PRINCESS: Framework for Comprehensive Detection and Phasing of SNP and SVs

Fritz Sedlazeck

Jan, 16, 2019

Emerging technologies improve genomes

BRIEF REPORT | Genetics inMedicine

Long-read genome sequencing identifies causal structural variation in a Mendelian disease

Jason D. Merker, MD, PhD[1,2], Aaron M. Wenger, PhD[3], Tam Sneddon, DPhil[2],

REVIEWS

COMPUTATIONAL TOOLS

Piercing the dark matter: bioinformatics of long-range sequencing and mapping

Fritz J. Sedlazeck[1], Hayan Lee[2], Charlotte A. Darby[3] and Michael C. Schatz[3,4]

CSH Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

Detection of GBA missense mutations and other variants using the Oxford Nanopore MinION

Melissa Leija-Salazar, Fritz J Sedlazeck, Katya Mokretar, Stephen Mullin, Marco Toffoli, Maria Athanasopoulou,

New Results

Multi-platform discovery of haplotype-resolved structural variation in human genomes

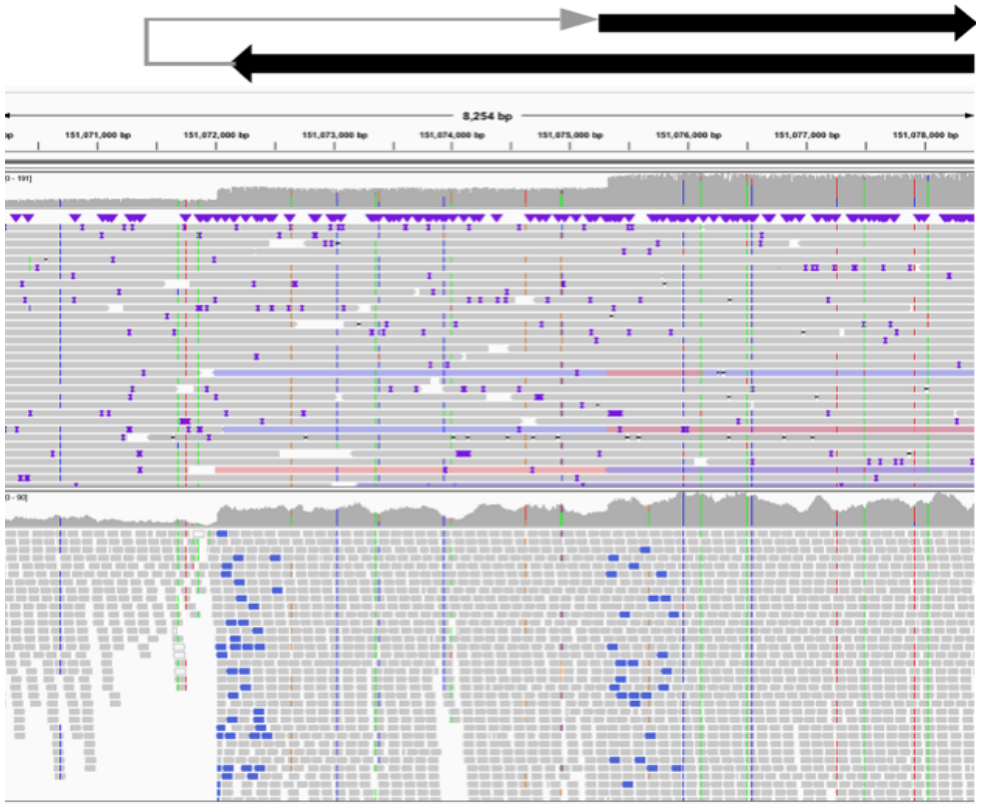
Mark J.P. Chaisson, Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J.

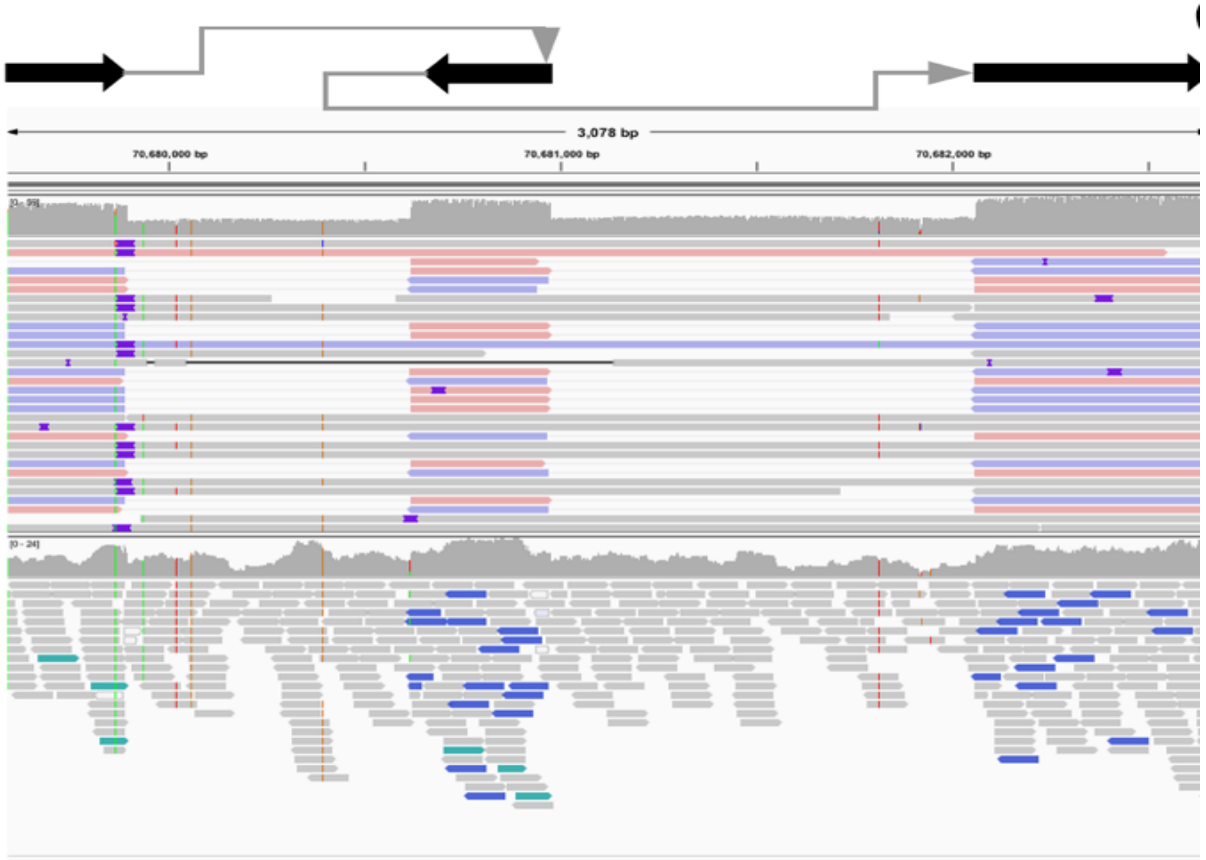
New Results

Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line

Maria Nattestad, Sara Goodwin, Karen Ng, Timour Baslan, Fritz J. Sedlazeck, Philipp Rescheneder, Tyler Garvin,

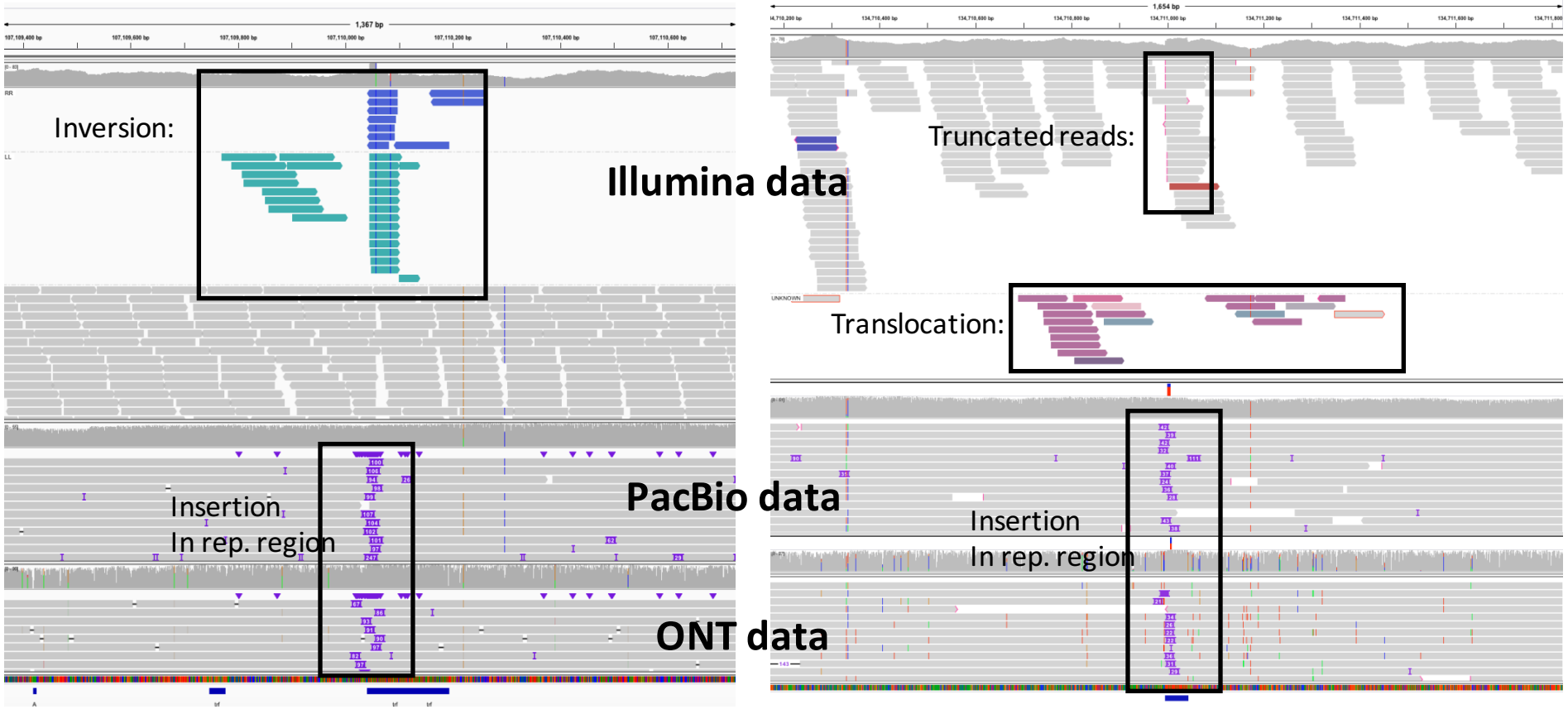
Identification of SVs with long reads





Sedlazeck et al. Nature Methods (2018)

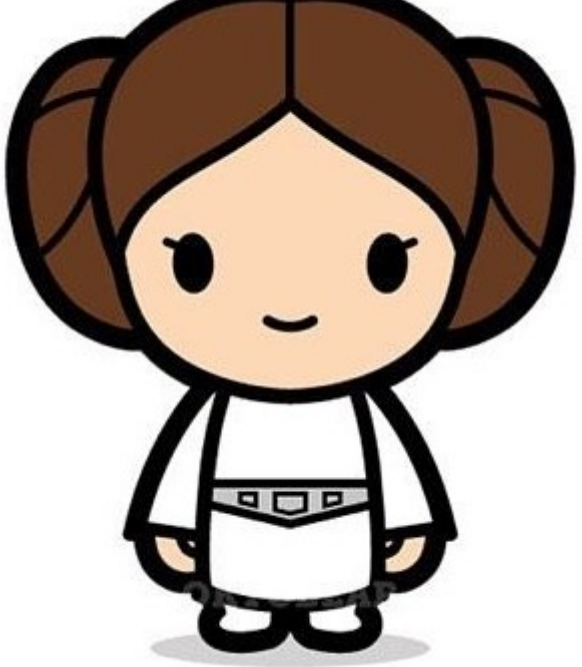
Short-read validation / False Discovery



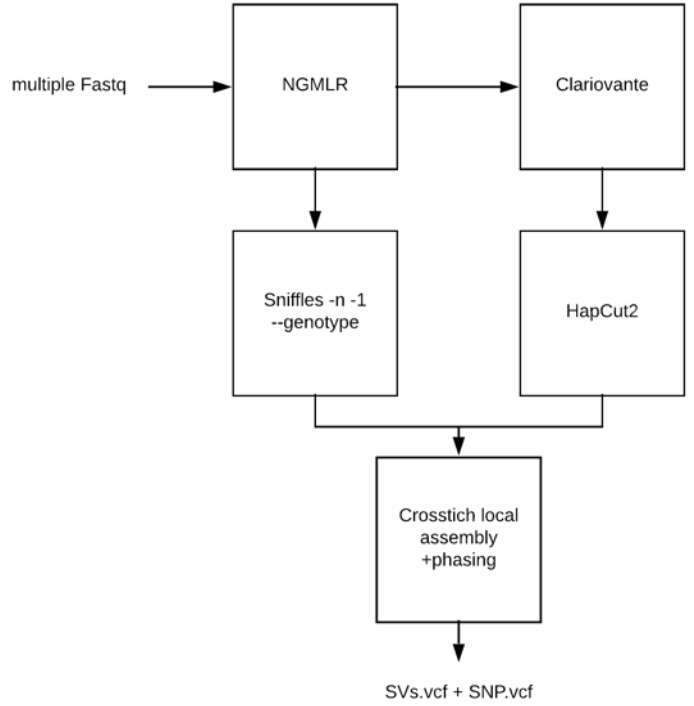
Sedlazeck et al. Nature Methods (2018)

Upcoming: Princess

- Snakemake pipeline to:
  - Call SNPs
  - Call SVs
  - Phase SNPs + SVs
- Includes helpful options
  - Parameter estimation
  - Parental phasing
  - Summary statistics and plots
    - Read length
    - Mapping statistics
    - Variation statistics
    - Phasing length statistics

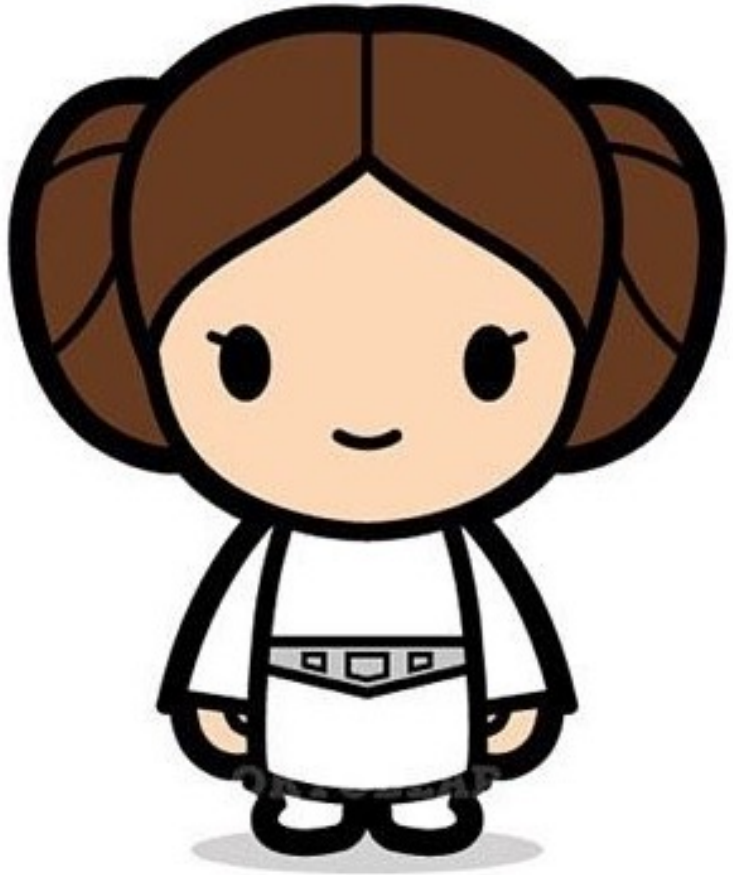


Medhat Mahmoud



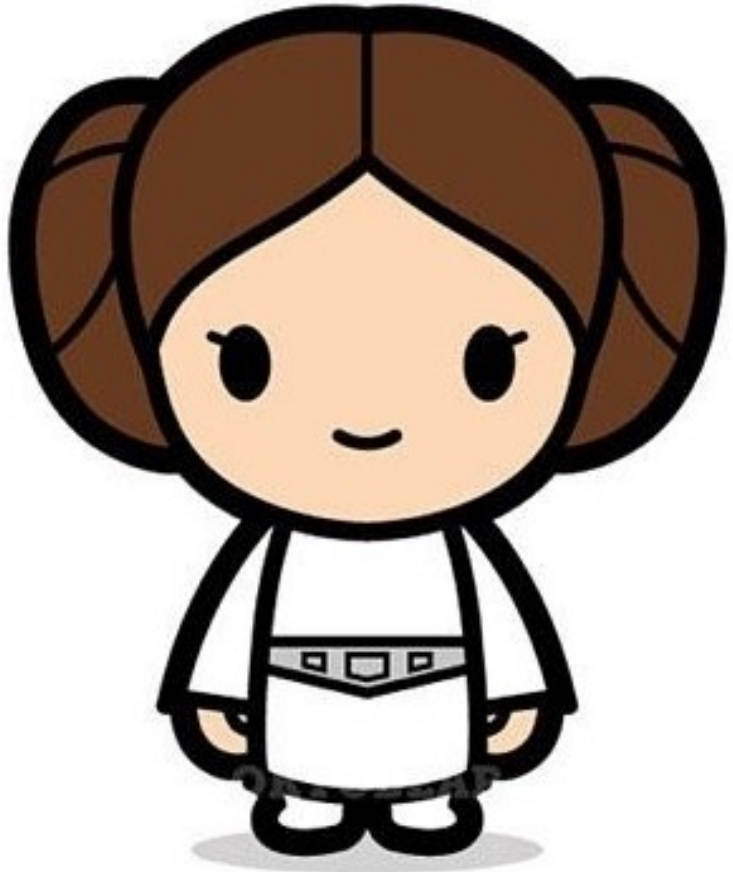
Princess

1. Mapping of long reads
2. SNP + SVs calling based on these reads
3. Phasing of SNP + SVs



Princess

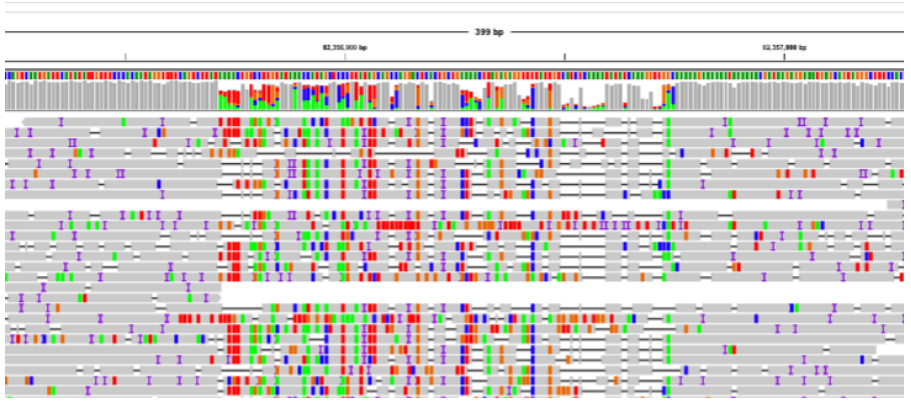
1. **Mapping of long reads**
2. SNP + SVs calling based on these reads
3. Phasing of SNP + SVs



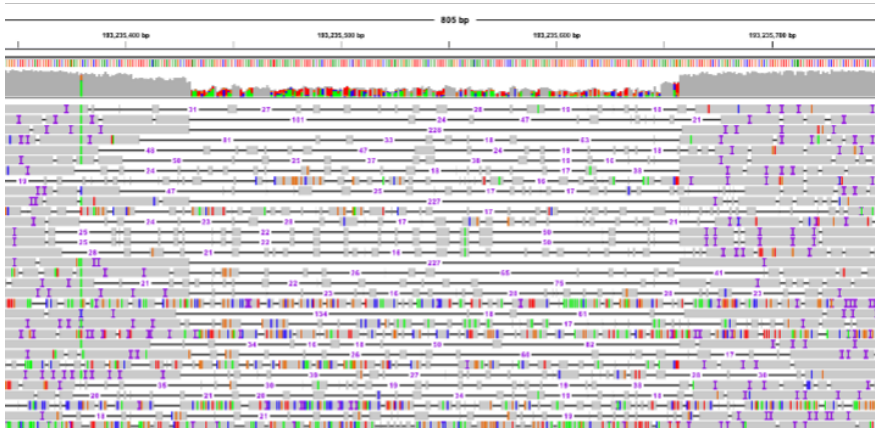
NGMLR

Philipp Rescheneder

1. Split the reads:
   - Translocations
   - Inversions
   - Duplications

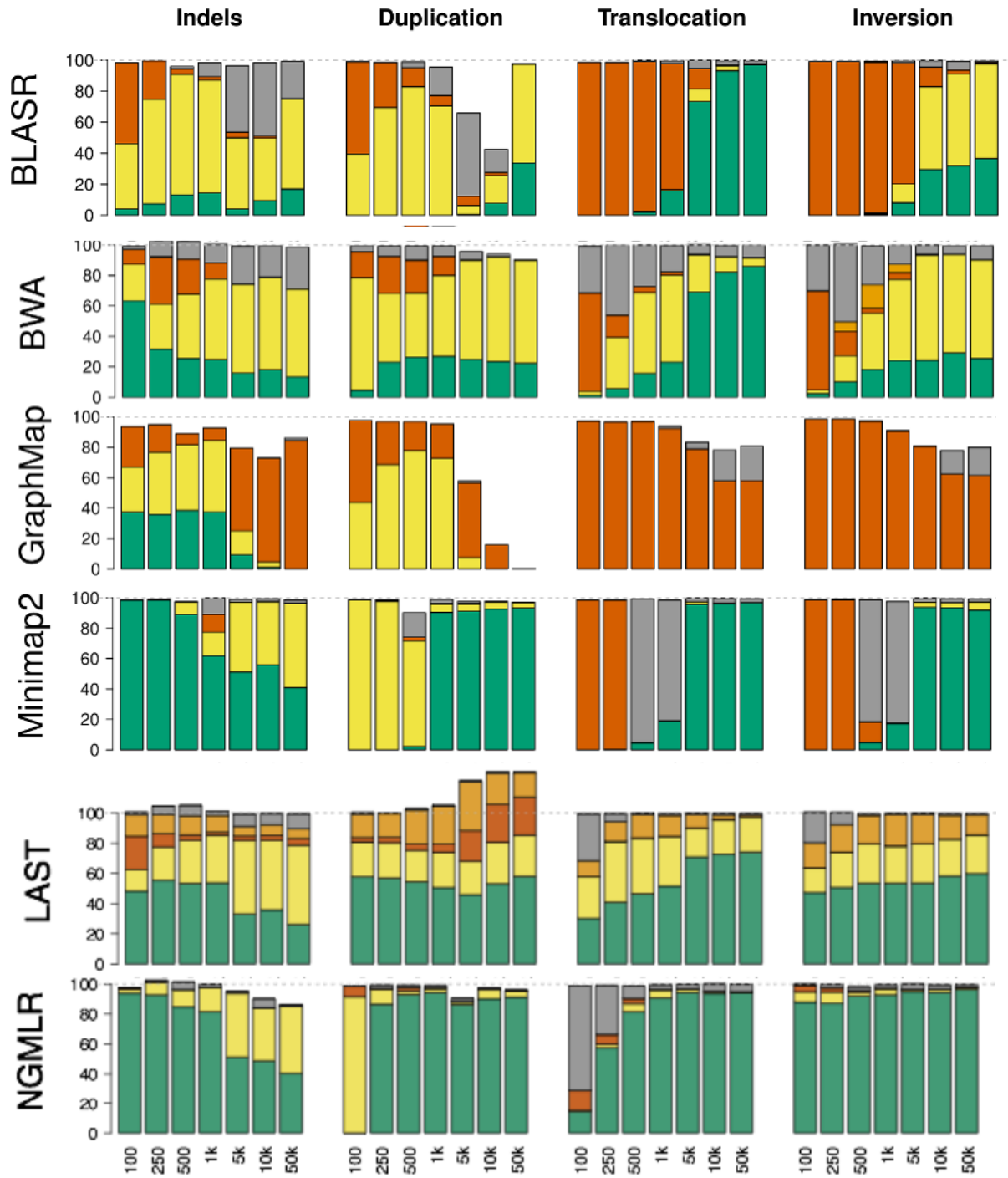


2. Improve alignment:
   - Insertions
   - Deletions



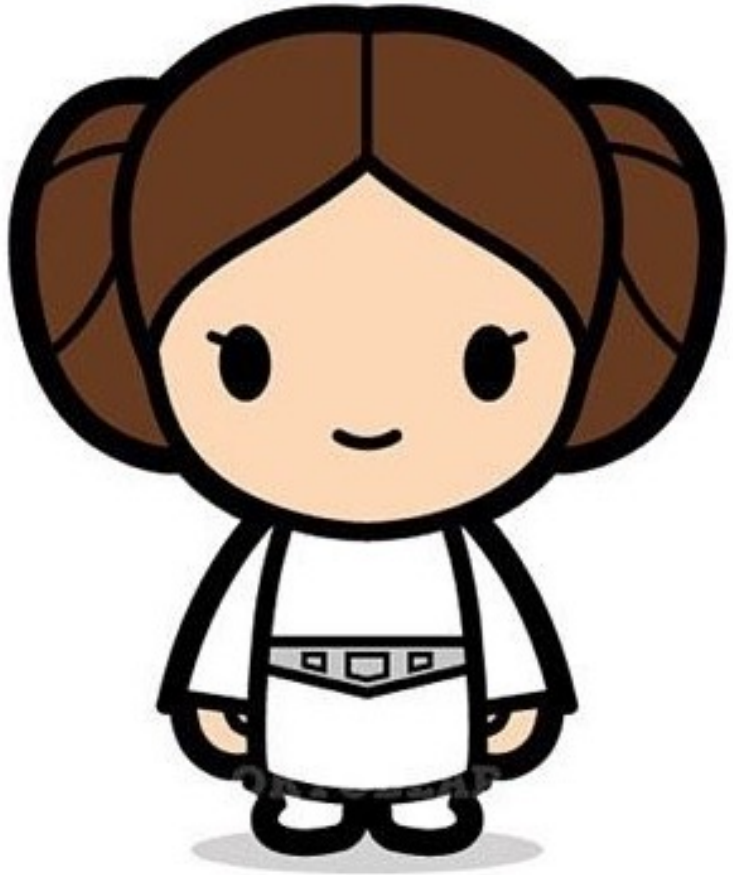
Mapping

- Precise
- Indicated
- Wrong
- Alignment stopped prior
- Fragmented

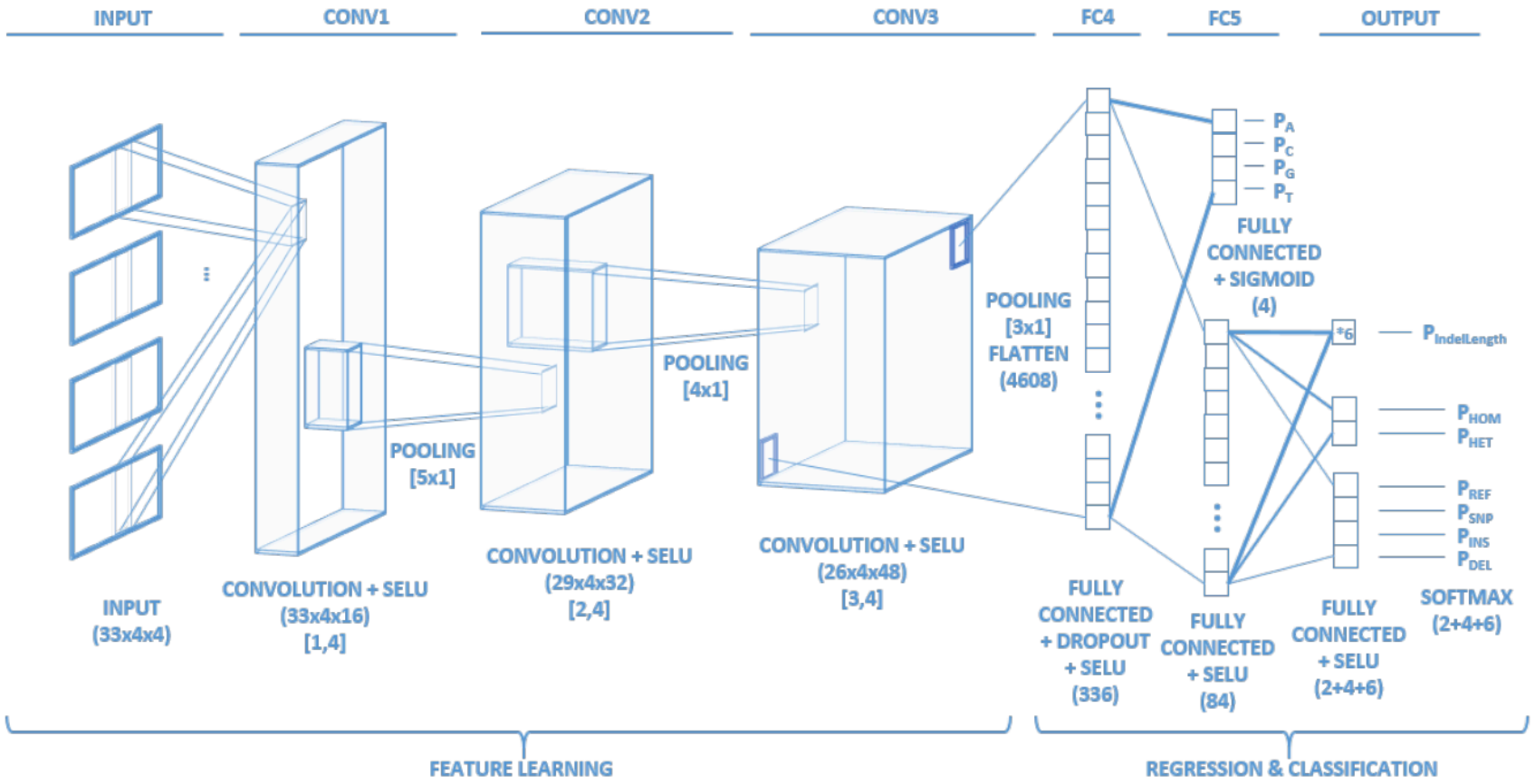


Princess

1. Mapping of long reads
2. **SNP + SVs calling based on these reads**
3. Phasing of SNP + SVs



Clairvoyante

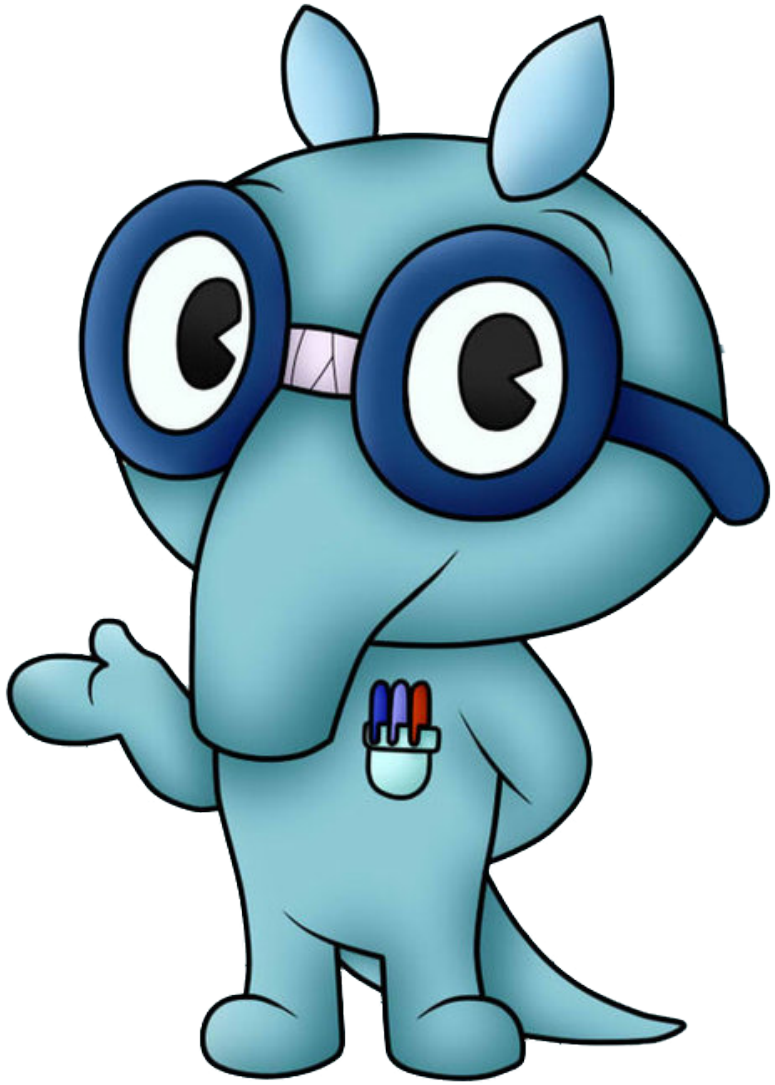


Ruibang Luo

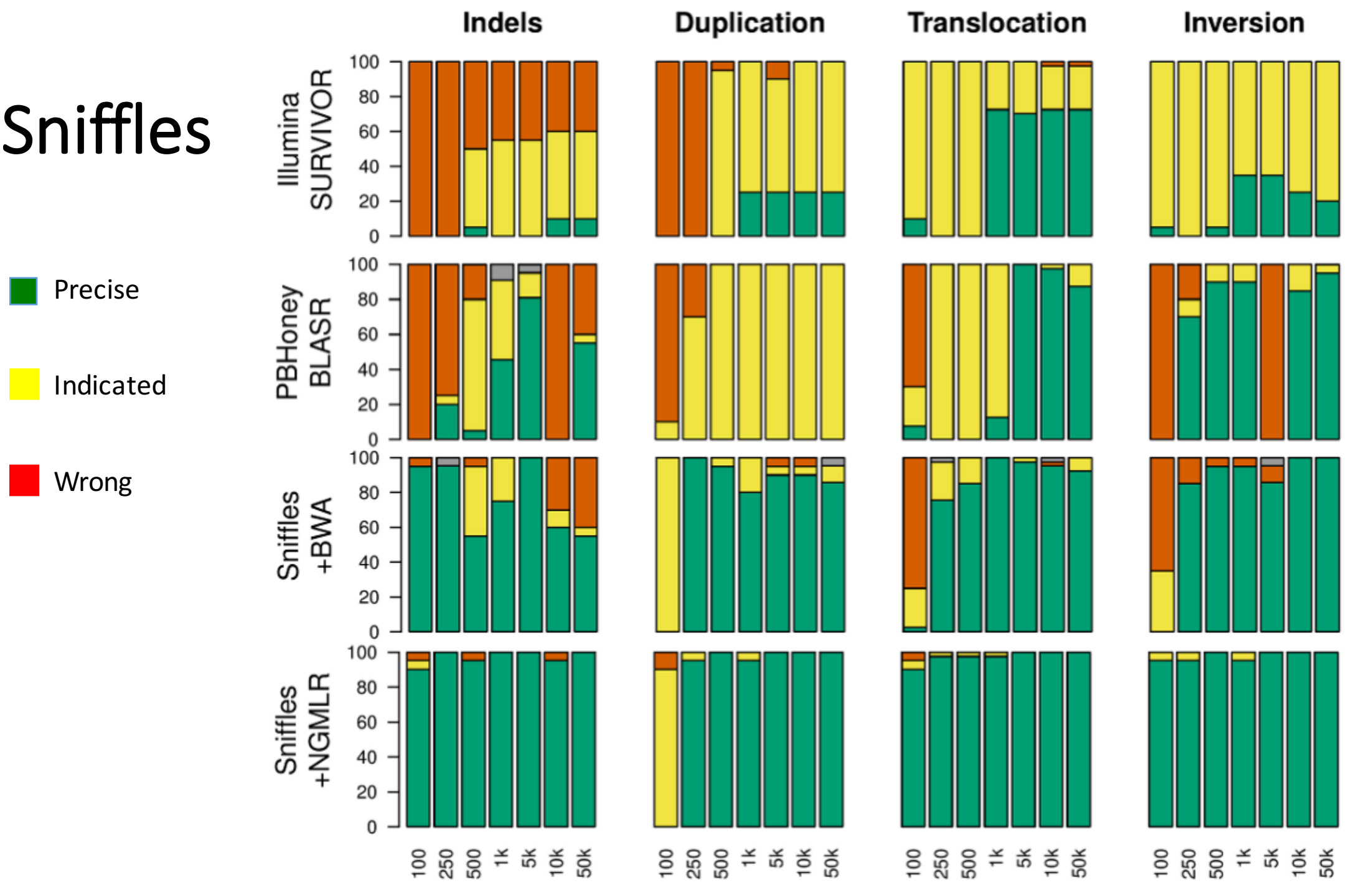
Sniffles

- Parameter estimation
- Analyzing:
  - split reads
  - alignment events
  - noisy regions
- Detect sequencing artifacts

- Optional:
  - Genotype estimation
  - Clustering/phasing of SVs
  - Sequence resolved indels



Sniffles


Precise
Indicated
Wrong
Indels
Duplication
Translocation
Inversion
Illumina SURVIVOR
PBHoney BLASR
Sniffles +BWA
Sniffles +NGMLR
0
20
40
60
80
100
100
250
500
1k
5k
10k
50k

Sniffles: HLA-F deletion

PacBio:
HS-1011

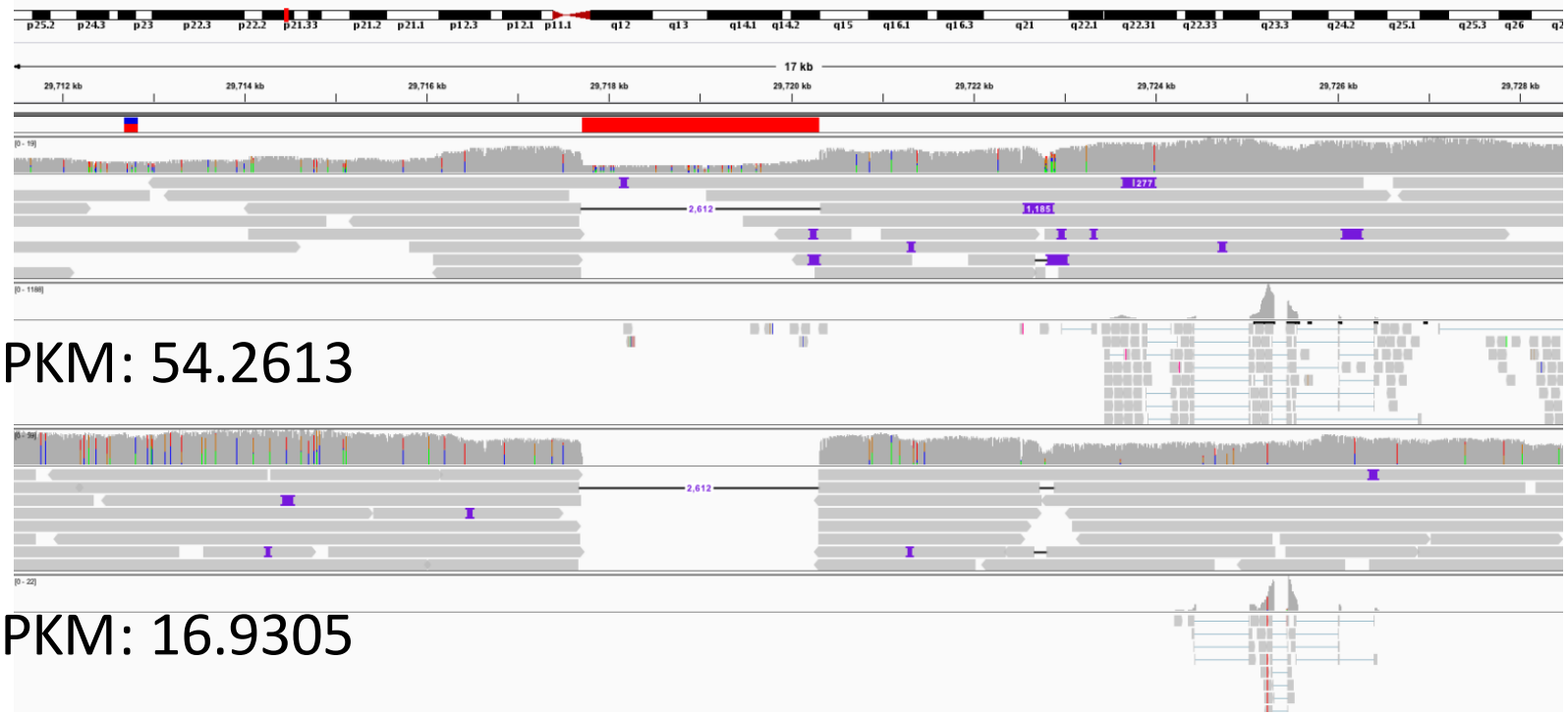
RNA-Seq
HS-1011

FPKM: 54.2613

PacBio:
NA12878

RNA-Seq
NA12878

FPKM: 16.9305

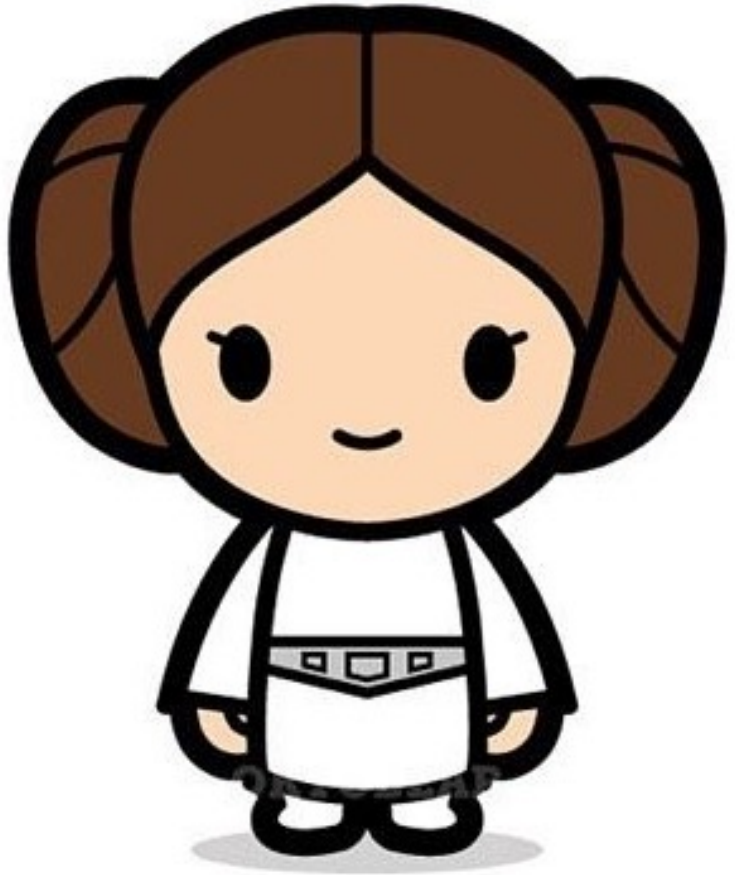


Princess

1. Mapping of long reads

2. SNP + SVs calling based on these reads

3. **Phasing of SNP + SVs**

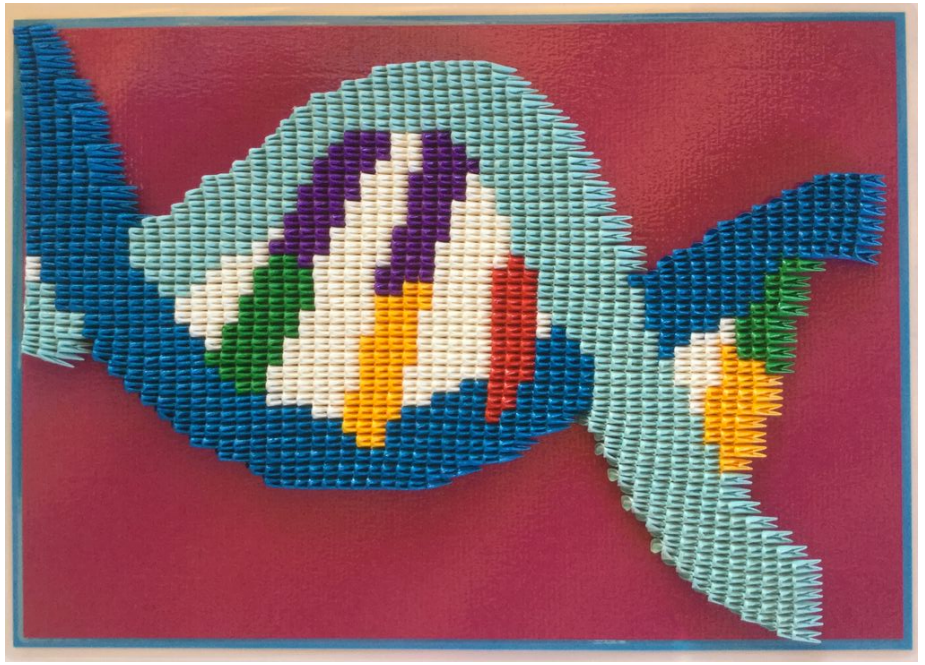


CrossStitch

Michael Kirsche

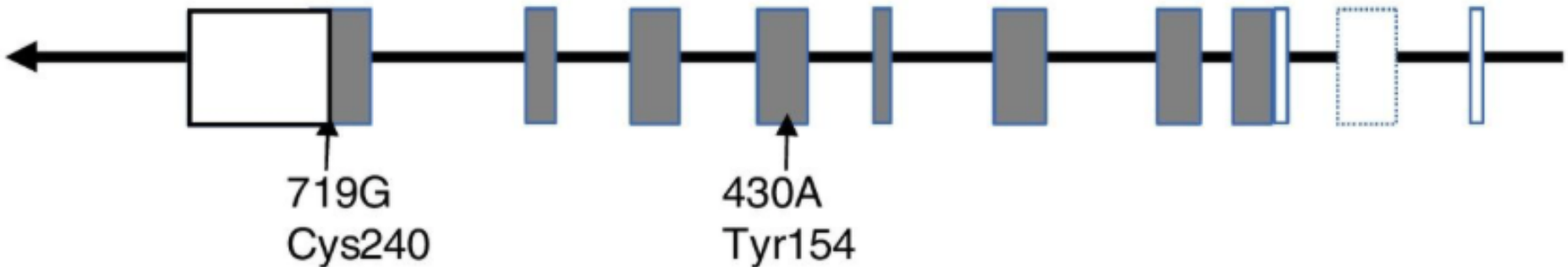
- Phasing:
  - Hapcut2 SNP phasing
  - Stitching phased SNPs + SVs calls

- Localized assembly
  - Improving insertion calls in Sniffles.

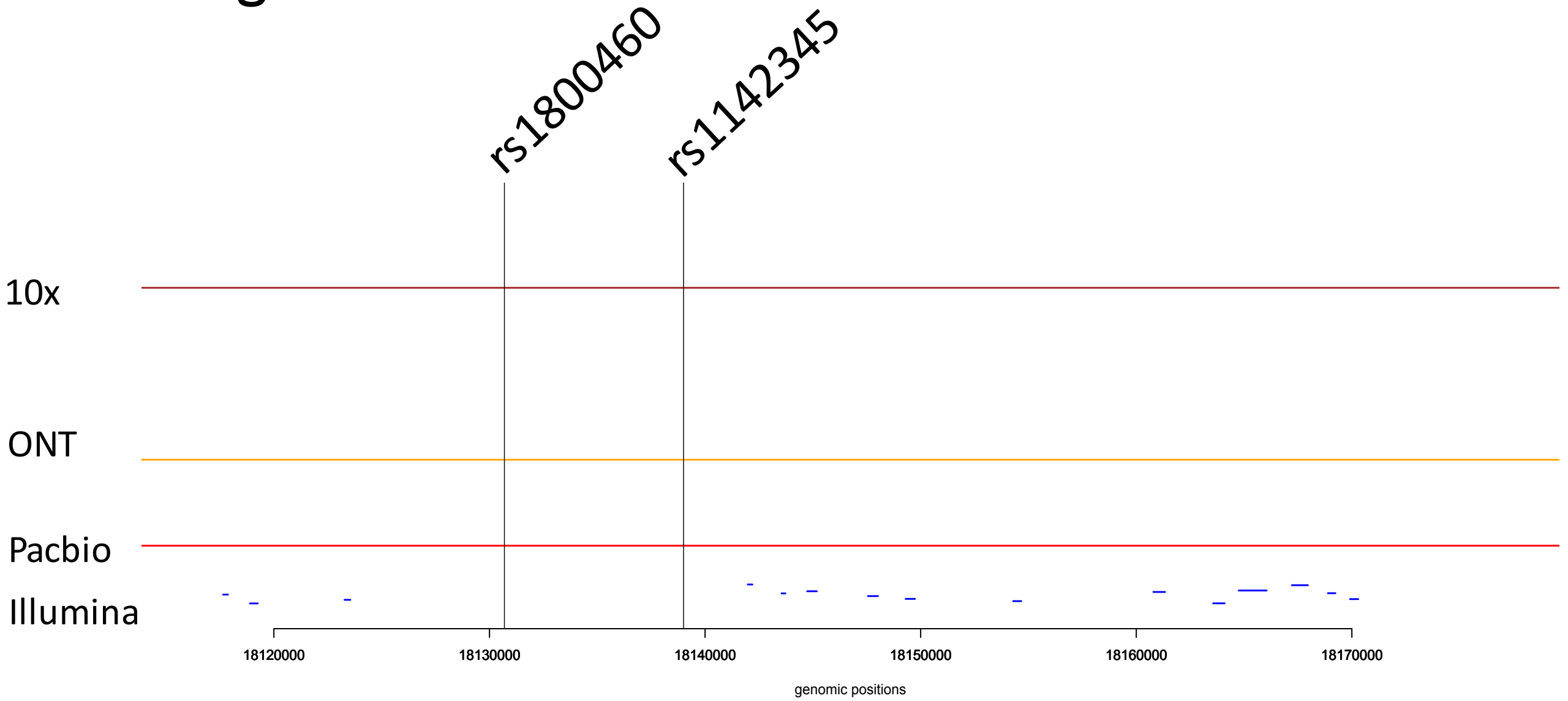


TPMT

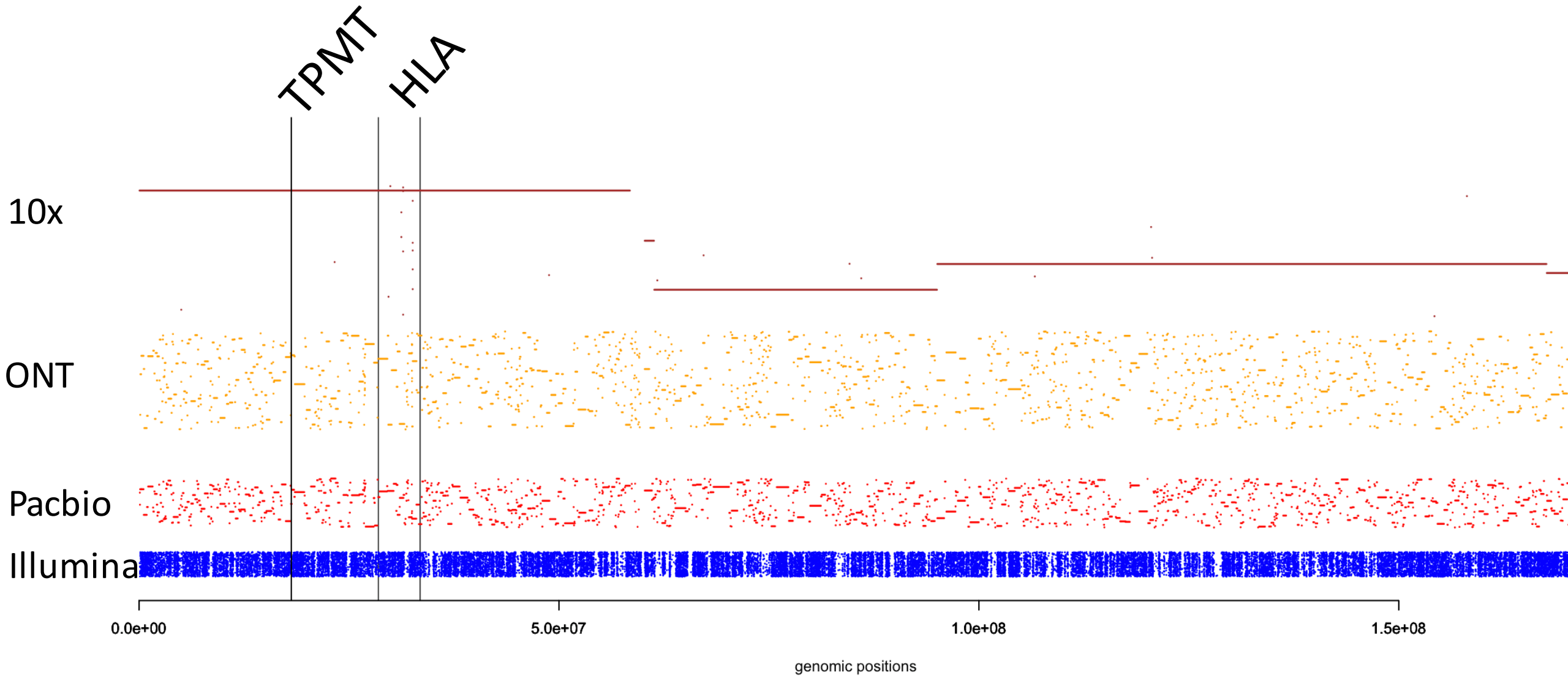
- TPMT gene - encodes the enzyme that metabolizes thiopurine drugs
- 3 drugs: azathioprine, mercaptopurine, thioguanine
  - Chemotherapy for pediatric leukemia
- Two variants, rs1800460 and rs1142345 (>8000 bases apart)
  - cis: TPMT*1/*3A diplotype (intermediate metabolizer)
  - trans: TPMT*3B/*3C diplotype (poor metabolizer)
  - *1/*3A diplotype is more common but less severe than *3B/*3C



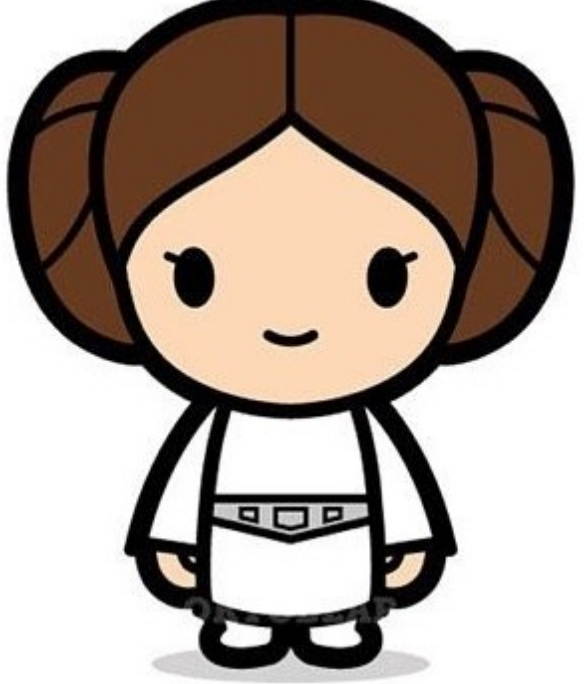
Phasing: TPMT



Phasing: Chr6



PRINCESS: Phase Resolution IN Comprehensive Small and Structural variants



- Focuses on high precision (95+).

- Will provide:
  - Parameter optimization
  - Provides QC and automatic parameter adaption.
  - Optional takes parental SNPs into account.

Data (HG002)	F-measure SVs	F-measure SNP
~17x long reads	79.19	46.8
~17x CCS	91.75	93.93

Acknowledgments

HGSC
HUMAN GENOME SEQUENCING CENTER

Medhat Helmi
Eric Venner
Adam Mansfield
David Murdock
William Salerno
Donna Muzny
Ginger Metcalf
Richard Gibbs

JOHNS HOPKINS
UNIVERSITY

Michael Schatz
Michael Kirsche
Schatz lab

香 港 大 學
THE UNIVERSITY OF HONG KONG

Ruibang Luo

NIH National Human Genome Research Institute