In pursuit of perfect genome sequencing

Michael Schatz

December 7, 2017

UMD Institute for Genome Sciences

In pursuit of perfect genome sequencing

1. Why “Perfect”?
2. What is “Perfect”?
3. How will we achieve it?
4. When will we achieve it?

In pursuit of perfect genome sequencing

1. **Why "Perfect"?**
2. What is "Perfect"?
3. How will we achieve it?
4. When will we achieve it?

The most wondrous map…

"Without a doubt, this is the most important, most wondrous map ever produced by humankind."

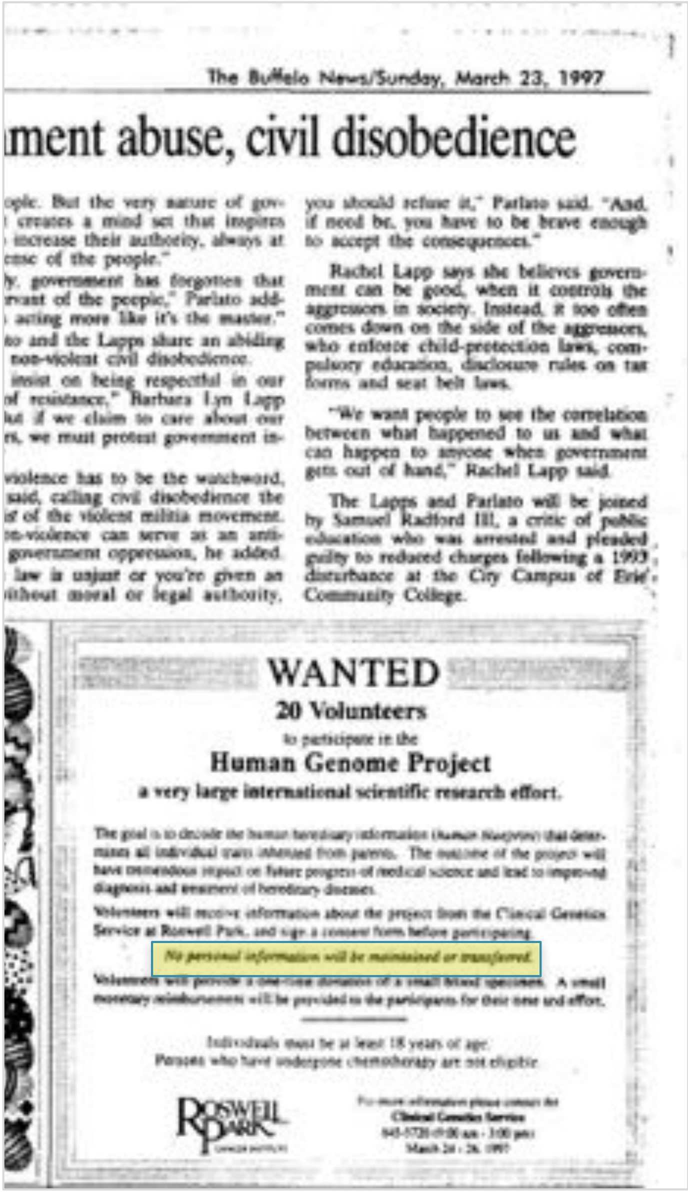
Bill Clinton
June 26, 2000

The most wondrous map…

"Without a doubt, this is the most important, most wondrous map ever produced by humankind."

Bill Clinton
June 26, 2000

Who is the reference human?



The Buffalo News/Sunday, March 23, 1997

ment abuse, civil disobedience

ople. But the very nature of gov-
t creates a mind set that inspires
increase their authority, always at
ense of the people."

y, government has forgotten that
rvant of the people," Parlato add-
acting more like it's the master."

to and the Lapps share an abiding
non-violent civil disobedience.

insist on being respectful in our
of resistance," Barbara Lyn Lapp
but if we claim to care about our
rs, we must protest government in-

violence has to be the watchword,
said, calling civil disobedience the
of the violent militia movement.
on-violence can serve as an anti-
government oppression, he added.
law is unjust or you're given an
ithout moral or legal authority,

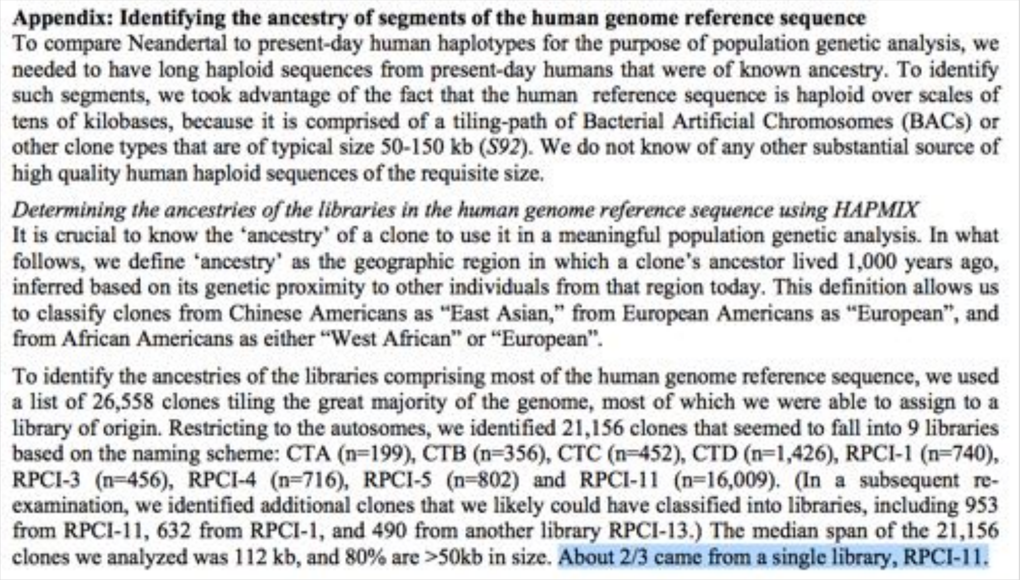
you should refuse it," Parlato said. "And, if need be, you have to be brave enough to accept the consequences."

Rachel Lapp says she believes government can be good, when it controls the aggressors in society. Instead, it too often comes down on the side of the aggressors, who enforce child-protection laws, compulsory education, disclosure rules on tax forms and seat belt laws.

"We want people to see the correlation between what happened to us and what can happen to anyone when government gets out of hand," Rachel Lapp said.

The Lapps and Parlato will be joined by Samuel Radford III, a critic of public education who was arrested and pleaded guilty to reduced charges following a 1993 disturbance at the City Campus of Erie Community College.

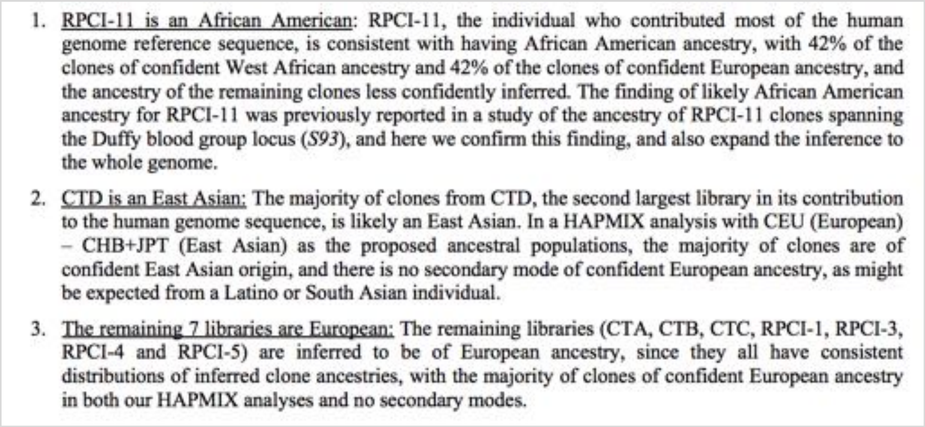
Pieter de Jong, RPCI



Appendix: Identifying the ancestry of segments of the human genome reference sequence
To compare Neandertal to present-day human haplotypes for the purpose of population genetic analysis, we needed to have long haploid sequences from present-day humans that were of known ancestry. To identify such segments, we took advantage of the fact that the human reference sequence is haploid over scales of tens of kilobases, because it is comprised of a tiling-path of Bacterial Artificial Chromosomes (BACs) or other clone types that are of typical size 50-150 kb (*S92*). We do not know of any other substantial source of high quality human haploid sequences of the requisite size.

Determining the ancestries of the libraries in the human genome reference sequence using HAPMIX
It is crucial to know the 'ancestry' of a clone to use it in a meaningful population genetic analysis. In what follows, we define 'ancestry' as the geographic region in which a clone's ancestor lived 1,000 years ago, inferred based on its genetic proximity to other individuals from that region today. This definition allows us to classify clones from Chinese Americans as "East Asian," from European Americans as "European", and from African Americans as either "West African" or "European".

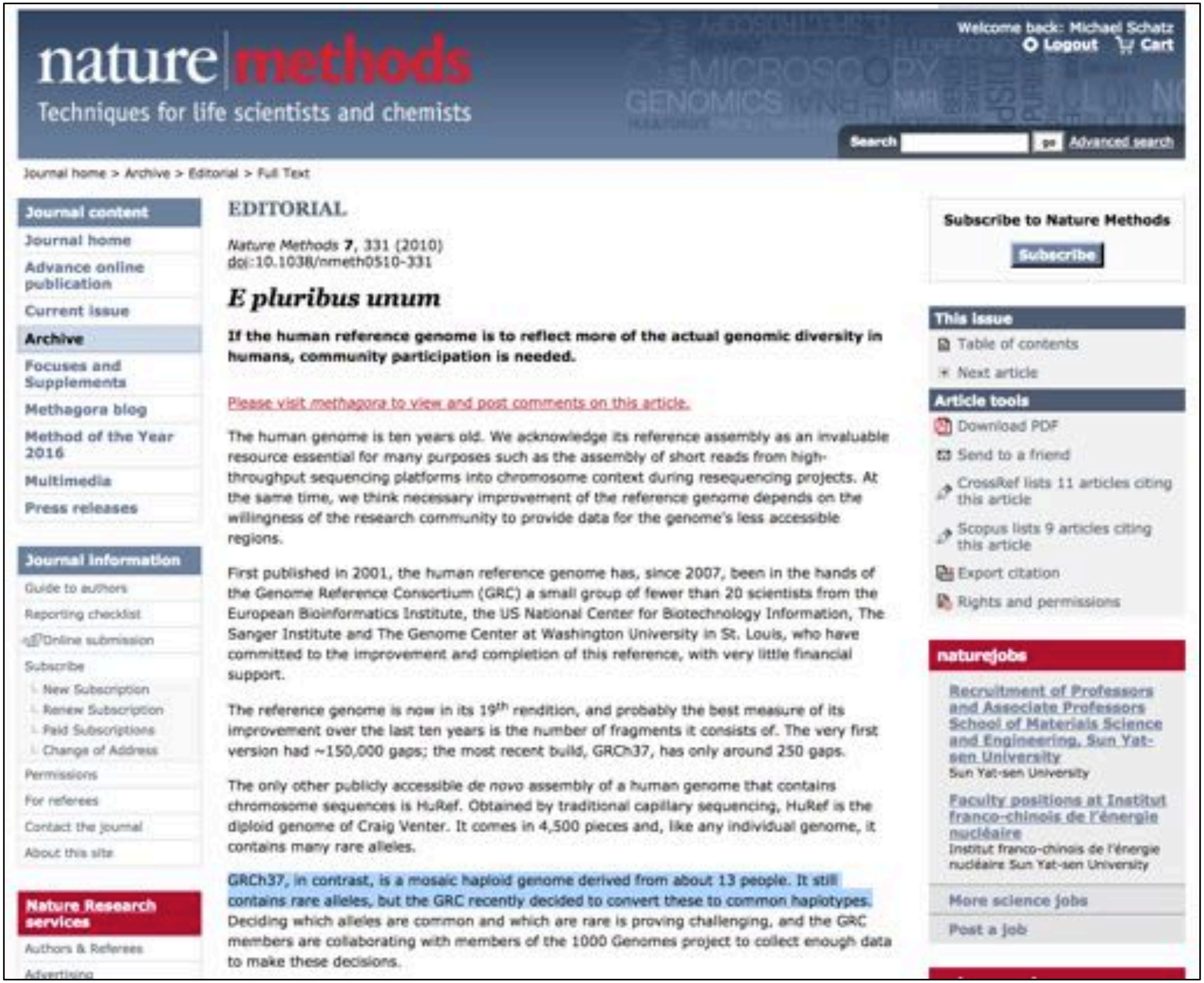
To identify the ancestries of the libraries comprising most of the human genome reference sequence, we used a list of 26,558 clones tiling the great majority of the genome, most of which we were able to assign to a library of origin. Restricting to the autosomes, we identified 21,156 clones that seemed to fall into 9 libraries based on the naming scheme: CTA (n=199), CTB (n=356), CTC (n=452), CTD (n=1,426), RPCI-1 (n=740), RPCI-3 (n=456), RPCI-4 (n=716), RPCI-5 (n=802) and RPCI-11 (n=16,009). (In a subsequent re-examination, we identified additional clones that we likely could have classified into libraries, including 953 from RPCI-11, 632 from RPCI-1, and 490 from another library RPCI-13.) The median span of the 21,156 clones we analyzed was 112 kb, and 80% are >50kb in size. About 2/3 came from a single library, RPCI-11.



1. RPCI-11 is an African American: RPCI-11, the individual who contributed most of the human genome reference sequence, is consistent with having African American ancestry, with 42% of the clones of confident West African ancestry and 42% of the clones of confident European ancestry, and the ancestry of the remaining clones less confidently inferred. The finding of likely African American ancestry for RPCI-11 was previously reported in a study of the ancestry of RPCI-11 clones spanning the Duffy blood group locus (*S93*), and here we confirm this finding, and also expand the inference to the whole genome.
2. CTD is an East Asian: The majority of clones from CTD, the second largest library in its contribution to the human genome sequence, is likely an East Asian. In a HAPMIX analysis with CEU (European) – CHB+JPT (East Asian) as the proposed ancestral populations, the majority of clones are of confident East Asian origin, and there is no secondary mode of confident European ancestry, as might be expected from a Latino or South Asian individual.
3. The remaining 7 libraries are European: The remaining libraries (CTA, CTB, CTC, RPCI-1, RPCI-3, RPCI-4 and RPCI-5) are inferred to be of European ancestry, since they all have consistent distributions of inferred clone ancestries, with the majority of clones of confident European ancestry in both our HAPMIX analyses and no secondary modes.

A Draft Sequence of the Neandertal Genome
Green et al (2010) Science. DOI: 10.1126/science.1188021
Supplemental Note 16 (pg 145-146)

Who is the reference human?



EDITORIAL

Nature Methods **7**, 331 (2010)
doi:10.1038/nmeth0510-331

E pluribus unum

If the human reference genome is to reflect more of the actual genomic diversity in humans, community participation is needed.

Please visit *methagora* to view and post comments on this article.

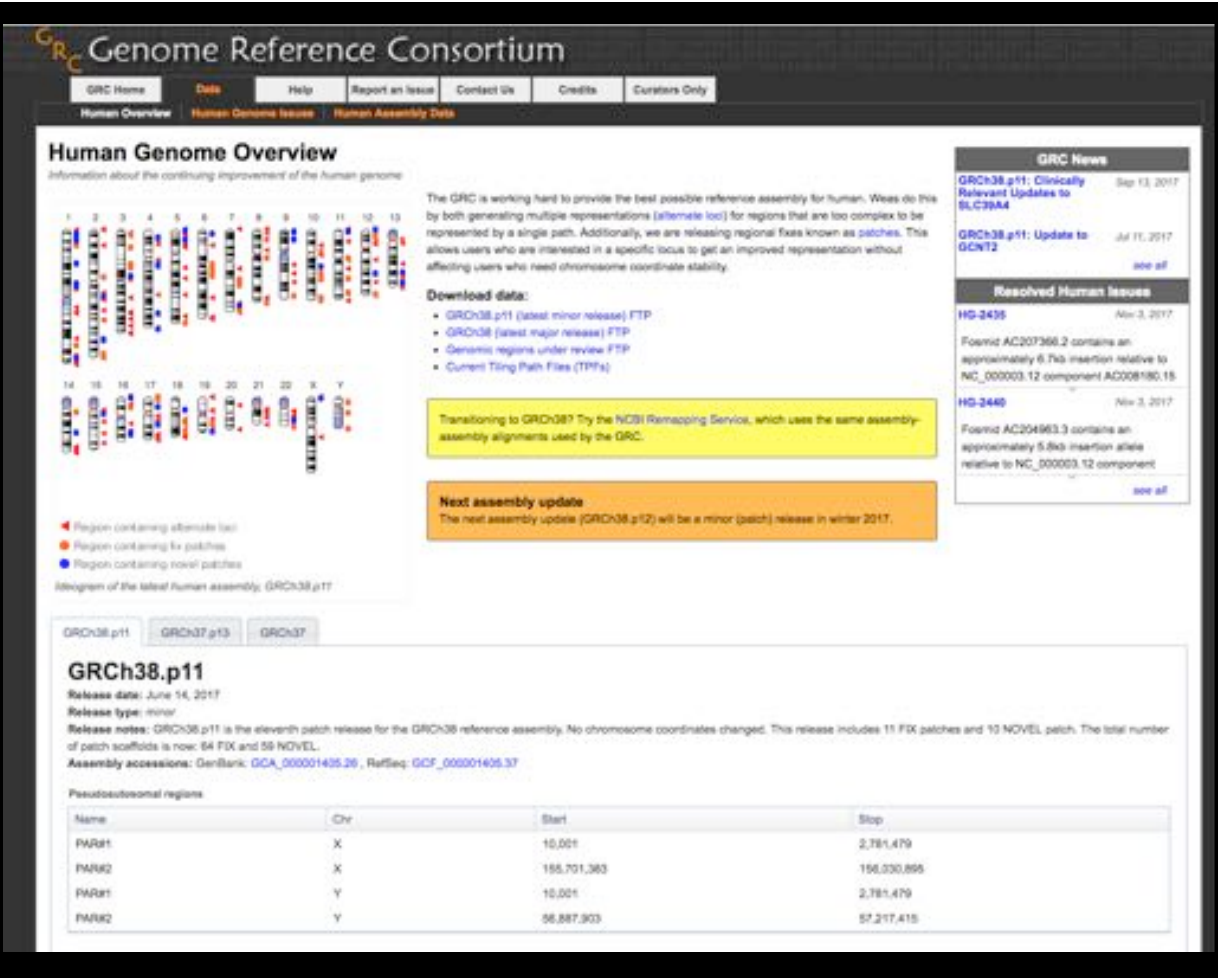
The human genome is ten years old. We acknowledge its reference assembly as an invaluable resource essential for many purposes such as the assembly of short reads from high-throughput sequencing platforms into chromosome context during resequencing projects. At the same time, we think necessary improvement of the reference genome depends on the willingness of the research community to provide data for the genome's less accessible regions.

First published in 2001, the human reference genome has, since 2007, been in the hands of the Genome Reference Consortium (GRC) a small group of fewer than 20 scientists from the European Bioinformatics Institute, the US National Center for Biotechnology Information, The Sanger Institute and The Genome Center at Washington University in St. Louis, who have committed to the improvement and completion of this reference, with very little financial support.

The reference genome is now in its 19th rendition, and probably the best measure of its improvement over the last ten years is the number of fragments it consists of. The very first version had ~150,000 gaps; the most recent build, GRCh37, has only around 250 gaps.

The only other publicly accessible *de novo* assembly of a human genome that contains chromosome sequences is HuRef. Obtained by traditional capillary sequencing, HuRef is the diploid genome of Craig Venter. It comes in 4,500 pieces and, like any individual genome, it contains many rare alleles.

GRCh37, in contrast, is a mosaic haploid genome derived from about 13 people. It still contains rare alleles, but the GRC recently decided to convert these to common haplotypes. Deciding which alleles are common and which are rare is proving challenging, and the GRC members are collaborating with members of the 1000 Genomes project to collect enough data to make these decisions.



Human Genome Overview

Information about the continuing improvement of the human genome

The GRC is working hard to provide the best possible reference assembly for human. Weas do this by both generating multiple representations (alternate loci) for regions that are too complex to be represented by a single path. Additionally, we are releasing regional fixes known as patches. This allows users who are interested in a specific locus to get an improved representation without affecting users who need chromosome coordinate stability.

Download data:

- GRCh38.p11 (latest minor release) FTP
- GRCh38 (latest major release) FTP
- Genomic regions under review FTP
- Current Tiling Path Files (TPFs)

Transitioning to GRCh38? Try the NCBI Remapping Service, which uses the same assembly-assembly alignments used by the GRC.

Next assembly update
The next assembly update (GRCh38.p12) will be a minor (patch) release in winter 2017.

◀ Region containing alternate loci
● Region containing fix patches
● Region containing novel patches

Ideogram of the latest human assembly, GRCh38.p11

GRC News

GRCh38.p11: Clinically Relevant Updates to SLC39A4 Sep 13, 2017

GRCh38.p11: Update to GCNT2 Jul 11, 2017

see all

Resolved Human Issues

HG-2435 Nov 3, 2017

Fosmid AC207366.2 contains an approximately 6.7kb insertion relative to NC_000003.12 component AC008180.15.

HG-2440 Nov 3, 2017

Fosmid AC204963.3 contains an approximately 5.8kb insertion allele relative to NC_000003.12 component

see all

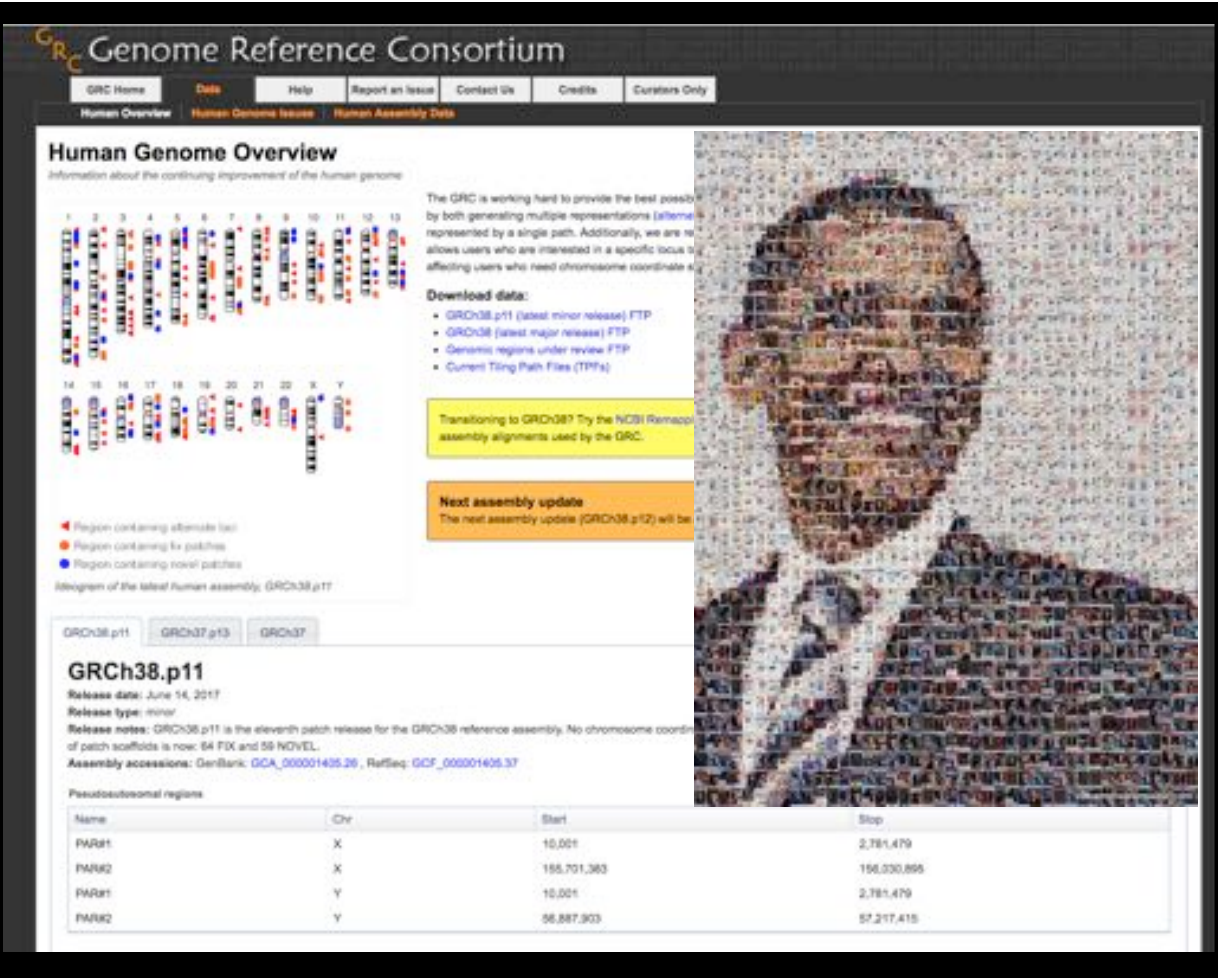
GRCh38.p11 | GRCh37.p13 | GRCh37

GRCh38.p11

Release date: June 14, 2017
Release type: minor
Release notes: GRCh38.p11 is the eleventh patch release for the GRCh38 reference assembly. No chromosome coordinates changed. This release includes 11 FIX patches and 10 NOVEL patch. The total number of patch scaffolds is now: 64 FIX and 59 NOVEL.
Assembly accessions: GenBank: GCA_000001405.26 , RefSeq: GCF_000001405.37

Pseudoautosomal regions

Name	Chr	Start	Stop
PAR#1	X	10,001	2,781,479
PAR#2	X	155,701,383	156,030,895
PAR#1	Y	10,001	2,781,479
PAR#2	Y	56,887,903	57,217,415



Human Genome Overview

Information about the continuing improvement of the human genome

The GRC is working hard to provide the best possib
by both generating multiple representations (alterna
represented by a single path. Additionally, we are re
allows users who are interested in a specific locus t
affecting users who need chromosome coordinate s

Download data:

- GRCh38.p11 (latest minor release) FTP
- GRCh38 (latest major release) FTP
- Genomic regions under review FTP
- Current Tiling Path Files (TPFs)

Transitioning to GRCh38? Try the NCBI Remapp
assembly alignments used by the GRC.

Next assembly update
The next assembly update (GRCh38.p12) will be

◀ Region containing alternate loci
● Region containing fix patches
● Region containing novel patches

Ideogram of the latest human assembly, GRCh38.p11

GRCh38.p11

Release date: June 14, 2017
Release type: minor
Release notes: GRCh38.p11 is the eleventh patch release for the GRCh38 reference assembly. No chromosome coordin
of patch scaffolds is now: 64 FIX and 59 NOVEL.
Assembly accessions: GenBank: GCA_000001405.26 , RefSeq: GCF_000001405.37

Pseudoautosomal regions

Name	Chr	Start	Stop
PAR#1	X	10,001	2,781,479
PAR#2	X	155,701,383	156,030,895
PAR#1	Y	10,001	2,781,479
PAR#2	Y	56,887,903	57,217,415

Importance of Personal Genomes

Current standard is to align your data to the “reference” human genome.

But the “reference” isn’t really the genome for *any* human and potentially biases the results in many ways:

- ***Genome:*** biased read mapping, causing false positive and false negative mutations
- ***Transcriptome***: mutations of splice sites, stop codons or branch point change gene models, CNVs modulate expression levels, gene fusions create new transcripts
- ***Epigenome***: *cis* versus *trans* effects, *allele-specific expression, allele-specific binding*

Same issues apply to most “reference” genomes

In pursuit of perfect genome sequencing

1. **Why “Perfect”?**

   Because it is important, complex, and personal

2. **What is “Perfect”?**

3. **How will we achieve it?**

4. **When will we achieve it?**

In pursuit of perfect genome sequencing

1. Why “Perfect”?

2. **What is “Perfect”?**

3. How will we achieve it?

4. When will we achieve it?

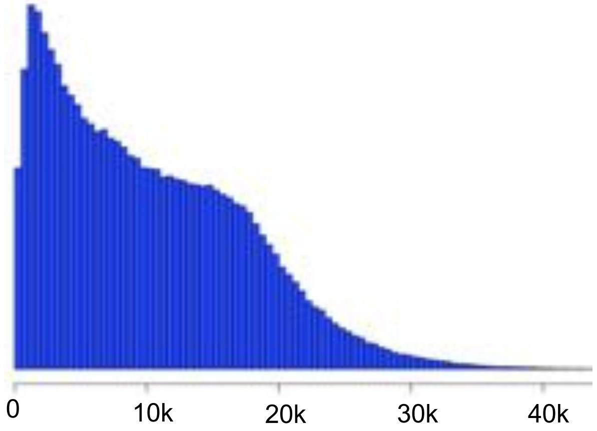
1. Correctness:

Is the genome faithfully represented?

1. Correctness:

Is the genome faithfully represented?

PacBio RS II



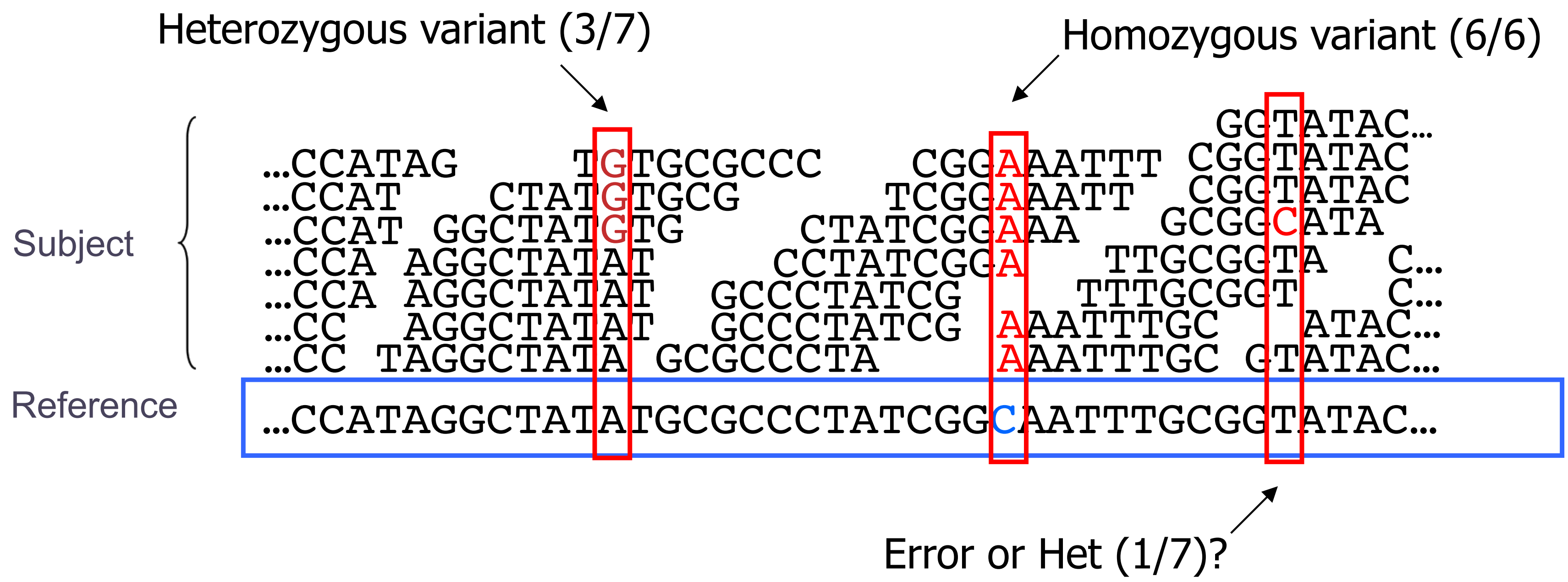
CSHL/PacBio

```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
||||||||||||||||||||||||| ||||||| |||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAAGGCGGCTAGG
| |||||| ||||||||||||| |||| | |||||| |||||| ||||||
A-TATAAATCAGTTGATCCATTAAGAA-AGAAACGC-AAAGGC-GCTAGG

CAACCTTGAATGTAATCGCACTTGAAGAACAAGATTTTATTCCGCGCCCG
| |||||| |||| ||  ||||||||||||||||||||||||||||||||
C-ACCTTG-ATGT-AT--CACTTGAAGAACAAGATTTTATTCCGCGCCCG

TAACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA
| ||||||| |||||||||||||| || ||    |||||||||| |||||
T-ACGAATC-AGATTCTGAAAACA-ATGAT----ACCTCCAAAAGCACAA

-AGGAGGGGAAAGGGGGGAATATCT-ATAAAAGATTACAAATTAGA-TGA
 ||||||   ||     |||||||| || |||||||||||||| || |||
GAGGAGG---AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA

ACT-AATTCACAATA-AATAACACTTTTA-ACAGAATTGAT-GGAA-GTT
||| ||||||||| | ||||||||||||| ||| ||||||| |||| |||
ACTAAATTCACAA-ATAATAACACTTTTAGACAAAATTGATGGGAAGGTT

TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA
|| ||||||||| ||||||| ||| |||| |||||| ||||| |||||||
TC-GAGAGATCC-AAACAAT-GGCGATCG-CTTTGACGTTACAAATCAAA
```

Sample of 100k reads aligned with BLASR requiring >100bp alignment

Average overall accuracy 83.7%: 11.5% insertions, 3.4% deletions, 1.4% mismatch

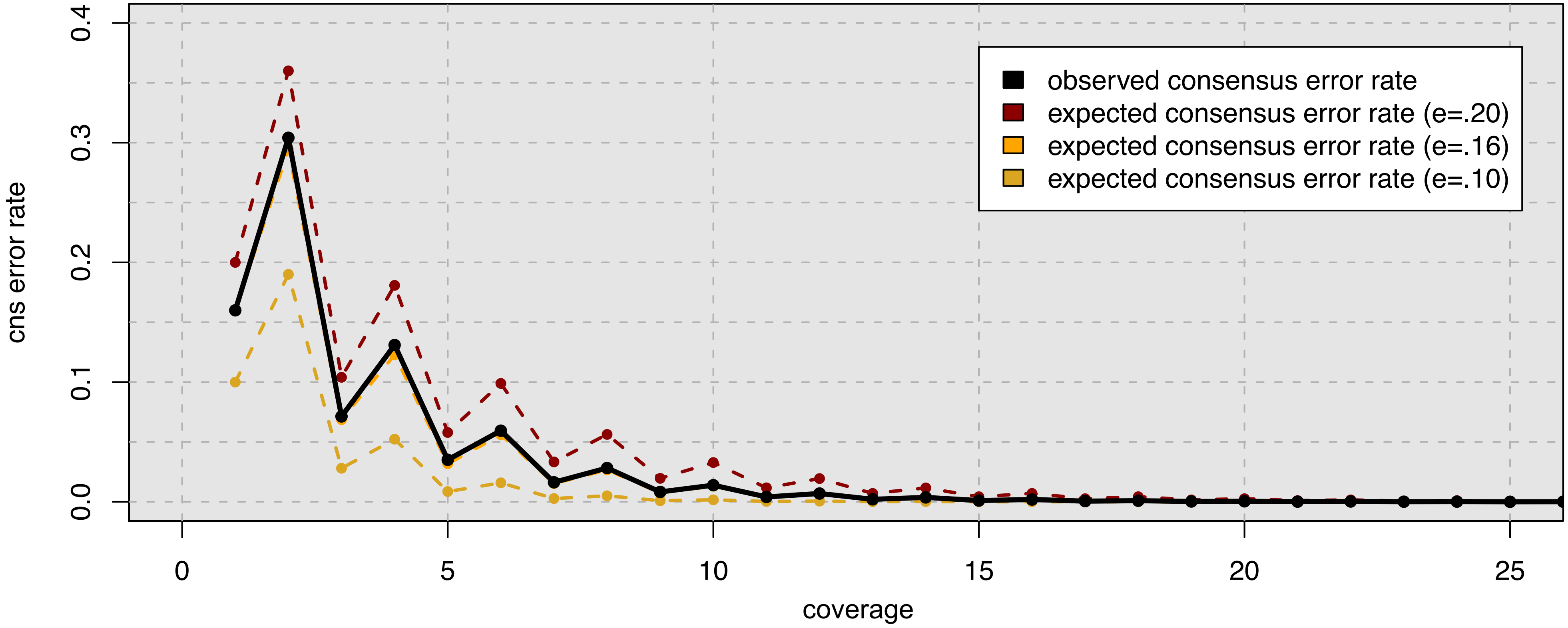
Genotyping Theory



```
                                    GGTATAC...
...CCATAG    TGTGCGCCC   CGGAAATTT CGGTATAC
...CCAT   CTATGTGCG     TCGGAAATT  CGGTATAC
...CCAT GGCTATGTG    CTATCGGAAA   GCGGCATA
...CCA AGGCTATAT    CCTATCGGA   TTGCGGTA  C...
...CCA AGGCTATAT  GCCCTATCG    TTTGCGGT   C...
...CC  AGGCTATAT  GCCCTATCG AAATTTGC   ATAC...
...CC TAGGCTATA GCGCCCTA    AAATTTGC GTATAC...
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
```

- If there were no sequencing errors, identifying SNPs would be trivial:
  - Any time a read disagrees with the reference, it must be a variant!
- A single read of many differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times
  - Use binomial test to evaluate prob. of heterozygosity vs. prob of error
  - Coverage (oversampling) is our main tool to improve accuracy

Consensus Accuracy and Coverage



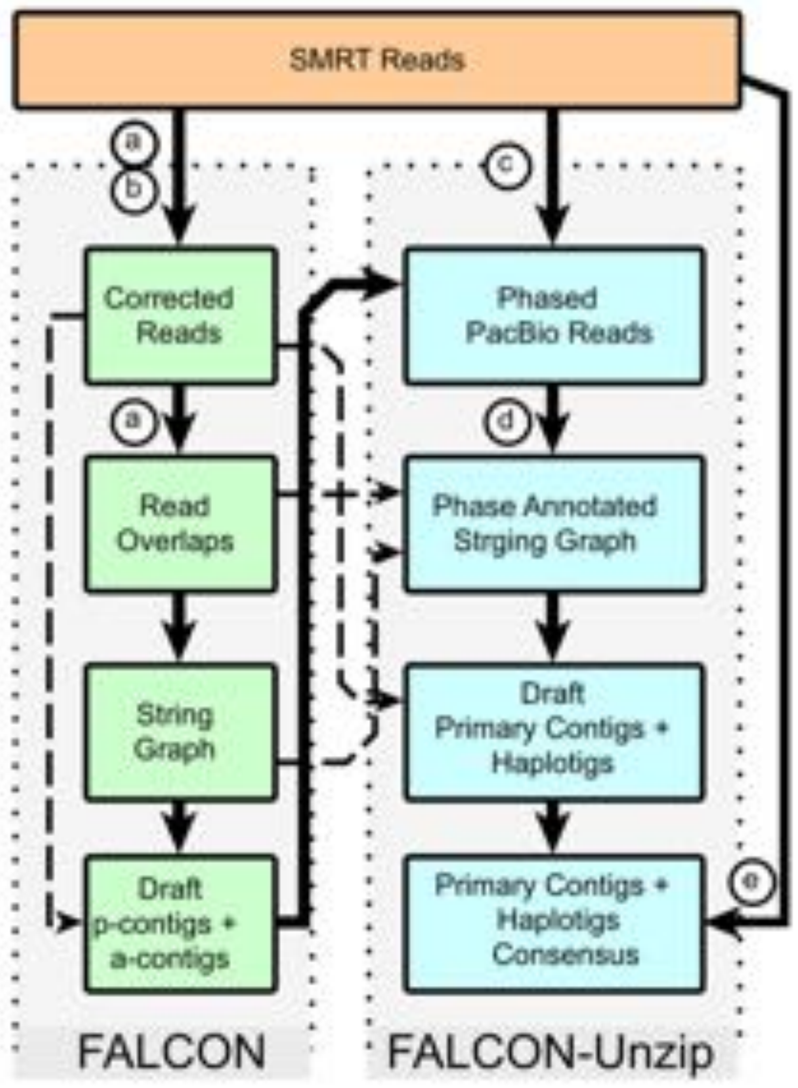
Coverage can overcome random errors

- Dashed: error model from binomial sampling
- Solid: observed accuracy

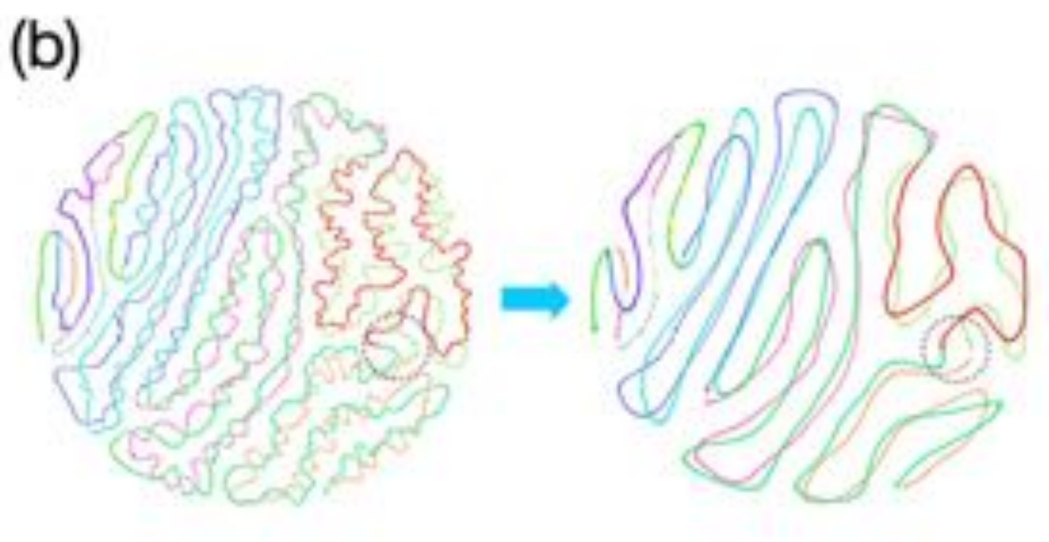
$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, Schatz et al (2012) *Nature Biotechnology. doi:10.1038/nbt.2280*

FALCON Accuracy


SMRT Reads
a
b
c
Corrected Reads
Phased PacBio Reads
a
d
Read Overlaps
Phase Annotated Strging Graph
String Graph
Draft Primary Contigs + Haplotigs
Draft p-contigs + a-contigs
Primary Contigs + Haplotigs Consensus
e
FALCON
FALCON-Unzip

"***The overall base-to-base concordance rate is about 99.99%*** (QV40 in Phred scale) in the F1 FALCON-Unzip assembly. The insertion and deletion (indel) concordances to the parental lines were lower (about QV40) than the SNP concordance rate (about QV50), with most residual errors concentrated in long homopolymer sequences"



Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing
Chin et al (2016) *Nature Methods. doi:10.1038/nmeth.4035.*

2. ***Completeness*:**

How much of the genome is present?

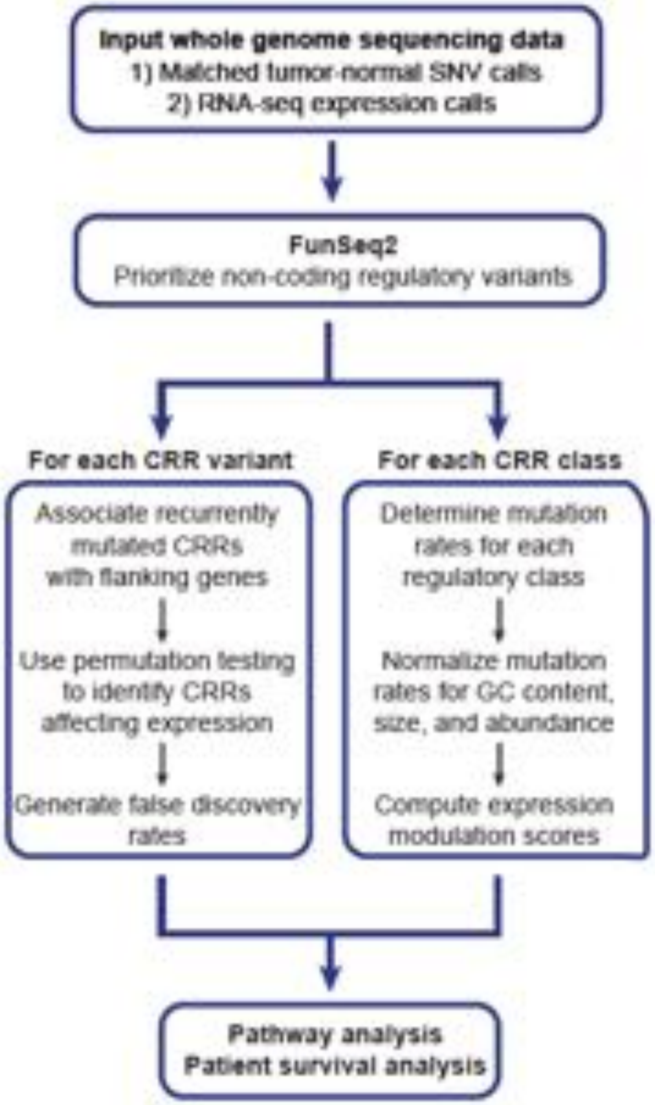
2. Completeness:

How much of the genome is present?

"88% of GWAS SNPs are intronic or intergenic of unknown function"

ENCODE Consortium (2012) Nature

Non-coding Somatic SNVs in PDAC



Input whole genome sequencing data
1) Matched tumor-normal SNV calls
2) RNA-seq expression calls

FunSeq2
Prioritize non-coding regulatory variants

For each CRR variant

Associate recurrently mutated CRRs with flanking genes

Use permutation testing to identify CRRs affecting expression

Generate false discovery rates

For each CRR class

Determine mutation rates for each regulatory class

Normalize mutation rates for GC content, size, and abundance

Compute expression modulation scores

Pathway analysis
Patient survival analysis

Coding alterations of PDAC are now fairly well established but non-coding mutations (NCMs) largely unexplored

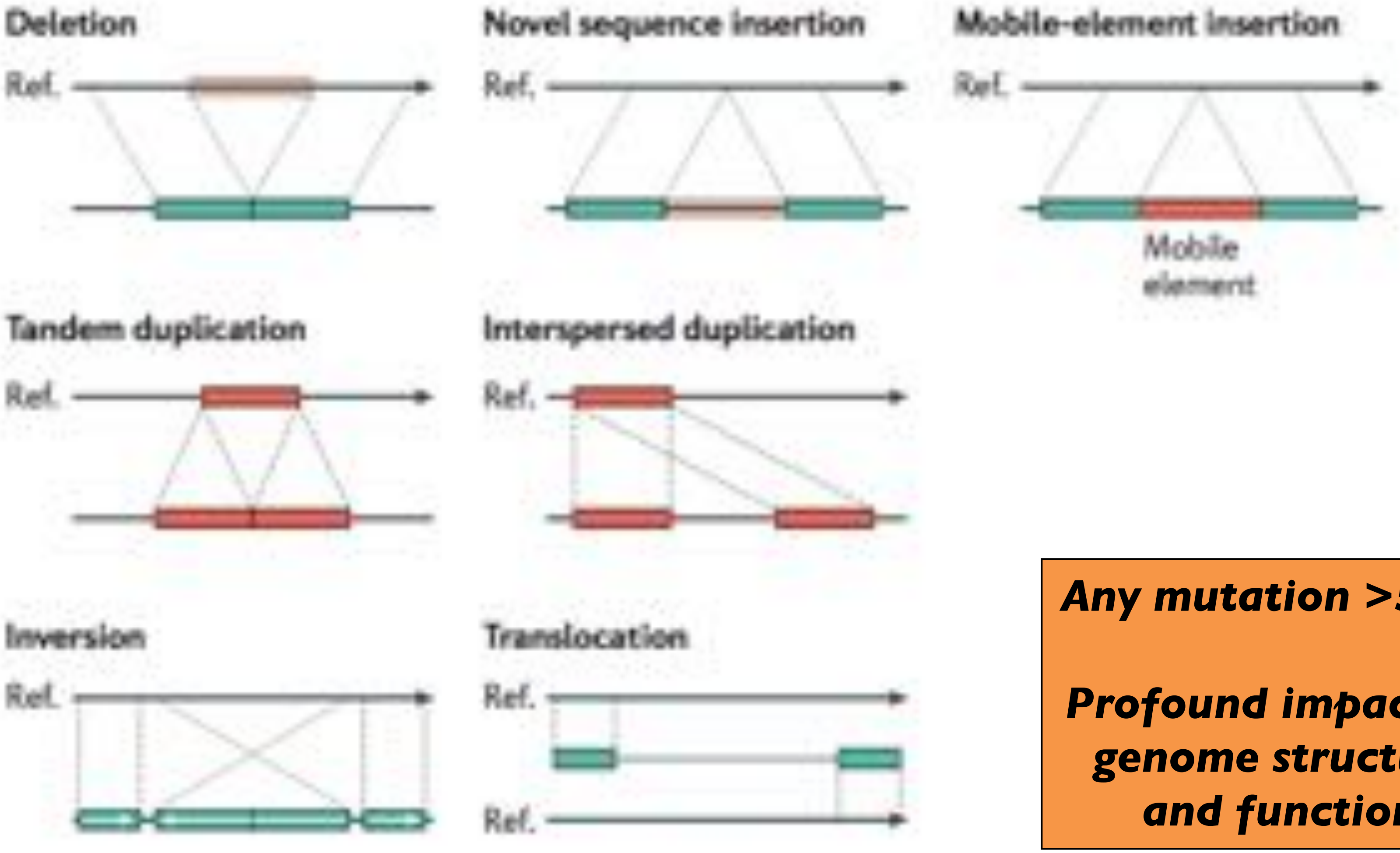
- Developed GECCO to analyze the thousands of somatic mutations observed from hundreds of tumors to find potential drivers of gene expression and pathogenesis

- NCMs are enriched in known and novel pathways
- NCMs correlate with changes in gene expression
- NCMs can demonstrably modulate gene expression
- NCMs correlate with novel clinical outcomes

NCMs are an important mechanism for tumor genome evolution

Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma
Feigin, M, Garvin, T et al. (2017) Nature Genetics. doi:10.1038/ng.3861

Structural Variations



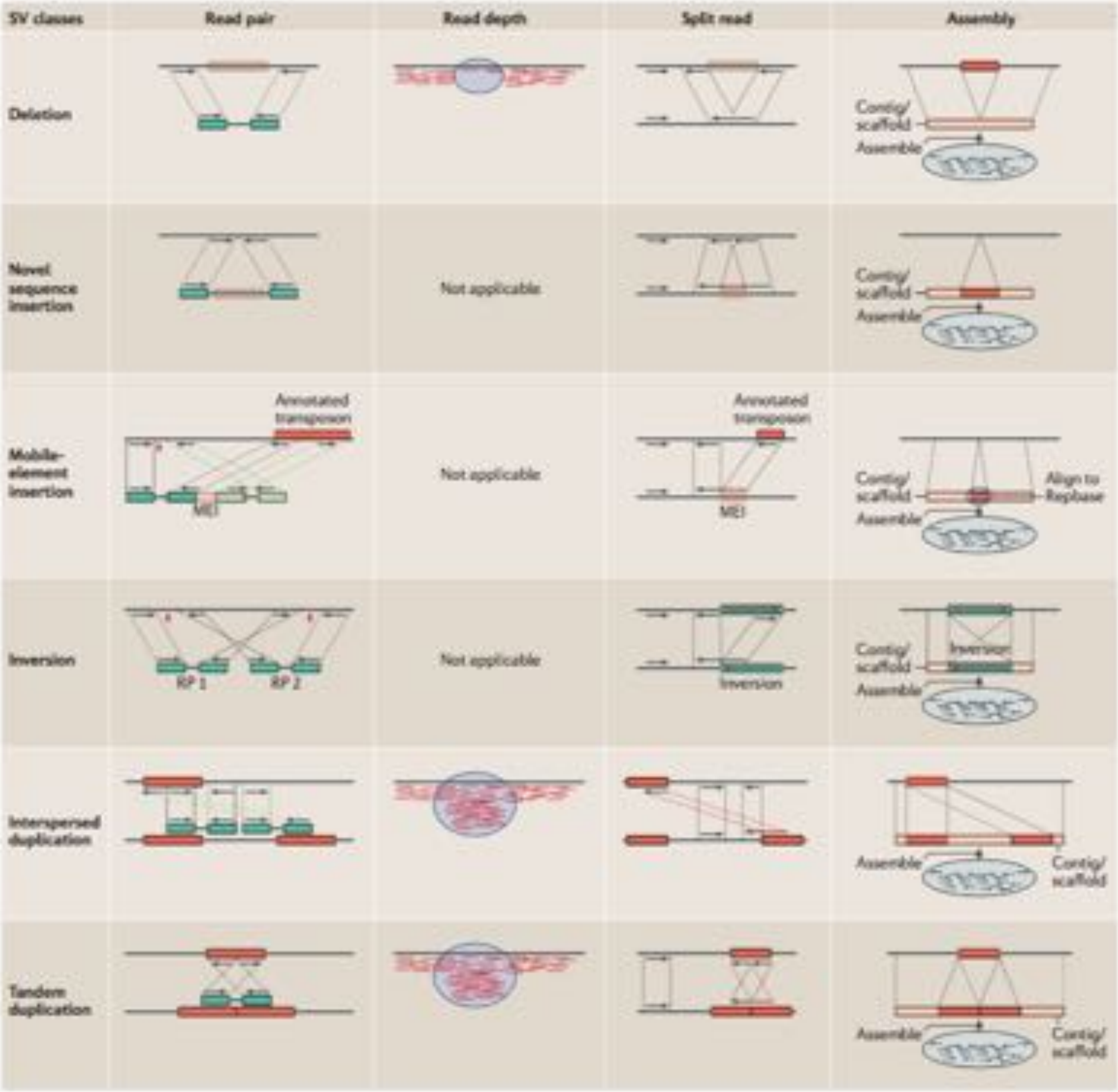
Any mutation >50bp

Profound impact on genome structure and function

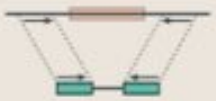
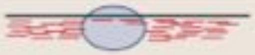
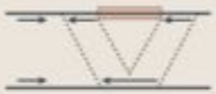
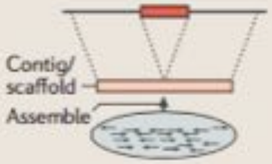
Genome structural variation discovery and genotyping

Alkan, C, Coe, BP, Eichler, EE (2011) *Nature Reviews Genetics*. May;12(5):363-76. doi: 10.1038/nrg2958.

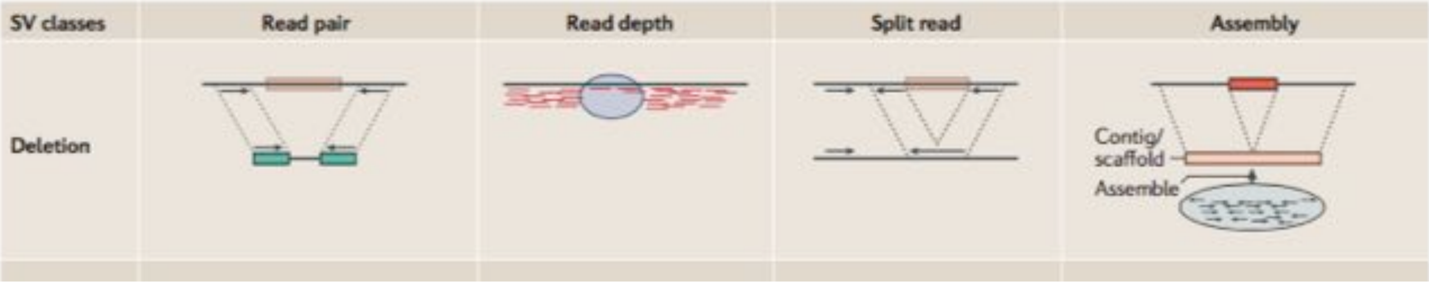
Structural Variation Sequence Signatures



Structural Variation Sequence Signatures

SV classes	Read pair	Read depth	Split read	Assembly
Deletion				

Structural Variation Sequence Signatures



PacBio Sequel

Oxford Nanopore MinION

Long Read Single Molecule Sequencing

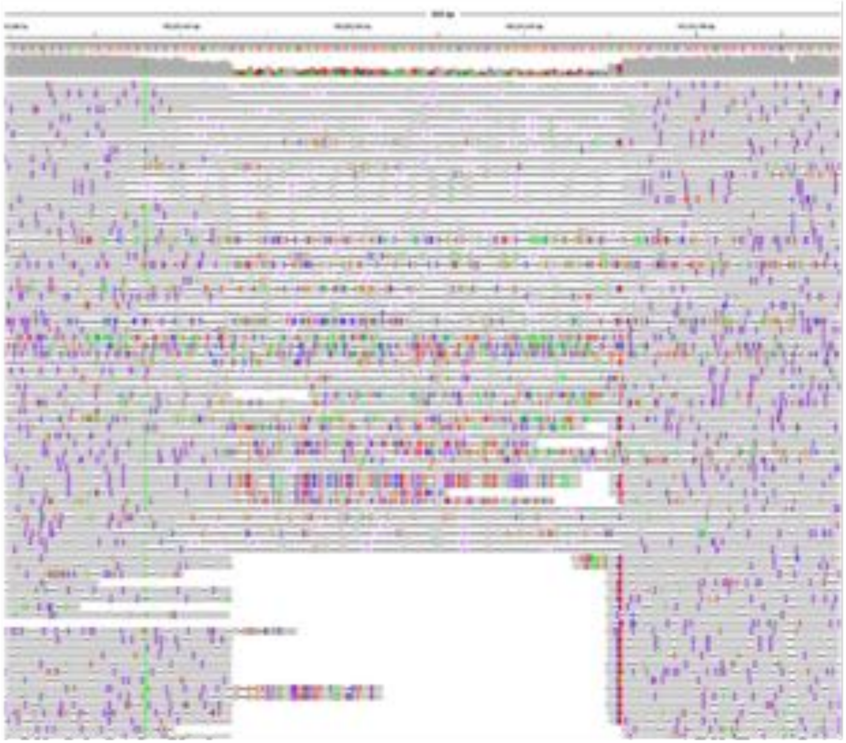
No Amplification Artifacts

Improved Mapping & De novo assemblies

Complete Genomes with all variant types

NGMLR + Sniffles

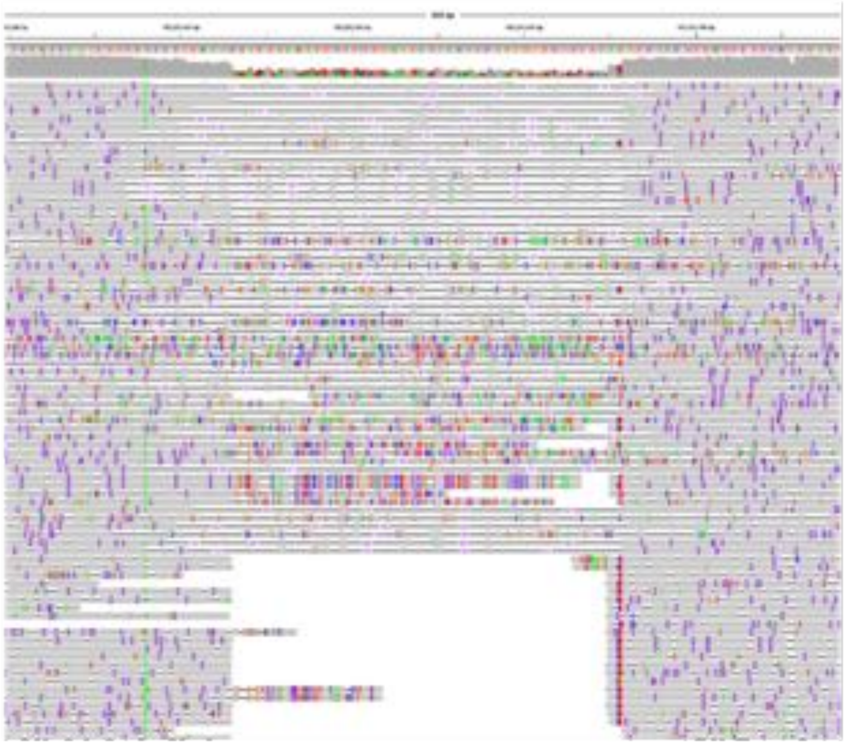
BWA-MEM:



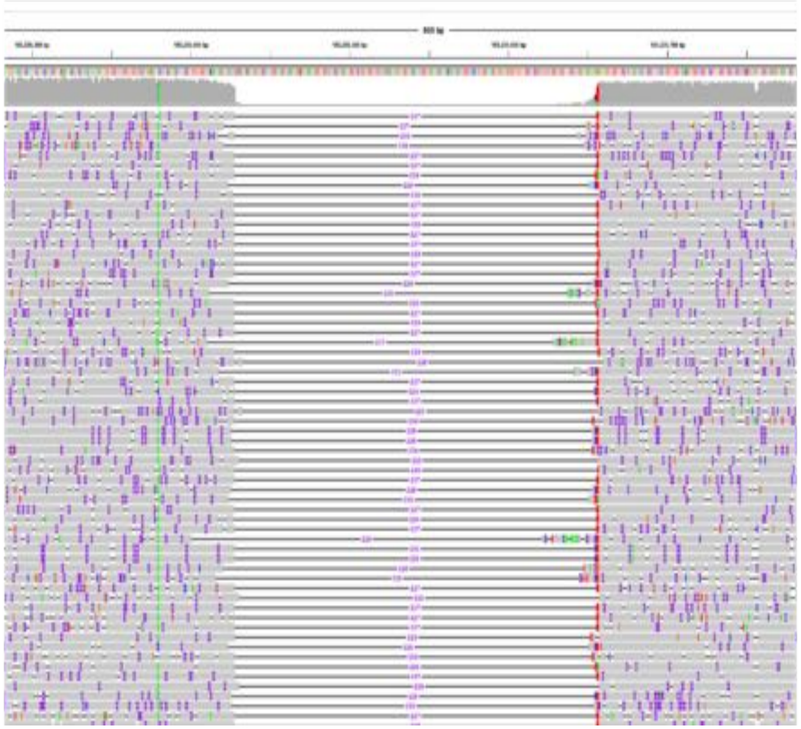
Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *bioRxiv https://doi.org/10.1101/169557*

NGMLR + Sniffles

BWA-MEM:



NGMLR:

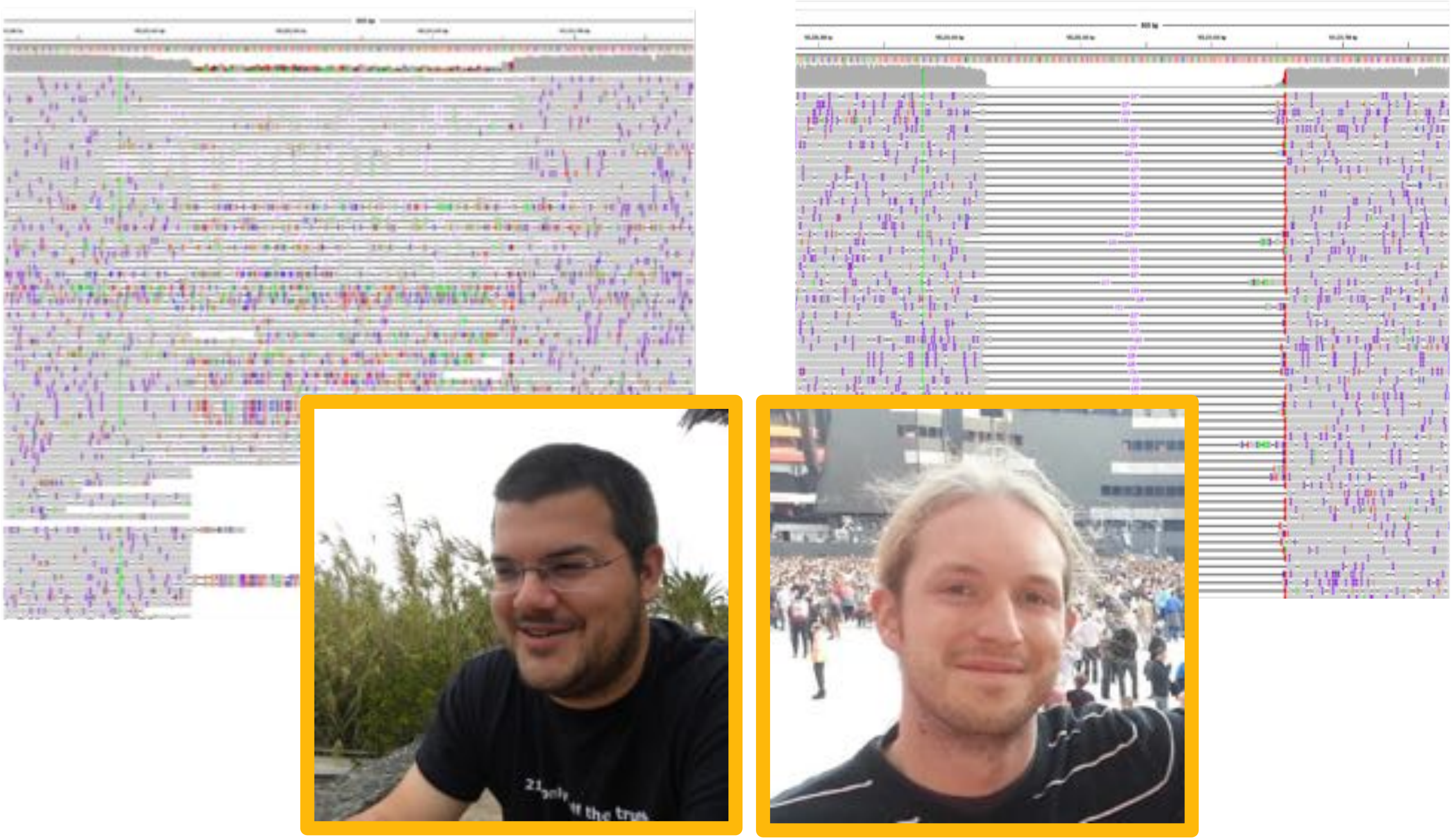


Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *bioRxiv https://doi.org/10.1101/169557*

NGMLR + Sniffles

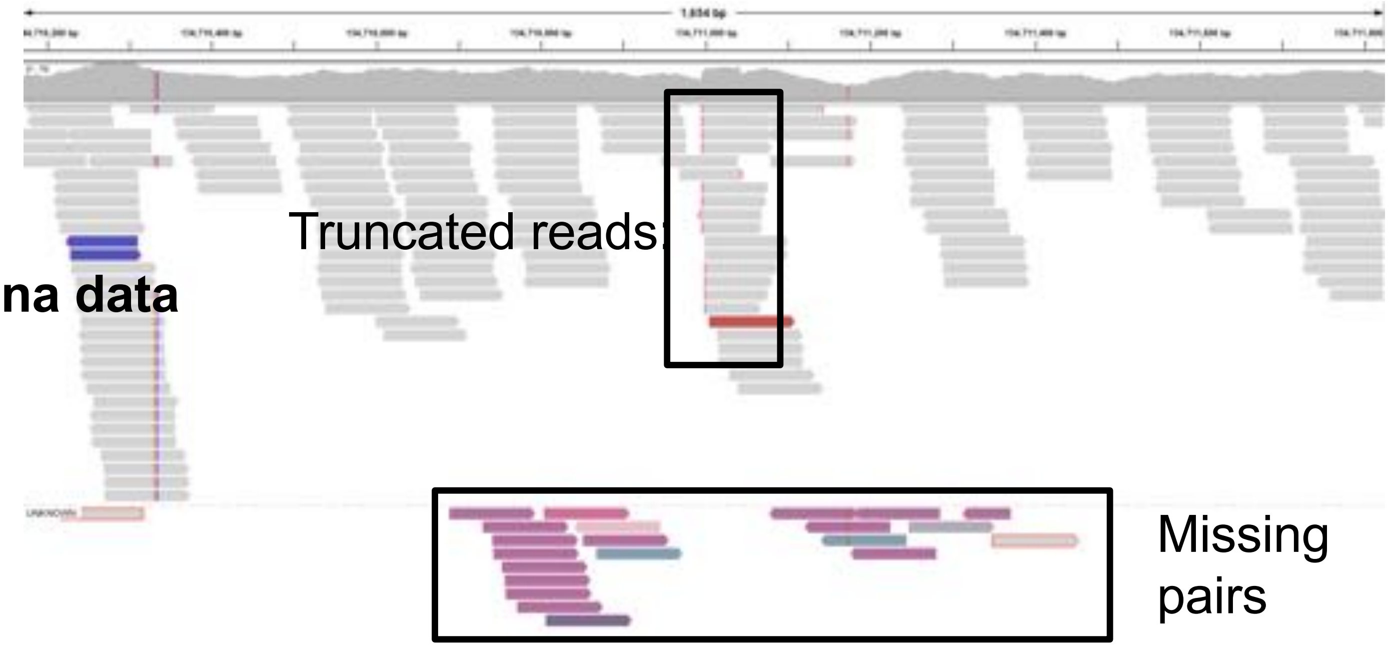
BWA-MEM:

NGMLR:



Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *bioRxiv https://doi.org/10.1101/169557*

No more false positives!



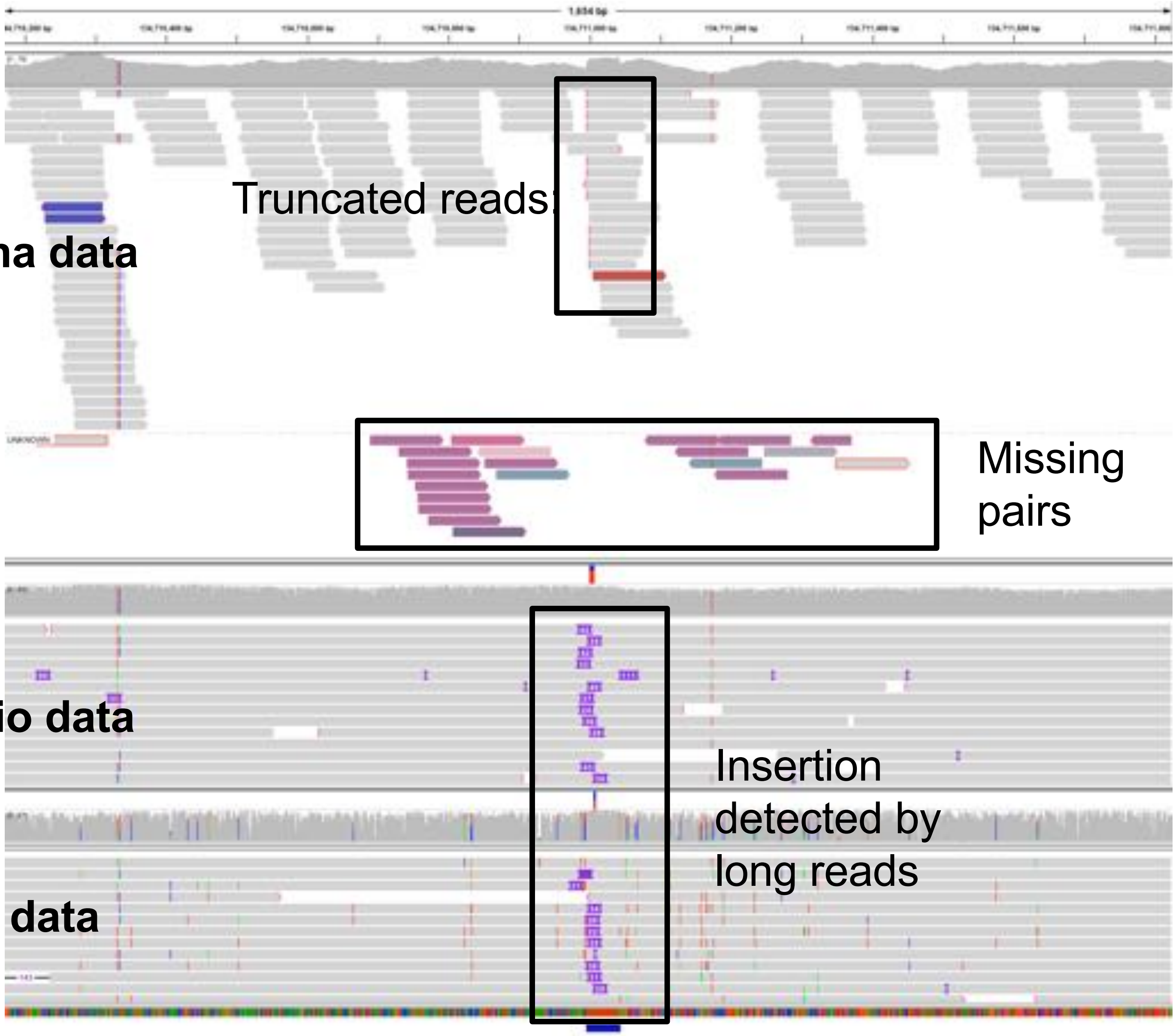
Truncated reads:

Illumina data

Missing pairs

Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *bioRxiv https://doi.org/10.1101/169557*

No more false positives!



Illumina data

Truncated reads:

Missing pairs

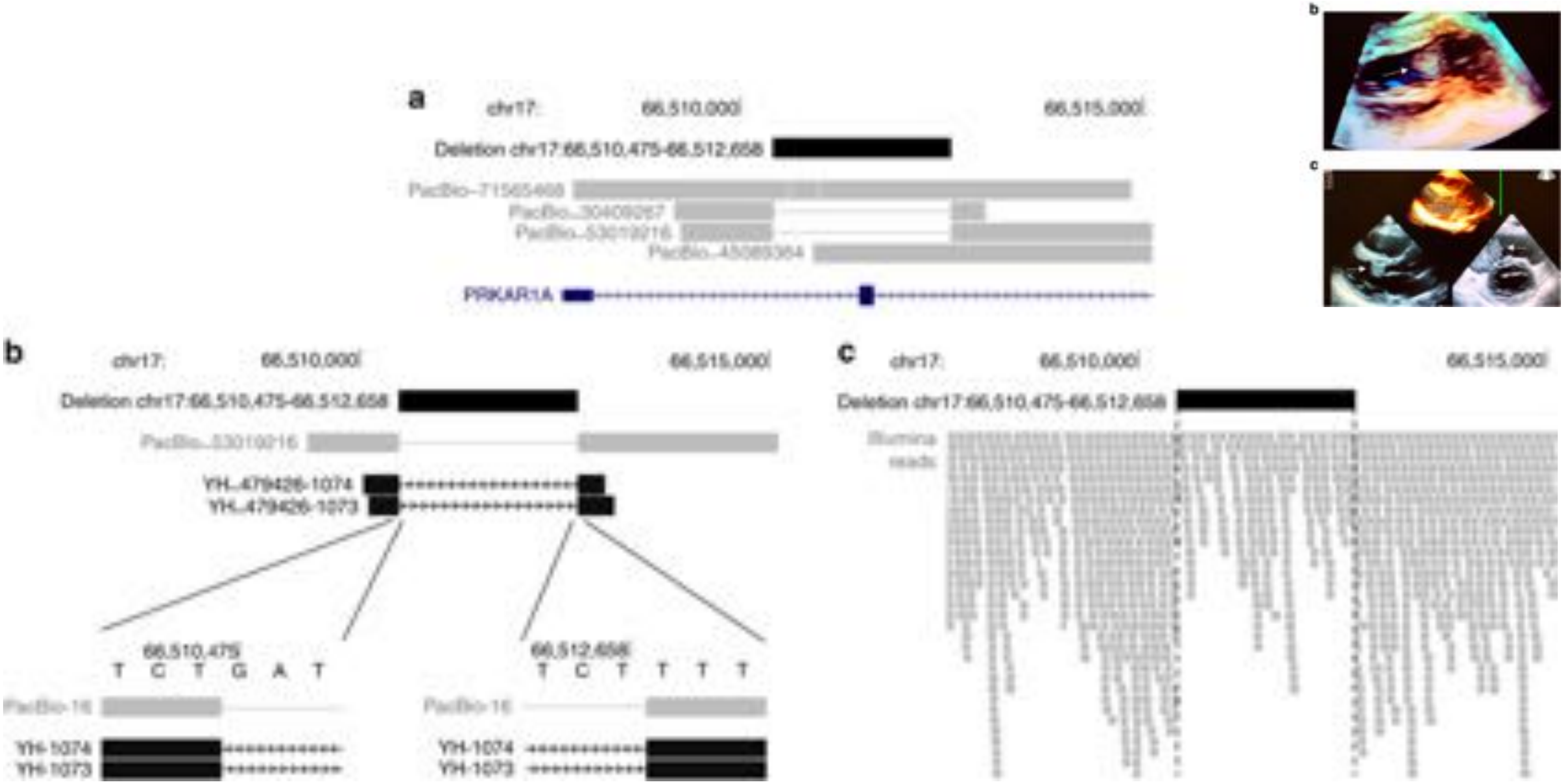
PacBio data

Insertion detected by long reads

ONT data

Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *bioRxiv https://doi.org/10.1101/169557*

Structural Variations in Mendelian Disease



Long-read genome sequencing identifies causal structural variation in a Mendelian disease
Merker et al (2017) *Genetics in Medicine. doi:10.1038/gim.2017.86*

Structural Variations in Breast Cancer

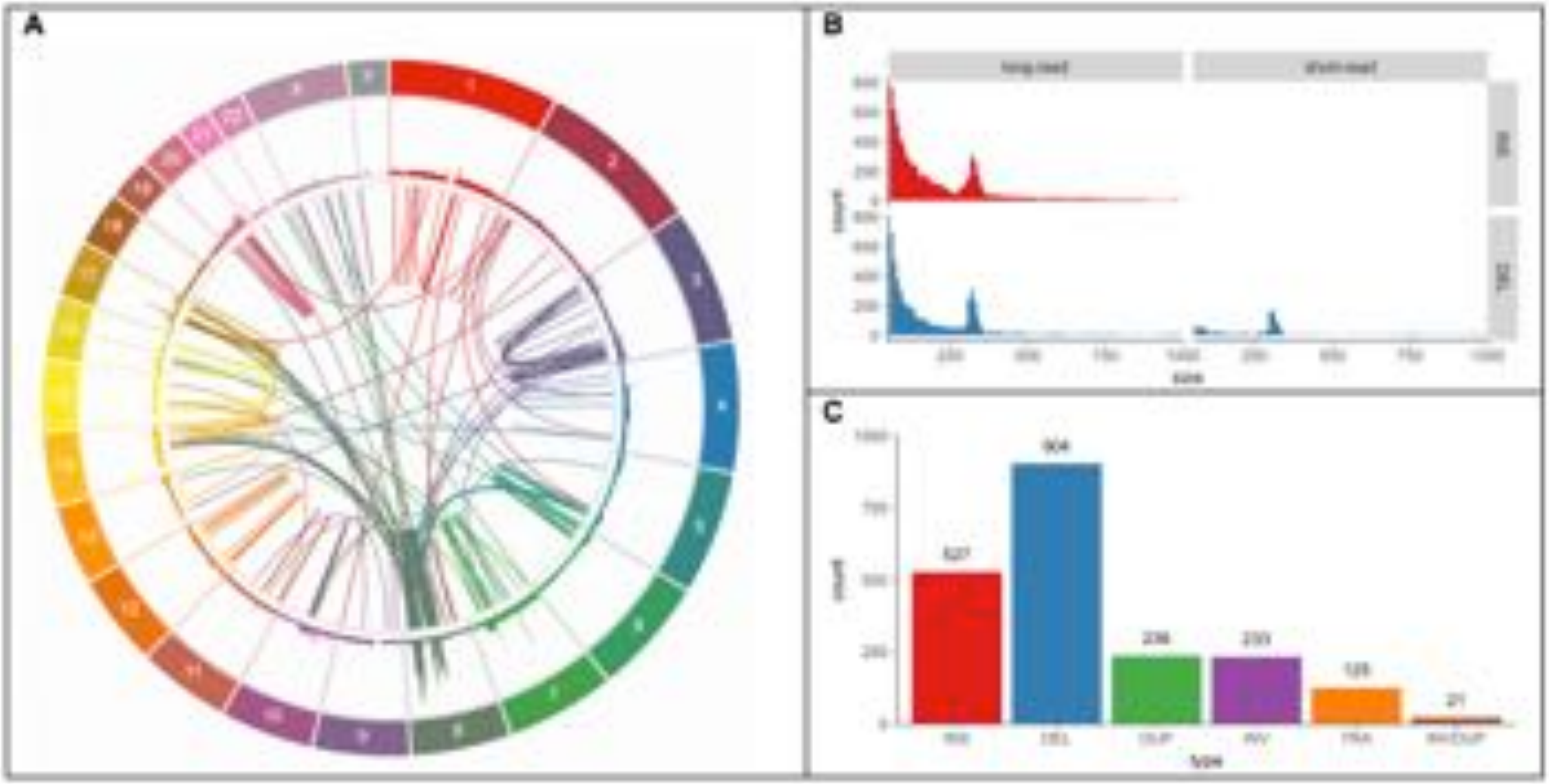


Figure 1 | Variants found in SK-BR-3 with PacBio long-read sequencing. (A) Circos plot showing long-range (larger than 10 kbp or interchromosomal) variants found by Sniffles from split-read alignments, with read coverage shown in the outer track. (B) Variant size histogram of deletions and insertions from size 50 bp up to 1 kbp found by long-read (Sniffles) and short-read (Survivor 2-caller consensus) variant-calling, showing similar size distributions for insertions and deletions from long reads but not for short reads where insertions are entirely missing. (C) Sniffles variant counts by type for variants above 1 kbp in size, including translocations and inverted duplications.

Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line
Nattestad, M et al (2017) bioRxiv https://doi.org/10.1101/174938

In pursuit of perfect genome sequencing

1. Why “Perfect”?

2. What is “Perfect”?

   100% Correct & 100% Complete

3. How will we achieve it?

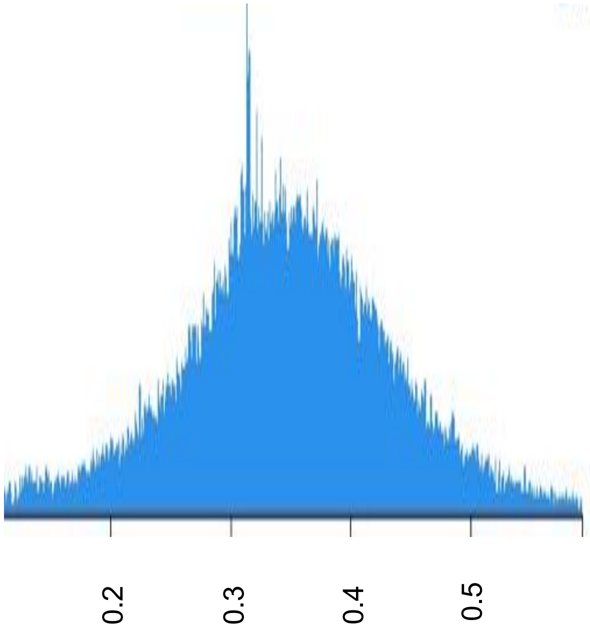
4. When will we achieve it?

In pursuit of perfect genome sequencing

1. Why "Perfect"?
2. What is "Perfect"?
3. **How will we achieve it?**
4. When will we achieve it?

Human Genome Sequencing Data

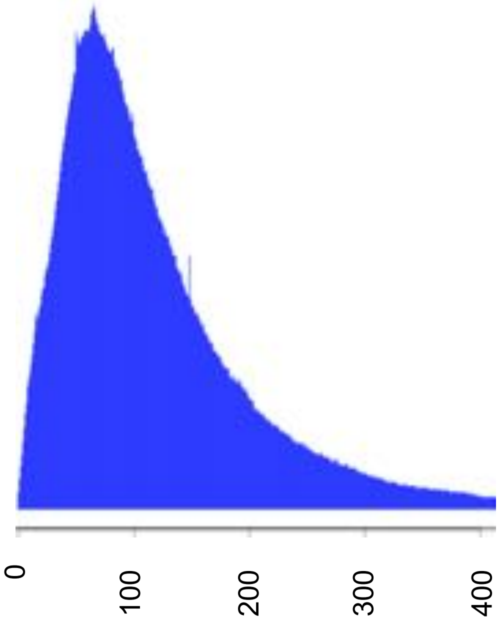
Illumina



Fragment Length (kbp)

60x Paired End
μ=350bp

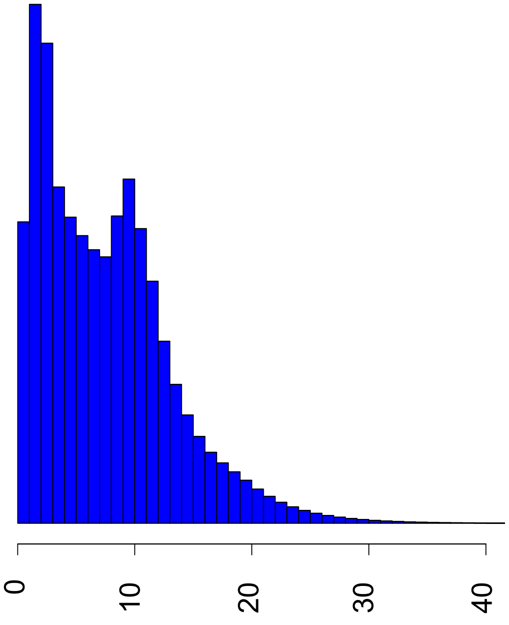
10X Genomics



Molecule Length (kbp)

35x Linked Reads
μ=117kbp

PacBio



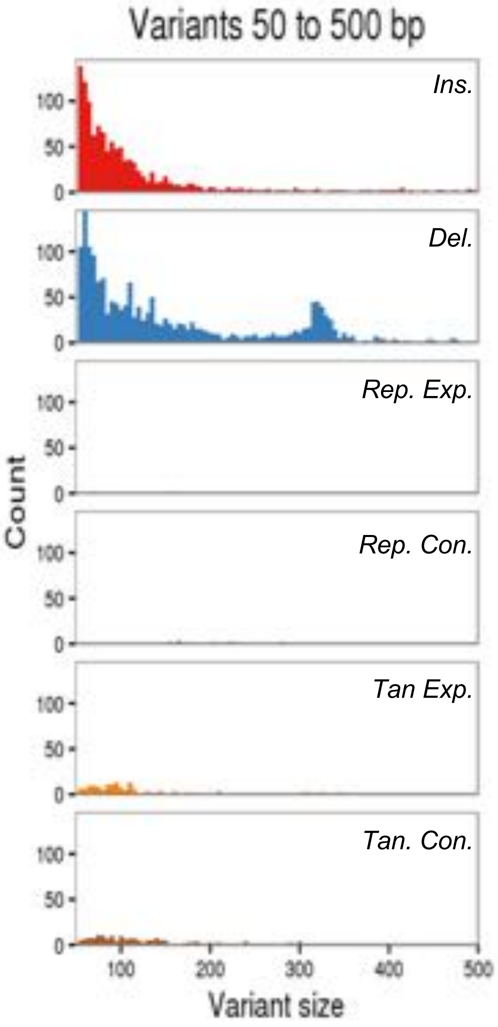
Read Length (kbp)

55x Long Reads
μ=7.5kbp

Missing Insertions from Short and Linked Read?

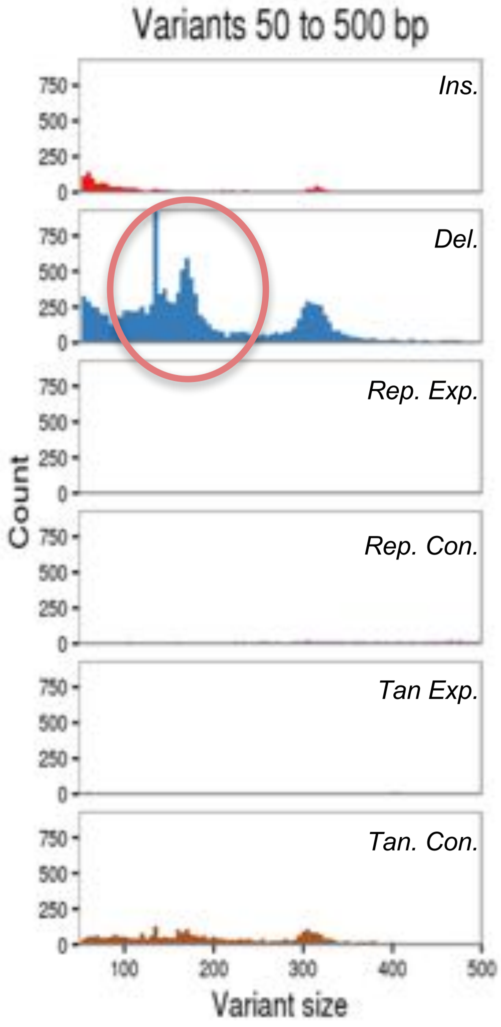
Illumina

13kbp contig N50



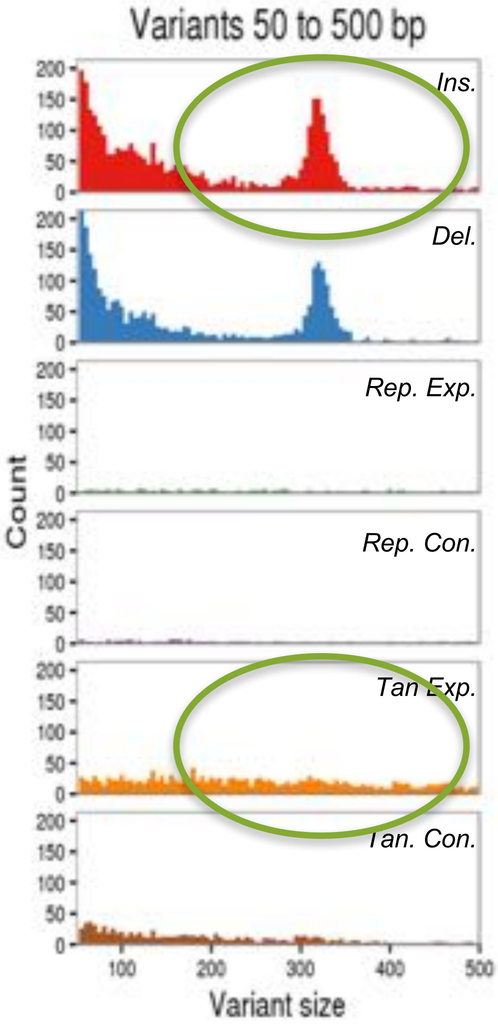
10X Genomics

50kbp contig N50



PacBio

7Mbp contig N50



Structural Variations Concordance

	Sniffles	Falcon	LongRanger	SuperNova	SURVIVOR2	MegaHit
Sniffles	17,139					
Falcon	7,857	12,241				
LongRanger	2,823	1,946	3,785			
SuperNova	3,394	2,837	1,486	18,862		
SURVIVOR2	3,291	2,163	2,274	1,646	6,631	
MegaHit	1,858	1,529	569	1,378	687	3,855

PacBio

10X Genomics

Illumina

Main Diagonal

- Calls per tool

Outer triplets

- Concordance by Technology

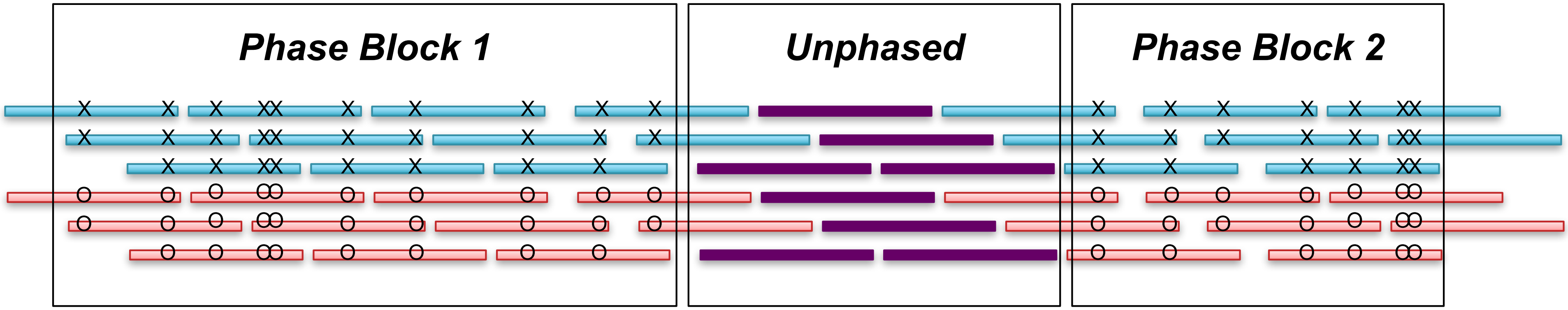
Inner triplets

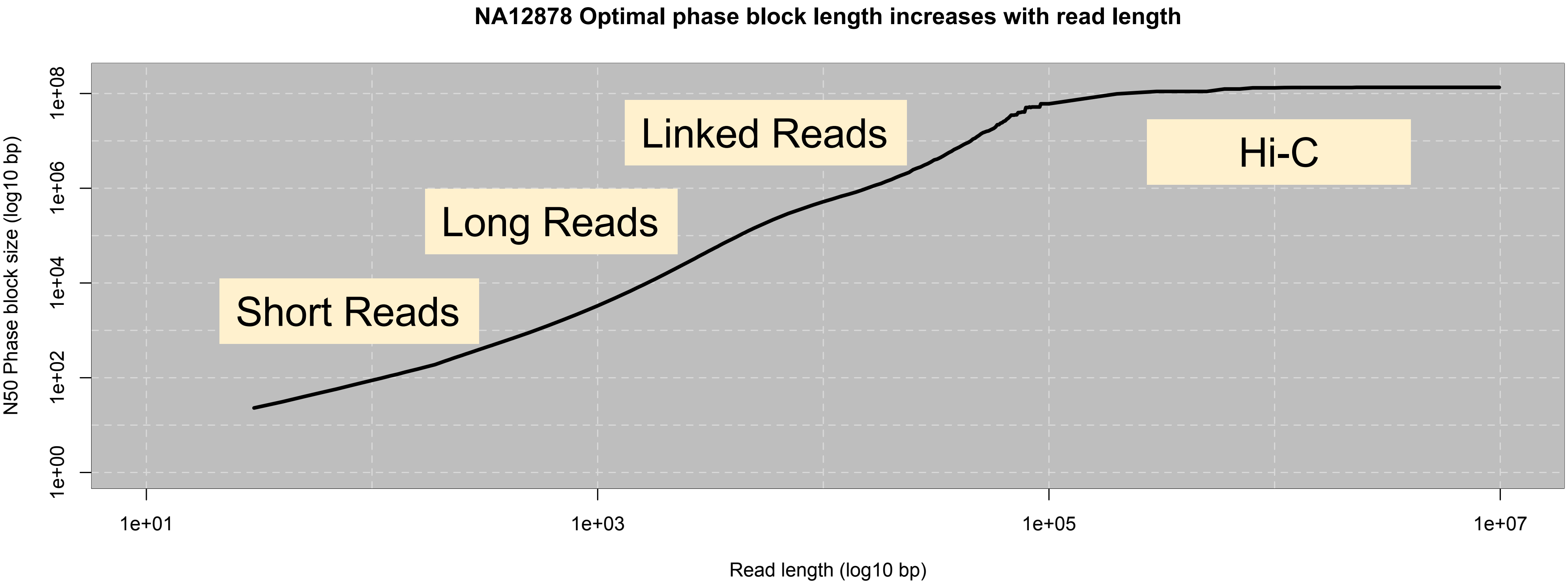
- Concordance by Assembly
- Concordance by Mappers

Overall:

- Lonnnnnnng reads give the most variants with the best concordance ☺

Phasing Results

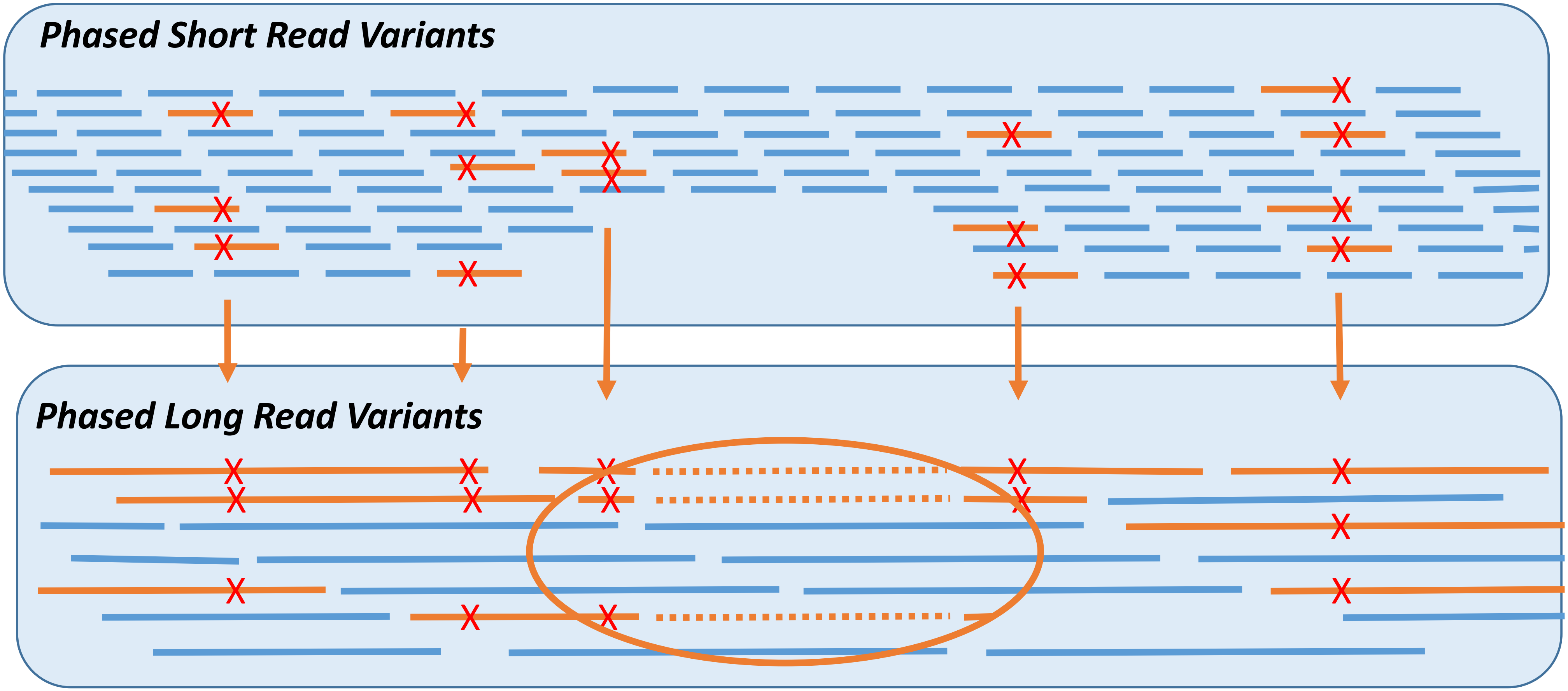




Piercing the dark matter: Bioinformatics for third generation sequencing
Sedlazeck *et al* (2017) *Under Review*

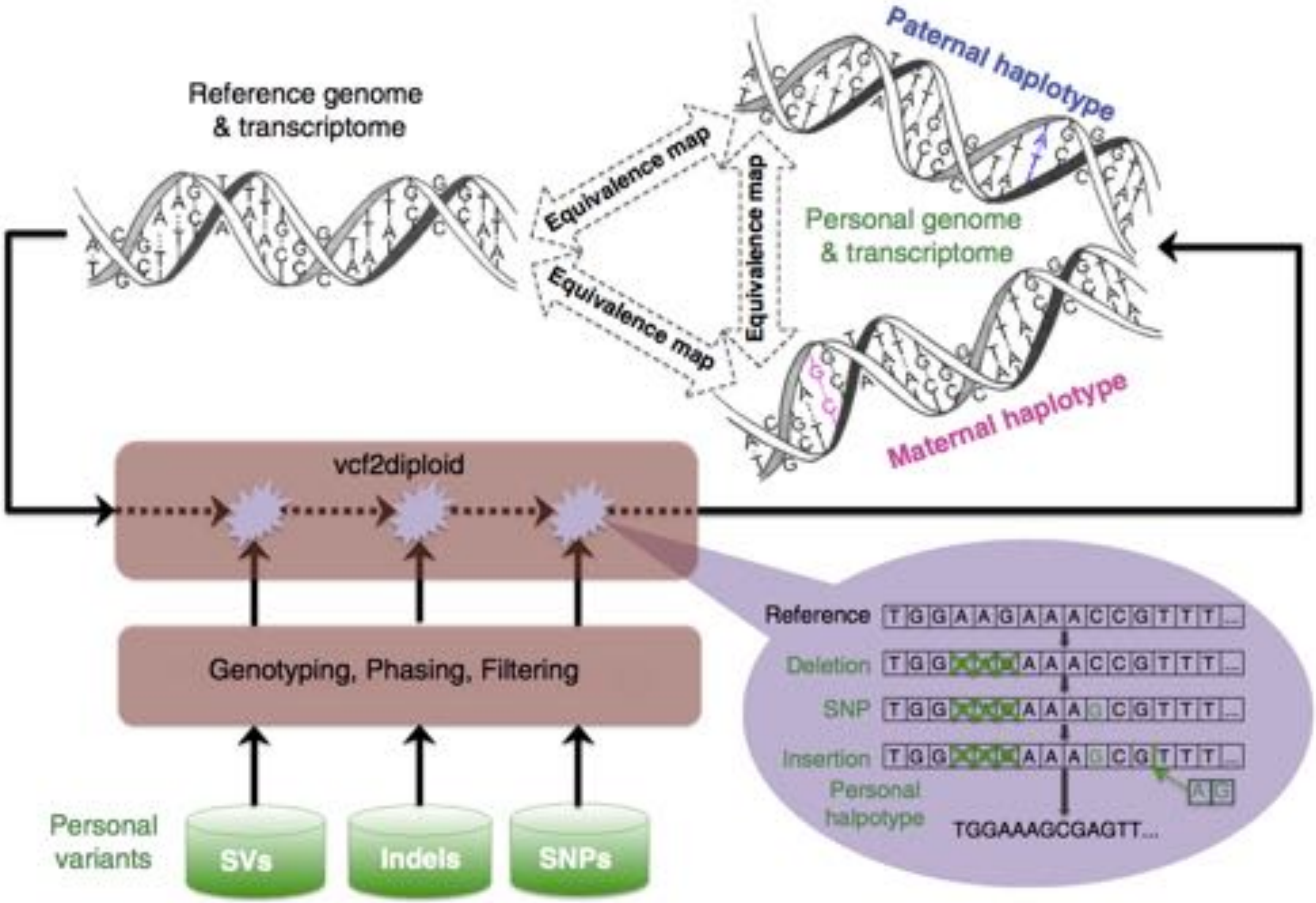
Hybrid Phasing of Structural Variations

Use the phased short read variants to phase the long reads
The phased long reads allow the SVs to be phased



Deletion must be on the orange haplotype!

Creating a "Perfect" Phased Diploid Genome



(J Rozowsky et al, 2011)

vcf2diploid inserts phased variants from a VCF file into the reference genome to create a pair of phased chromosome fasta files

In pursuit of perfect genome sequencing

1. Why "Perfect"?

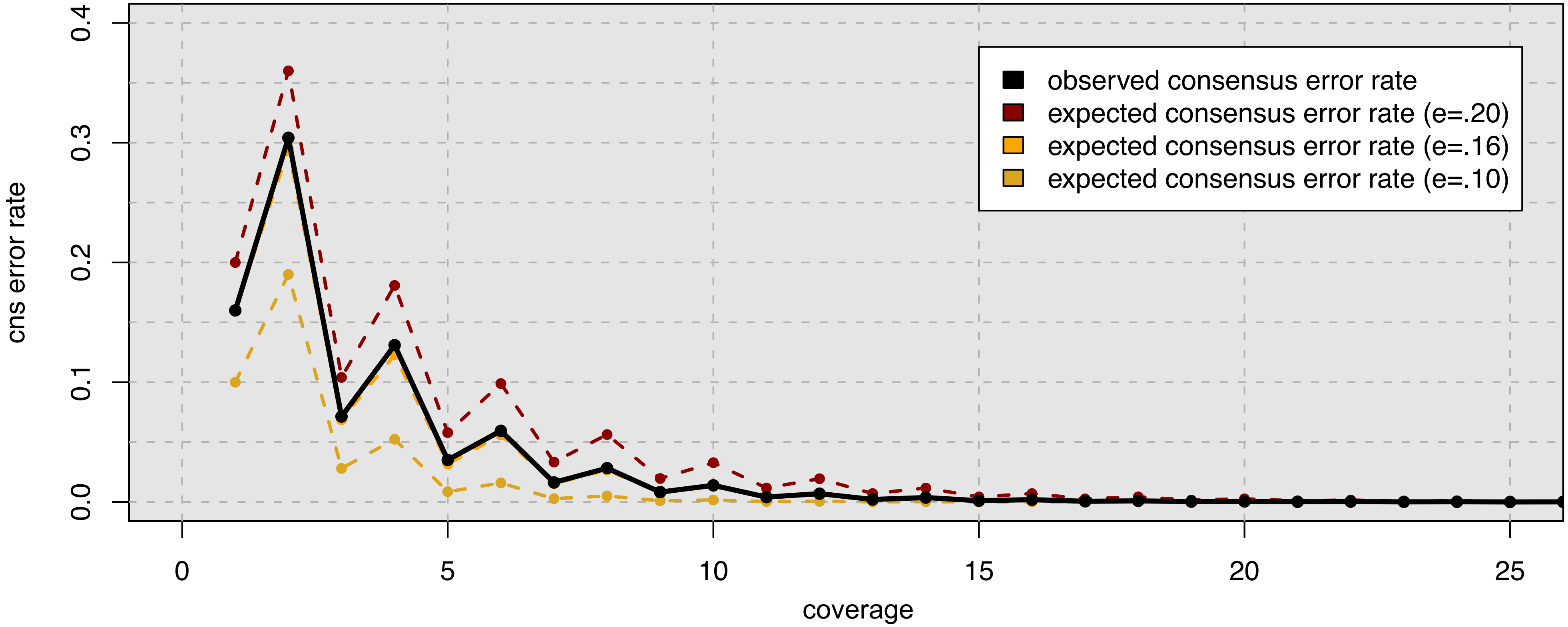
2. What is "Perfect"?

3. **How will we achieve it?**
   Lonnnnnng reads + Loooooong mates :-)
4. When will we achieve it?

In pursuit of perfect genome sequencing

1. Why “Perfect”?
2. What is “Perfect”?
3. How will we achieve it?
4. **When will we achieve it?**

Consensus Accuracy and Coverage



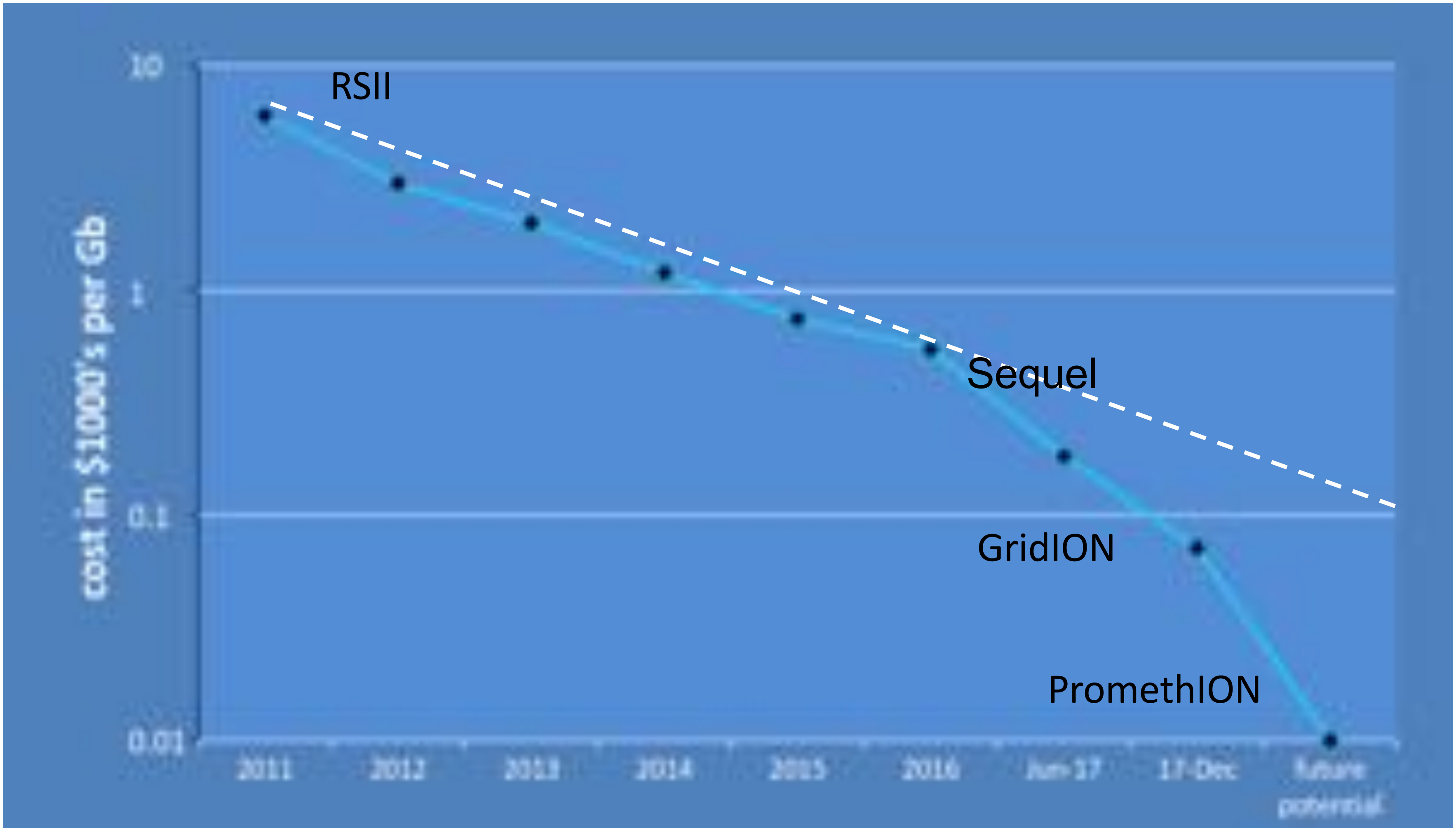
Coverage can overcome random errors

- Dashed: error model from binomial sampling
- Solid: observed accuracy

$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

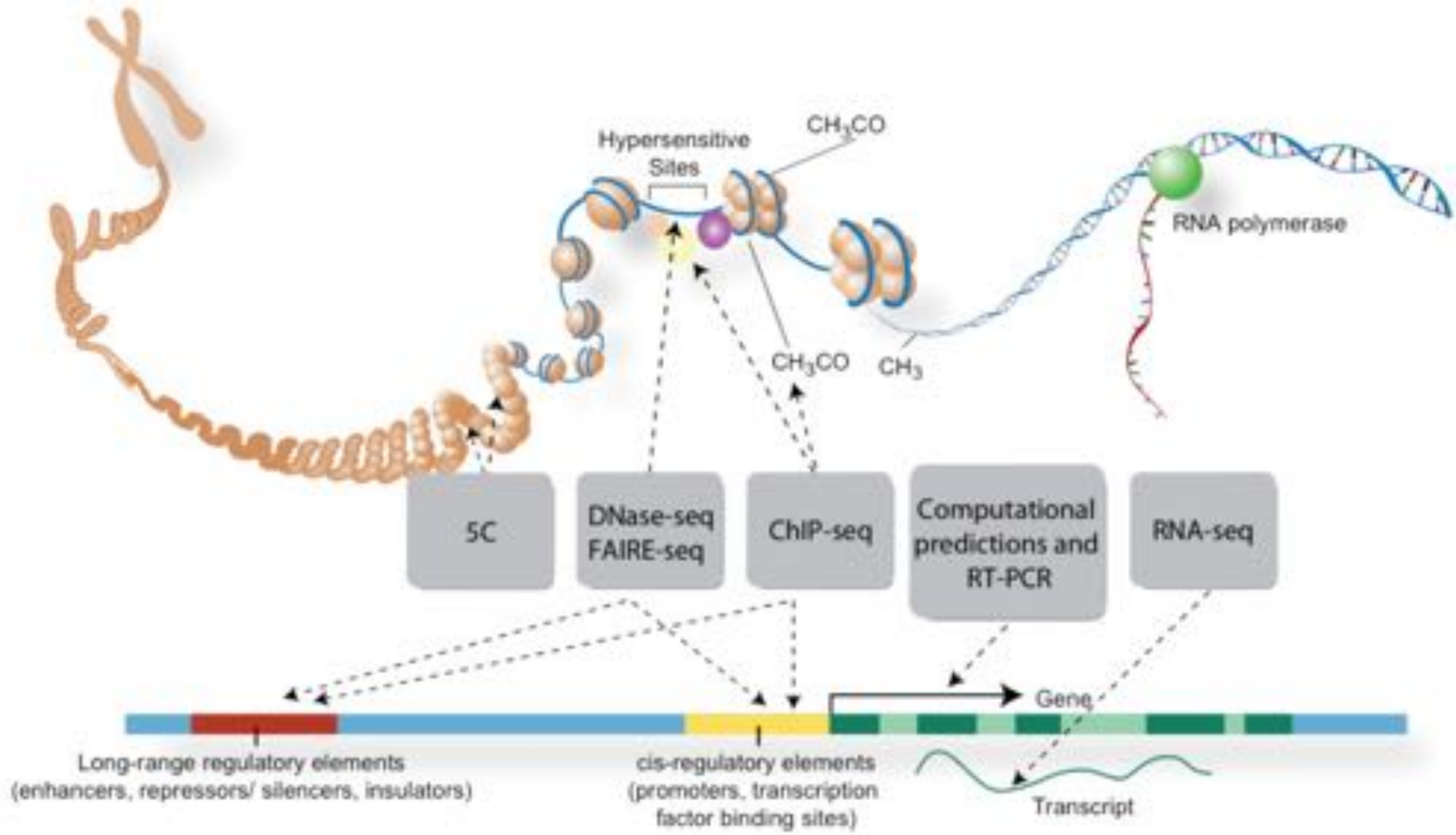
Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren et al (2012) *Nature Biotechnology. doi:10.1038/nbt.2280*

Costs for Long Read Sequencing



Sara Goodwin, CSHL

“Perfect” Genome Projects


Hypersensitive Sites
CH$_3$CO
RNA polymerase
CH$_3$CO
CH$_3$
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

ENCODE + CancerCODE

Illumina + PacBio/ONT + 10X
RNA-seq, ChipSeq, Hi-C, etc
4 healthy + 10 Organoids

MaizeCode

Illumina + PacBio/ONT + 10X
RNA-seq, ChipSeq, MNase-seq
2 maize + 2 teosinte

Tomato Diversity

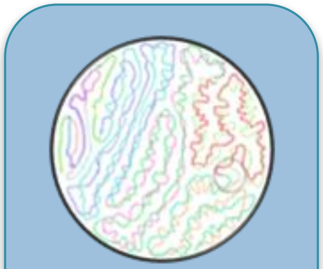
PacBio/ONT + 10x
RNA-seq
50 accessions

In pursuit of perfect genome sequencing

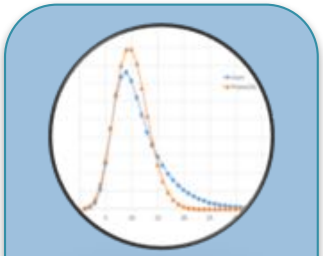
- ***Strive for Perfection: 100% Correct and 100% Complete***
  - The key for perfect genomes is lonnnnnnnnng reads ☺
  - Expect new insights on the causes of diseases, forces of evolution

- ***Multiple sequencing technologies & approaches needed***
  - *PacBio*: Best Resolution of SVs
  - *10X/HIC*: Best Phasing
  - *De novo*: Best Resolution of small SVs
  - *Mapping*: Best resolution of large SVs

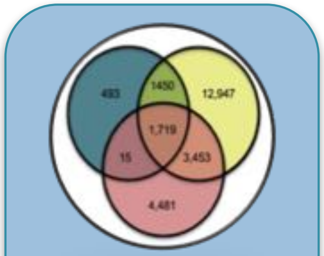
- ***We have just begun to explore the universe of variants present***
  - Tens of thousands of SVs per person, many megabases of variation
  - Also need to push these ideas into single cell and population scale analysis



FALCON

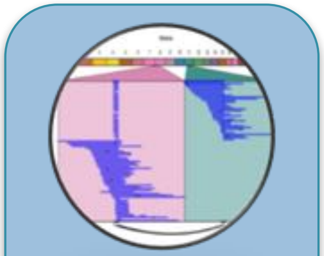


LRSim

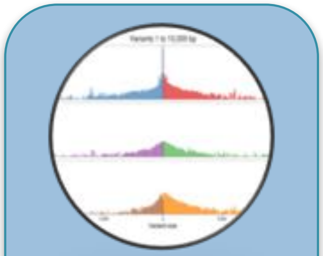


SURVIVOR

NGM+Sniffles



Ribbon



Assemblytics

http://schatz-lab.org

Acknowledgements

Schatz Lab
Mike Alonge
Amelia Bateman
Charlotte Darby
Han Fang
Tyler Gavin
Michael Kirsche
Sam Kovaka
Laurent Luo
Maria Nattestad
Srividya Ramakrishnan
T. Rhyker Ranallo-Benavide

Baylor Medicine
Fritz Sedlazeck

University of Vienna
Arndt von Haeseler
Philipp Rescheneder

CSHL
Gingeras Lab
Jackson Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab

SBU
Skiena Lab
Patro Lab

GRC
Roderic Guido
Alessandra Breschi
Anna Vlasova

JHU
Battle Lab
Langmead Lab
Leek Lab
Salzberg Lab
Taylor Lab
Timp Lab
Wheelan Lab

Cornell
Susan McCouch
Lyza Maron
Mark Wright

OICR
John McPherson
Karen Ng
Timothy Beck
Yogi Sundaravadanam

NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

ALFRED P. SLOAN FOUNDATION

Thank you!

@mike_schatz