sensitivity and 1-specificity

`reliability.env`
: Reliability of the assessment of quantitative exposure

Value

A dataframe of two columns:

`environ.orig`
: Input data (true data)

`environ.new`
: Observed data

Author(s)

Amadou Gaye

See Also

sim.env.data, misclassify

Examples

```
# load a dataframe that contains environmental exposure data
data(init.data)
# get observed data by adding some error to the initial data
true.env.data  <-  init.data$env1.U
x  <-  make.obs.env(true.env.data)
observed.env.data  <-  x$environ.new
```

---

make.obs.geno Adds some error to genotype data

Description

Simulates errors and adds it to the true data to obtain observed data. The alleles simulated by the function *sim.geno.data* are randomly misclassified and used to form new genotypes that represent the observed genotypes.

Usage

```
make.obs.geno(allele1, allele2, error.1.0=0.05, error.0.1=0.05,
is.add=0, MAF=0.1)
```

Arguments

`allele1`
: Allele A

`allele2`
: Allele B

`error.1.0`
: 1 to 0 misclassification rate

`error.0.1`
: 0 to 1 misclassification rate

`is.add`
: Genetic model: Set to 0 for binary model and 1 for additive model

`MAF`
: Minor Allele frequency

Value

A dataframe containing the following columns:

`genotyp.U`
: observed genotypes

`allele.A.orig`
: true A alleles

`allele.A.new`
: observed A alleles

`allele.B.orig`
: true B alleles

`allele.B.new`
: observed B alleles

Author(s)

Amadou Gaye

See Also

sim.geno.data

Examples

```
# Example 1
#
# simulate genotypes for two independent genetic variants
geno  <-  sim.geno.data(10000,0.1,0)
allele.A  <-  geno$allele.A
allele.B  <-  geno$allele.B
# randomly misclassify the above simulated alleles and form
# observed genotypes
obs.geno  <-  make.obs.geno(allele.A, allele.B, 0.05, 0.05, 0,
0)
```

---

make.posdef Turns a matrix into a positive definite one

Description

Computes the nearest positive definite matrix of a real symmetric matrix.

Usage

```
make.posdef(matrix, tolerance = 1e-06)
```

Arguments

`matrix`

Input matrix

`tolerance`

A constant

Value

A positive-definite matrix

Author(s)

Amadou Gaye

References

N.J. Higham, 1988 Computing a nearest symmetric positive semidefinite matrix, Linear Algebra Appl. vol. 103, pp.103 118

See Also

is.posdef

Examples

```
# a non positive definite matrix
mat  <-  matrix(c(0.9999924,0.9999924,0.9999924,0.9999924),2,2)
# make the matrix positive definite
mat.new  <-  make.posdef(mat, 0.000001)
```

---

misclassify Adds some misclassification error to binary data

Description

Introduces some misclassification in binary data. The number of values misclassified is determined by the misclassification rates.

Usage

```
misclassify(binary.vector, error.1.0 = 0.05, error.0.1 = 0.05)
```

Arguments

`binary.vector`
: A vector binary values

`error.1.0`
: 1 to 0 misclassification rate

`error.0.1`
: 0 to 1 misclassification rate

Value

`A.new`
: A binary vector

Author(s)

Amadou Gaye

Examples

```
# Example 1
#
# simulate a vector of 0s and 1s
v1  <-  rbinom(100,1,0.4)
# 1 to 0 misclassification rate
error.1.0  <-  0.2
# 0 to 1 misclassification rate
error.0.1  <-  0.2
# randomly misclassify the vector v
v1.new  <-  misclassify(v1, error.1.0, error.0.1)
# Example 2
#
# simulate a vector of 2s and 3s
v2  <-  ifelse(runif(100, 0, 1) <= 0.4, 2,3)
# 1 to 0 misclassification rate
error.1.0  <-  0.2
# 0 to 1 misclassification rate
model.power.calc 23
error.0.1  <-  0.2
# randomly misclassify the vector v
v2.new  <-  misclassify(v2, error.1.0, error.0.1
```

---

model.power.cacl Calculates the theoretical power

Description

Computes the power of the study from the set sample size and desired power.

Usage

```
model.power.calc(is.interaction = 0, pval = 1e-04,
mean.model.z.geno, mean.model.z.env, mean.model.z.int)
```

Arguments

`is.interaction`
: Type of interaction; 1 for gene-environment, 2 for gene-gene and 3 for environment-environmentinteraction

`pval`
: Cut-off p-value defining statistical significance

`mean.model.z.geno`
: mean z-statistics of the two genetic determinants

`mean.model.z.env`
: mean z-statistics of the environmental determinants

`mean.model.z.int`
: mean z-statistics of the interaction term

Value

A list containing:

`model.power.geno1`
: Theoretical power under the main effect of genetic variant 1

`model.power.geno2`
: Theoretical power under the main effect of genetic variant 2

`model.power.env1`
: Theoretical power under the main effect environment 1

`model.power.env2`
: Theoretical power under the main effect environment 2

`model.power.int`
: Theoretical power under Interaction

Author(s)

Amadou Gaye

See Also

empirical.power.calc

Examples

```
# sets the model
is.interaction  <-  0 # no interaction
# cut-off p-value
pval  <-  1e-04
# mean z-statistics for the effects of the two genetic variants
mean.model.z.geno  <-  c(0.269, 0.268)
# mean z-statistics for environmental exposures
mean.model.z.env  <-  c(2.508, 2.512)
# mean z-statistics for the interaction part
mean.model.z.int  <-  NA
# calculate the power reached under the initial sample size and
```

```
# desired power
power  <-
model.power.calc(is.interaction,pval,mean.model.z.geno,
mean.model.z.env,mean.model.z.int)
```

---

obs.data Simulated genotypes, phenotypes and environmental exposure data

---

Description
A table of simulated observed data. The number of rows represents the number of cases and controls (if the outcome is binary) or the number of subjects (if the outcome is continuous).

Usage

```
data(obs.data)
```

Format
A data frame with 10000 observations for the following11 variable:

`id`
: Scenario id

`cc.U`
: Phenotypes

`geno1.U`
: Genotypes for genetic variant 1

`geno2.U`
: Genotypes for genetic variant 2

`allele.A1`
: Alelle A of genetic variant 1

`allele.B1`
: Alelle B of genetic variant 1

`allele.A2`
: Alelle A of genetic variant 2

`allele.B2`
: Alelle B of genetic variant 2

`env1.U`
: Data for environment 1

`env2.U`
: Data for environment 2

`int.U`
: Data for the interaction term

Examples

```
data(obs.data)
```

---

regr.analysis carries out regression analysis

Description

Fits a conventional unconditional logistic regression model with a binary or continuous phenotype as outcome and the genetic, environmental, interaction determinants as covariates.

Usage

```
regr.analysis(is.interaction = 0, pheno.model = 0, sim.df)
```

Arguments

`is.interaction`

Effect model: main effects=0, Gene-Environment interaction=1, Gene-Gene interaction=2 and Environment-Environment interaction=3

`pheno.model`

Type of the outcome; 0 for binary and 1 for continuous

`sim.df`

A dataframe that contains covariates and outcome data

Value

A vector containing the beta, standard-error and z-statistic of each of the covariates

Author(s)

Amadou Gaye

Examples

```
# load a table containing covariates and binary outcome data
data(obs.data)
# binary outcome
pheno.model  <-  0
# is there an interaction
is.interaction  <-  0
# regression analysis
glm.estimates  <-  regr.analysis(is.interaction, pheno.model,
obs.data)
```

---

samplsize.calc Calculates the sample size required to achieved the desired power

---

Description

Estimates by how much the simulated study size needs to be inflated or shrank in order to obtain the specified level of power. The ratio of z- statistic required for desired power to mean model z-statistic obtained indicates the relative change in standard error required. This corresponds to relative change on scale of square root of sample size.

Usage

```
samplsize.calc(numcases = 2000, numcontrols = 8000, num.subjects
= 500, pheno.model = 0, is.interaction = 0, pval = 1e-04, power
= 0.8, mean.model.z.geno,
mean.model.z.env, mean.model.z.int)
```

Arguments

`numcases`
: Number of cases when outcome is binary

`numcontrols`
: Number of controls when outcome is binary

`num.subjects`
: Number of subjects when outcome is continuous

`pheno.model`
: Outcome type, 0 for binary and 1 for continuous

`is.interaction`
: Effect model: main effects=0, Gene-Environment interaction=1, Gene-Gene interaction=2 and Environment-Enviroment interaction=3

`pval`
: Cut-off p-value defining statistical significance

`power`
: Desired power

`mean.model.z.geno`
: Ratio of mean beta estimate over mean se estimate for genetic main effect

`mean.model.z.env`
: Ratio of mean beta estimate over mean se estimate for environment main effect

`mean.model.z.int`
: Ratio of mean beta estimate over mean se estimate for Interaction effect

Value

A list containing:

`numcases.required.geno1`
: Number of cases required to achieve the desired power under genetic variant 1main effect

`numcontrols.required.geno1`
: Number of controls required to achieve the desired power under genetic variant1 main effect

`numcases.required.geno2`
: Number of cases required to achieve the desired power under genetic variant 2main effect

`numcontrols.required.geno2`
: Number of controls required to achieve the desired power under genetic variant2 main effect

`numcases.required.env1`
: Number of cases required to achieve the desired power under environment 1main effect

`numcontrols.required.env1`
: Number of controls required to achieve the desired power under environment 1main effect

`numcases.required.env2`
: Number of cases required to achieve the desired power under environment 2main effect

`numcontrols.required.env2`
: Number of controls required to achieve the desired power under environment 2main effect

`numcases.required.int`
: Number of cases required to achieve the desired power under interaction effect

`numcontrols.required.int`
: Number of controls required to achieve the desired power under interaction effect

`numsubjects.required.geno1`
: Number of subjects required to achieve the desired power under genetic variant1 main effect

`numsubjects.required.geno2`
: Number of subjects required to achieve the desired power under genetic variant2 main effect

`numsubjects.required.env1`
: Number of subjects required to achieve the desired power under environment 1main effect

`numsubjects.required.env2`
: Number of subjects required to achieve the desired power under environment 2main effect

`numsubjects.required.int`
: Number of subjects required to achieve the desired power under interaction

Author(s)

Amadou Gaye

See Also

empirical.power.calc, model.power.calc

Examples

```
# Example 1: sample size calculation for a binary outcome
# set outcome type to binary
pheno.model  <-  0
# set the model
is.interaction  <-  0 # no interaction
# cut-off p-value
pval  <-  1e-04
# desired power
power  <-  0.80
# mean z-statistics for genetic variants
mean.model.z.geno  <-  c(0.269, 0.268)
# mean z-statistics for environmental exposures
mean.model.z.env  <-  c(2.508, 2.512)
# mean z-statistics for the interaction part
mean.model.z.int  <-  NA
# calculate the sample size required under each model
sample.sizes.required  <-  samplsize.calc(2000,8000, NA,
pheno.model, is.interaction,pval, power,
mean.model.z.geno,mean.model.z.env,
mean.model.z.int)
# Example 2: sample size calculation for a continuous outcome
# set outcome type to binary
pheno.model  <-  1
# set the model
is.interaction  <-  0 # no interaction
# cut-off p-value
pval  <-  1e-04
# desired power
power  <-  0.80
# mean z-statistics for genetic variants
mean.model.z.geno  <-  c(4.890, 4.872)
# mean z-statistics for environmental exposures
mean.model.z.env  <-  c(2.508, 2.512)
# mean z-statistics for the interaction part
mean.model.z.int  <-  NA
# calculate the sample size required under each model
sample.sizes.required  <-  samplsize.calc(NA,NA, 1000,
pheno.model,
is.interaction,pval, power, mean.model.z.geno,mean.model.z.env,
mean.model.z.int)
```

Sim.CC.data Simulates cases and controls

Description

Generates affected and non-affected subjects until the set sample size is achieved.

Usage

```
sim.CC.data(num.obs = 20000, numcases = 2000, numcontrols =
8000,allowed.sample.size = 2e+07, is.interaction = 0,
```

```
disease.prev = 0.1, MAF = c(0.1, 0.1), is.add = c(0, 0),
R.target = 0.7, LD = 0, cov.mat.req, display = FALSE,
or.geno = c(1.5, 1.5), env.expo = c(0, 0), env.mean.lowlm =
c(3.3, 3.3), env.stdev.uplm = c(1, 1), env.prev = c(0.1, 0.1),
or.env = c(1.5, 1.5), skewness = c(0, 0), or.int = 1.8,
sigma.subject = 12.36, pheno.error = c(0, 0))
```

Arguments

`num.obs`
: Number of observations to generate per iteration

`numcases`
: Number of cases to simulate

`numcontrols`
: Number of controls to simulate

`allowed.sample.size`
: Maximum number of observations allowed

`is.interaction`
: Effect model: main effects=0, Gene-Environment interaction=1, Gene-Gene interaction=2 and Environment-Enviroment interaction=3

`disease.prev`
: Prevalence of the binary outcome

`MAF`
: Minor allele frequencies of the genetic variants is.add Genetic models of the variants: 0 for binary model and 1 for additive model R.target Correlation coefficient required if the alleles of the two genetic variants are inLD

`LD`
: Sets independence or LD between the two genetic variants: 0 for independence and 1 for LD

`cov.mat.req`
: The covariance matrix required to generate 2 genetic variants in LD

`display`
: If TRUE, a summary is printed on screen

`or.geno`
: Odds ratios of the genetic variants

`env.expo`
: Models of the environmental exposures

`env.mean.lowlm`
: Mean under quantitative-normal model and lower limit under quantitative-uniform model

`env.stdev.uplm`

Standard deviation under quantitative-normal model and upper limit under quantitative uniform model

`env.prev`
: Prevalences of the environmental determinants or.env Odds ratios of the environmental determinants

`skewness`
: Determines skewness under quantitative-normal model; right-skewed distribution if set to a positive value and left-skewed when set to a negative value

`or.int`
: Odds ration of the interaction

`sigma.subject`
: Baseline odds ratio for subject on 95 percent population centile versus 5 percentile. This parameter reflects the heterogeneity in disease risk arising from determinants that have not been measured or have not been included in the model

`pheno.error`
: Phenotype misclassification rates

Value

A matrix that contains 11 variables

Author(s)

Amadou Gaye

See Also

sim.geno.data, sim.LDgeno.data, sim.env.data, sim.subject.data, sim.pheno.bin

Examples

```
# number of cases
numcases  <-  2000
# number of controls
numcontrols  <-  8000
# main effect model
is.interaction  <-  0
# generate cases and controls untill the set number of cases,
# controls and sample size is achieved
sim.matrix  <-  sim.CC.data(10000, numcases,numcontrols,
20000000,is.interaction, 0.1, c(0.1,0.1), c(0,0), 0.7, 0,
cov.mat.req=NULL, FALSE, c(1.5,1.5), c(0,0),c(3.3,3.3), c(1,1),
c(0.1,0.1), c(1.5,1.5), c(0,0),1.8,12.36,c(0,0))
```

---

sim.env.data Simulates cases and controls

Description

Generates data for a binary, quantitative-normal, or quantitative-uniform environmental determinant.

Usage

```
sim.env.data(num.obs = 20000, env.expo = 0, env.mean.lowlm =
3.3,env.stdev.uplm = 5, env.prev = 0.1, skewness = 0)
```

Arguments

`num.obs`
: Number of observations to simulate

`env.expo`
: Model of the exposure: binary=0, quantitative-normal=1, quantitative-uniform=2

`env.mean.lowlm`
: Mean under quantitative-normal model and lower limit under quantitative-uniformmodel

`env.stdev.uplm`
: Standard deviation under quantitative-normal model and upper limit under quantitative-uniform model

`env.prev`
: Prevalence of the environmental exposure

`skewness`
: Determines skewness under quantitative-normal model; right-skewed distributionfor positive values, left-skewed for negative value. The default is 0 (nonskewed).

Value

A vector of continuous or binary values

Author(s)

Amadou Gaye

See Also

make.obs.env

Examples

```
# Generate data for a binary enviromental exposure
env.data  <-  sim.env.data(1000, 0, 3.3, 5, 0.1, 0)
```

---

sim.env.sesp Estimates sensitivity and specificity for environmental exposure assessment

---

Description

Generates environmental exposure status and tabulates truly exposed versus observed exposed to determine the number of false positives and false negatives which are then used to compute sensitivity and specificity.

Usage

```
sim.env.sesp(seed.val = 333333, prevalence.exp = 0.5,
reliability = 0.8)
```

Arguments

`seed.val`
: Seed value

`prevalence.exp`
: Prevalence of the exposure

`reliability`
: Reliability of the environmental exposure assessment

Details

The function uses reliability as estimate of a standardized error (1-reliability) and generates exposure data with and without error. The prevalence of the exposure represents the probability ina quantile function to determine the threshold of truly exposed in the observed data (data without error) and in the true data (data with error).

Value

A vector of two values:

`sensitivity`
: Estimated sensitivity of environmental exposure assessment

`specificity`
: Estimated specificity of environmental exposure assessment

Author(s)

Amadou Gaye

See Also

sim.geno.sesp

Examples

```
# simulate sensitivity and specificity
env.sens.spec  <-  sim.env.sesp(333, 0.2, 0.8)
```

---

sim.geno.data Simulates genotypes for a genetic variant

Description

Generates two alleles and combines them to form the genotype of a SNP under a binary or additive genetic model.

Usage

```
sim.geno.data(num.obs = 20000, MAF = 0.1, is.add = 0)
```

Arguments

`num.obs`
Number of observations to simulate

`MAF`
Minor allele frequency of the variant

`is.add`
Genetic model of the variant

Value

A dataframe that contains the following variables:

`allele.A`
Major allele

`allele.B`
Minor allele

`geno.U`
Genotype

Author(s)

Amadou Gaye

See Also

sim.LDsnps, sim.LDgeno.data

Examples

```
# simulate genotypes for a binary SNP with a MAF of 0.1
geno  <-  sim.geno.data(num.obs = 10000, 0.1, 0)
```

---

sim.geno.sesp Estimates sensitivity and specificity for allele assessment

---

Description

Generates the sensitivity and specificity values to assess the genotype of an unobserved variant based on the knowledge of an observed variant in LD with the unobserved one.

Usage

```
sim.geno.sesp(seed.val = 333333, prevalence.exp = 0.5, R2.target
= 0.8)
```

Arguments

`seed.val`
Seed value

`prevalence.exp`
Prevalence of the observed genetic variant

`R2.target`
: Measure of LD between the observed and the unobserved variant

Value

A vector of two values:

`sensitivity.mid`
: Estimated sensitivity

`specificity.mid`
: Simulated specificity

Author(s)

Amadou Gaye

See Also

sim.env.sesp

Examples

```
# simulate sensitivity and specificity
geno.sens.spec  <-  sim.geno.sesp(333333,0.5,0.8)
```

---

sim.interact.data Generates data for the interaction term

Description

Computes the interaction term for the pre-specified interaction model.

Usage

```
sim.interact.data(geno1.U, geno2.U, env1.U, env2.U,
is.interaction = 1)
```

Arguments

`geno1.U`
: Genotype data for genetic variant 1

`geno2.U`
: Genotype data for genetic variant 2

`env1.U`
: Exposure data for environment 1

`env2.U`
: Exposure data for environment 2

`is.interaction`
: Effect model: main effects=0, Gene-Environment interaction=1, Gene-Gene interaction=2 and Environment-Environment interaction=3

Value

A numerical vector

Author(s)

Amadou Gaye

Examples

```
# genotypic data
geno1  <- sim.geno.data(10000, 0.1, 0)
geno2  <- sim.geno.data(10000, 0.2, 0)
# Environmental exposure data
env1  <-  sim.env.data(1000, 0, 3.3, 5, 0.1, 0)
env2  <-  sim.env.data(1000, 0, 3.3, 5, 0.2, 0)
# interaction data for a gene-environment interaction model
int.data  <-  sim.interact.data(geno1, geno2, env1, env2, 1)
```

---

sim.LDgeno.data Simulates genotypes for two genetic variants in LD

Description

Generates alleles of two SNPs in LD and uses these alleles to form the genotypes of the genetic variants. Each variant can be binary or additive.

Usage

```
sim.LDgeno.data(num.obs = 20000, MAF = c(0.1, 0.1), is.add =
c(0, 0), R.target = 0.7, cov.mat.req, display = FALSE)
```

Arguments

`num.obs`
: Number of observations to simulate

`MAF`
: Minor allele frequencies of the two variants

`is.add`
: Models of the two variants

`R.target`
: Correlation coefficient, desired level of LD

`cov.mat.req`
: Covariance matrix required required to achieved the desired LD

`display`
: If TRUE, a summary is printed on screen

Value

A dataframe that contains the following variables:

`allele.A1`
: Major allele of variant 1

`allele.B1`
: Minor allele of variant 1

`geno1.U`
: Genotype of variant 1

`allele.A2`
: Major allele of variant 2

`allele.B2`
: Minor allele of variant 2

`geno2.U`
: Genotype of variant 2

Author(s)

Amadou Gaye

See Also

sim.LDsnps, sim.geno.data

Examples

```
# desired LD
R.target  <-  0.8
# MAFs of the two genetic variants
MAFs  <-  c(0.1,0.1)
# the covariance matrix required to generate 2 variants with the
# desired LD
cor.mat  <-  matrix(c(1,R.target,R.target,1),2,2) # cor. matrix
cov.mat.req  <-  make.cov.mat(cor.mat, c(1-MAFs[1], 1-MAFs[2]))
# if the required covariance matrix is not positive-definite get
# the nearest positive-definite matrix (tolerance = 1e-06)
if(!is.posdef(cov.mat.req, 0.000001)){
cov.mat.req  <-  make.posdef(cov.mat.req, 0.000001)
}
# generate genotypes for two gentic variants in LD
LDgeno  <-  sim.LDgeno.data(10000, c(0.1, 0.1), c(0, 0), 0.8,
cov.mat.req, TRUE)
```

---

sim.LDsnps Simulates alleles for two biallelic SNPs in linkage disequilibrium

---

Description

Generates alleles data for pre-specified alleles frequencies. The covariance matrix required to achieve the desired LD is computed and used to produce a random vector from a bivariate normal distribution.

Usage

```
sim.LDsnps(num.obs, maf.snp1 = 0.1, maf.snp2 = 0.1, R.target =
0.7, cov.mat.req, display = FALSE)
```

Arguments

`num.obs`
: Number of observations to simulate

`maf.snp1`
: Minor allele frequency of the first snp

`maf.snp2`

Minor allele frequency of the second snp

`R.target`
Correlation coefficient, desired level of LD

`cov.mat.req`
Covariance matrix required to achieved the desired LD

`display`
If TRUE, a summary is printed on screen

Value

A dataframe of two variables where the rows represent haplotypes

`snp1.allele`
allele data for the first snp

`snp2.allele`
allele data for the second snp

Author(s)

Amadou Gaye

References

Montana, G. 2005, HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients., Bioinformatics, vol. 21 (23), pp.4309-4311.

See Also

sim.LDgeno.data

Examples

```
# desired LD
R.target  <-  0.8
# MAF of the first snp
maf.1  <-  0.1
# MAF of the second snp
maf.2  <-  0.1
# the covariance matrix required to achieve the desired LD
cor.mat  <-  matrix(c(1,R.target,R.target,1),2,2) # cor. matrix
cov.mat.req  <-  make.cov.mat(cor.mat, c(1-maf.1, 1-maf.2))
# if the required covariance matrix is not positive-definite get
# the nearest positive-definite matrix (tolerance = 1e-06)
if(!is.posdef(cov.mat.req, 0.000001)){
cov.mat.req  <-  make.posdef(cov.mat.req, 0.000001)
}
# generate allele data
alleles  <-  sim.LDsnps(10000, maf.1, maf.2, R.target,
cov.mat.req,
TRUE)
```

---

sim.LDsnps Simulates alleles for two biallelic SNPs in linkage disequilibrium

---

Description

Uses the effects data and the odds-ratios of genetic, environmental and eventually interaction determinants to construct a linear predictor (LP). The probability of disease, obtained through logistic transformation of the LP, is used to generate a binomially distributed outcome (the true phenotypes).The vector of true phenotypes is then randomly misclassified to obtain the observed phenotypes. The level of misclassification is given by the sensitivity and specificity of the phenotype assessment.

Usage

```
sim.pheno.bin(num.obs = 20000, is.interaction = 0, disease.prev
= 0.1, geno1.U, geno2.U, env1.U, env2.U, int.U, subject.effect,
or.geno = c(1.5, 1.5), or.env = c(1.5, 1.5), or.int = 1.8,
pheno.error = c(0, 0))
```

Arguments

`num.obs`
: Number of observations to simulate

`is.interaction`
: Effect model: main effects=0, Gene-Environment interaction=1, Gene-Gene interaction=2 and Environment-Enviroment interaction=3

`disease.prev`
: Prevalence of the binary outcome

`geno1.U`
: Genotype data for genetic variant 1

`geno2.U`
: Genotype data for genetic variant 2

`env1.U`
: Exposure data for environment 1

`env2.U`
: Exposure data for environment 2

`int.U`
: Interaction effect data

`subject.effect`
: Subject effect data, reflects the heterogeneity in baseline disease risk

`or.geno`
: Odds ratios of the two genetic variants

`or.env`
: Odds ratios of the two environments

`or.int`
: Odds ratio of the interaction

`pheno.error`

Misclassification rates in phenotype assessment

Value
A dataframe of two columns:

`pheno.original`
True phenotypes

`environ.new`
Observed phenotypes

Author(s)
Amadou Gaye

See Also
misclassify

Examples

```
# simulate genotype data for 2 genetic variants
geno1.data  <-  sim.geno.data(10000, 0.1, 0)
geno1  <-  geno1.data$geno.U
geno2.data  <-  sim.geno.data(10000, 0.2, 0)
geno2  <-  geno2.data$geno.U
# simulate environmental measures for two environments
sim.pheno.qtl 39
env1  <-  sim.env.data(10000, 0, 3.3, 5, 0.1, 0)
env2  <-  sim.env.data(10000, 0, 3.8, 5, 0.1, 0)
# generate interaction effect data
int  <-  sim.interact.data(geno1, geno2, env1, env2, 1)
# simulate subject effect data
subject.effect  <-  sim.subject.data(10000, 12.36)
# generate phenotypes
pheno.data  <-  sim.pheno.bin(10000, 1, 0.1, geno1, geno2, env1,
env2, int, subject.effect, c(1.5,1.5), c(1.5,1.5),
1.8, c(0,0))
true.pheno  <-  pheno.data$pheno.original
observed.pheno  <-  pheno.data$pheno.U
```

---

sim.pheno.qtl Simulates continuous outcome data

Description
Uses the effects data of the determinants to construct a linear predictor (LP). The outcome is normally distributed variable generated with a mean equal to LP and a standard deviation of 1. Some error is then added to the simulated outcome to obtained the observed outcome.

Usage

```
sim.pheno.qtl(num.subjects = 10000, is.interaction = 0, geno1.U,
geno2.U, env1.U, env2.U, int.U, geno.efkt = c(0.25, 0.25),
env.efkt = c(0.25, 0.25), int.efkt = 0.5, reliability.pheno)
```

Arguments

`num.subjects`
: Number of subjects to simulate

`is.interaction`
: Effect model: main effects=0, Gene-Environment interaction=1, Gene-Gene interaction=2 and Environment-Environment interaction=3

`geno1.U`
: Genotypes for genetic variant 1

`geno2.U`
: Genotypes for genetic variant 2

`env1.U`
: Exposure data for environment 1

`env2.U`
: Exposure data for environment 2

`int.U`
: Interaction effect data

`geno.efkt`
: Effects of the genetic variants

`env.efkt`
: Effects of the environmental determinants

`int.efkt`
: Effect of the interaction term

`reliability.pheno`
: Reliability of the phenotype assessment

Value

A dataframe of two columns:

`pheno.original`
: True phenotypes

`environ.new`
: Observed phenotypes

Author(s)

Amadou Gaye

Examples

```
# simulate genotype data for 2 genes
geno1.data  <-  sim.geno.data(10000, 0.1, 0)
geno1  <-  geno1.data$geno.U
geno2.data  <-  sim.geno.data(10000, 0.2, 0)
```

```
geno2  <-  geno2.data$geno.U
# simulate environmental measures for two environments
env1  <-  sim.env.data(10000, 0, 3.3, 5, 0.1, 0)
env2  <-  sim.env.data(10000, 0, 3.8, 5, 0.1, 0)
# generate interaction effect data
int  <-  sim.interact.data(geno1, geno2, env1, env2, 1)
# generate phenotypes
pheno.data  <-  sim.pheno.qtl(10000, 1, geno1, geno2, env1,
env2,int, c(0.25, 0.25), c(0.25, 0.25), 0.5, 0.9)
true.pheno  <-  pheno.data$pheno.original
observed.pheno  <-  pheno.data$pheno.U
```

---

sim.QTL.data Simulates subjects for continuous outcome

Description

Generates the specified number of subjects using functions *sim.geno.data*or *sim.LDgeno.data*,*sim.env.data*and*sim.pheno.bin*

Usage

```
sim.QTL.data(num.subjects, is.interaction = 0, MAF = c(0.1,
0.1), is.add = c(0, 0), R.target = 0.7, LD = 0, cov.mat.req,
display = FALSE, geno.efkt = c(0.25, 0.25), env.expo = c(0, 0),
env.mean.lowlm = c(3.3, 3.3), env.stdev.uplm = c(1, 1),env.prev
= c(0.1, 0.1), env.efkt = c(0.25, 0.25), skewness = c(0, 0),
int.efkt = 0.5, reliability.pheno=0.9)
```

Arguments

`num.subjects`
: Number of subjects to simulate

`is.interaction`
: Effect model: main effects=0, Gene-Environment interaction=1, Gene-Gene interaction=2 and Environment-Environment interaction=3

`MAF`
: Minor allele frequencies of the two genetic variants

`is.add`
: Genetic models of the two variants: 0 for binary and 1 for additive

`R.target`
: Correlation coefficient required if the alleles of the two genes are in linkage disequilibrium

`LD`
: Are the alleles of the two genes in LD (0 for no LD and 1 for LD

`cov.mat.req`
: Covariance matrix required to generate 2 genetic variants in LD

`display`
: If TRUE, a summary is printed on screen

`geno.efkt`
: Effects of the genetic variants

`env.expo`
: Models of the environmental exposures

`env.mean.lowlm`
: Mean under quantitative-normal model and lower limit under quantitative-uniformmodel

`env.stdev.uplm`
: Standard deviation under quantitative-normal model and upper limit under quantitative-uniformmodel

`env.prev`
: Prevalences of the environmental exposures

`env.efkt`
: Effects of the environmental determinants

`skewness`
: Determines skewness under quantitative-normal model; right-skewed distributionif set to a positive value and left-skewed when set to a negative value.

`int.efkt`
: Effect of the interaction

`reliability.pheno`
: Reliability of the phenotype assessment

Value

A matrix that has 11 variables

Author(s)

Amadou Gaye

See Also

sim.geno.data, sim.LDgeno.data, sim.env.data and sim.pheno.qtl

Examples

```
# number of subjects
num.subjects  <-  500
# main effect model
is.interaction  <-  0
# generate cases and controls untill the set number of cases,
# controls and sample size is achieved
sim.QTL.data(num.subjects, is.interaction, c(0.1, 0.1), c(0, 0),
0.7, 0,cov.mat.req=NULL, FALSE, c(0.25, 0.25), c(0, 0), c(3.3,
3.3), c(1, 1), c(0.1, 0.1),c(0.25, 0.25), c(0, 0), 0.5, 0.9)
```

---

sim.sesp.params Table of parameters to simulate sensitivity and specificity values

Description

Parameters stored in this table are used to simulate sensitivity and specificity values for genotype and environmental measures assessment.

Usage

```
data(sim.sesp.params)
```

Format

A table that contains fiveconstants

`seed.val`
: Seed value

`prev.expo.geno`
: Prevalence of the genotype

`prev.expo.env`
: Prevalence of the environmental exposure

`R2.target`
: Squared correlation coefficient to compute sensitivity and specificity for the genetic determinant

`reliability`
: Reliability to compute sensitivity and specificity for the genetic determinant

Examples

```
data(sim.sesp.params)
```

---

sim.subject.data Simulates the individual effect related to heterogeneity in disease risk

---

Description

The variation in baseline disease risk is assumed to be normally distributed on the logistic scale. If this parameter is set to 10, the implication is that a 'high risk' subject (someone at the upper 95 percentile of population risk) is, all else being equal, at 10 times the odds of developing disease compared to someone else who is at 'low risk' (at the lower 5 percentile of population risk).

Usage

```
sim.subject.data(num.obs = 20000, sigma.subject = 12.36)
```

Arguments

`num.obs`
: Number of observations to simulate

`sigma.subject`
: Baseline odds ratio for subject on 95 percent population centile versus 5 percentile. This parameter reflects the heterogeneity in disease risk arising from determinants that have not been measured or have not been included in the model

Value

A numerical vector

Author(s)

Amadou Gaye

Examples

```
# generate subject effect data with a baseline OR of 10
subject.effect  <-  sim.subject.data(20000, 10)
```

---

skew.rnorm Allows to generate right or left-skewed normal distributed data

---

Description

This function allows one to set a level of skewness for normally distributed data.

Usage

```
skew.rnorm(num.obs = 20000, mean = 0, sd = 1, skewness = 0)
```

Arguments

`num.obs`
: Number of observations to simulate

`mean`
: Statistical mean

`sd`
: Standard deviation

`skewness`
: Determines the direction and level of skewness; right-skewed for positive value,left-skewed for negative value. The default is 0 (non skewed)

Value

A numerical vector

Author(s)

Amadou Gaye

References

Azzalini, A. 1985, A class of distributions which includes the normal ones., Scandinavian Journal of Statistics, vol. 12, pp171-178.

See Also

sim.env.data

Examples

```
some.data <-  skew.rnorm(20000, 0, 1, 0)
```

Appendix 2

The graphs in Figure 38 are plots of the accuracy of CNV calls by CNV size and number of probes. These plots show that the conclusions are similar to those made from Figure 29 even when a finer breakdown of CNV length was chosen.

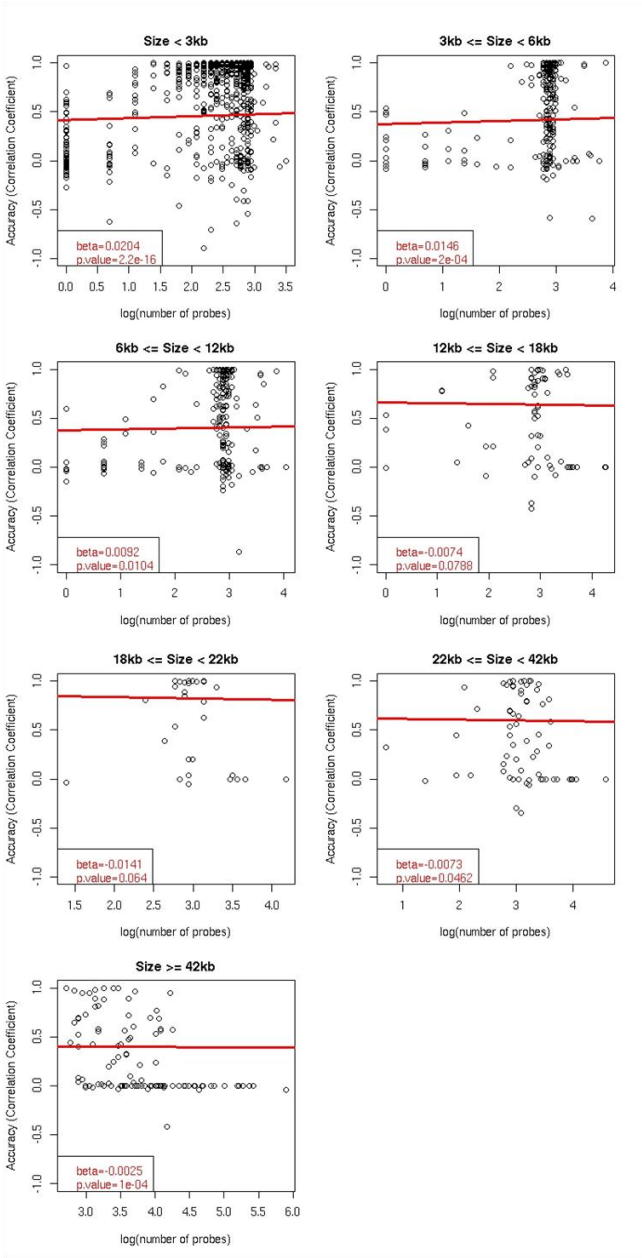


Figure 38: Plots of number of probes against accuracy.
The accuracy was assessed as the correlation between copy numbers from aCGH and 1.2M for seven CNV size categories. The red line across each plot represents the line of best fit. The beta and p.values at the bottom of each plot represent respectively the gradient of the line of best fit and the significance of the test of association between accuracy and number of probe for the given size category.

Appendix 3

Figure 39 and Table 100 show the distribution of accuracy by minor allele frequency when finer MAF intervals, than those on Table 91 and Figure 31 in section 5.3.3.3, are used. The conclusion is similar to that derived from Figure 31 and Table 91 where the MAF was divided into two categories (≥ 0.05 and < 0.05)

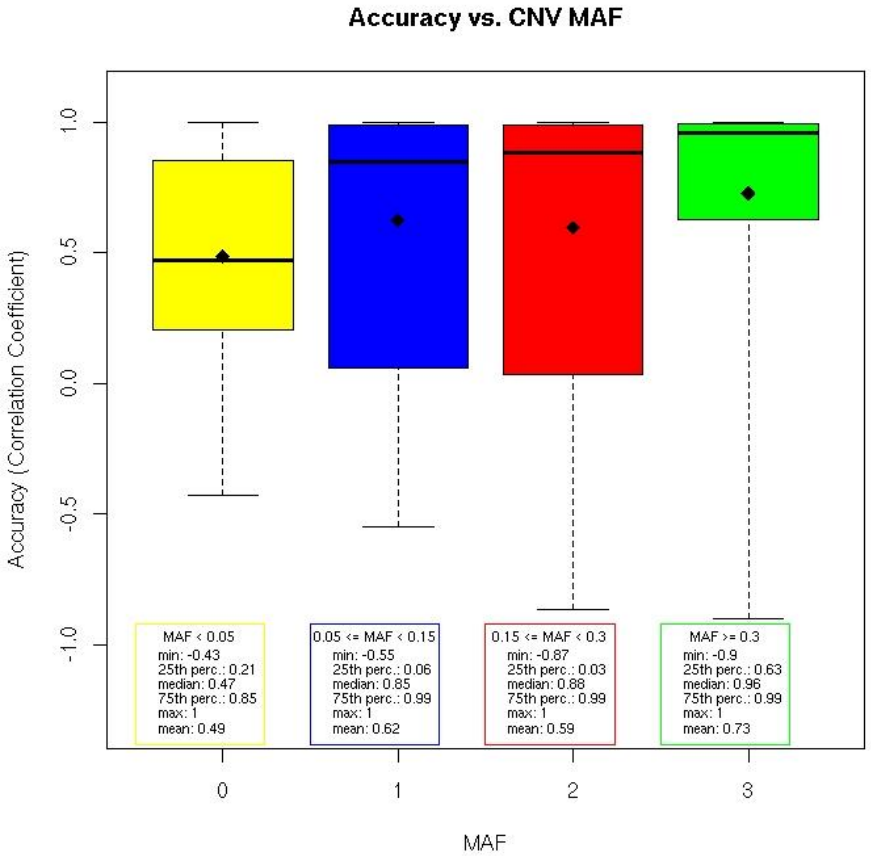


Figure 39: Plot of accuracy by CNV MAF.
The summary statistics that explain the proportion of data points within each part of the boxplots are in the boxes at the bottom of the plot. The black lines on the boxplots are the medians and the black diamonds.

Correlation (r)	MAF < 0.05	0.05 ≤ MAF < 0.15	0.15 ≤ MAF < 0.3	MAF ≥ 0.3
≥ 0.8	115 (28%)	198 (56%)	244 (55%)	215 (66%)
0.5 ≤ r < 0.8	83	43	40	42
< 0.5	215	115	163	70
Total	413	158	203	112

Table 100: CNV count by MAF and level of accuracy and level of accuracy.
The accuracy was assessed as the correlation between copy numbers from aCGH and 1.2M. The percentages represent the corresponding proportions in the MAF category.

Appendix 4

This appendix explains how the CNV genotype dose for each individual was calculated in the process of determining the accuracy of CNV calls using genotype dose.

The accuracy of CNV calls from SNP platforms (Illumina 1.2M, 660, 610, and 300) was evaluated by comparing copy numbers measured from the gold standard (aCGH data) with those measured from the SNP platforms. The copy numbers used for these comparisons were assigned based on the highest posterior probability (the other posterior probabilities were not taken into account).

I undertook a sensitivity analysis as a check for the copy number assignment approach described above. In this analysis the comparison between aCGH and 1.2M was based on CNV genotype dose for each individual where the CNV dose, for each of the two arrays, for a CNV with genotypes 0,1 and 2 was as follows:

$$\boldsymbol{CNV_{DOSE}} = [\boldsymbol{CNV_{GENO.0}} \times \boldsymbol{P}(\boldsymbol{CNV_{GENO.0}})] + [\boldsymbol{CNV_{GENO.1}} \times \boldsymbol{P}(\boldsymbol{CNV_{GENO.1}})] + [\boldsymbol{CNV_{GENO.2}} \times \boldsymbol{P}(\boldsymbol{CNV_{GENO.2}})]$$

Where:

$P(CNV_{GENO.0})$ was the posterior probability of CNV genotype 0,

$P(CNV_{GENO.1})$ was the posterior probability of CNV genotype 1,

$P(CNV_{GENO.2})$ was the posterior probability of CNV genotype 2.

The dose took into account of all the posterior probabilities. There was little difference between the results of the comparison that used the CNV dose strategy and those of the comparison that used the 'highest posterior probability' strategy as shown on the Q-Q plot in Figure 40.

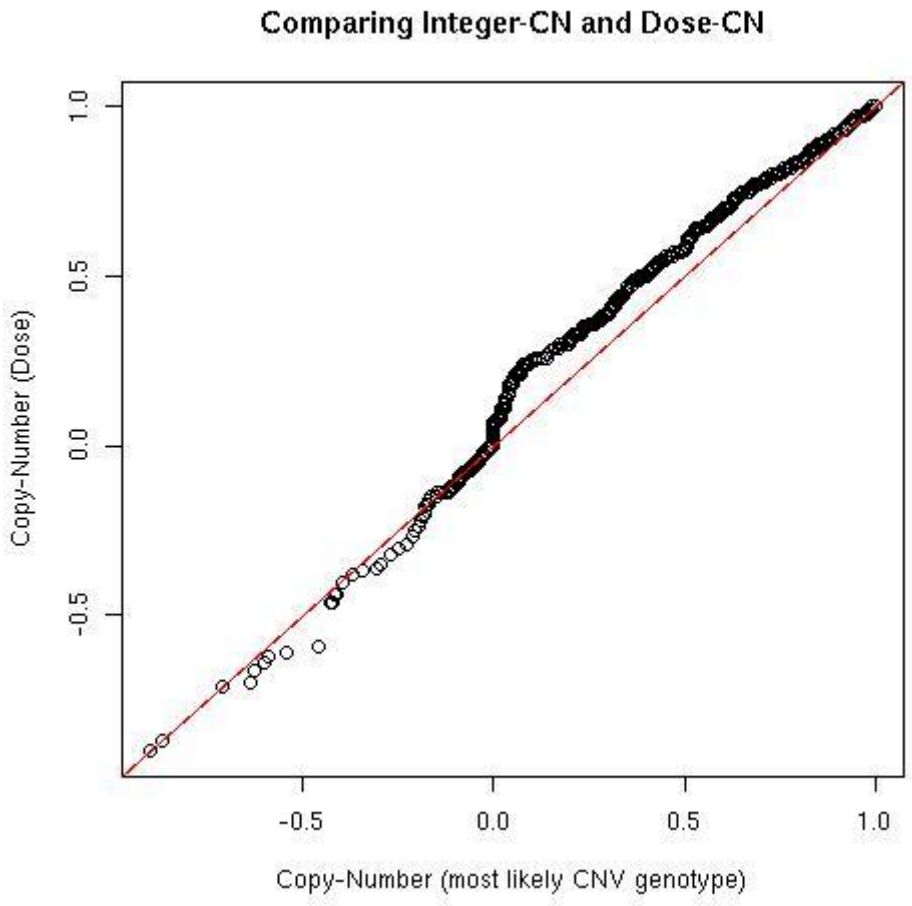


Figure 40: Q-Q plot that compares correlation coefficients obtained.
This plot compares the correlation coefficients obtained when the CNV dose strategy was used to those obtained when the highest posterior probability determined the CNV genotype. This comparison was done across the 1610 CNVs for which there was a gold standard measure of the genotype.

Appendix 5

In this appendix I cross-tabulated the accuracy of CNV calls by CNV characteristics to determine what characteristic actually drives the accuracy of calls. I also used the figures in the below tables to carry out a chi-squared test of independence (see Table 110) which showed that the characteristics were nearly all correlated.

Accuracy of calls by CNV size and type

CNV Type	CNV size	r ≥ 0.8	r < 0.8	Total count
Duplication	size < 3kb	125 (59 %)	88 (41 %)	213
	3kb ≤ Size < 22kb	63 (46 %)	73 (54 %)	136
	Size ≥ 22kb	12 (17 %)	60 (83 %)	72
				421
Duplication / Deletion	Size < 3kb	86 (83 %)	17 (17 %)	103
	3kb ≤ Size < 22kb	53 (79 %)	14 (21 %)	67
	Size ≥ 22kb	5 (16 %)	26 (84 %)	31
				201
Deletion	Size < 3kb	230 (53 %)	206 (47 %)	436
	3kb ≤ Size < 22kb	188 (47 %)	215 (53 %)	403
	Size ≥ 22kb	15 (32 %)	32 (68 %)	47
				886

Table 101: CNV counts by accuracy and CNV length stratified by CNV type. The percentages represent the proportion obtained by dividing the count by the total count on the same row.

Accuracy of calls by CNV MAF and type

CNV Type	MAF	r ≥ 0.8	r < 0.8	Total
Duplication	< 0.05	14 (25 %)	43 (75 %)	57
	≥ 0.05	181 (53 %)	160 (47 %)	341
Of the 421 CNVs in this category, 23 had > 3 classes (no MAF calculated), 421-823=				**398**
Duplication / Deletion	< 0.05	3 (23 %)	10 (77 %)	13
	≥ 0.05	139 (80 %)	34 (20 %)	173
Of the 201 CNVs in this category, 15 had > 3 classes (no MAF calculated), 201-15 =				**186**
Deletion	< 0.05	86 (32 %)	181 (68 %)	267
	≥ 0.05	335 (57 %)	255 (43 %)	590
Of the 886 CNVs in this category, 29 had > 3 classes (no MAF calculated), 886-29 =				**857**

Table 102: CNV counts by accuracy and MAF stratified by CNV type. The percentages represent the proportion obtained by dividing the count by the total count on the same row.

Accuracy of calls by CNV level of LD with HapMap SNP and CNV type

CNV Type	LD	r ≥ 0.8	r < 0.8	Total count
Duplication	LD < 0.25	26 (28 %)	68 (72 %)	94
	0.25 ≤ LD < 0.5	18 (35 %)	34 (65 %)	52
	0.5 ≤ LD < 0.75	18 (44 %)	23 (56 %)	41
	LD ≥ 0.75	138 (61 %)	87 (39 %)	225
Of the 421 CNVs in this category, 9 had no LD info available, 421-9 =				**412**
Duplication / Deletion	LD < 0.25	12 (34 %)	23 (66 %)	35
	0.25 ≤ LD < 0.5	19 (61 %)	12 (39 %)	31
	0.5 ≤ LD < 0.75	11 (69 %)	5 (31 %)	16
	LD ≥ 0.75	102 (86 %)	16 (14 %)	118
Of the 201 CNVs in this category, 3 had no LD info available, 200-1 =				**200**
Deletion	LD < 0.25	76 (31 %)	166 (69 %)	242
	0.25 ≤ LD < 0.5	55 (51 %)	53 (49 %)	108
	0.5 ≤ LD < 0.75	42 (53 %)	37 (47 %)	79
	LD ≥ 0.75	260 (58 %)	185 (42 %)	445
Of the 886 CNVs in this category, 12 had no LD info available, 886-12 =				**874**

Table 103: CNV counts by accuracy and level of LD stratified by CNV type. The percentages represent the proportion obtained by dividing the count by the total count on the same row.

Accuracy of calls by CNV size and number of CNV classes

Copy number classes	CNV size	r ≥ 0.8	r < 0.8	Total count
2	size < 3kb	58 (28 %)	147 (72 %)	205
	3kb ≤ Size < 22kb	75 (24 %)	232 (76 %)	307
	Size ≥ 22kb	10 (15 %)	56 (85 %)	66
				578
3	Size < 3kb	383 (69 %)	173 (31 %)	556
	3kb ≤ Size < 22kb	224 (68 %)	106 (32 %)	330
	Size ≥ 22kb	22 (28 %)	57 (72 %)	79
				965
>3	Size < 3kb	3 (38 %)	5 (62 %)	8
	3kb ≤ Size < 22kb	13 (52 %)	12 (48 %)	25
	Size ≥ 22kb	3 (9 %)	31 (91 %)	34
				67

Table 104: CNV counts by accuracy and CNV size stratified by number of CNV classes. The percentages represent the proportion obtained by dividing the count by the total count on the same row.

Accuracy of calls by level of LD with HapMap SNP and number of CNV classes

Copy number classes	LD	r ≥ 0.8	r < 0.8	Total count
2	LD < 0.25	66 (25 %)	198 (75 %)	264
	0.25 ≤ LD < 0.5	17 (21 %)	65 (79 %)	82
	0.5 ≤ LD < 0.75	5 (12 %)	37 (88 %)	42
	LD ≥ 0.75	55 (33 %)	110 (67 %)	165
Of the 578 CNVs in this category, 25 had no LD info available, 578-25 =				**553**
3	LD < 0.25	51 (44 %)	66 (56 %)	117
	0.25 ≤ LD < 0.5	71 (61 %)	46 (39 %)	117
	0.5 ≤ LD < 0.75	62 (66 %)	32 (34 %)	94
	LD ≥ 0.75	445 (70 %)	189 (30 %)	634
Of the 965 CNVs in this category, 3 had no LD info available, 965-3 =				**962**
> 3	LD < 0.25	7 (16 %)	37 (84 %)	44
	0.25 ≤ LD < 0.5	5 (42 %)	7 (58 %)	12
	0.5 ≤ LD < 0.75	5 (100 %)	0 (0 %)	5
	LD ≥ 0.75	2 (40 %)	3 (60 %)	5
Of the 67 CNVs in this category, 1 had no LD info available, 67-1 =				**66**

Table 105: CNV counts by accuracy and LD stratified by number of CNV classes. The percentages represent the proportion obtained by dividing the count by the total count on the same row.

Accuracy of calls by CNV MAF and number of CNV classes

Copy number classes	MAF	r ≥ 0.8	r < 0.8	Total count
2	< 0.05	109 (28 %)	287 (72 %)	396
	≥ 0.05	34 (19 %)	148 (81 %)	182
				578
3	< 0.05	6 (35 %)	11 (65 %)	17
	≥ 0.05	623 (66 %)	325 (34 %)	948
				965
> 3	No MAF calculated for this category of CNV			**67**

Table 106: CNV counts by accuracy and MAF stratified by number of CNV classes. The percentages represent the proportion obtained by dividing the count by the total count on the same row.

Accuracy of calls by level of LD with HapMap SNP and CNV MAF

CNV MAF	LD	r ≥ 0.8	r < 0.8	Total count
< 0.05	LD < 0.25	62 (31 %)	141 (69 %)	203
	0.25 ≤ LD < 0.5	12 (20 %)	47 (80 %)	59
	0.5 ≤ LD < 0.75	5 (19 %)	22 (81 %)	27
	LD ≥ 0.75	36 (32 %)	76 (68 %)	112
Of the 413 CNVs in this category, 12 had no LD info available, 413-12 =				**401**
≥ 0.05	LD < 0.25	55 (31 %)	123 (69 %)	178
	0.25 ≤ LD < 0.5	76 (54 %)	64 (46 %)	140
	0.5 ≤ LD < 0.75	62 (57 %)	47 (43 %)	109
	LD ≥ 0.75	464 (68 %)	223 (32 %)	687
Of the 1130 CNVs in this category, 3 had no LD info available, 1130-16 =				**1114**

Table 107: CNV counts by accuracy and level of LD stratified by CNV frequency. The percentages represent the proportion obtained by dividing the count by the total count on the same row.

Accuracy of calls by level of LD with HapMap SNP and CNV size

CNV Size	LD	r >= 0.8	r < 0.8	Total count
size < 3kb	LD < 0.25	43 (38 %)	70 (62 %)	113
	0.25 ≤ LD < 0.5	49 (52 %)	46 (48 %)	95
	0.5 ≤ LD < 0.75	47 (63 %)	28 (37 %)	75
	LD ≥ 0.75	305 (64 %)	173 (36 %)	478
Of the 769 CNVs in this category, 8 had no LD info available, 769-8 =				**761**
3kb ≤ Size < 22kb	LD < 0.25	70 (32 %)	150 (68 %)	220
	0.25 ≤ LD < 0.5	40 (44 %)	50 (56 %)	90
	0.5 ≤ LD < 0.75	21 (38 %)	35 (62 %)	56
	LD ≥ 0.75	181 (64 %)	104 (36 %)	285
Of the 662 CNVs in this category, 11 had no LD info available, 662-11 =				**651**
Size ≥ 22kb	LD < 0.25	11 (12 %)	81 (88 %)	92
	0.25 ≤ LD < 0.5	4 (15 %)	22 (85 %)	26
	0.5 ≤ LD < 0.75	4 (40 %)	6 (60 %)	10
	LD ≥ 0.75	16 (39 %)	25 (61 %)	41
Of the 179 CNVs in this category, 10 had no LD info available, 179-10 =				**169**

Table 108: CNV counts by accuracy and level of LD stratified by CNV length. The percentages represent the proportion obtained by dividing the count by the total count on the same row.

Accuracy of calls by CNVMAF and CNV size

CNV Size	MAF	r ≥ 0.8	r < 0.8	Total
size < 3kb	< 0.05	38 (27 %)	104 (73 %)	142
	≥ 0.05	403 (65 %)	216 (35 %)	619
Of the 769 CNVs in this category, 8 had > 3 classes (no MAF calculated), 769-8 =				**761**
3kb ≤ Size < 22kb	< 0.05	69 (31 %)	155 (69 %)	224
	≥ 0.05	230 (56 %)	183 (44 %)	413
Of the 662 CNVs in this category, 25 had > 3 classes (no MAF calculated), 662-25 =				**637**
Size ≥ 22kb	< 0.05	8 (17 %)	39 (83 %)	47
	≥ 0.05	24 (24 %)	74 (76 %)	98
Of the 179 CNVs in this category, 34 had > 3 classes (no MAF calculated), 179-34 =				**145**

Table 109: CNV counts by accuracy and CNV frequency stratified by CNV length. The percentages represent the proportion obtained by dividing the count by the total count on the same row.

Table 110 contains the results of the chi-squared tests of independence carried out to assess the correlation between the five CNV characteristics cross-tabulated in this appendix. The counts of CNVs measured with an accuracy r ≥ 0.8 in Table 93 and Table 101 to Table 109, were used for the tests.

	Size	Type	Number of classes	LD
Type	statistic = 9.92 p.value = 4.18E-002 df = 4			
Number of classes	statistic = 35.64 p.value = 3.43E-007 df = 4	statistic = 22.54 p.value = 1.56E-004 df = 4		
LD	statistic = 33.62 p.value = 7.96E-006 df = 6	statistic = 12 p.value = 6.20E-002 df = 6	statistic = 155.94 p.value = 4.29E-031 df = 6	
MAF	statistic = 32.08 p.value = 1.08E-007 df = 2	statistic = 39.52 p.value = 2.62E-009 df = 2	statistic = 514.77 p.value = 5.80E-114 df = 1	statistic = 160.33 p.value = 1.56E-034 df = 3

Table 110: Results of the chi-squared tests.
The table reports the results of the chi-squared tests of independence between CNV characteristics; df represents the degrees of freedom. All the tests were statistically significant, at a cut-off p.value of 0.05, except for the test between CNV type and LD.