Genomic variation detection in Atlantic cod using long reads

Lex Nederbragt

Centre for Evolutionary and Ecological Synthesis

Norwegian Sequencing Centre

University of Oslo

OK

Who am I

NORWEGIAN SEQUENCING CENTRE

HiSeq 2500 from Illumina

MiSeq from Illumina

PacBio RS II from
Pacific Biosciences

HiSeq X five from Illumina

Sequel from
Pacific Biosciences

Who am I

NORWEGIAN SEQUENCING CENTRE

CEES

Centre for Ecological and Evolutionary Synthesis

@lexnederbragt

flxlexblog.wordpress.com

Acknowledgements

University of Oslo

NORWEGIAN SEQUENCING CENTRE

Sequencing team NSC

CEES

Centre for Ecological and Evolutionary Synthesis

Ole Kristian Tøressen

Bastiaan Star

William Brynildsen

Kjetill Jakobsen

Sissel Jentoft

Cod genome group

Jason Miller, JCVI

J. Craig Venter INSTITUTE

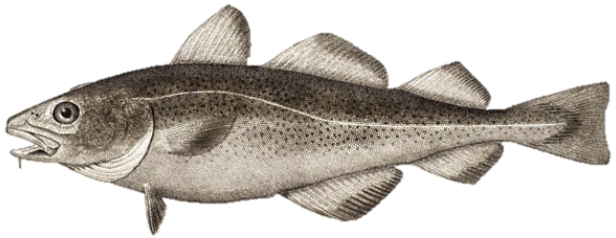
Pacific Biosciences

PACIFIC BIOSCIENCES

CIGENE

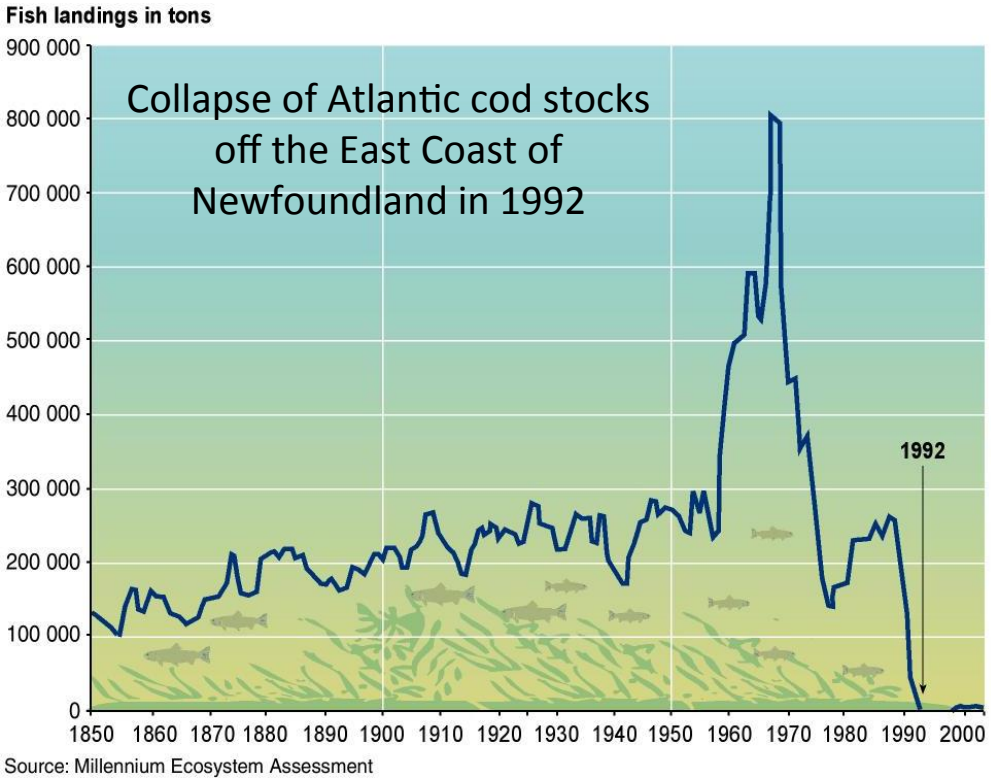
Centre for Integrative Genetics

Atlantic cod



Atlantic cod

Exploitation of natural stocks



Credit: Philippe Rekacewicz, Emmanuelle Bournay, UNEP/GRID-Arendal, http://www.grida.no/

Atlantic cod

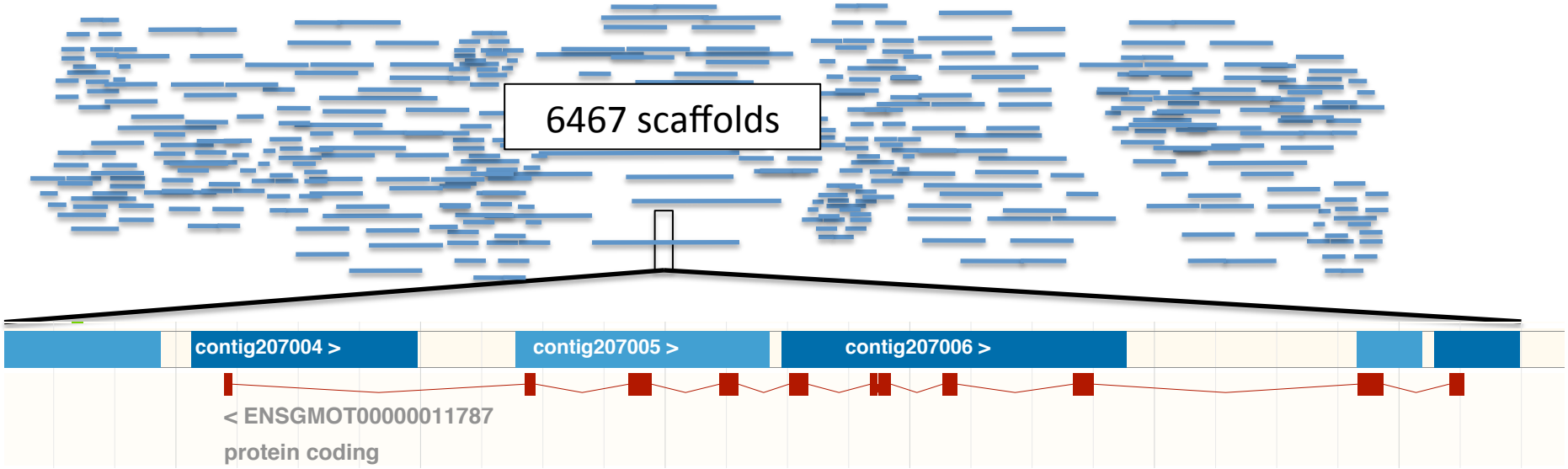
Fisheries

Aquaculture

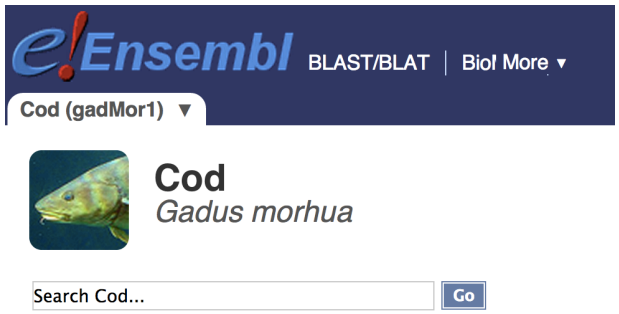
Wikimedia commons, user Slaunger, nofima.no

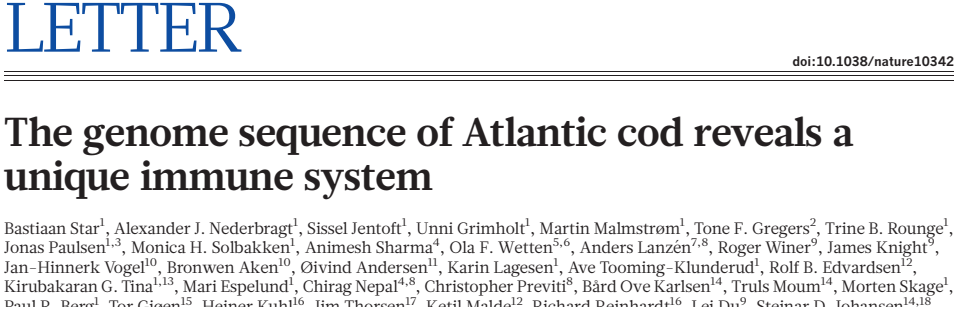
Cod phase 1: The results

gadMor1



35% gap bases





LETTER

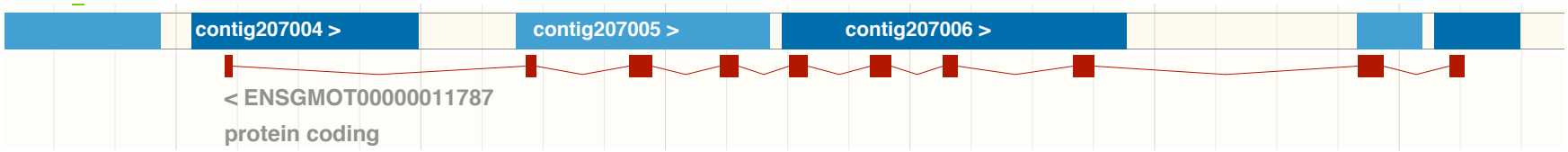
doi:10.1038/nature10342

The genome sequence of Atlantic cod reveals a unique immune system

Bastiaan Star[1], Alexander J. Nederbragt[1], Sissel Jentoft[1], Unni Grimholt[1], Martin Malmstrøm[1], Tone F. Gregers[2], Trine B. Rounge[1], Jonas Paulsen[1,3], Monica H. Solbakken[1], Animesh Sharma[4], Ola F. Wetten[5,6], Anders Lanzén[7,8], Roger Winer[9], James Knight[9], Jan-Hinnerk Vogel[10], Bronwen Aken[10], Øivind Andersen[11], Karin Lagesen[1], Ave Tooming-Klunderud[1], Rolf B. Edvardsen[12], Kirubakaran G. Tina[1,13], Mari Espelund[1], Chirag Nepal[4,8], Christopher Previti[8], Bård Ove Karlsen[14], Truls Moum[14], Morten Skage[1],

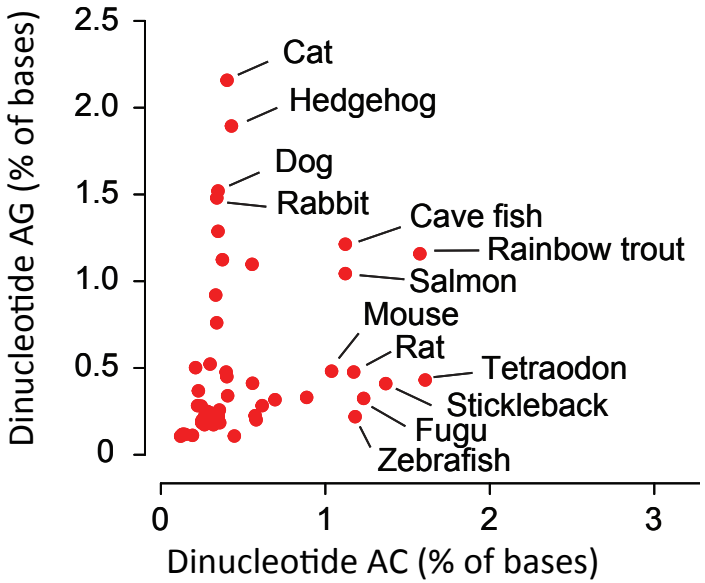
gadMor1: Consequences

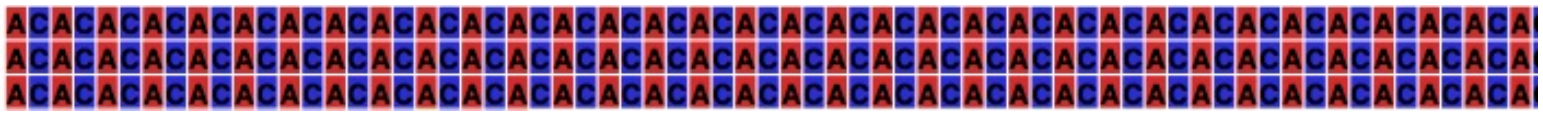
Gaps in genic regions -> lost exons?



Cod genome: repeats

Percentage of bases in 63 vertebrate genomes classified as dinucleotide tandem repeats

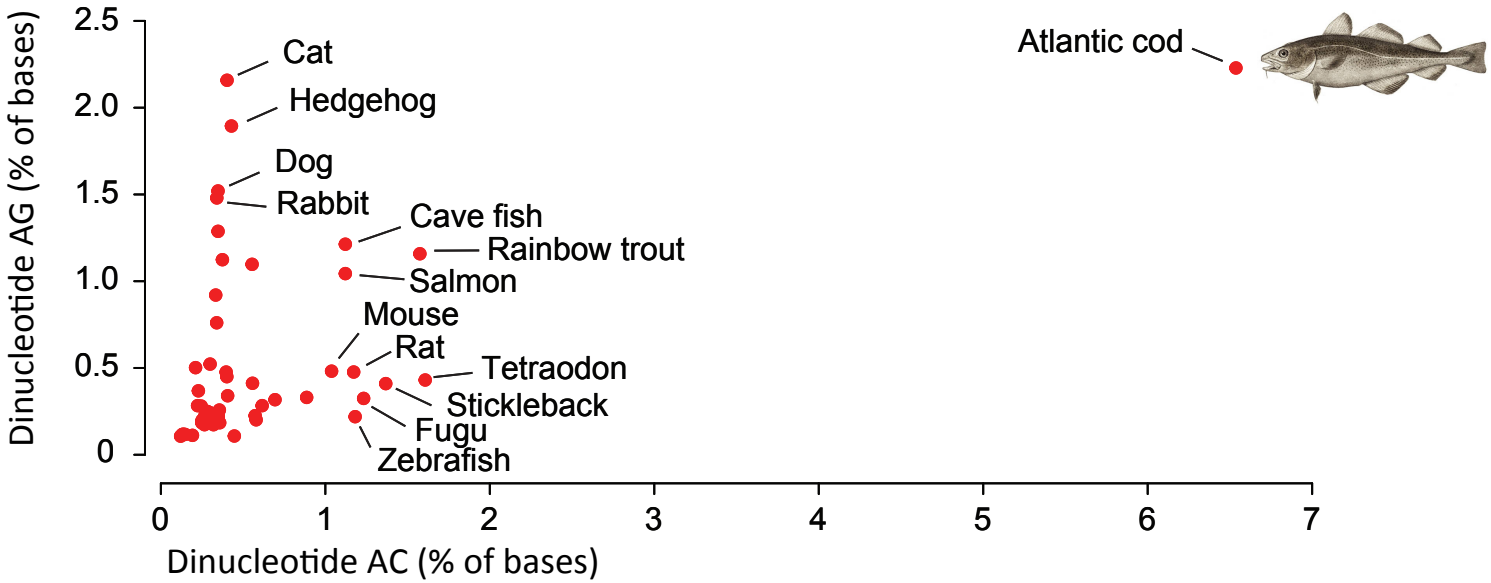


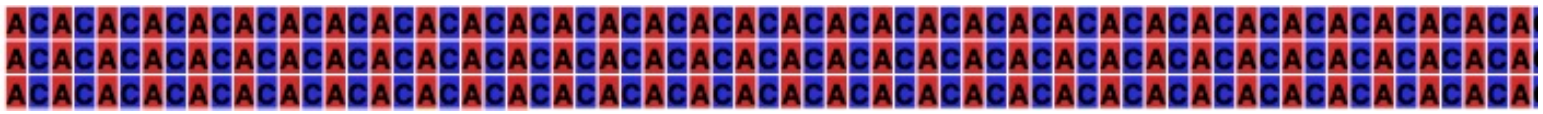


Bastiaan Star

Cod genome: repeats

Percentage of bases in 63 vertebrate genomes classified as dinucleotide tandem repeats





Bastiaan Star

Cod: phase 2

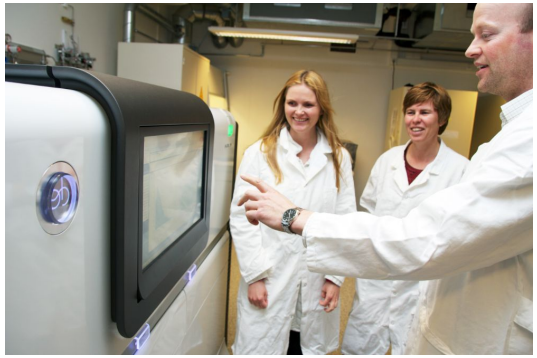
Existing data

<u>454 sequencing</u>
~40 x shotgun
Many mate pair libs

New data

<u>Illumina sequencing</u>
Paired end >200x
Mate Pair 5kb >100x

New data



<u>PacBio sequencing</u>
~16x coverage
C2, C2XL and XLXL data
3.6 Kbp average read length

Cod2: assemblies

Newbler: 454 + BAC ends

Celera: 454 + Illumina

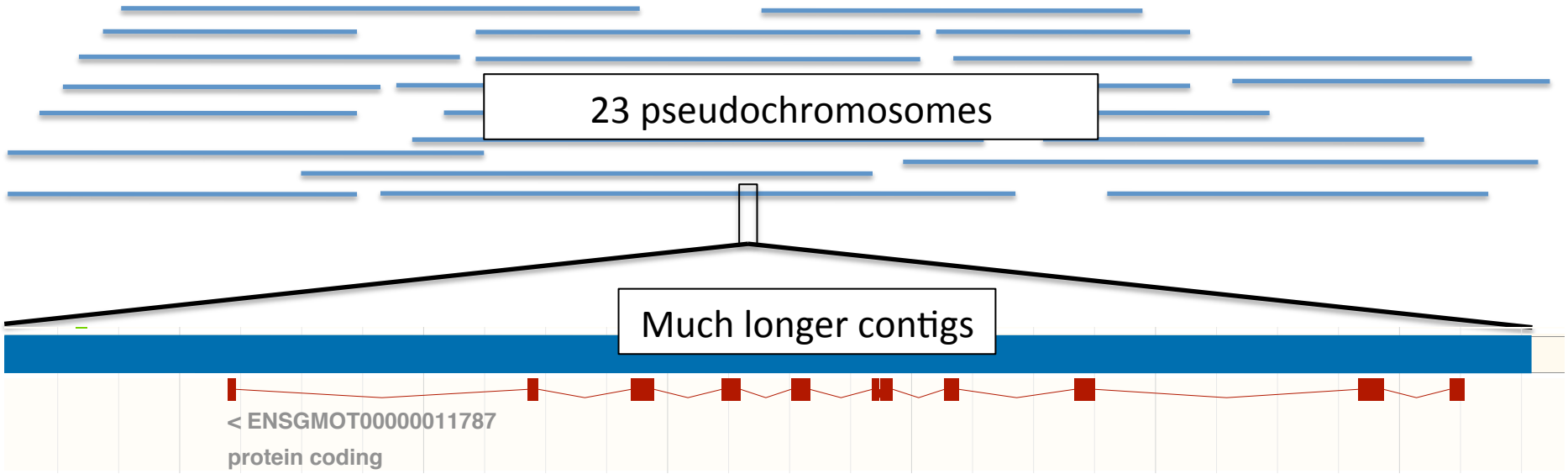
Celera: 454 + raw (!) PacBio

ALLPATHS-LG: Illumina

gadMor2

Ole Kristian Tørresen

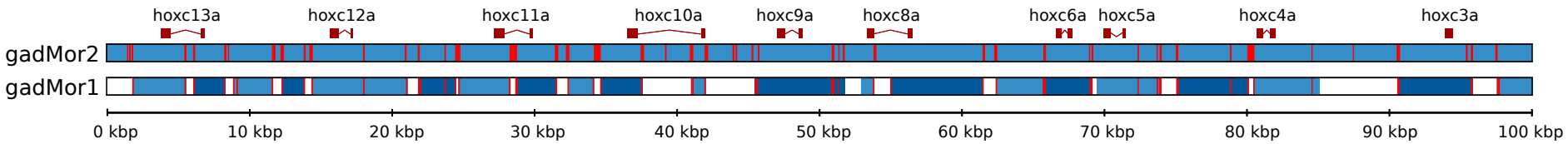
gadMor2

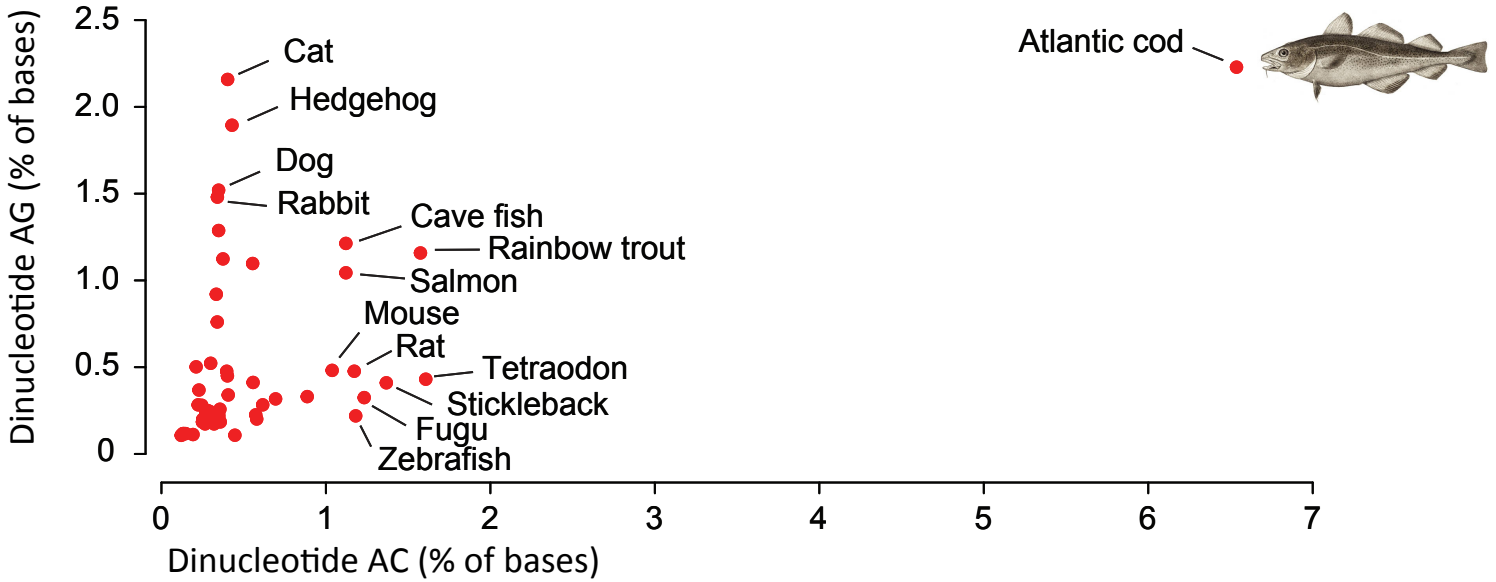


Below 2% gap bases

Ole Kristian Tørresen

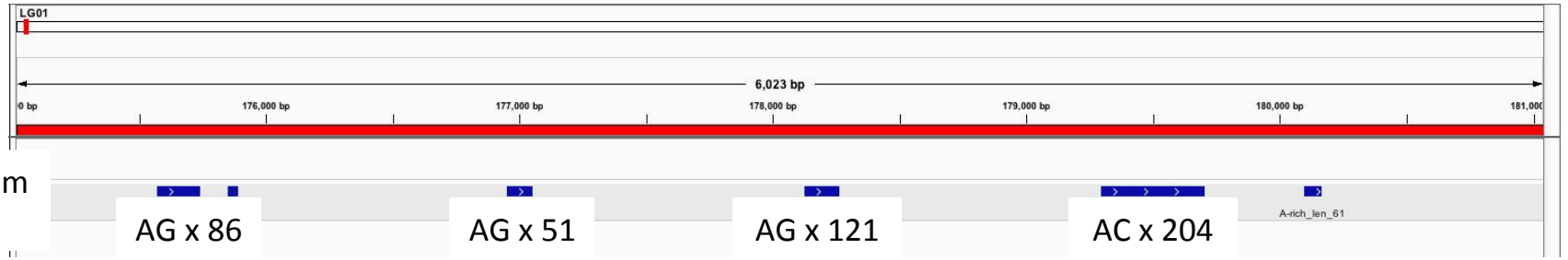
Cod genome: repeats





Exploring the genome

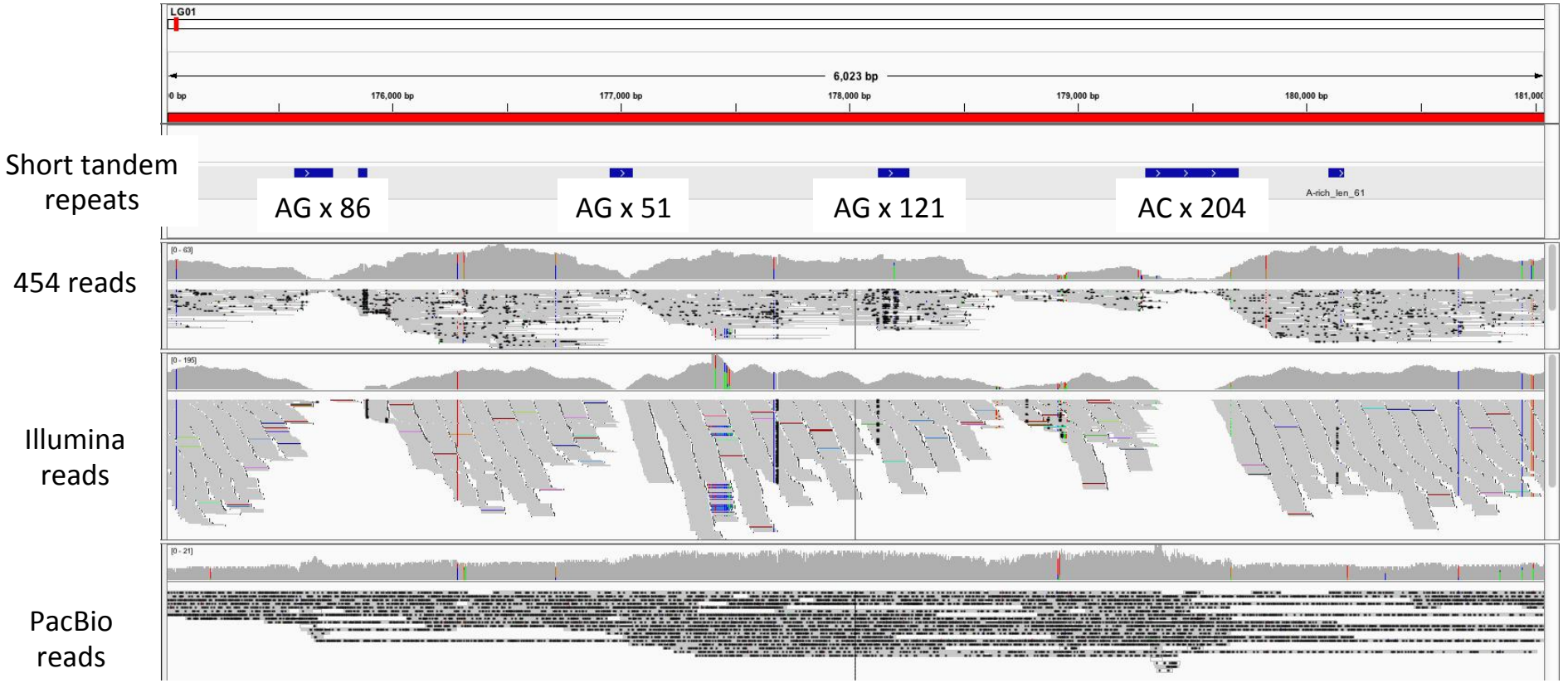
Short tandem repeats



Ole Kristian Tørresen

William Brynildsen

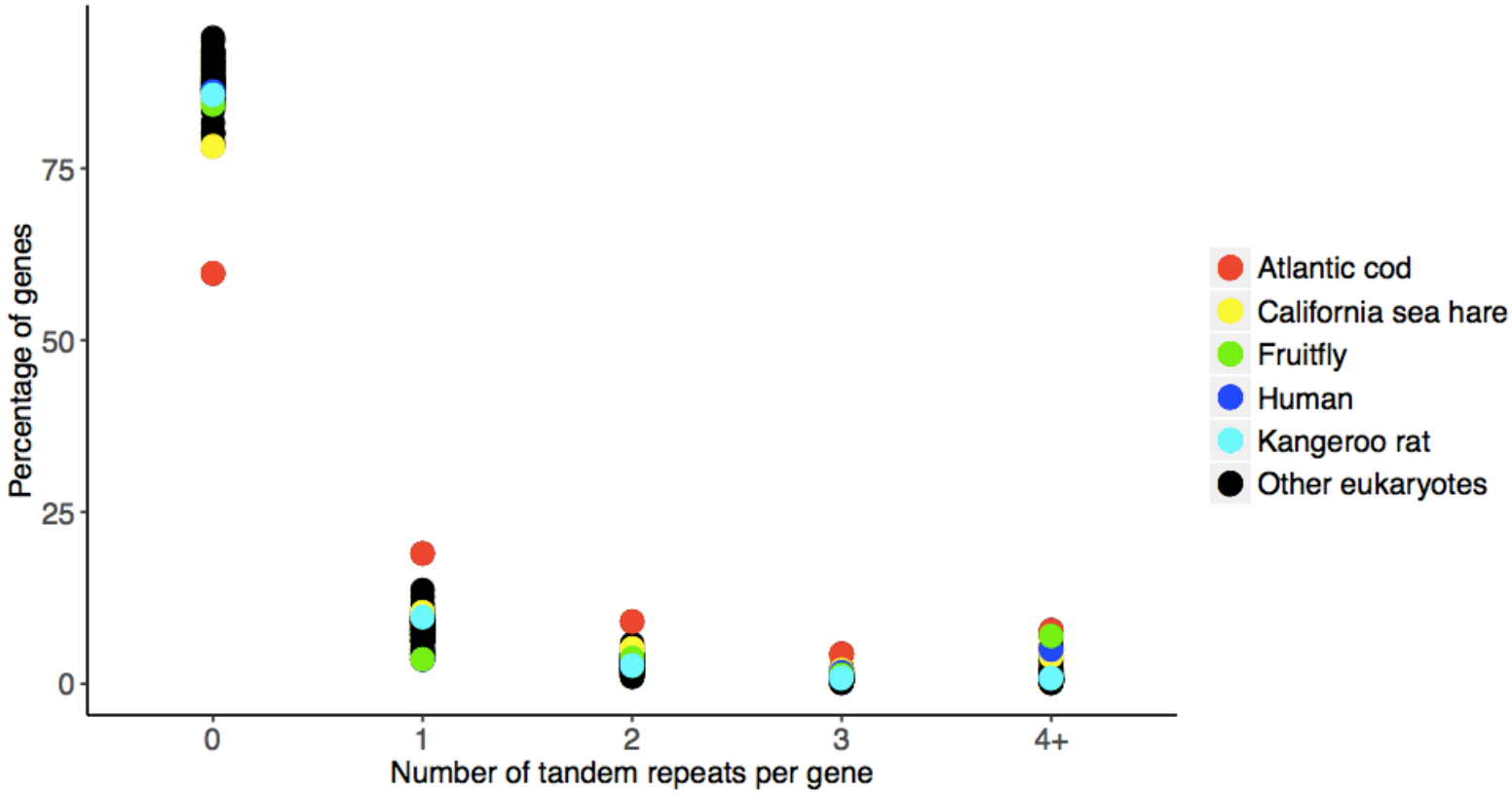
Exploring the genome



Ole Kristian Tørresen

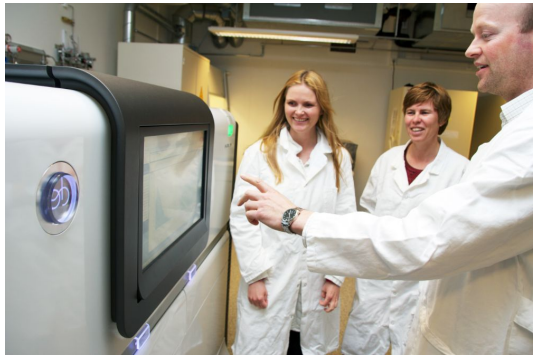
William Brynildsen

STRs


Percentage of genes
75
50
25
0
0
1
2
3
4+
Number of tandem repeats per gene
Atlantic cod
California sea hare
Fruitfly
Human
Kangeroo rat
Other eukaryotes

gadMor2

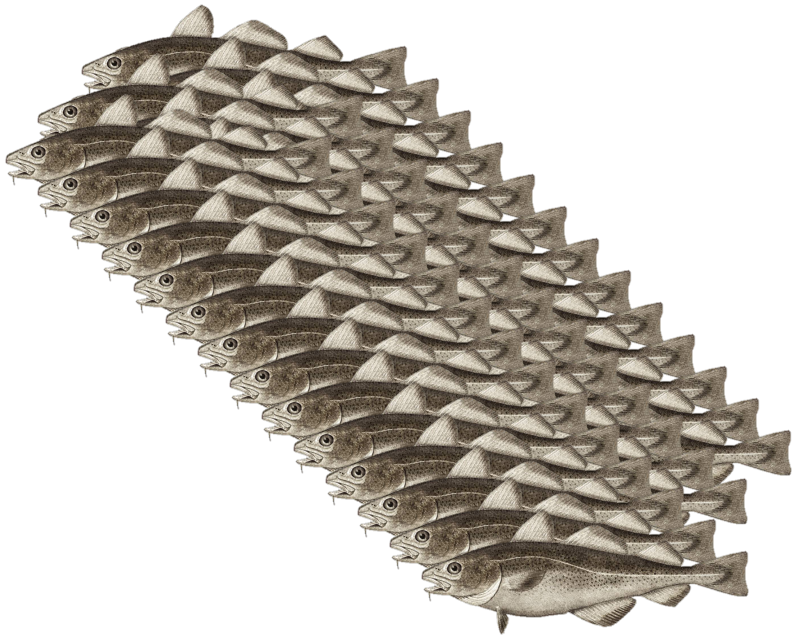
PacBio to the rescue



AquaGenome project

Adaptation in Atlantic cod

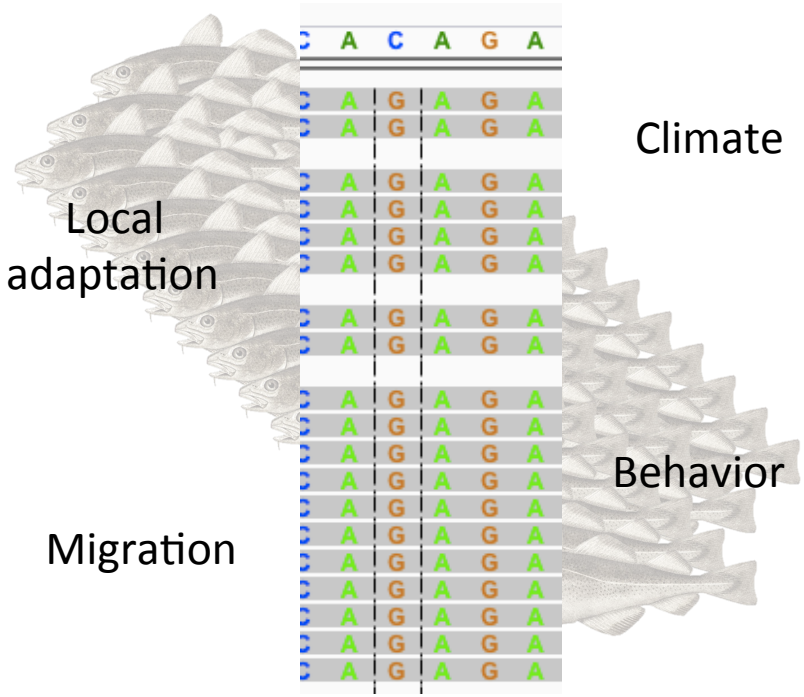
THE AQUA GENOME PROJECT



AquaGenome project

Adaptation in Atlantic cod

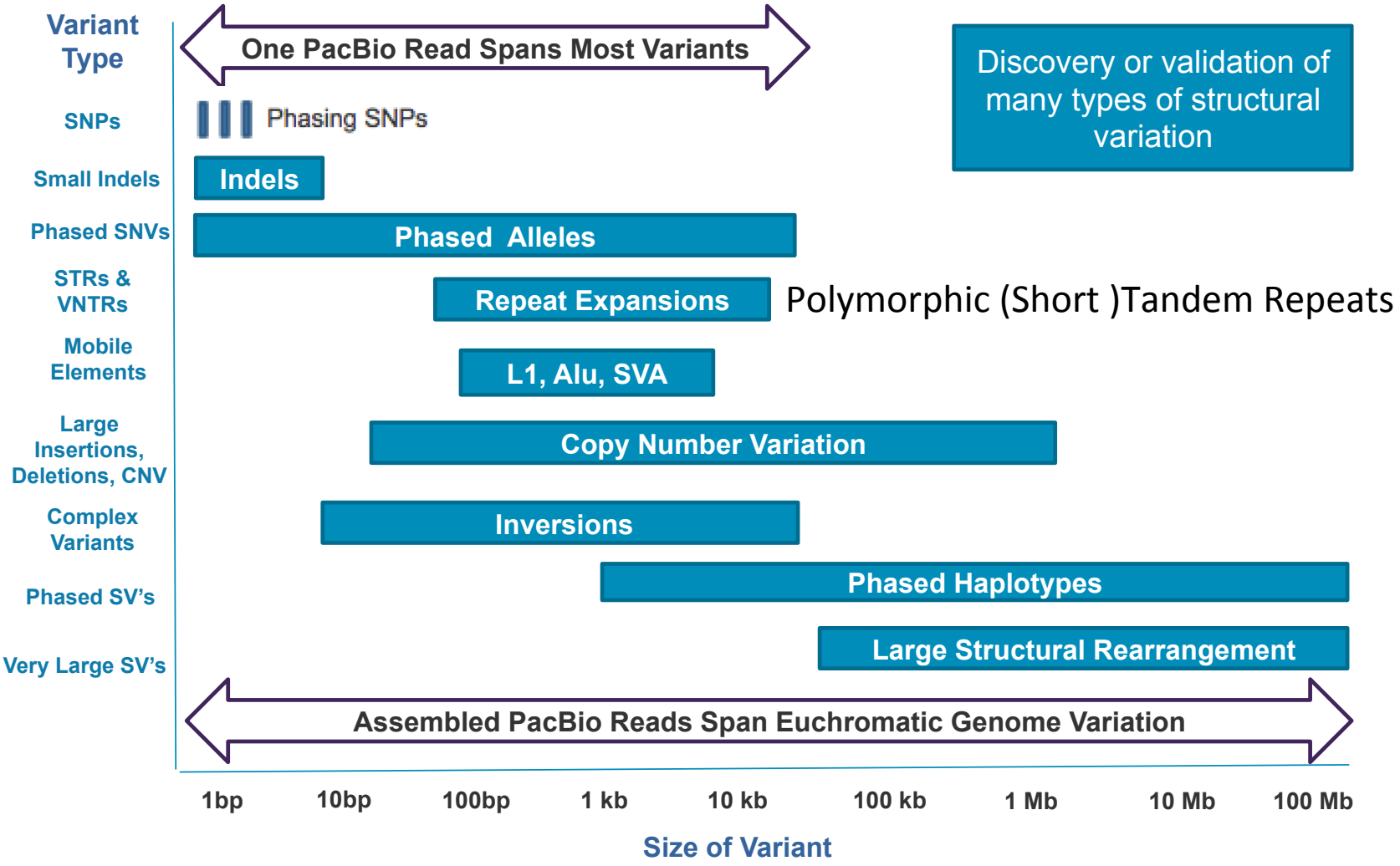
THE AQUA GENOME PROJECT



AquaGenome project

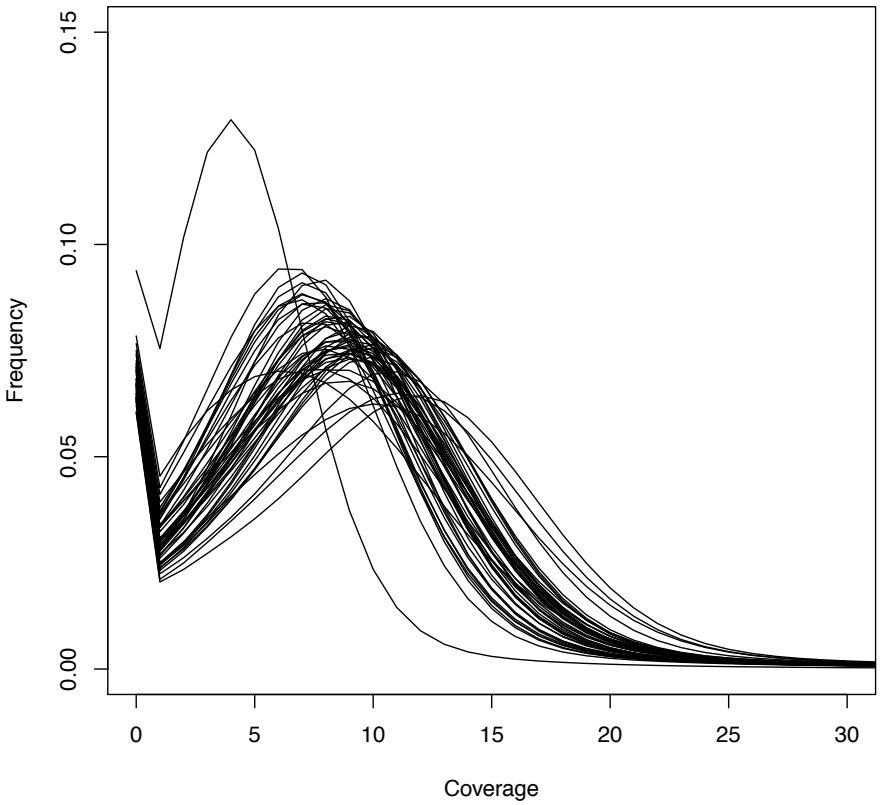
The usual suspect

STRUCTURAL VARIATION DETECTION



Courtesy of Pacific Biosciences

The data



Individual used for reference genome

Instrument	Chem.	Coverage
PacBio RSII	P5C3	2.3x
PacBio RSII	P6C4	26.8x

Individual from different population 2

Instrument	Chem.	Coverage
PacBio RSII	P6C4	24.2x

Doing bioinformatics

Step	Outcome
Find tools	multiple tools

Doing bioinformatics

Step	Outcome
Find tools	multiple tools
Install tools	fewer tools

Doing bioinformatics

Step	Outcome
Find tools	multiple tools
Install tools	fewer tools
Understand tools	ready to run tools

Doing bioinformatics

Step	Outcome
Find tools	multiple tools
Install tools	fewer tools
Understand tools	ready to run tools
Run tools	many result files

Doing bioinformatics

Step	Outcome
Find tools	multiple tools
Install tools	fewer tools
Understand tools	ready to run tools
Run tools	many result files
Filter, filter, filter	fewer results

Doing bioinformatics

Step	Outcome
Find tools	multiple tools
Install tools	fewer tools
Understand tools	ready to run tools
Run tools	many result files
Filter, filter, filter	fewer results
Integrate results	publishable results

Analysis

Assembly
- canu
- falcon

Mapping
- Blasr
- Bwa-mem

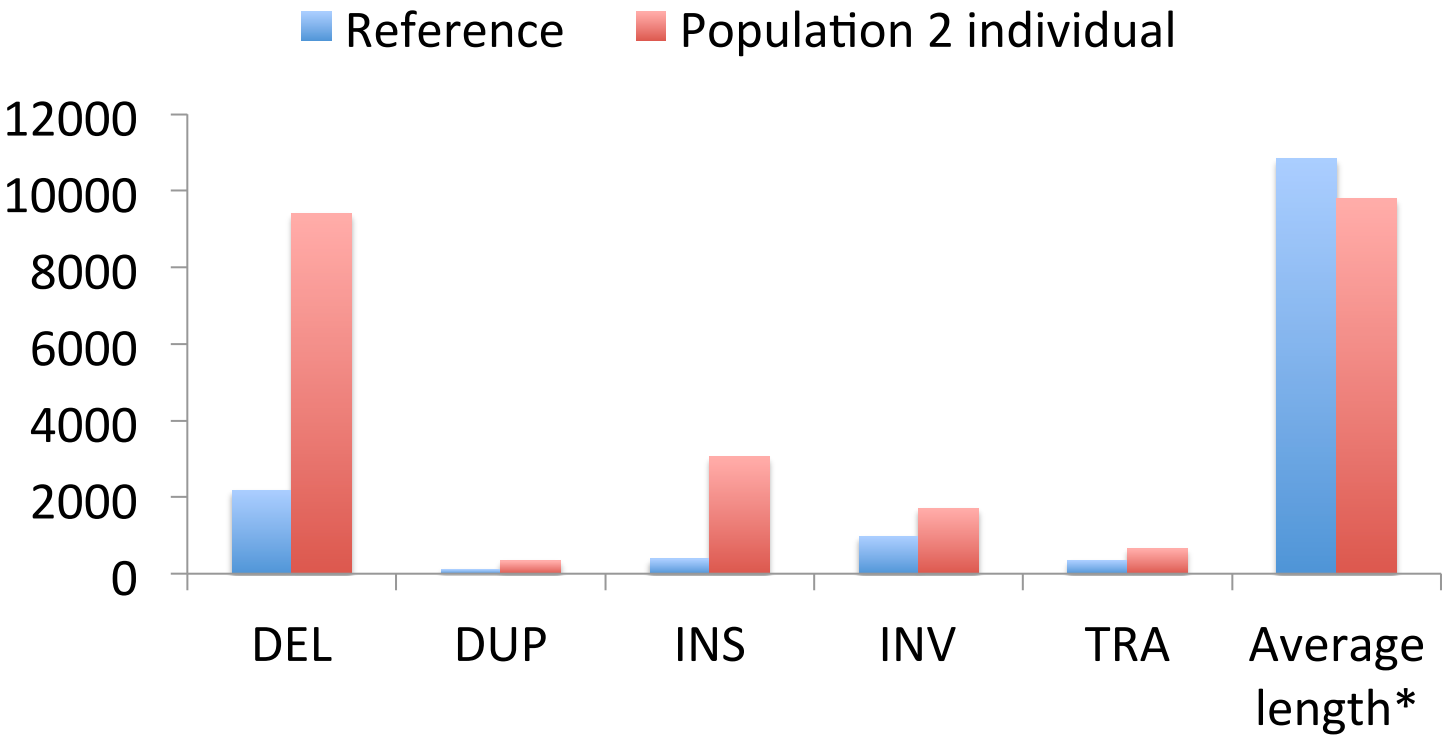
Structural variations
- Bridgemapper
- Sniffles
- PBHoney
- pacmonstr
- Parliament?

Base modifications
- AgIn

Integration
&
Interpretation

Sniffles

Note: before filtering and validation



*Excluding translocations

Analysis

Assembly
- canu
- falcon

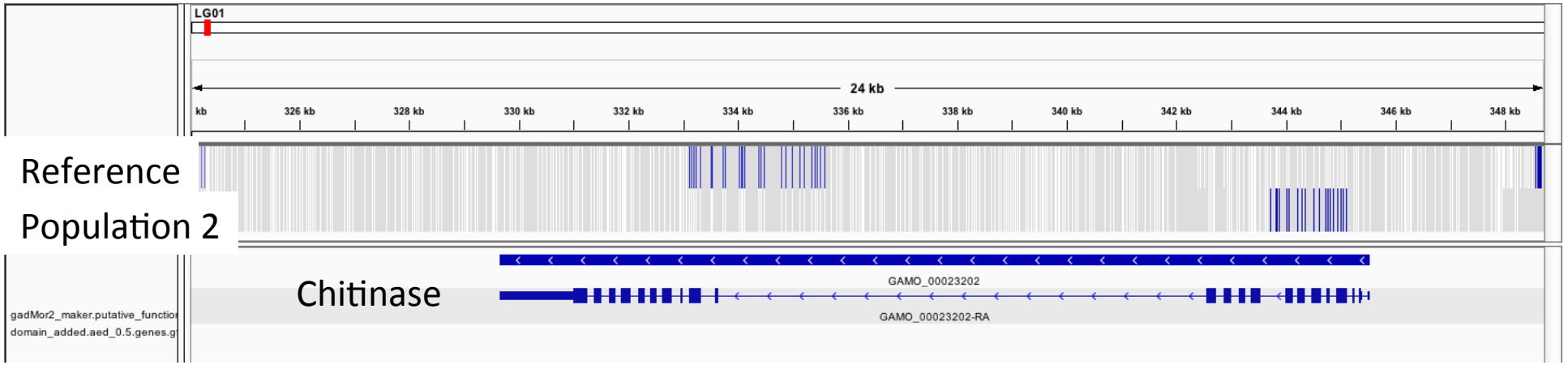
Mapping
- Blasr
- Bwa-mem

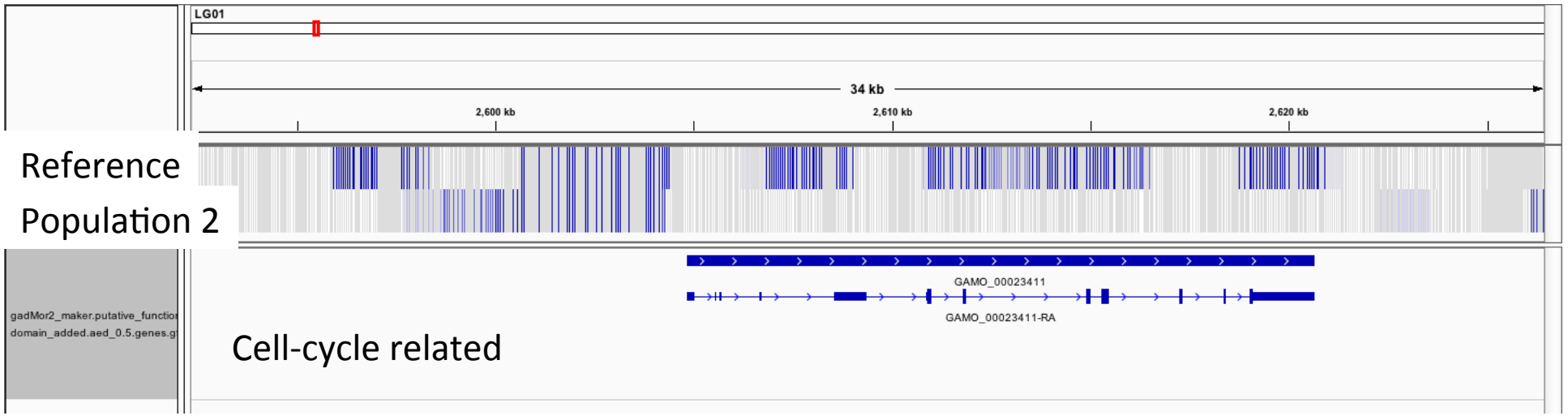
Structural variations
- Bridgemapper
- Sniffles
- PBHoney
- pacmonstr
- Parliament?

Base modifications
- AgIn

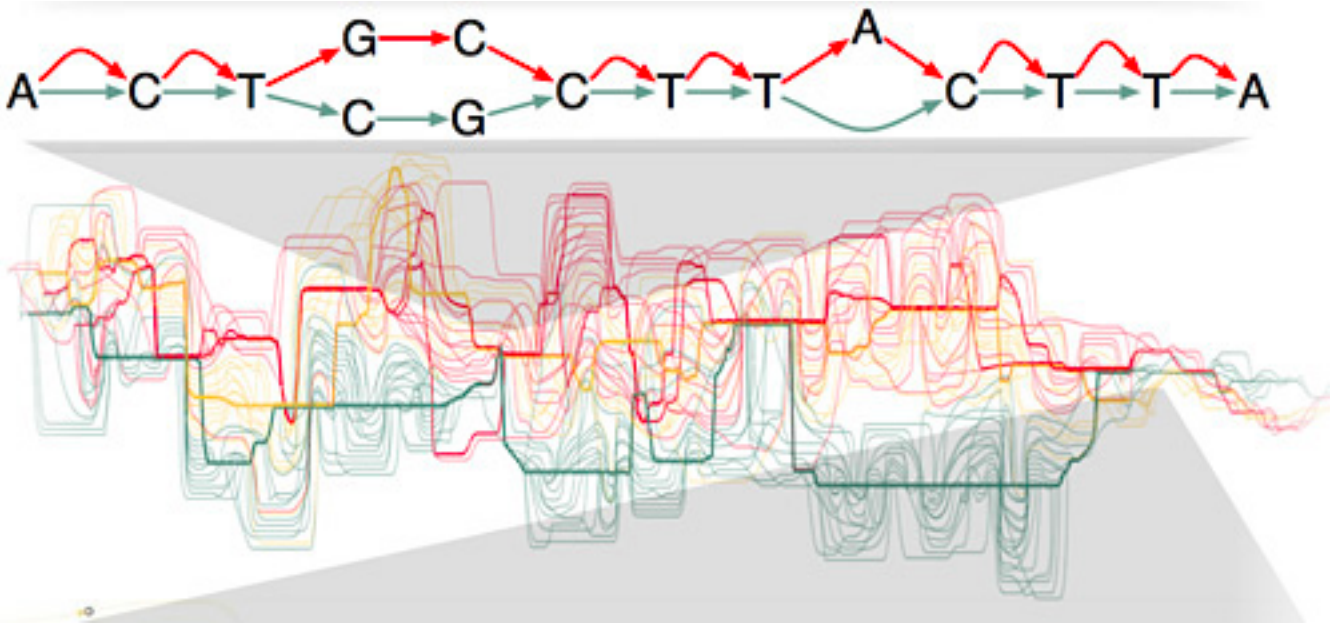
Integration
&
Interpretation

Base modification





THE reference?



Analysis

Assembly
- canu
- falcon

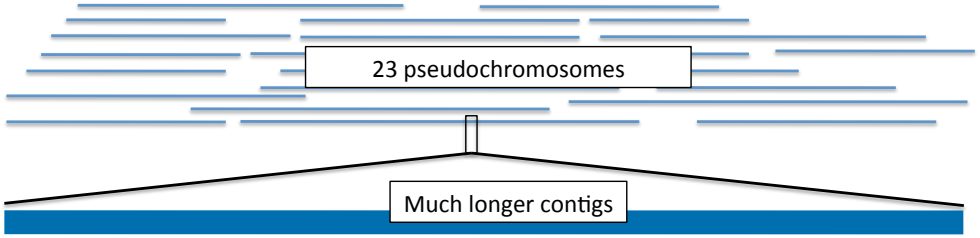
Mapping
- Blasr
- Bwa-mem

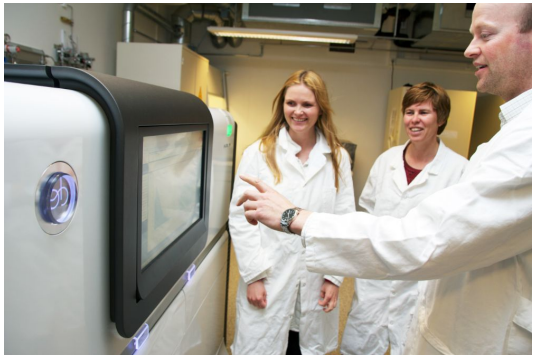
Structural variations
- Bridgemapper
- Sniffles
- PBHoney
- pacmonstr
- Parliament?

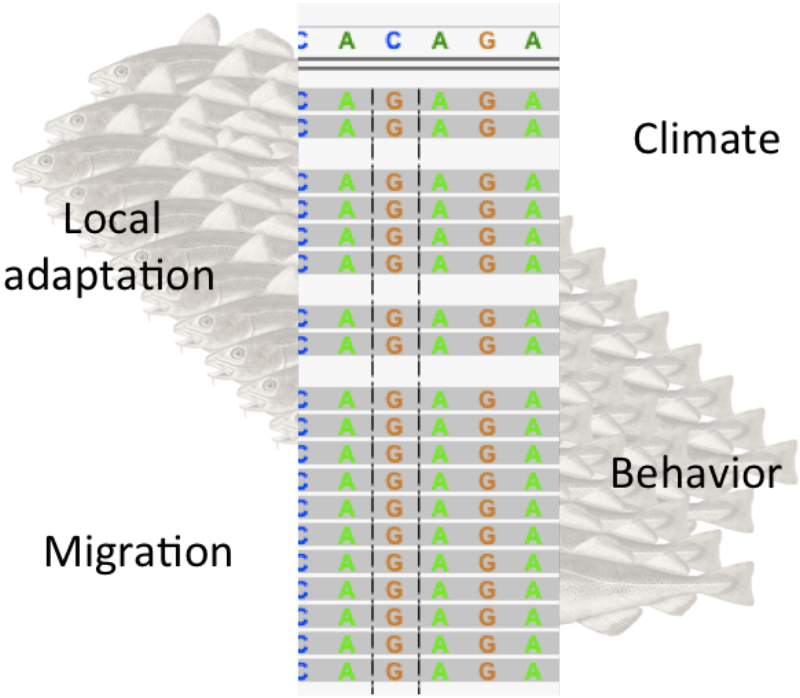
Base modifications
- AgIn

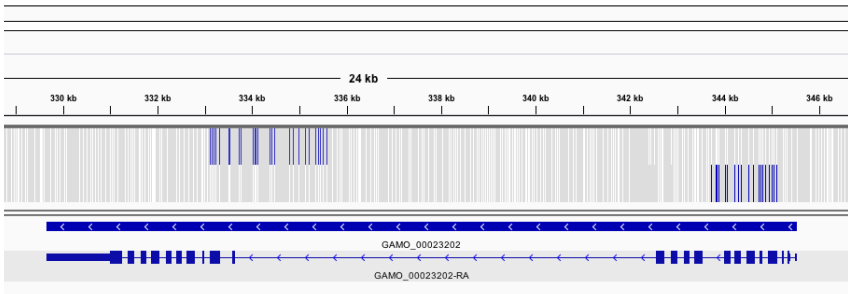
Integration
&
Interpretation

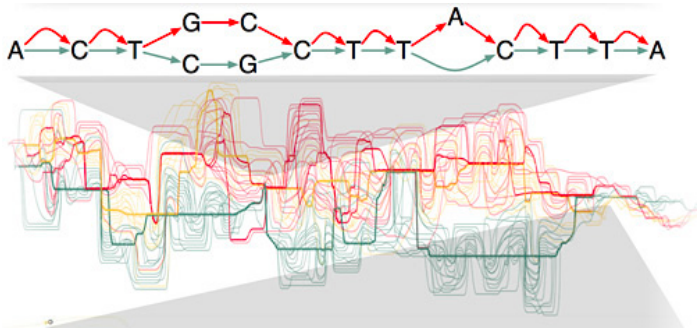
Conclusion


23 pseudochromosomes
Much longer contigs

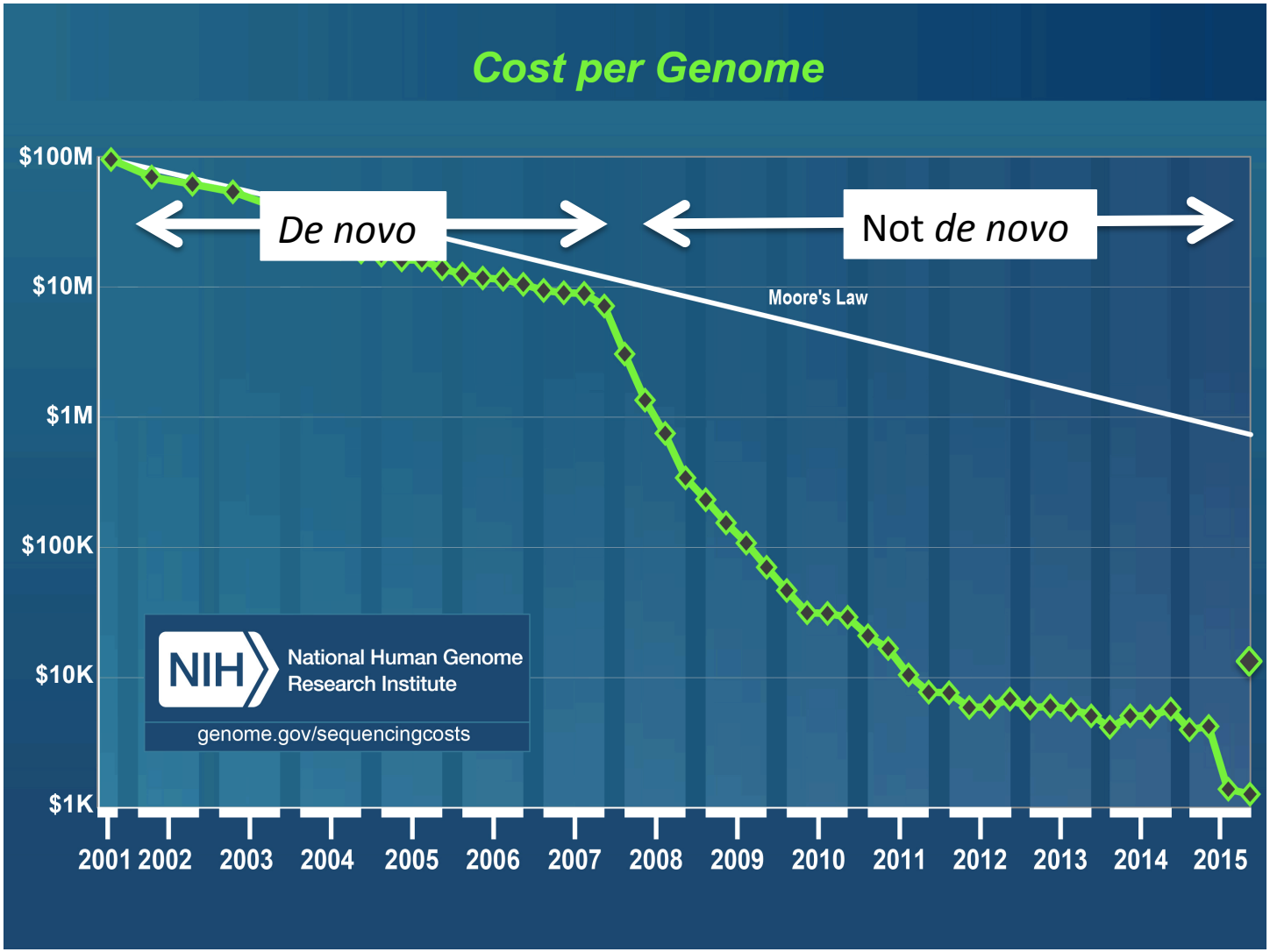



Local adaptation
Climate
Behavior
Migration


24 kb
330 kb
332 kb
334 kb
336 kb
338 kb
340 kb
342 kb
344 kb
346 kb
GAMO_00023202
GAMO_00023202-RA



To sequence a genome



50x pacbio *de novo*

www.genome.gov/sequencingcostsdata [accessed May 2016]

https://www.etsy.com/listing/90627510/in-cod-we-trust-sign