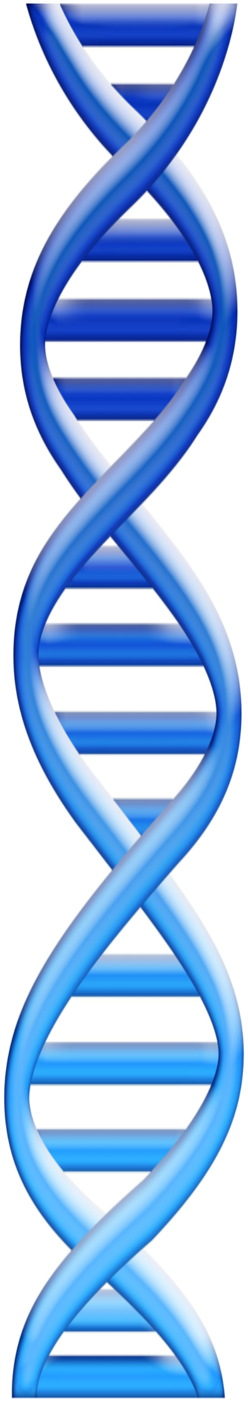
In pursuit of perfect genome sequencing

Michael Schatz

June 28, 2017

PacBio Users Meeting



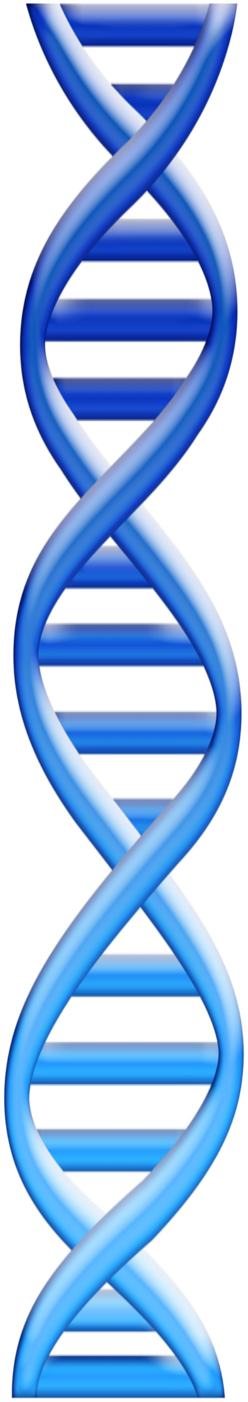
In pursuit of perfect genome sequencing

1. Why “Perfect”?

2. What is “Perfect”?

3. How will we achieve it?

4. When will we achieve it?

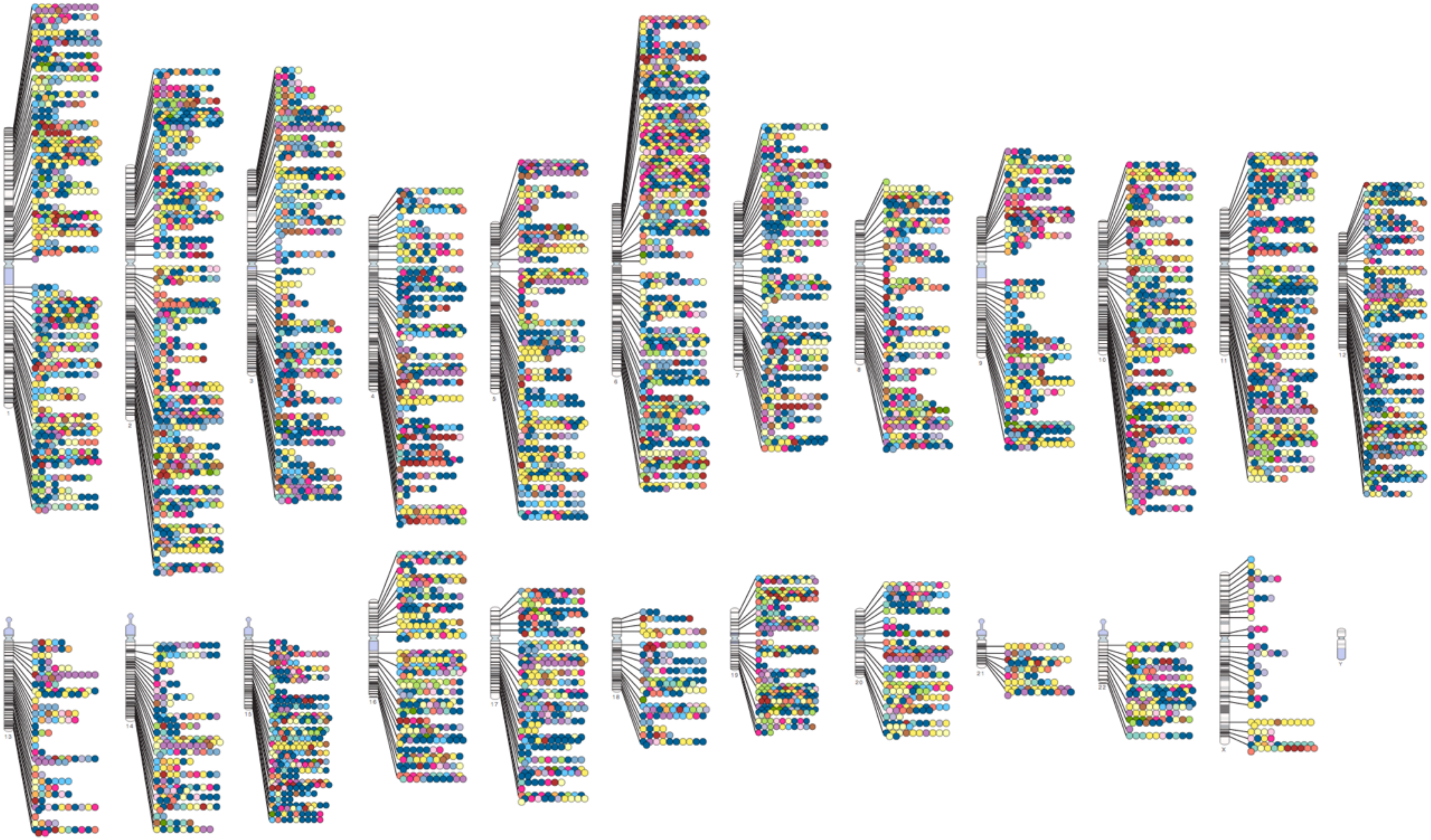


In pursuit of perfect genome sequencing

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

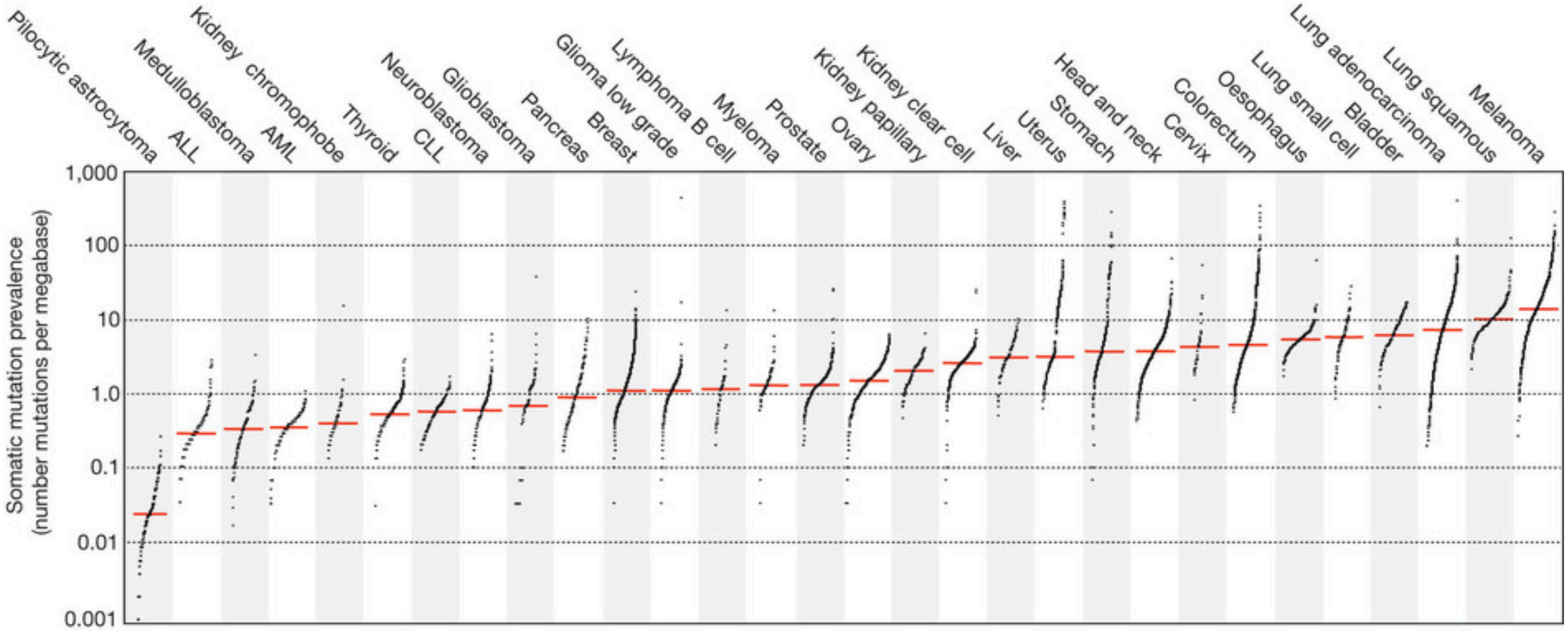
Genetic Origins of Human Diversity

GWAS Catalog contains 33,674 unique SNP-trait associations.
However, most traits remain only partially explained or not at all



http://www.ebi.ac.uk/gwas/diagram

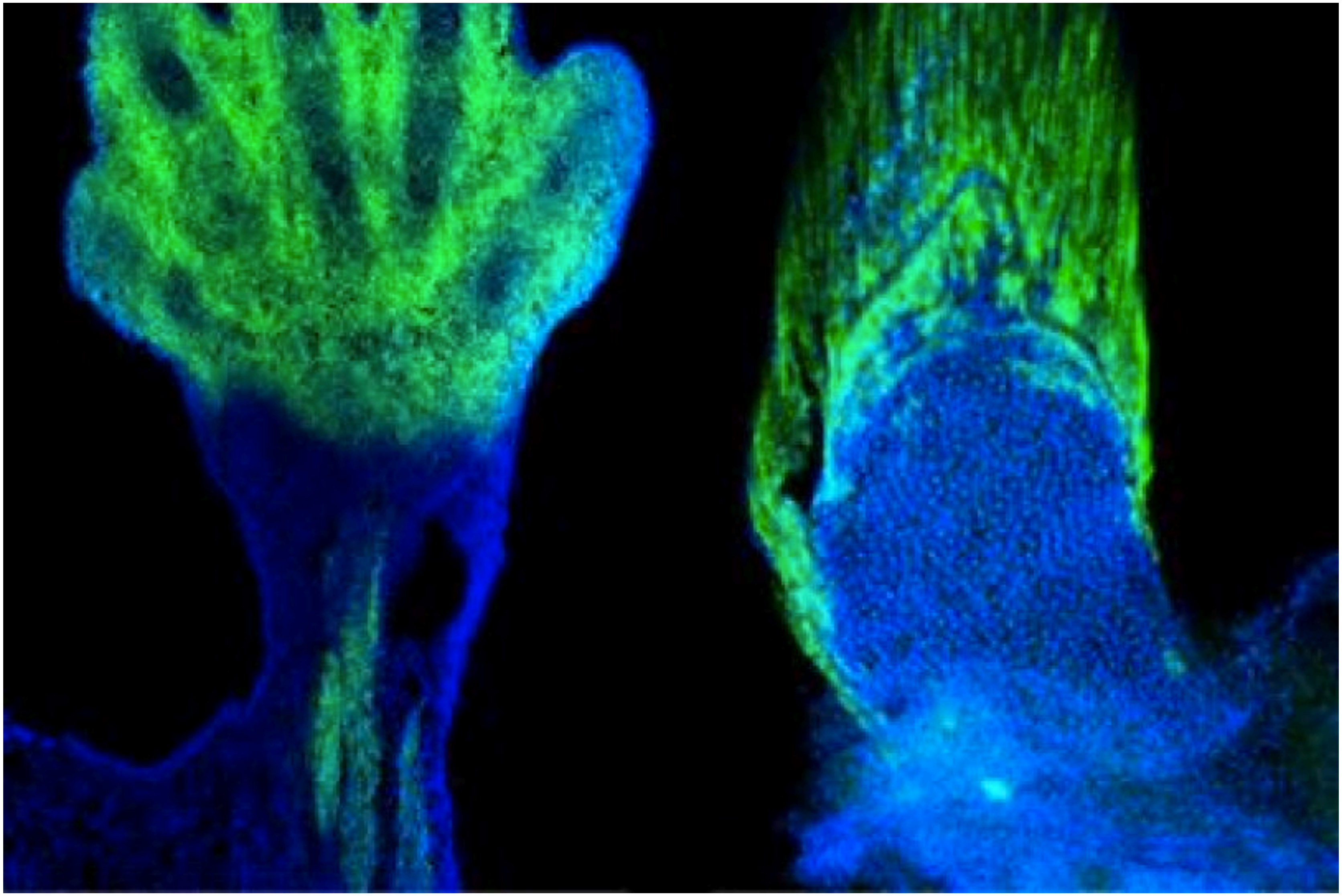
Somatic Mutations In Cancer



Signatures of mutational processes in human cancer

Alexandrov et al (2013) *Nature*. doi:10.1038/nature12477

Mammalian Evolution



Digits and fin rays share common developmental histories
Nakamura et al (2016) *Nature*. 537, 225–228. doi:10.1038/nature19322

“Needles in a stack of needles”

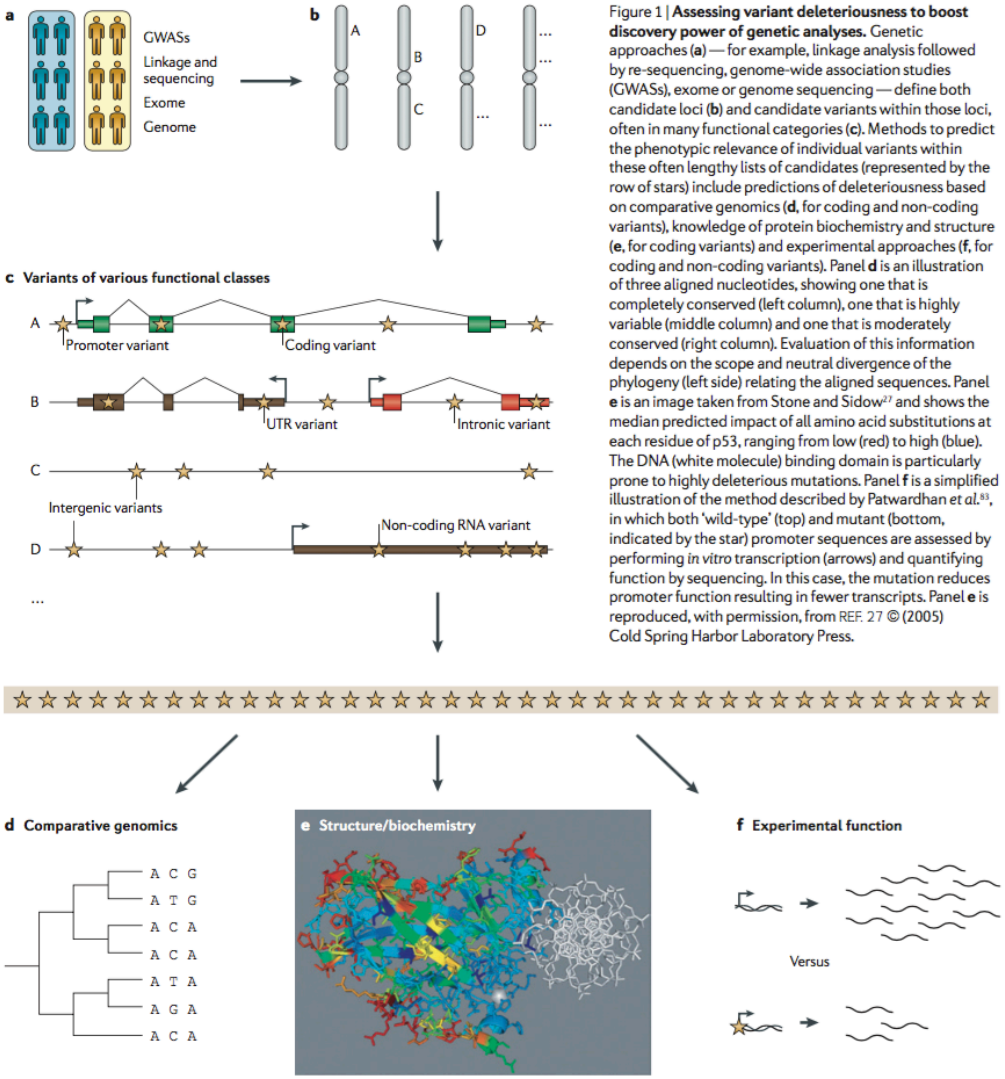
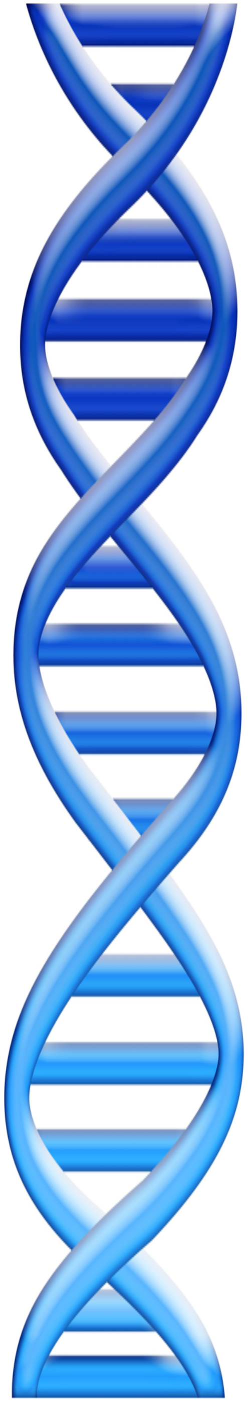


Figure 1 | **Assessing variant deleteriousness to boost discovery power of genetic analyses.** Genetic approaches (**a**) — for example, linkage analysis followed by re-sequencing, genome-wide association studies (GWASs), exome or genome sequencing — define both candidate loci (**b**) and candidate variants within those loci, often in many functional categories (**c**). Methods to predict the phenotypic relevance of individual variants within these often lengthy lists of candidates (represented by the row of stars) include predictions of deleteriousness based on comparative genomics (**d**, for coding and non-coding variants), knowledge of protein biochemistry and structure (**e**, for coding variants) and experimental approaches (**f**, for coding and non-coding variants). Panel **d** is an illustration of three aligned nucleotides, showing one that is completely conserved (left column), one that is highly variable (middle column) and one that is moderately conserved (right column). Evaluation of this information depends on the scope and neutral divergence of the phylogeny (left side) relating the aligned sequences. Panel **e** is an image taken from Stone and Sidow[27] and shows the median predicted impact of all amino acid substitutions at each residue of p53, ranging from low (red) to high (blue). The DNA (white molecule) binding domain is particularly prone to highly deleterious mutations. Panel **f** is a simplified illustration of the method described by Patwardhan *et al.*[83], in which both 'wild-type' (top) and mutant (bottom, indicated by the star) promoter sequences are assessed by performing *in vitro* transcription (arrows) and quantifying function by sequencing. In this case, the mutation reduces promoter function resulting in fewer transcripts. Panel **e** is reproduced, with permission, from REF. 27 © (2005) Cold Spring Harbor Laboratory Press.

Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data
Cooper & Shendure (2011) Nature Reviews Genetics.



In pursuit of perfect genome sequencing

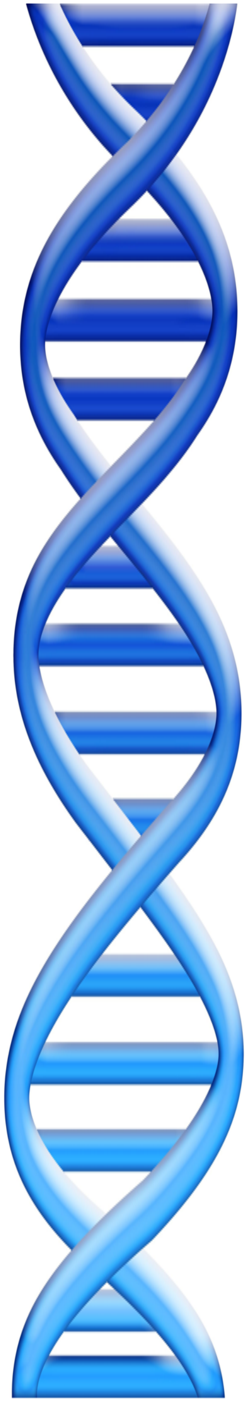
1. **Why "Perfect"?**

   Because it is important, complex, and diffuse

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?

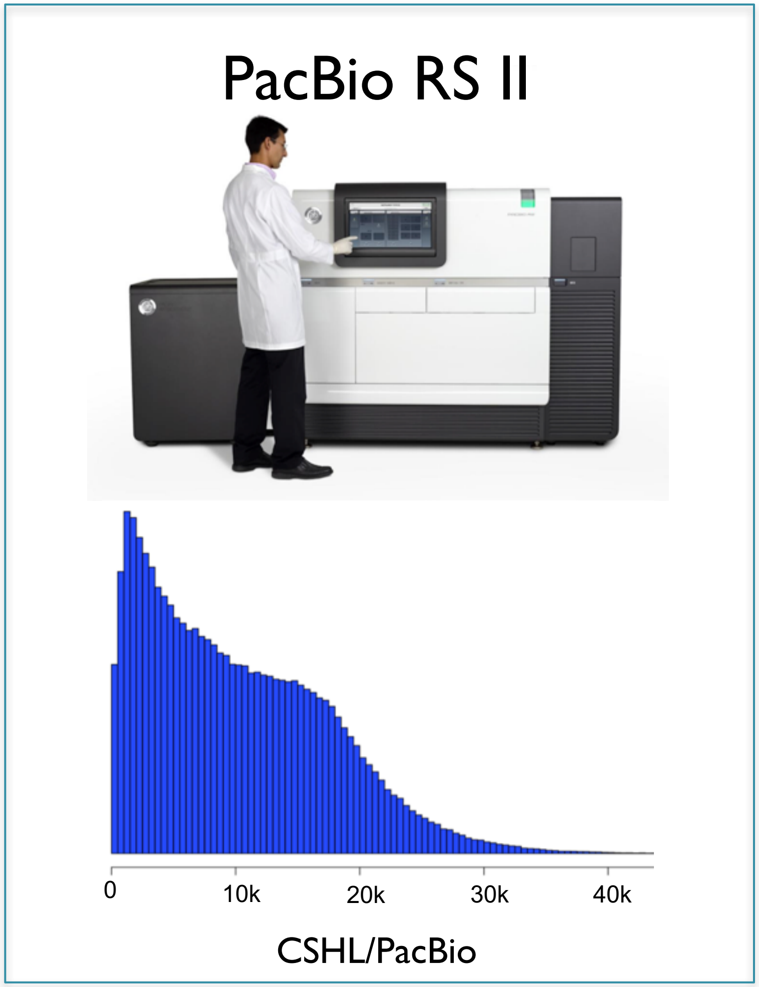
4. When will we achieve it?

1. Correctness:

Is the genome faithfully represented?

1. Correctness:

Is the genome faithfully represented?



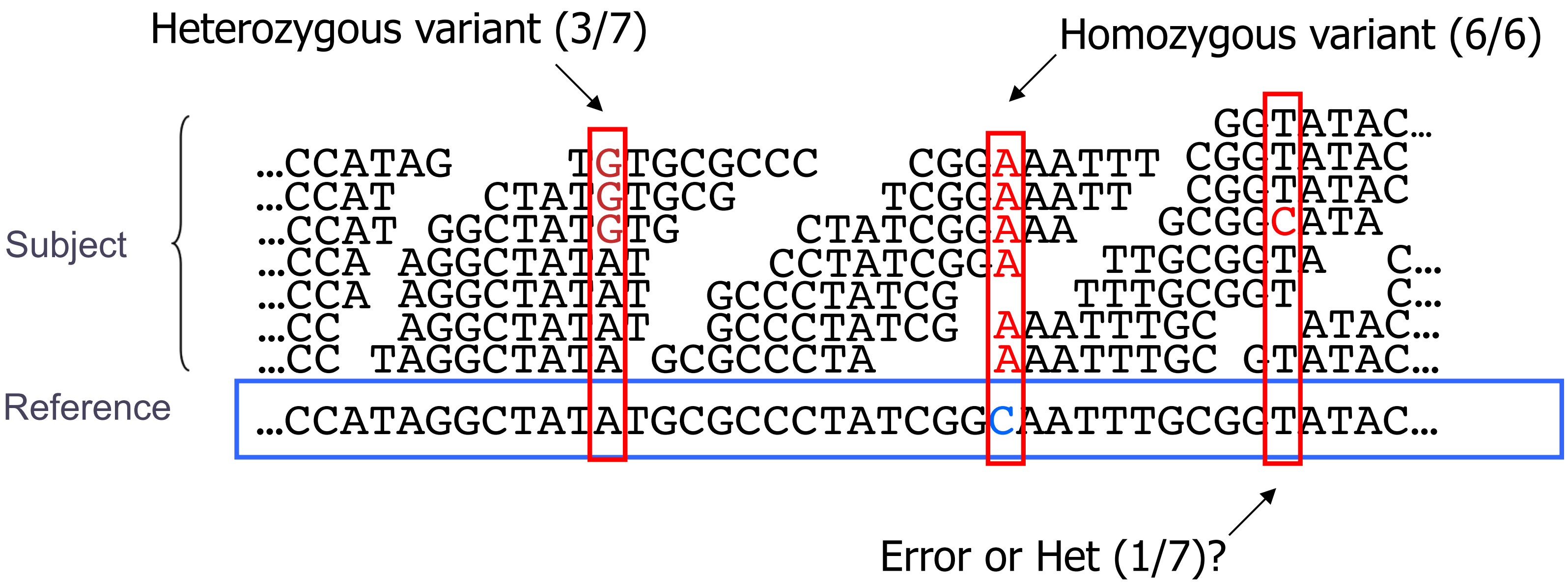
PacBio RS II

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

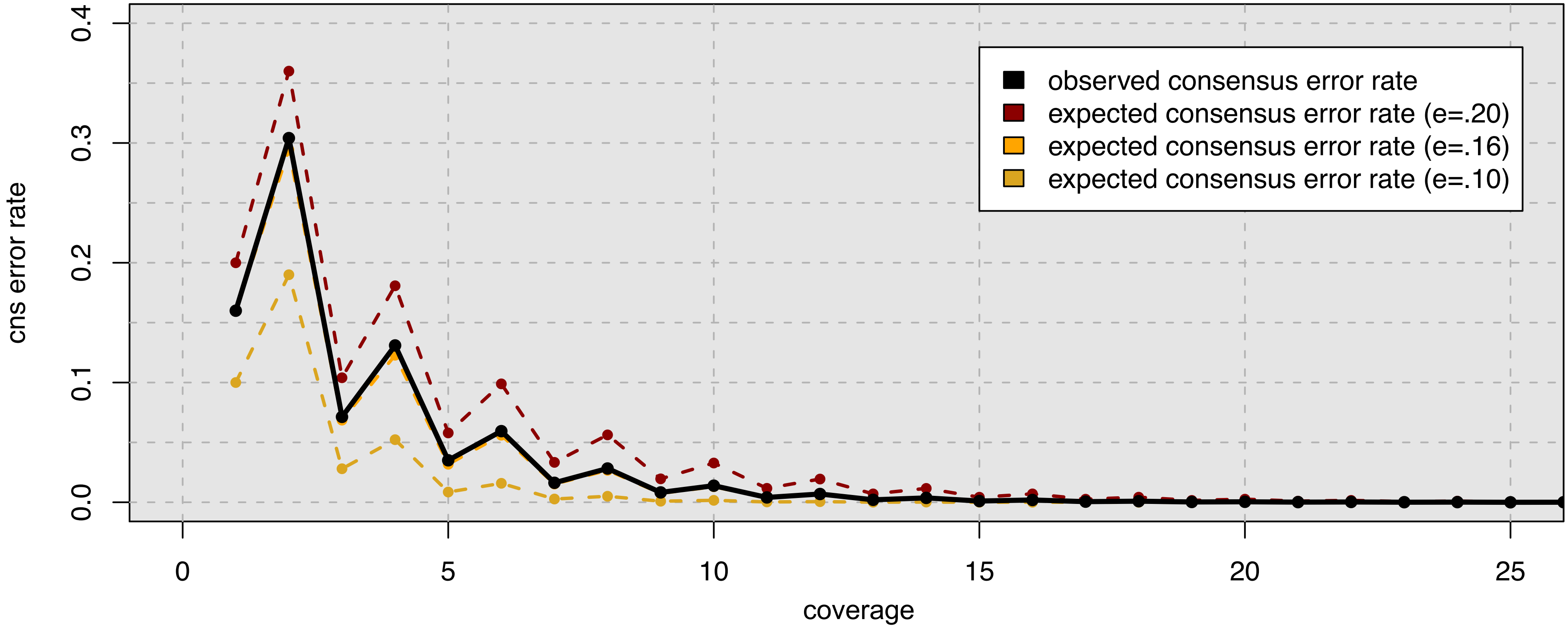


```
                                    GGTATAC...
...CCATAG    TGTGCGCCC   CGGAAATTT CGGTATAC
...CCAT   CTATGTGCG     TCGGAAATT  CGGTATAC
...CCAT GGCTATGTG    CTATCGGAAA   GCGGCATA
...CCA AGGCTATAT    CCTATCGGA   TTGCGGTA  C...
...CCA AGGCTATAT  GCCCTATCG    TTTGCGGT   C...
...CC  AGGCTATAT  GCCCTATCG 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



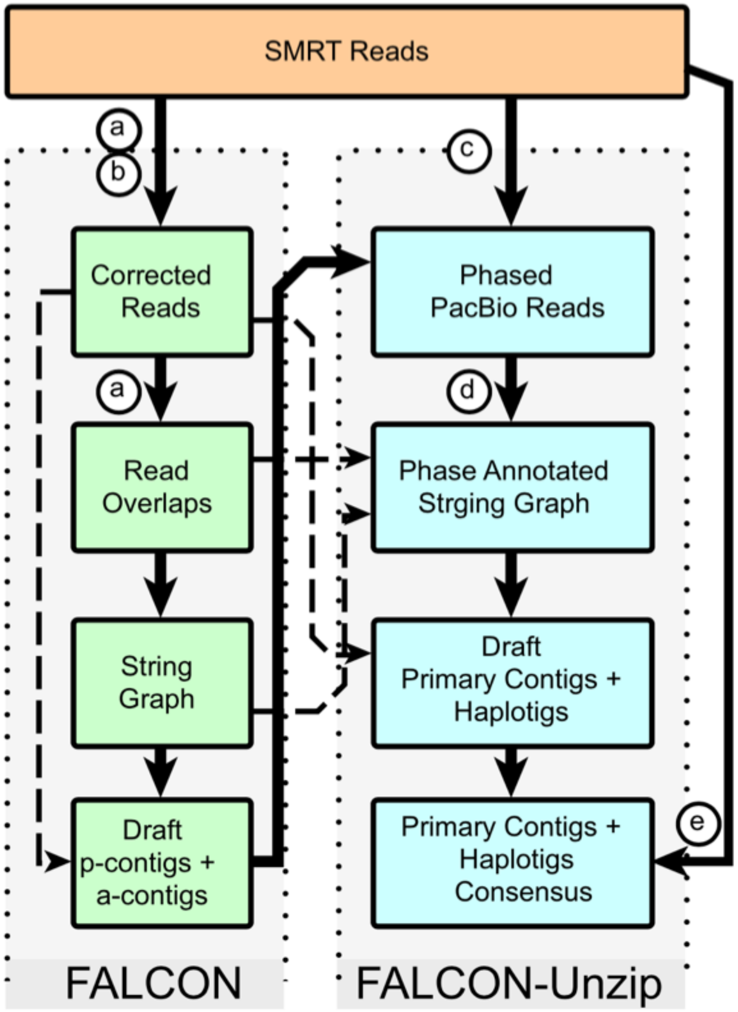
Coverage can overcome random errors

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

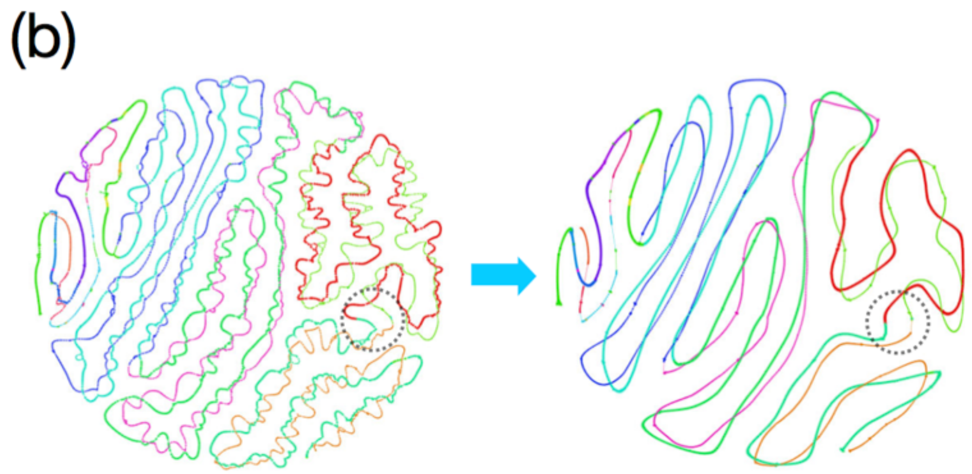
$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren 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"



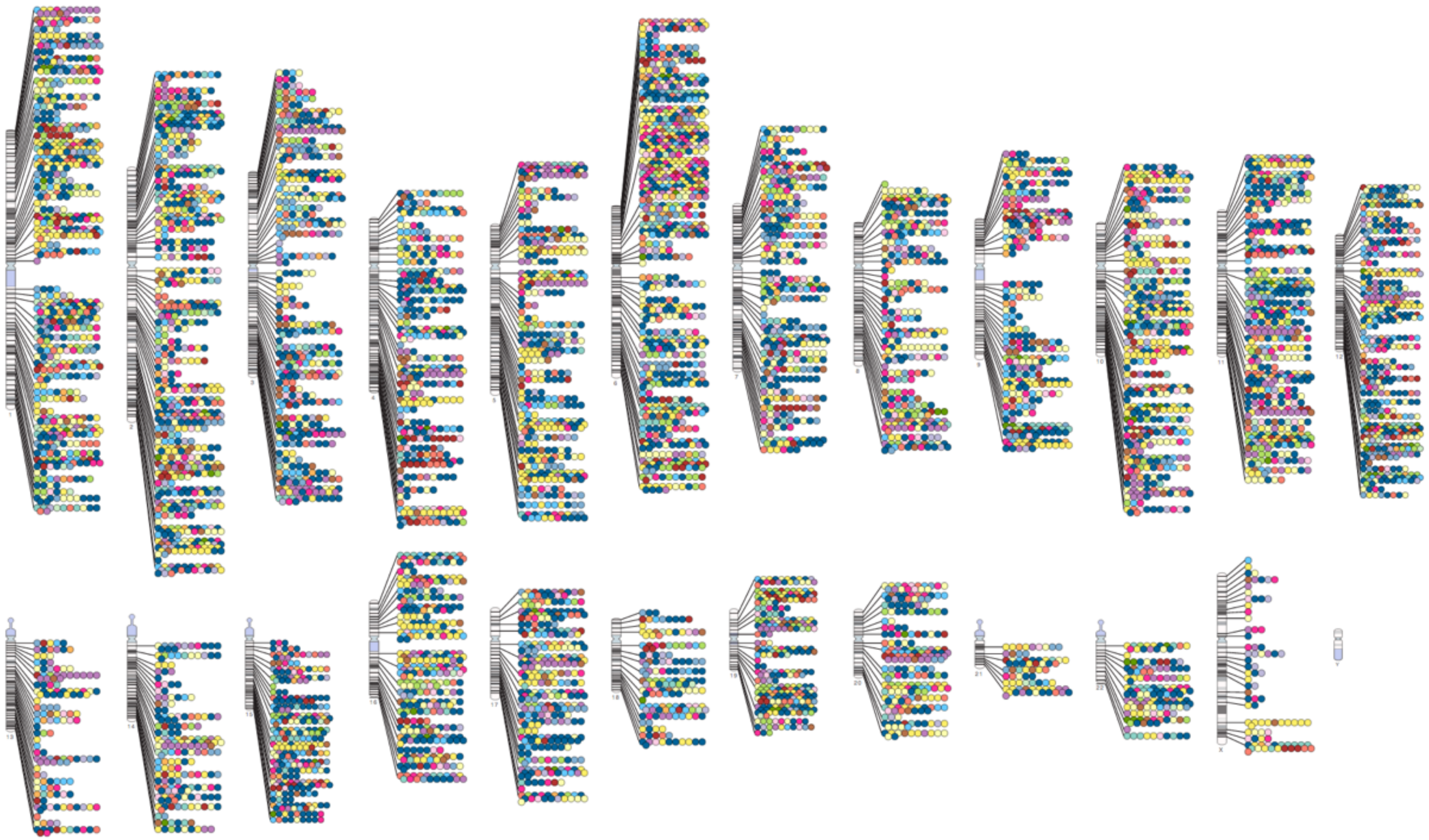
Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing
Chin et al (2016) *Nature Methods. doi:10.1038/nmeth.4035.*

2. Completeness:

How much of the genome is present?

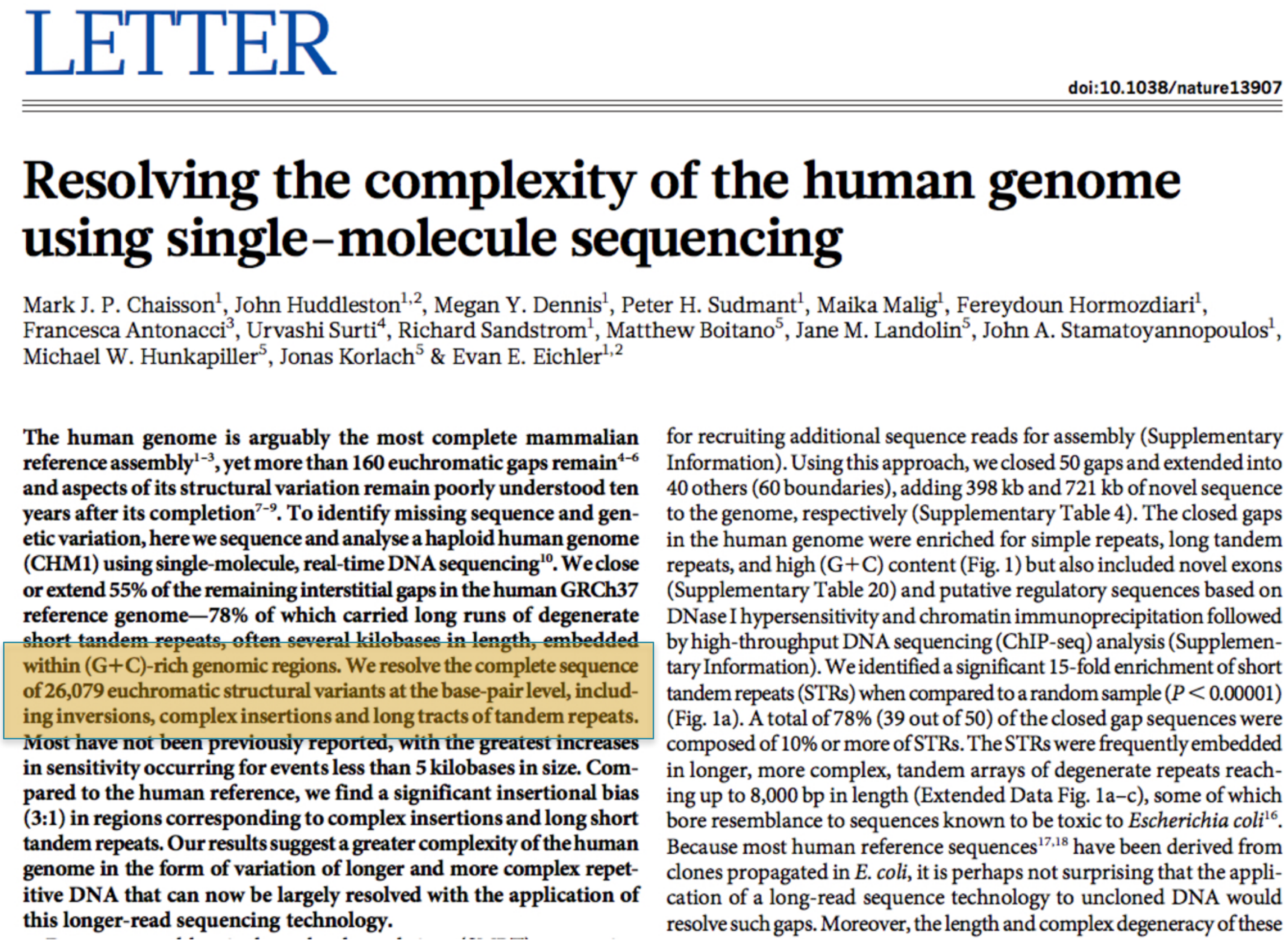
2. Completeness:

How much of the genome is present?



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

ENCODE Consortium (2012)



LETTER

doi:10.1038/nature13907

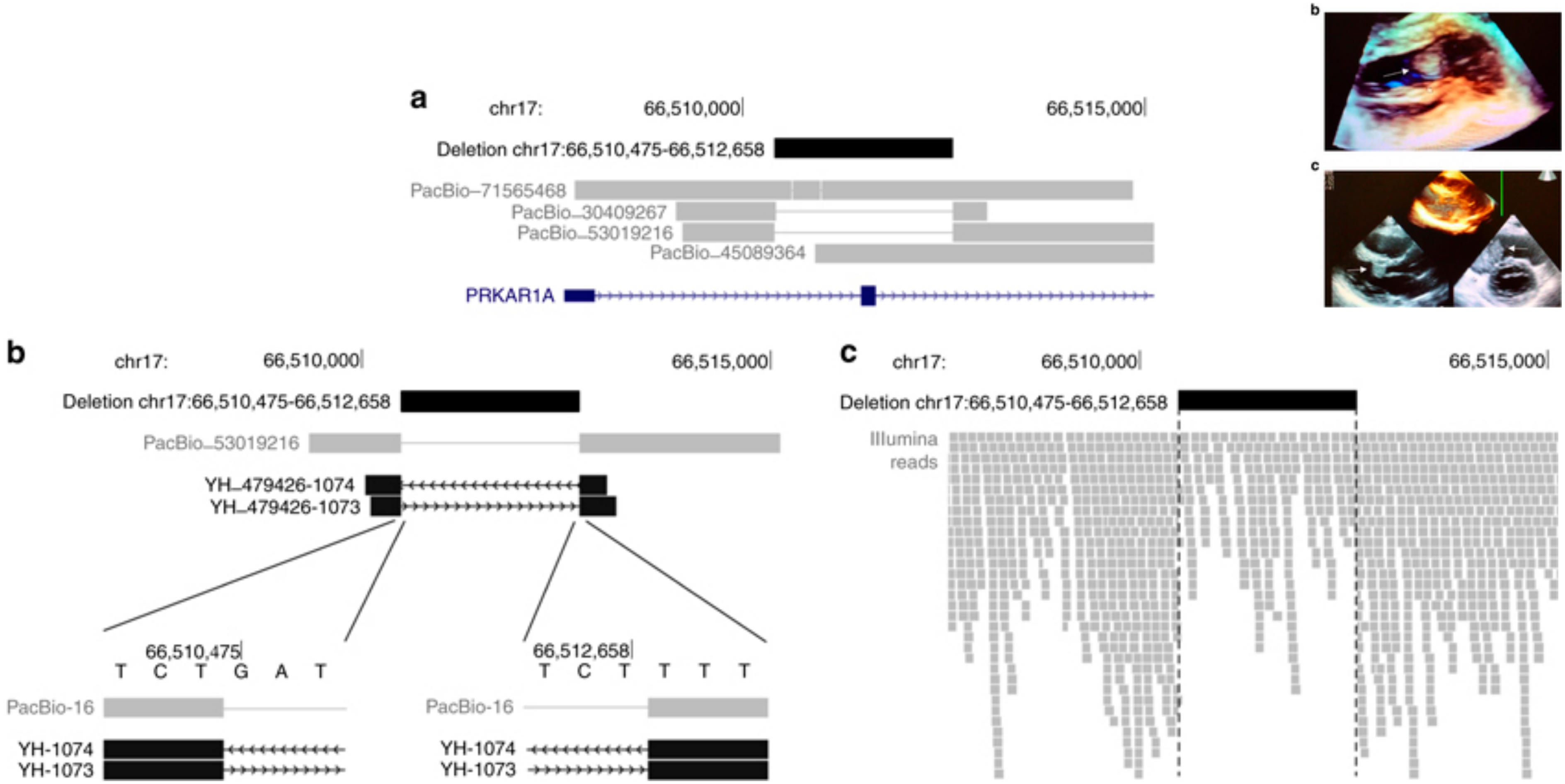
Resolving the complexity of the human genome using single-molecule sequencing

Mark J. P. Chaisson[1], John Huddleston[1,2], Megan Y. Dennis[1], Peter H. Sudmant[1], Maika Malig[1], Fereydoun Hormozdiari[1], Francesca Antonacci[3], Urvashi Surti[4], Richard Sandstrom[1], Matthew Boitano[5], Jane M. Landolin[5], John A. Stamatoyannopoulos[1], Michael W. Hunkapiller[5], Jonas Korlach[5] & Evan E. Eichler[1,2]

The human genome is arguably the most complete mammalian reference assembly[1–3], yet more than 160 euchromatic gaps remain[4–6] and aspects of its structural variation remain poorly understood ten years after its completion[7–9]. To identify missing sequence and genetic variation, here we sequence and analyse a haploid human genome (CHM1) using single-molecule, real-time DNA sequencing[10]. We close or extend 55% of the remaining interstitial gaps in the human GRCh37 reference genome—78% of which carried long runs of degenerate short tandem repeats, often several kilobases in length, embedded within (G+C)-rich genomic regions. We resolve the complete sequence of 26,079 euchromatic structural variants at the base-pair level, including inversions, complex insertions and long tracts of tandem repeats. Most have not been previously reported, with the greatest increases in sensitivity occurring for events less than 5 kilobases in size. Compared to the human reference, we find a significant insertional bias (3:1) in regions corresponding to complex insertions and long short tandem repeats. Our results suggest a greater complexity of the human genome in the form of variation of longer and more complex repetitive DNA that can now be largely resolved with the application of this longer-read sequencing technology.

for recruiting additional sequence reads for assembly (Supplementary Information). Using this approach, we closed 50 gaps and extended into 40 others (60 boundaries), adding 398 kb and 721 kb of novel sequence to the genome, respectively (Supplementary Table 4). The closed gaps in the human genome were enriched for simple repeats, long tandem repeats, and high (G+C) content (Fig. 1) but also included novel exons (Supplementary Table 20) and putative regulatory sequences based on DNase I hypersensitivity and chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) analysis (Supplementary Information). We identified a significant 15-fold enrichment of short tandem repeats (STRs) when compared to a random sample ($P < 0.00001$) (Fig. 1a). A total of 78% (39 out of 50) of the closed gap sequences were composed of 10% or more of STRs. The STRs were frequently embedded in longer, more complex, tandem arrays of degenerate repeats reaching up to 8,000 bp in length (Extended Data Fig. 1a–c), some of which bore resemblance to sequences known to be toxic to *Escherichia coli*[16]. Because most human reference sequences[17,18] have been derived from clones propagated in *E. coli*, it is perhaps not surprising that the application of a long-read sequence technology to uncloned DNA would resolve such gaps. Moreover, the length and complex degeneracy of these

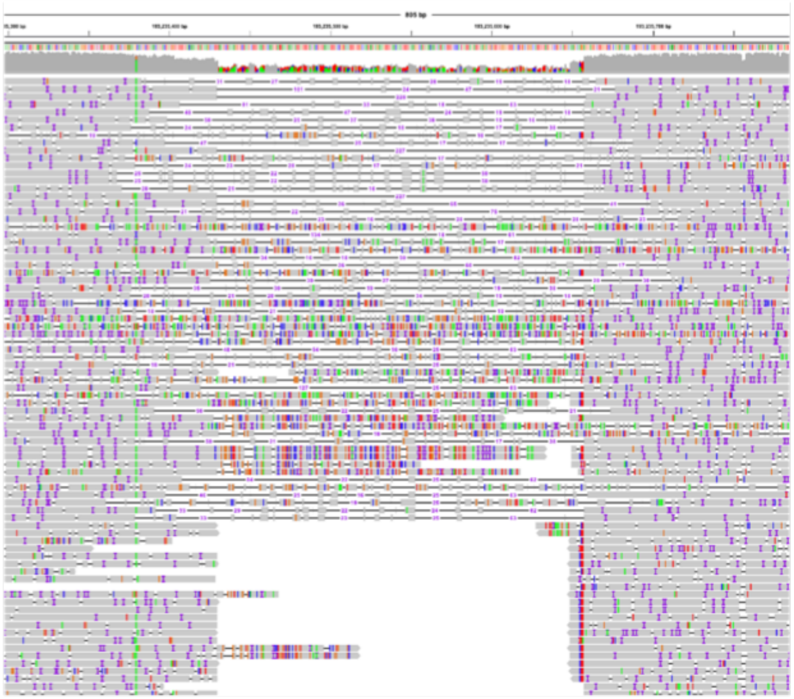
Structural Variations in Human Disease



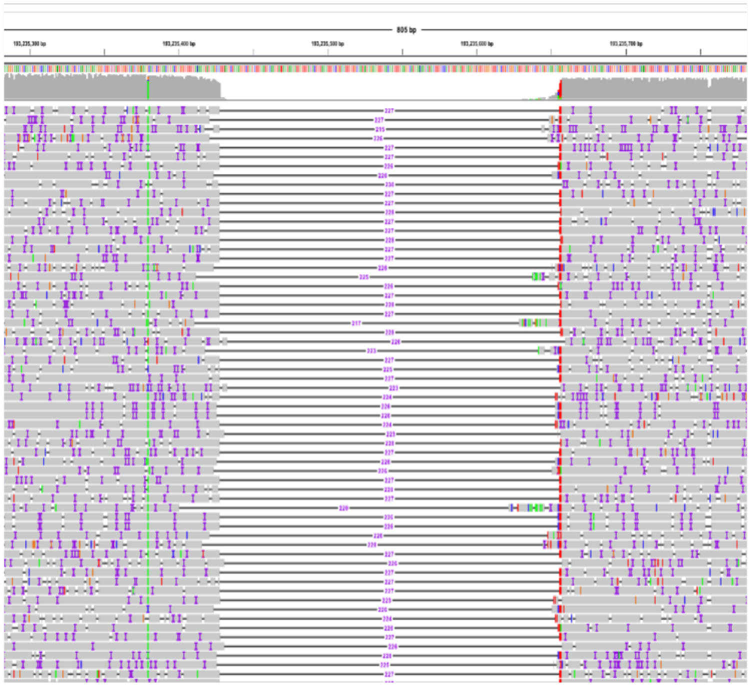
Long-read genome sequencing identifies causal structural variation in a Mendelian disease
Merker et al (2017) *Genetics in Medicine. doi:10.1038/gim.2017.86*

NGMLR + Sniffles

BWA-MEM:



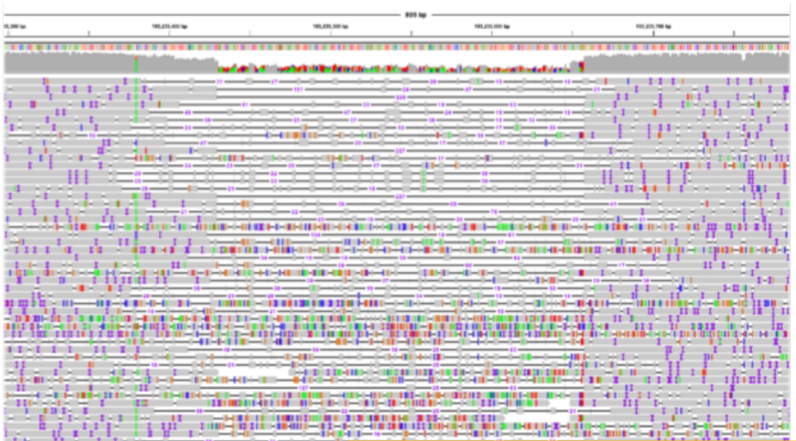
NGMLR:



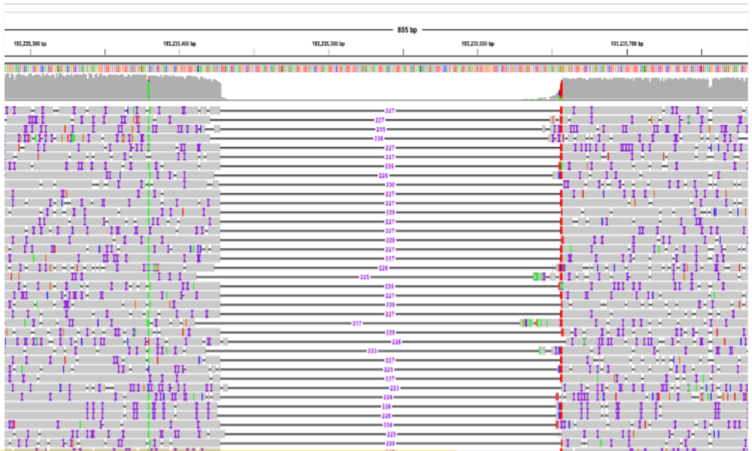
Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *In preparation*

NGMLR + Sniffles

BWA-MEM:



NGMLR:

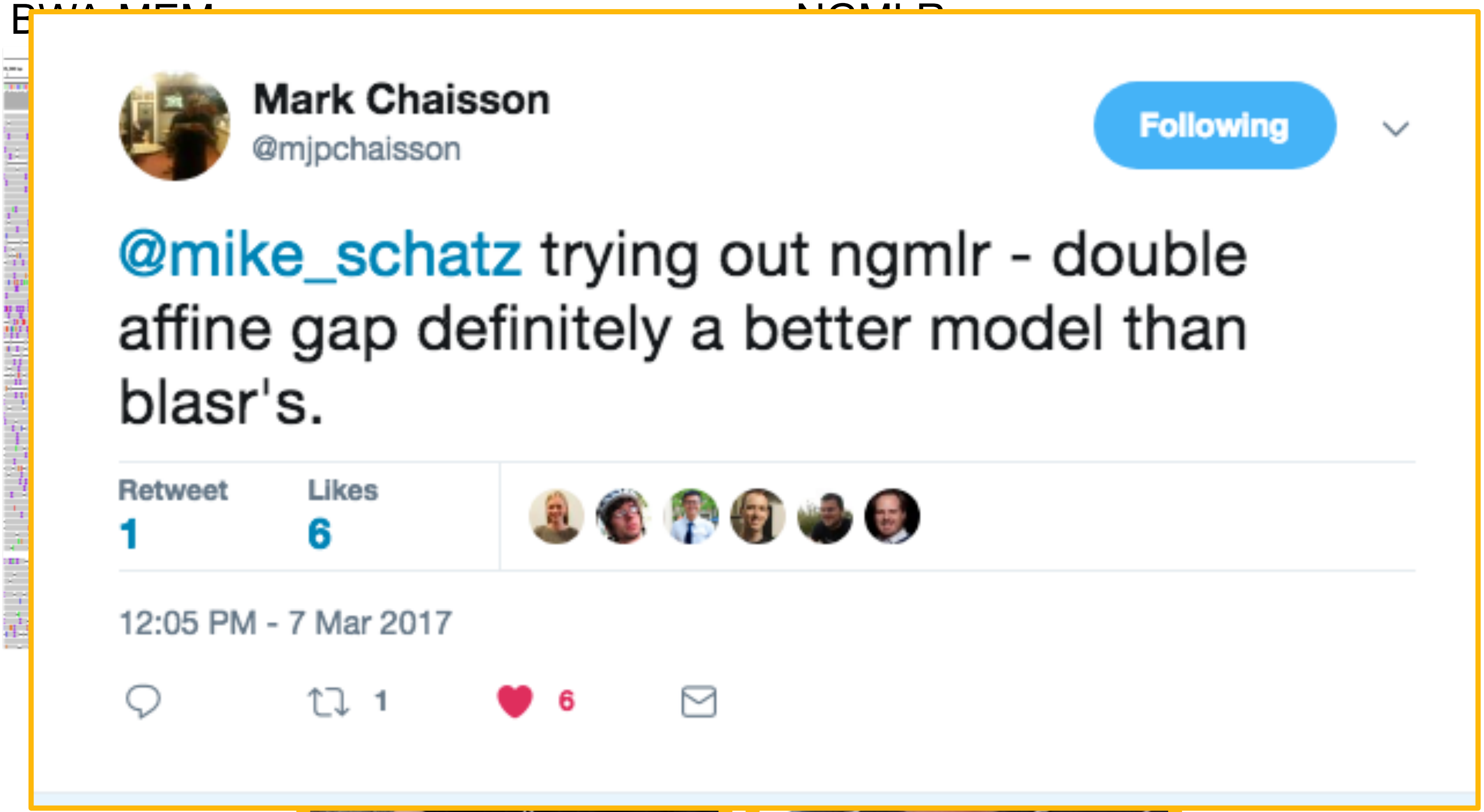


Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *In preparation*

NGMLR + Sniffles

BWA-MEM

NGMLR

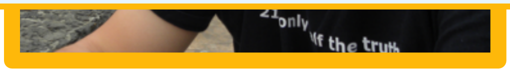


Mark Chaisson
@mjpchaisson

@mike_schatz trying out ngmlr - double affine gap definitely a better model than blasr's.

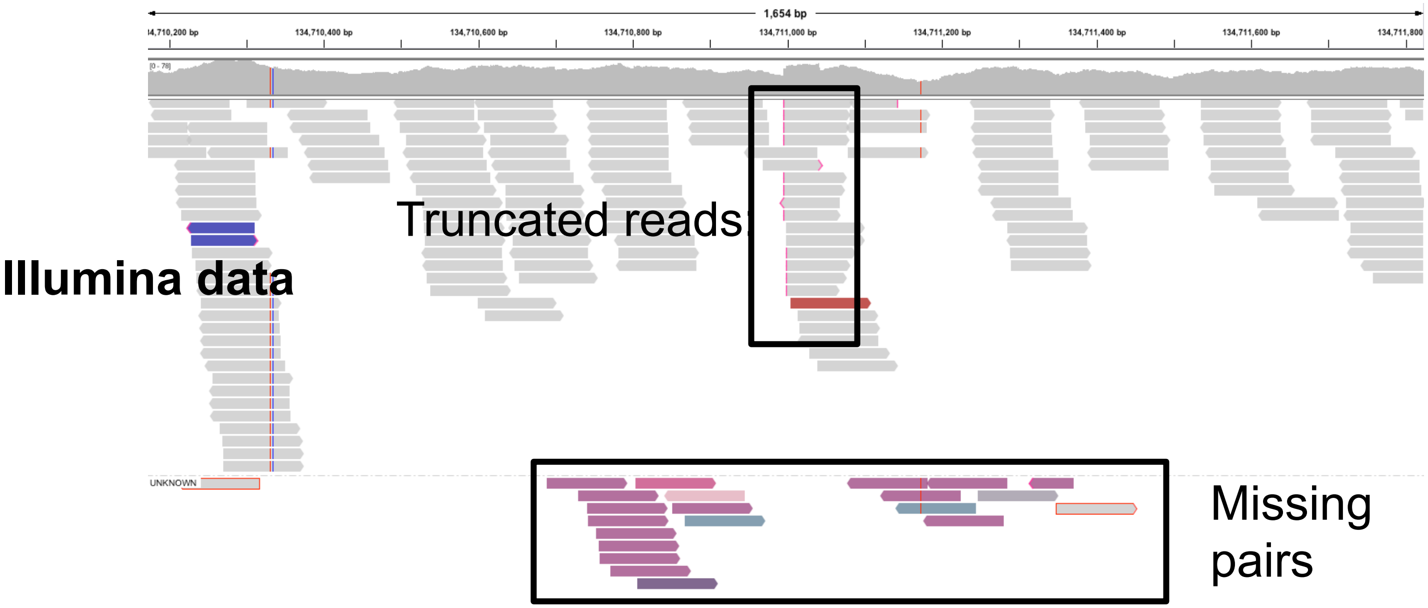
Retweet 1 Likes 6

12:05 PM - 7 Mar 2017



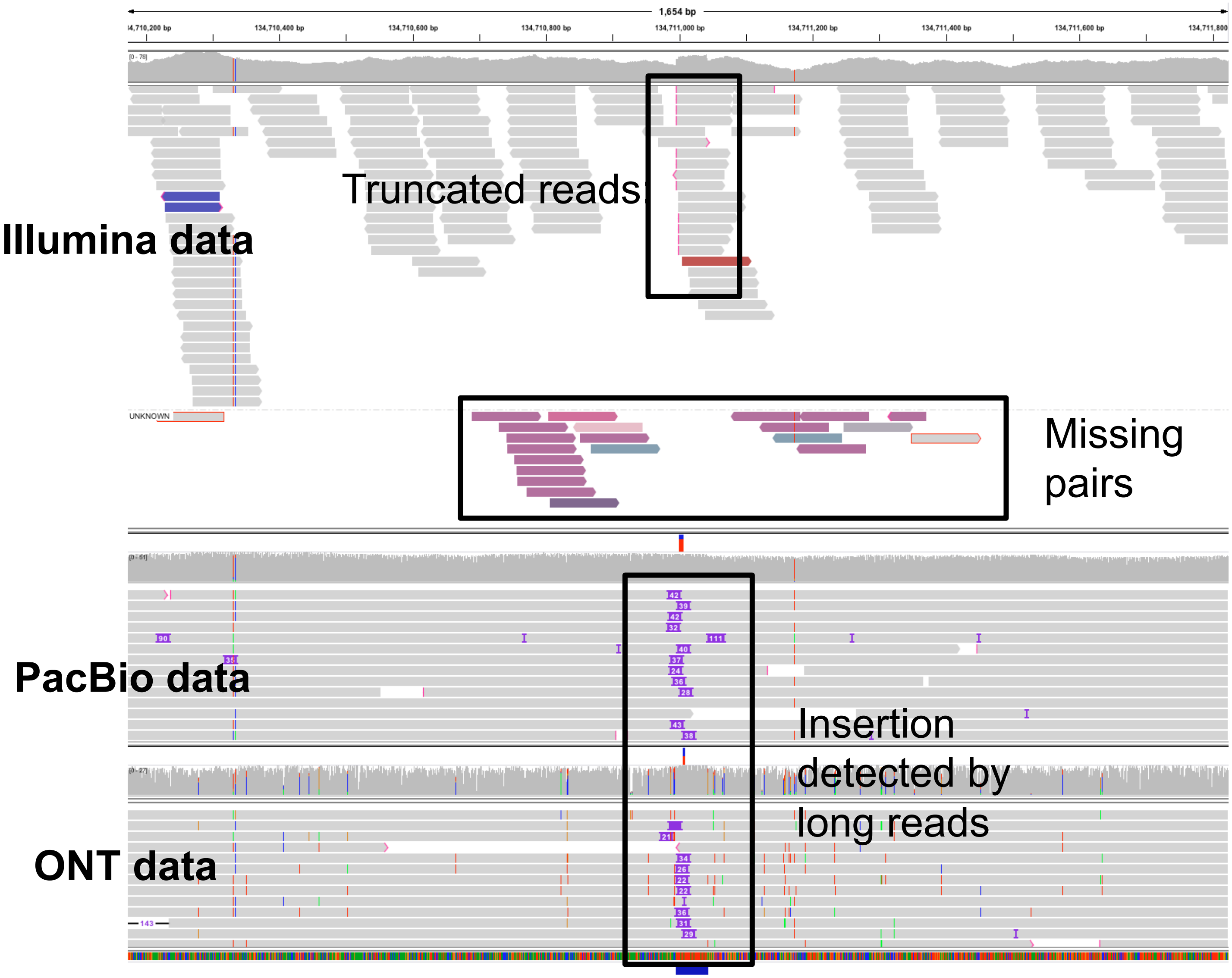
Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *In preparation*

No more false positives!



Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *In preparation*

No more false positives!



Illumina data

Truncated reads

Missing pairs

PacBio data

Insertion detected by long reads

ONT data

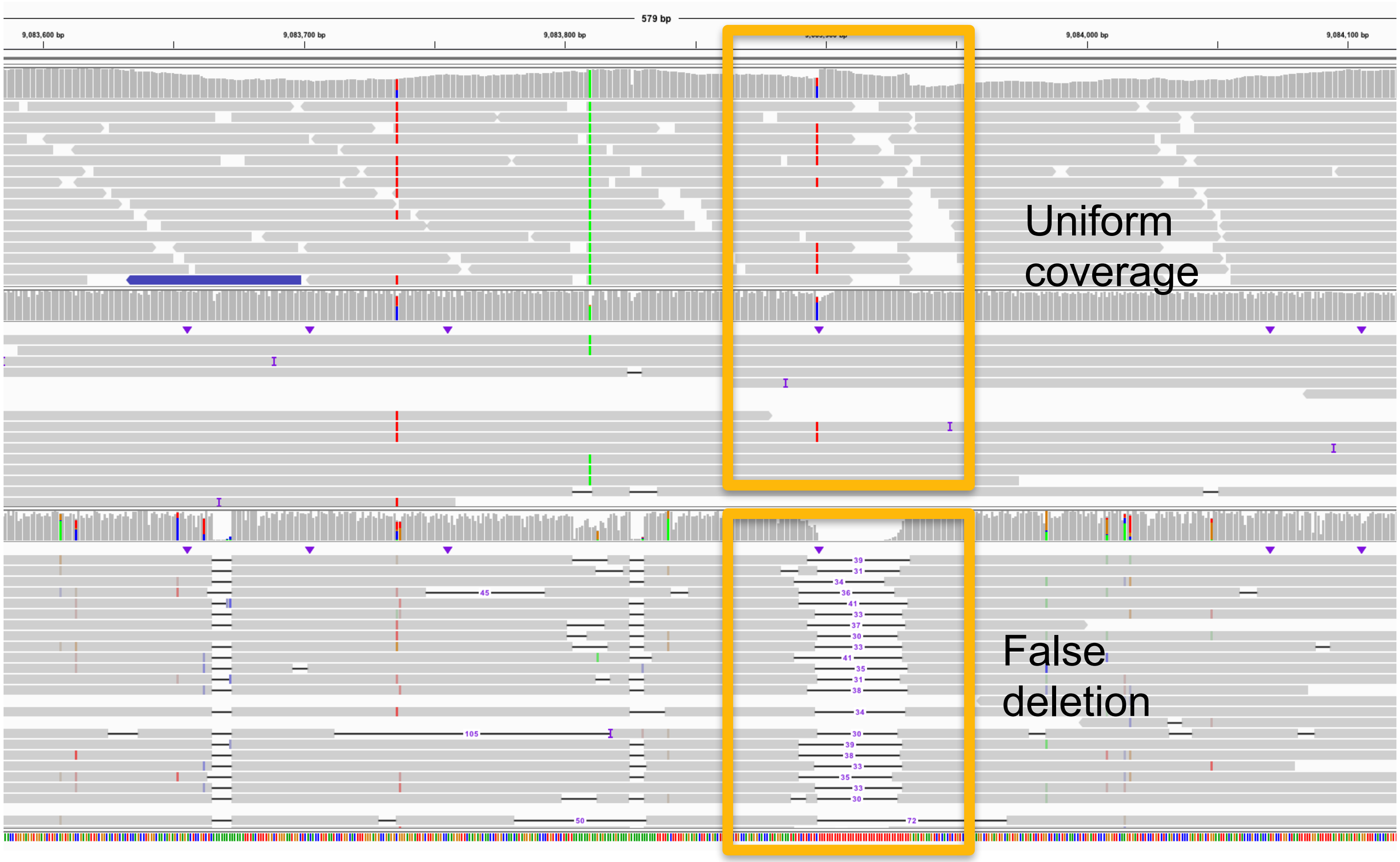
Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *In preparation*

No more false positives!

Illumina data

PacBio data

ONT data



Uniform coverage

False deletion

Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *In preparation*

3. Contiguity

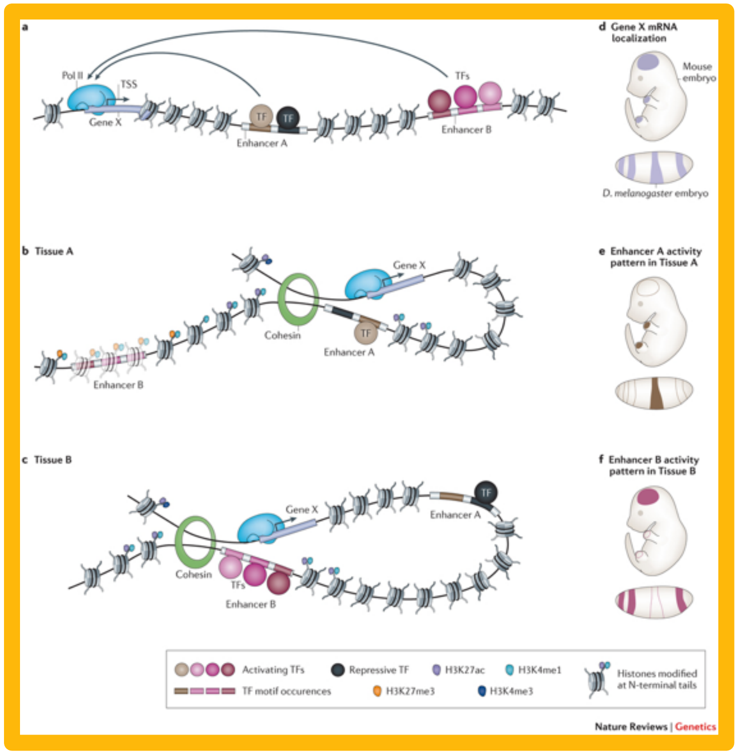
How much context is available?

3. Contiguity

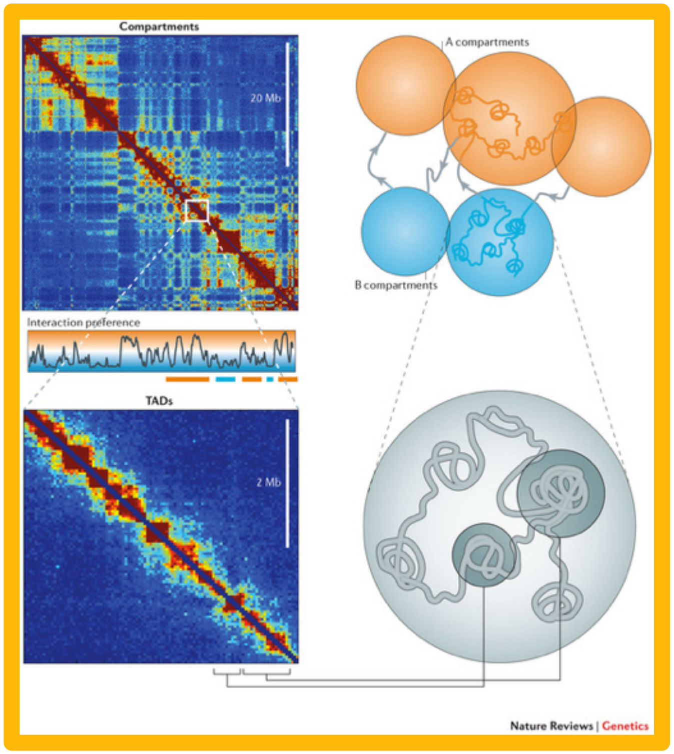
How much context is available?

If you have 99% completeness, are you missing 1% of every gene or are the missing sequences localized to certain regions?

How far can you go until you hit a gap in resolution?

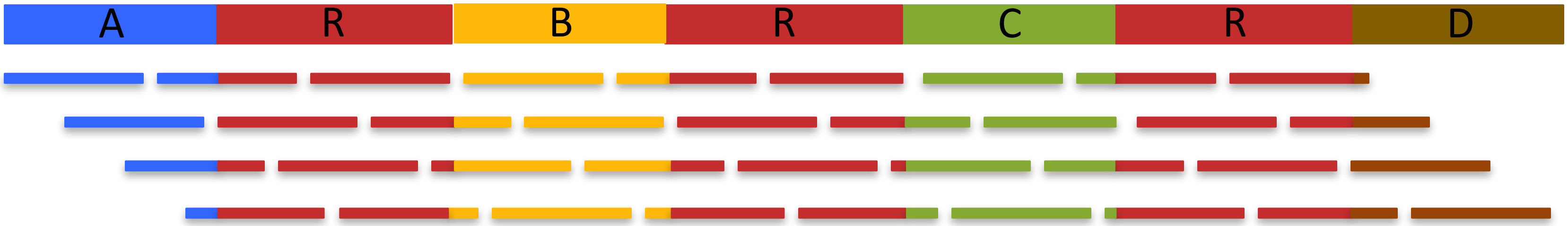


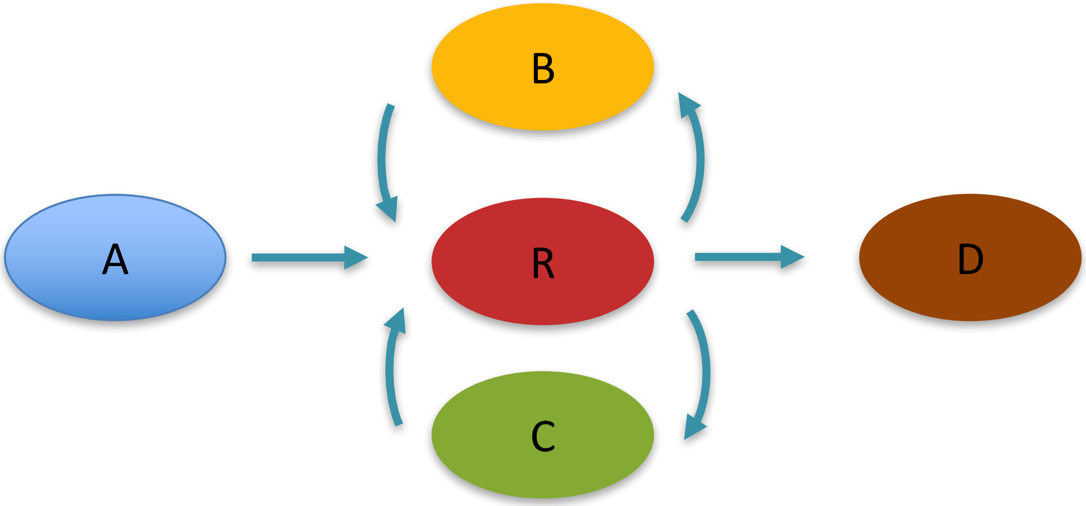
Nature Reviews | Genetics



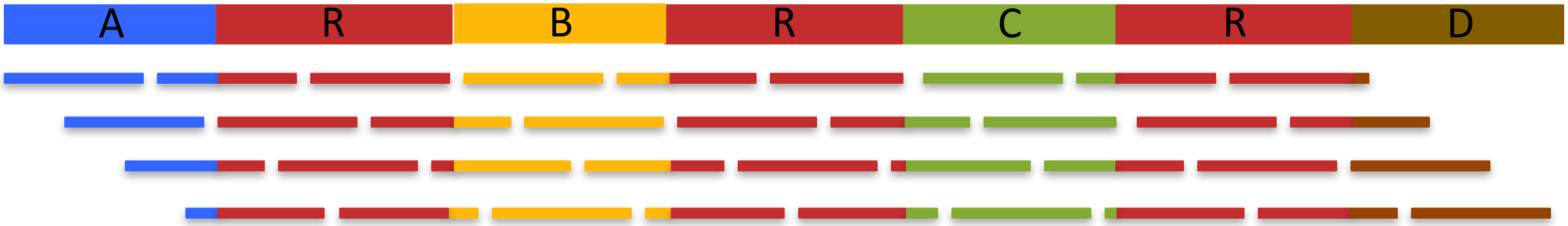
Nature Reviews | Genetics

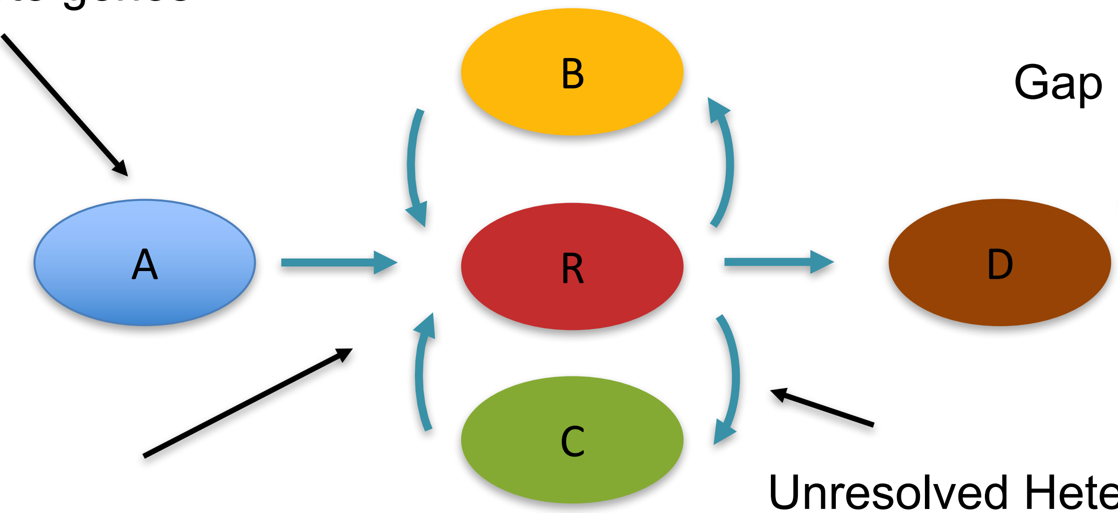
Assembly Complexity





Assembly Complexity





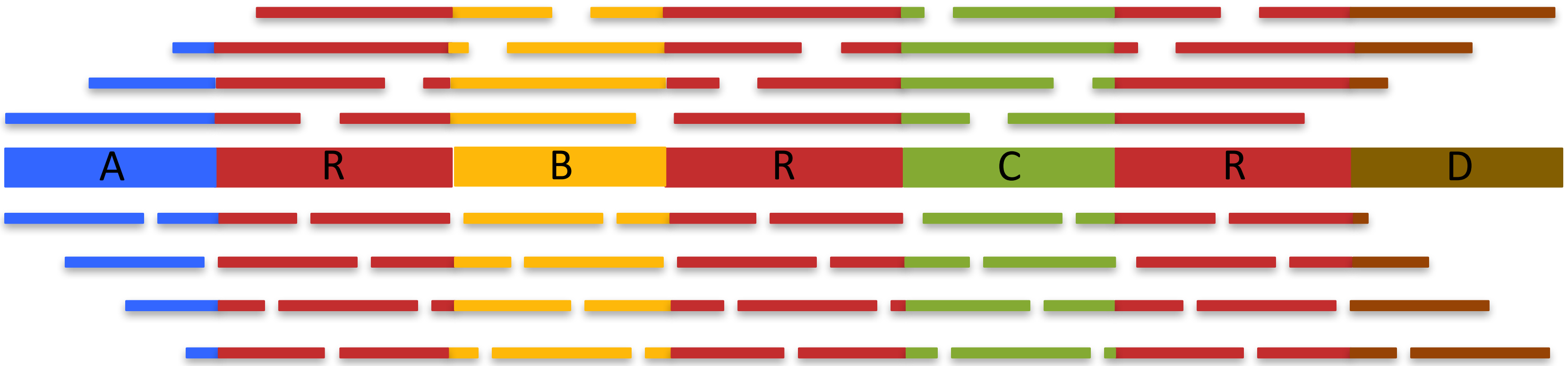
Short contigs & Incomplete genes

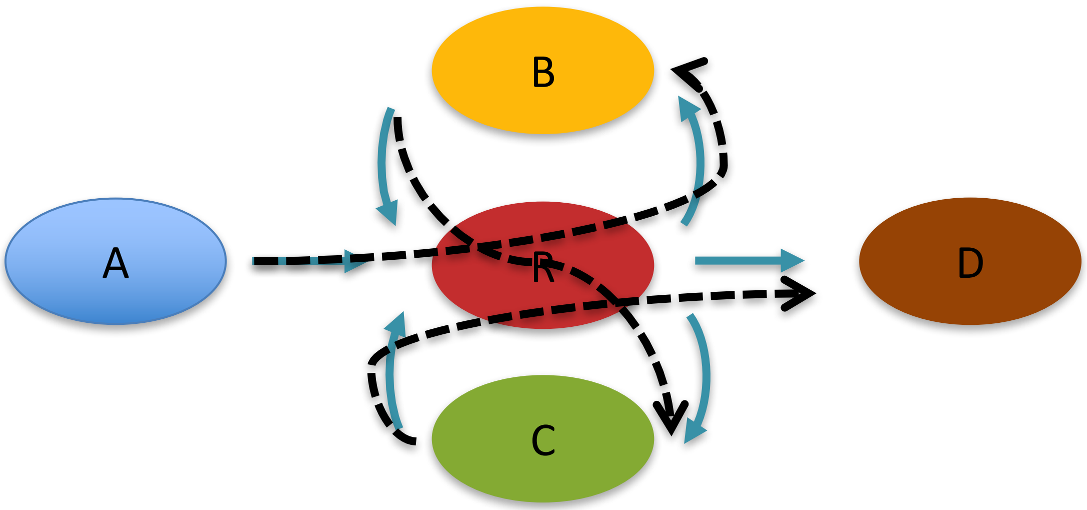
Gap in coverage

Unresolved Heterozygosity

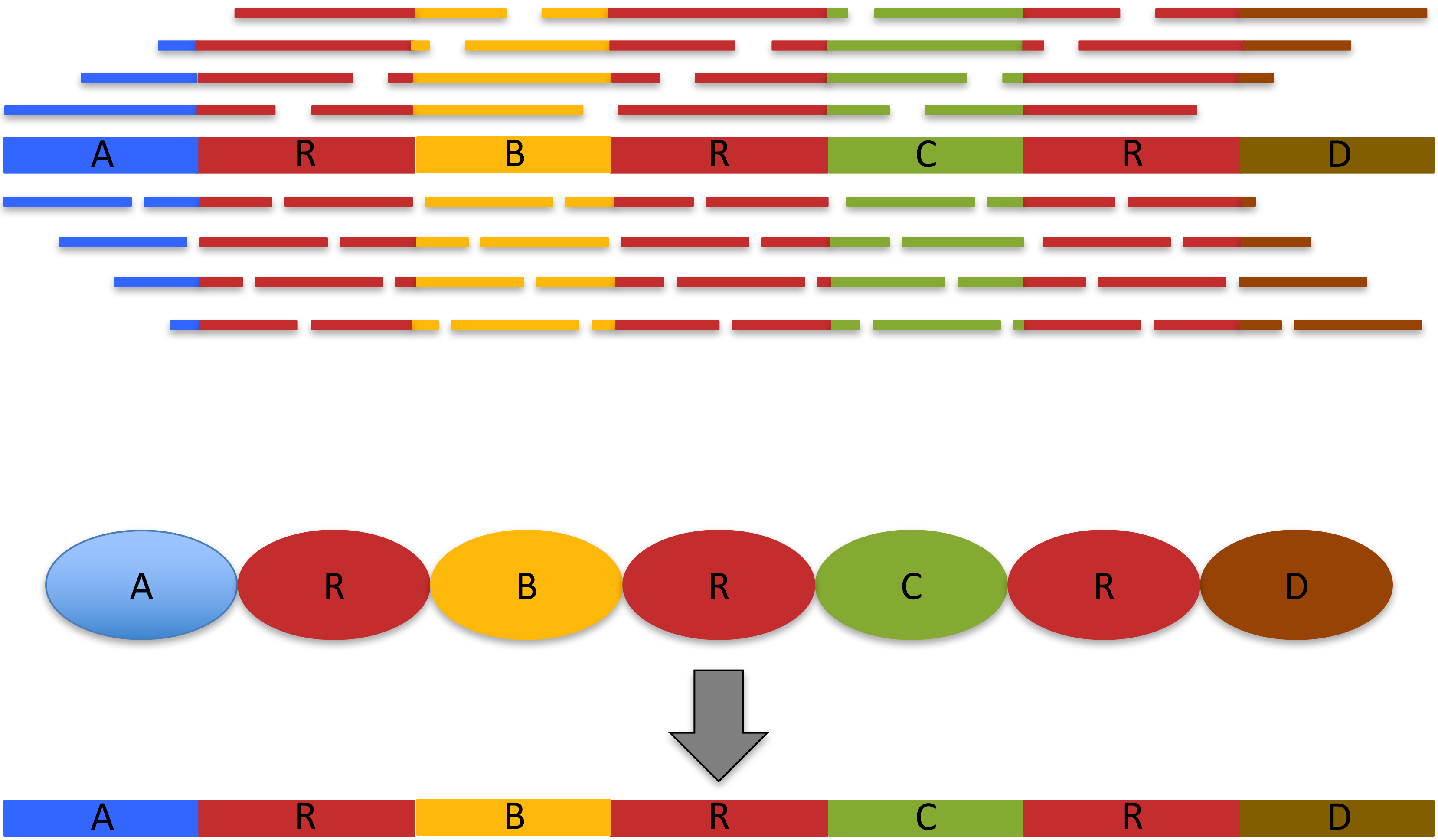
Variants hidden in repeats

Assembly Complexity



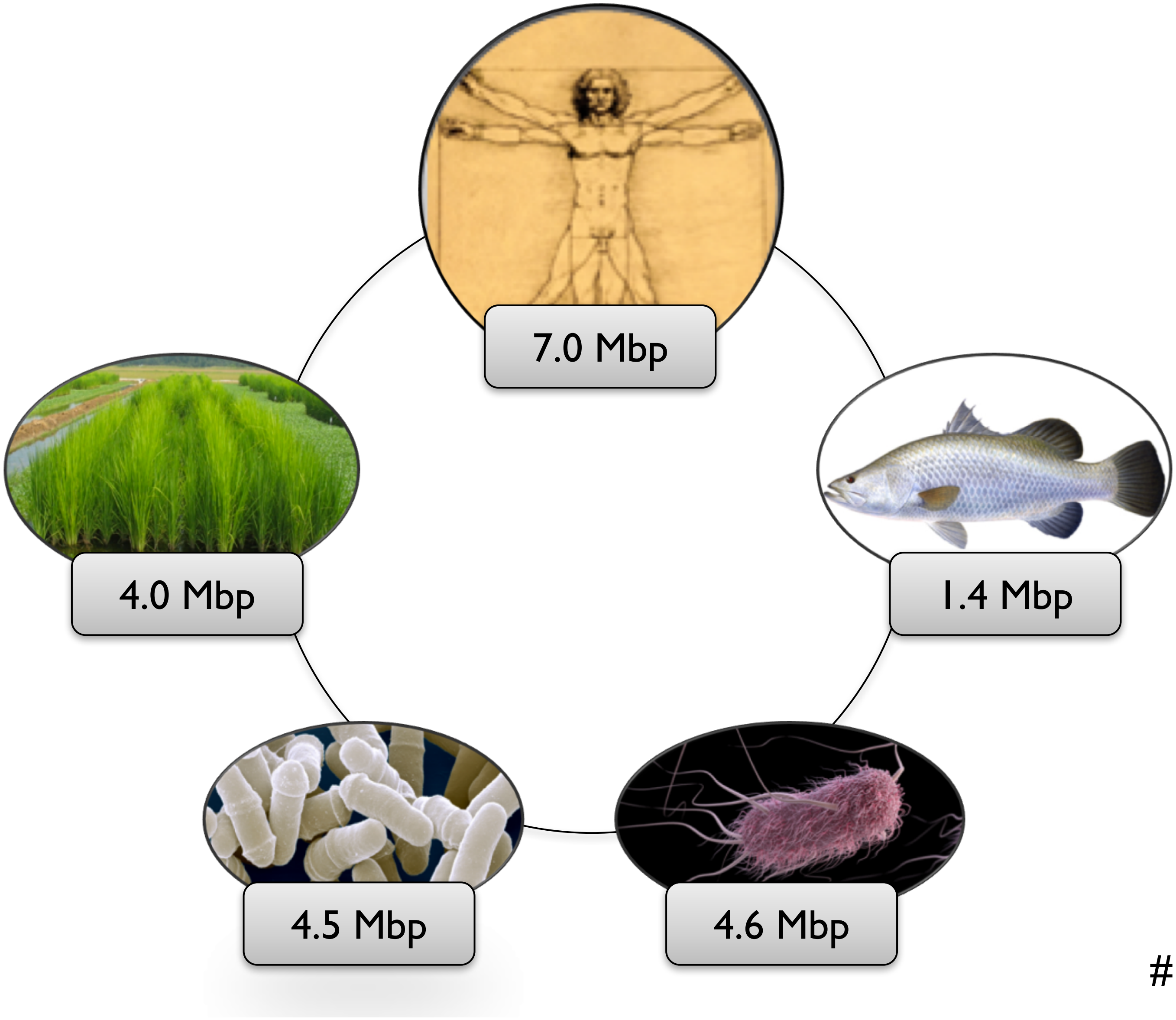


Assembly Complexity



The advantages of SMRT sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology.* 14:405

(A few) Recent PacBio Assemblies



7.0 Mbp

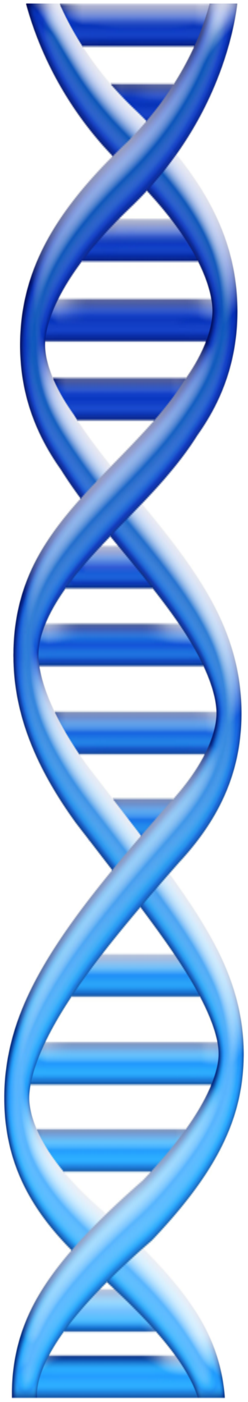
4.0 Mbp

1.4 Mbp

4.5 Mbp

4.6 Mbp

#1mbctgclub



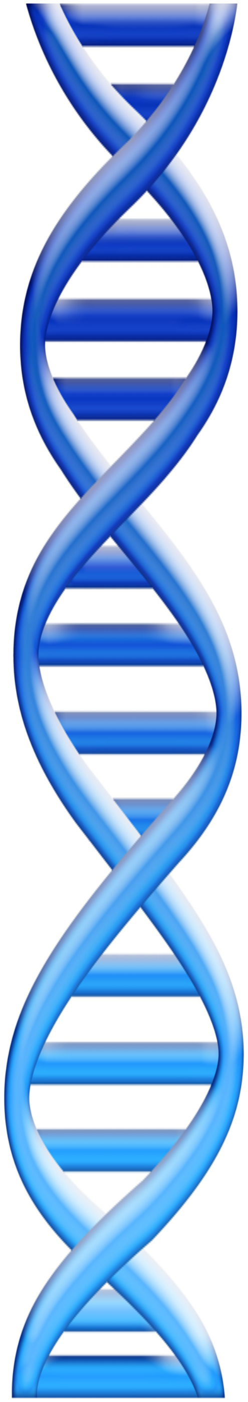
In pursuit of perfect genome sequencing

1. Why "Perfect"?

2. **What is "Perfect"?**
   100% Correct, Complete, & Contiguous

3. How will we achieve it?

4. When will we achieve it?

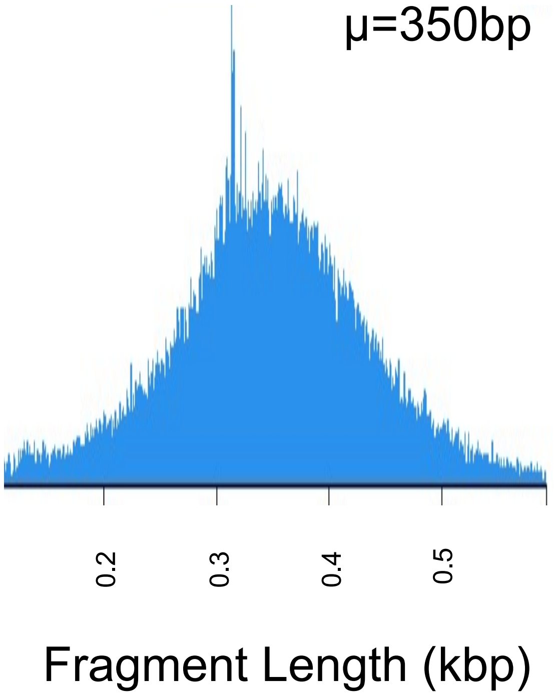


In pursuit of perfect genome sequencing

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

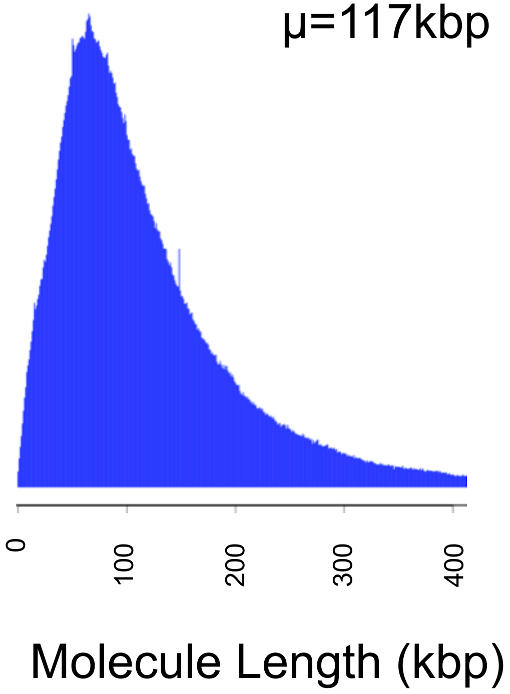
Genomic Sequencing Data

Illumina



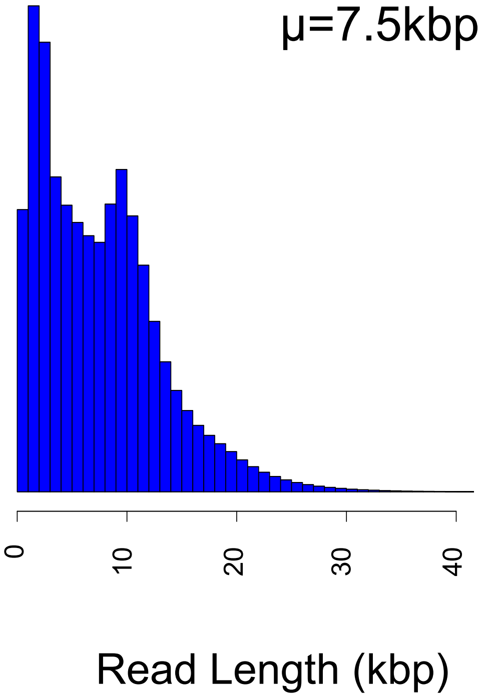
60x Paired End
All 4 samples

10X Genomics



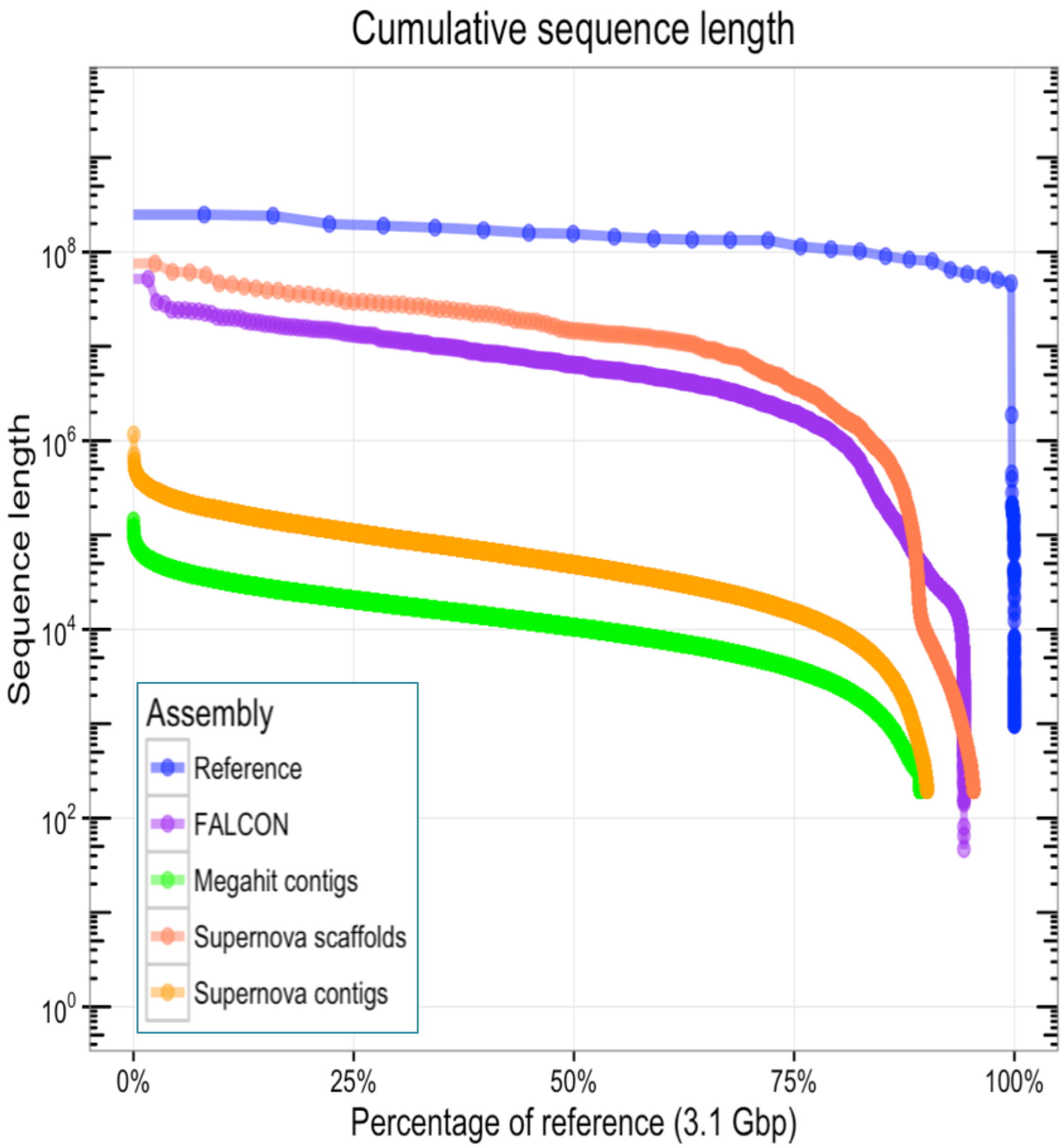
35x Linked Reads
All 4 samples

PacBio



55x Long Reads
****Only ENC-002***

Assembly Contiguity



GRC38 Reference

- **Includes alt sequences**

10X Genomics/SuperNova

- 21 Mbp scaffold N50
- 162 Mbp in scaffold gaps

PacBio/Falcon-unzip

- 7.0 Mbp contig N50

10X Genomics/Supernova

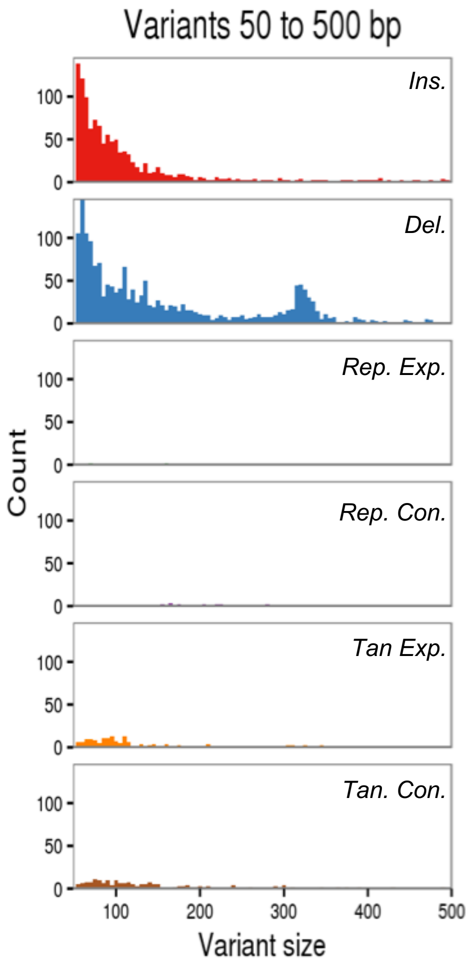
- 50 kbp contig N50

Illumina/MegaHit

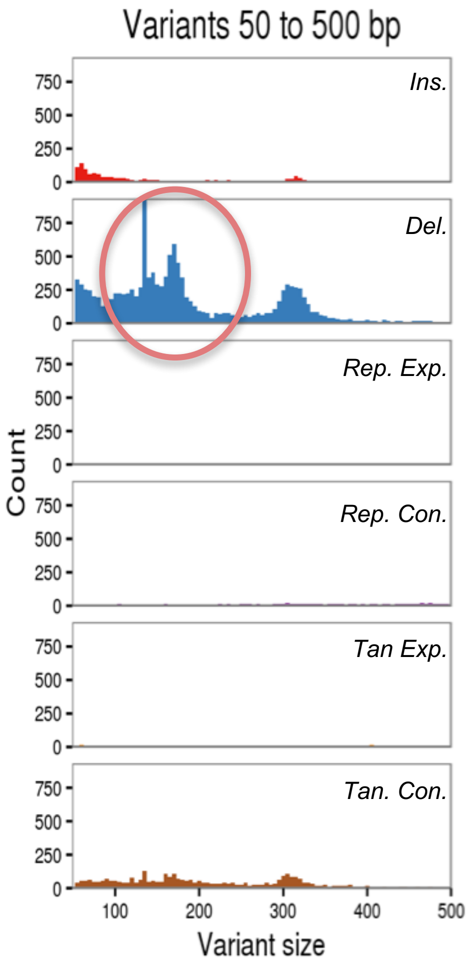
- 13 kbp contig N50

Missing Insertions from Short and Linked Read?

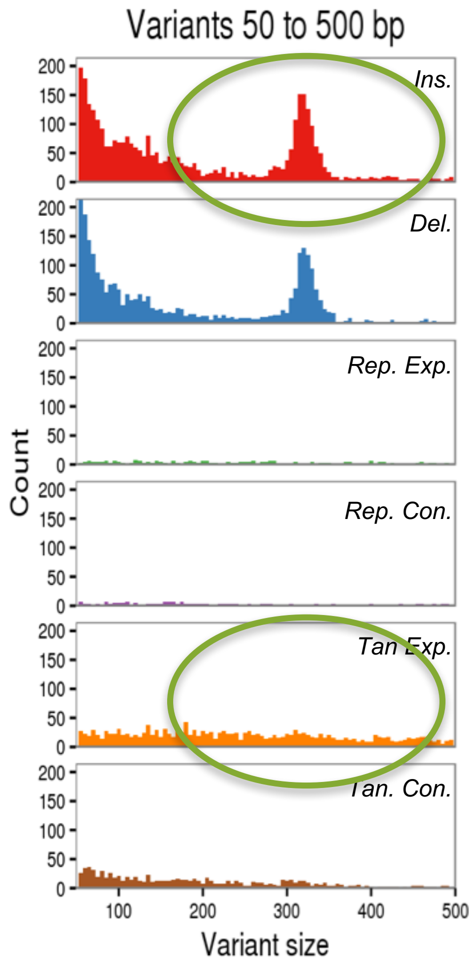
Illumina



10X Genomics



PacBio



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

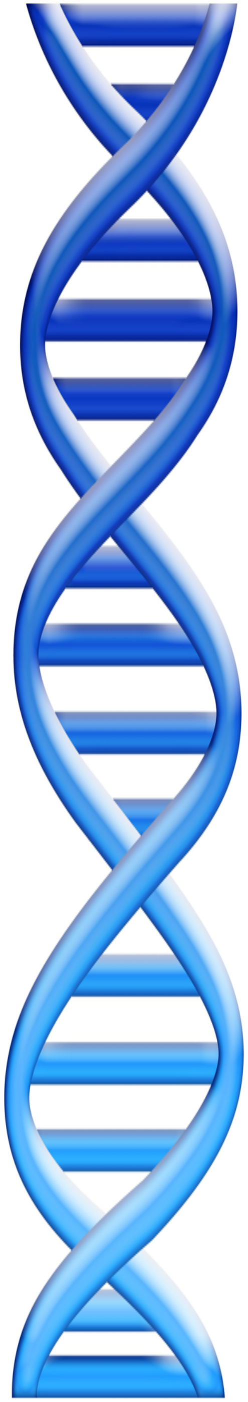
- Concordance by Technology

Inner triplets

- Concordance by Assembly
- Concordance by Mappers

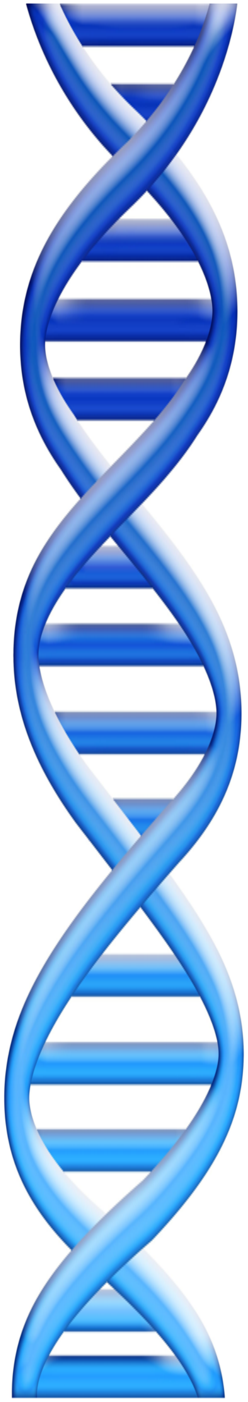
Overall:

- Lonnnnnnng reads give the best concordance ☺



In pursuit of perfect genome sequencing

1. Why "Perfect"?
2. What is "Perfect"?
3. **How will we achieve it?**
   Lonnnnnng reads :-)
4. When will we achieve it?



In pursuit of perfect genome sequencing

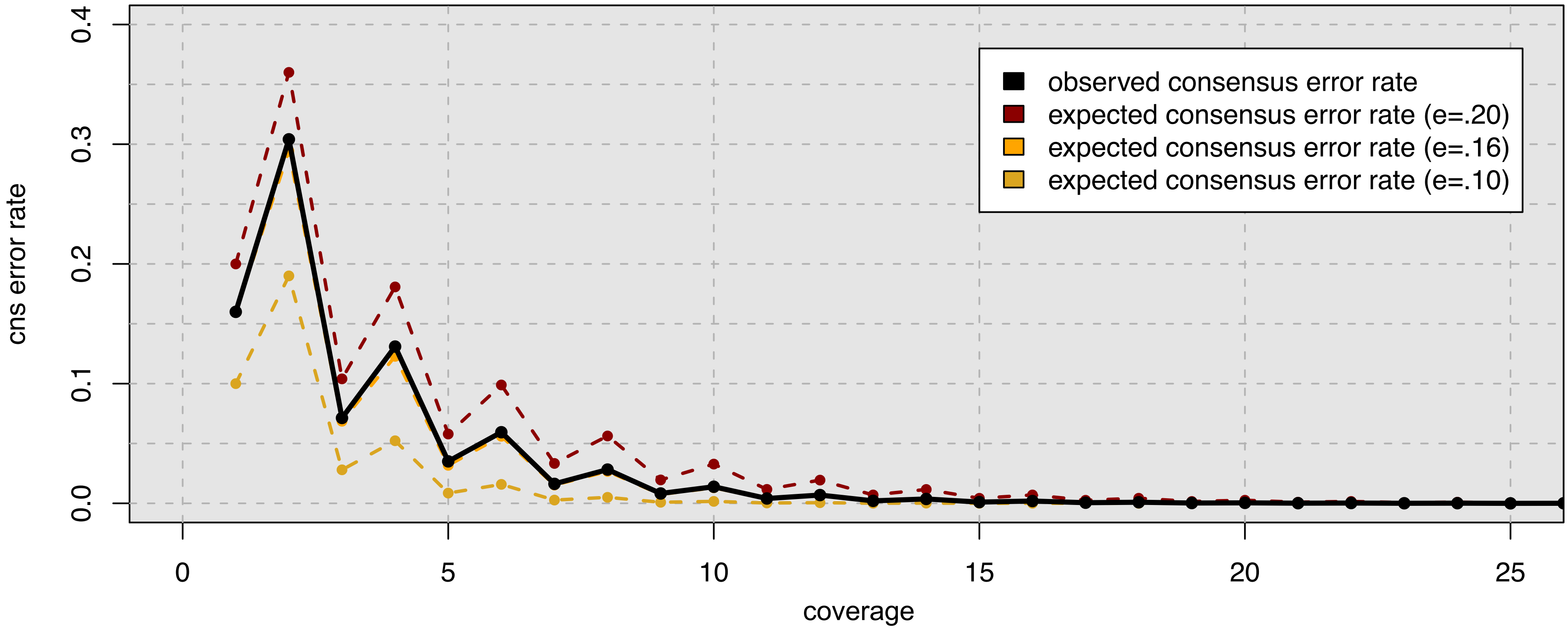
1. Why “Perfect”?

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*

PacBio Roadmap

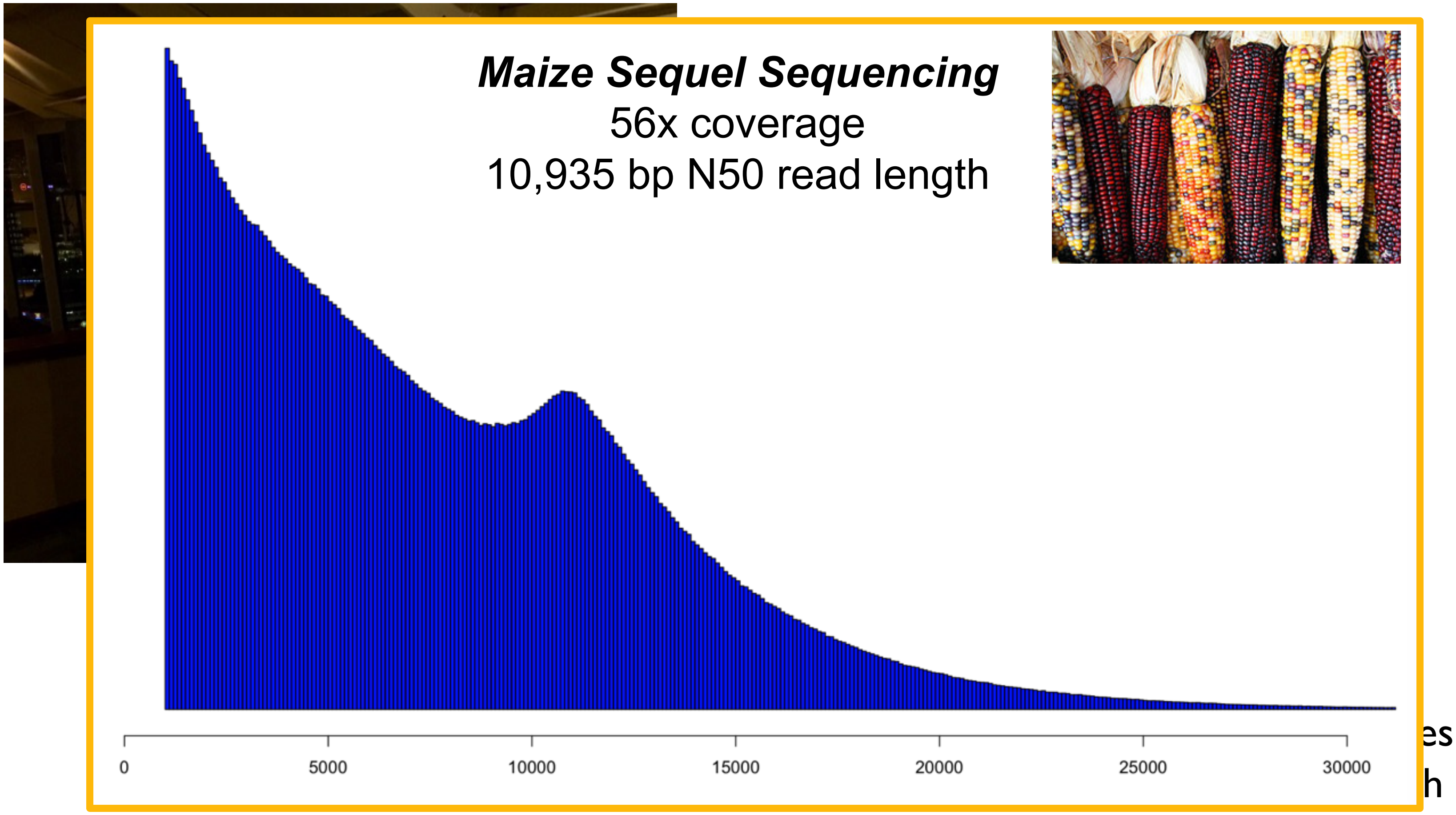
PacBio Sequel

$350k instrument cost
~$30k / human @ 50x

SMRTcell v2

1M Zero Mode Waveguides
~15kb average read length

PacBio Roadmap



Maize Sequel Sequencing
56x coverage
10,935 bp N50 read length

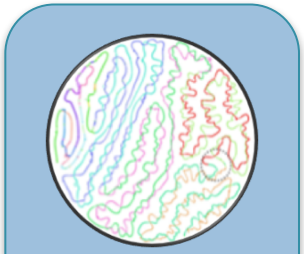
es
h

In pursuit of perfect genome sequencing

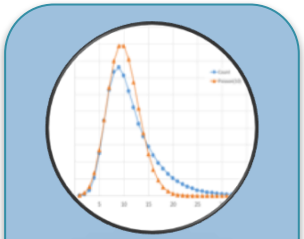
- ***Three C's of Genome Quality: Correctness, Completeness & Contiguity***
  - The key for perfect genomes is lonnnnnnnnng reads ☺
  - Expect new insights on the causes of diseases, forces of evolution

- ***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



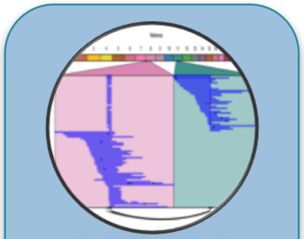
FALCON



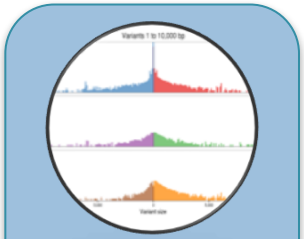
LRSim

SURVIVOR

NGM+Sniffles



Ribbon



Assemblytics

http://schatz-lab.org

Acknowledgements

Schatz Lab
Charlotte Darby
Han Fang
Tyler Gavin
James Gurtowski
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
Langmead Lab
Salzberg 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

Oxford NANOPORE Technologies

PACIFIC BIOSCIENCES®

Thank you!

@mike_schatz

Looking for a postdoc?
http://schatz-lab.org/apply/