Lecture 9. 3rd Generation Sequencing

Michael Schatz

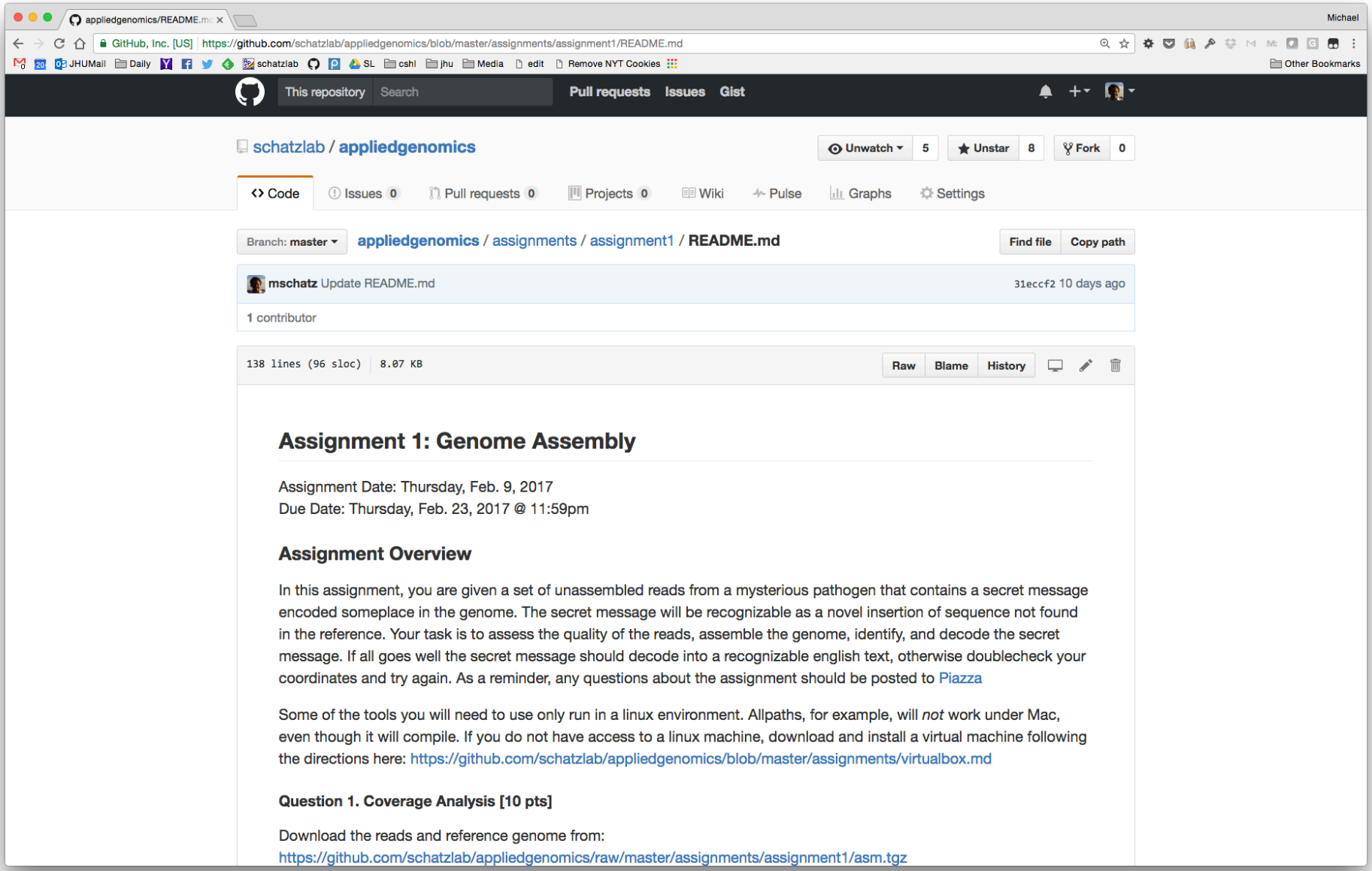
Feb 28, 2017

JHU 600.649: Applied Comparative Genomics

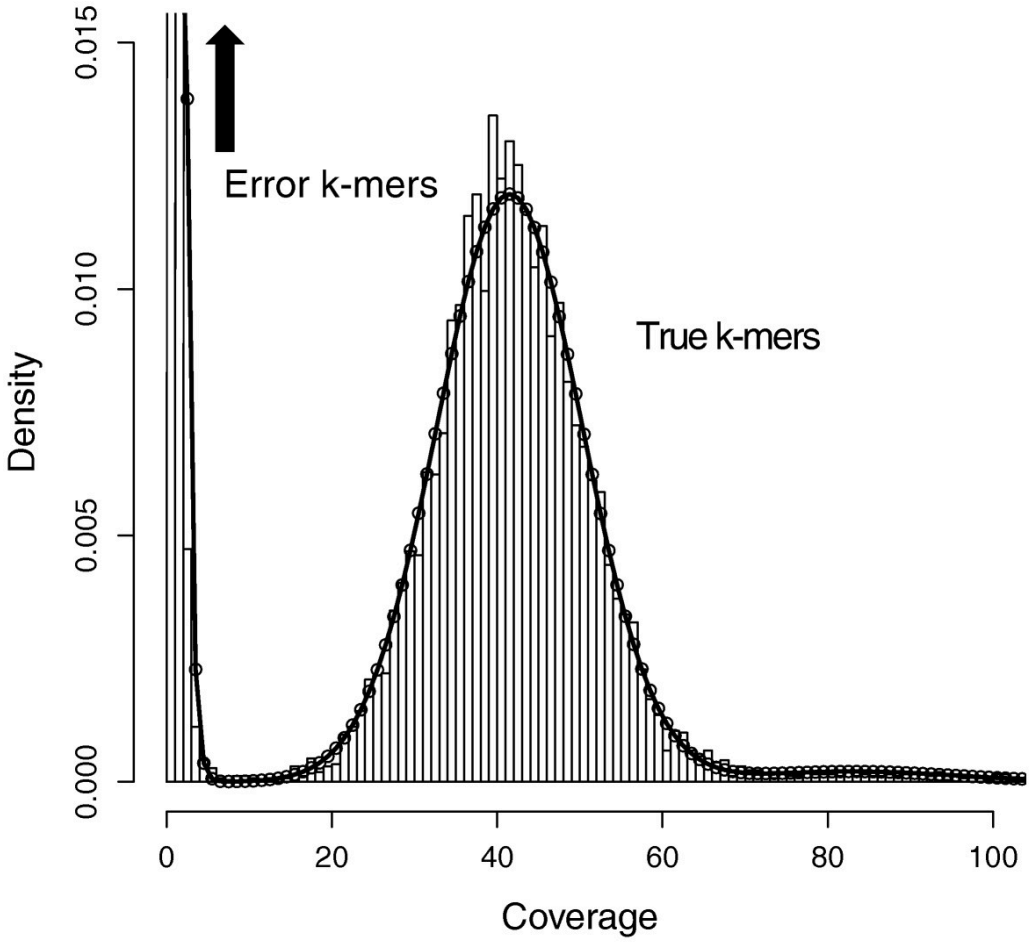
Part 0:
Assignment 1 Review

Assignment 1: Due Thursday @ 11:59pm

Email PDF to: jhuappliedgenomics@gmail.com



Quake: Quality-aware detection and correction of sequencing errors



Reference-free approach for correcting sequencing errors

1. Scan reads, count #occurrences of all k-mers using Jellyfish
2. Analyze k-mer profile to find local minimum between error k-mers (occur < ~5 times) and trusted k-mers (occur > 5 times)
3. For each untrusted k-mer in a read, search for minimum # of substitutions to become trusted

Quake: quality-aware detection and correction of sequencing errors
Kelley, DR, Schatz, MC, Salzberg, SL (2010) Genome Biology 11:R116

Heterozygous Kmer counting

Sequencing read from homologous chromosome 1A

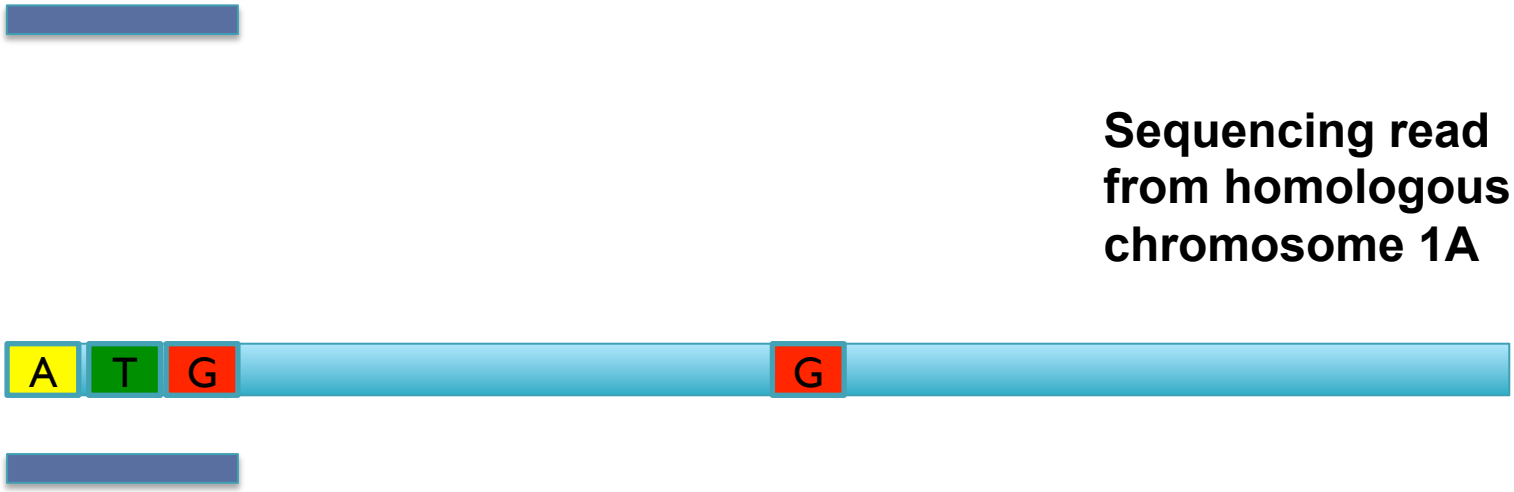
Sequencing read from homologous chromosome 1B

Heterozygous Kmer counting

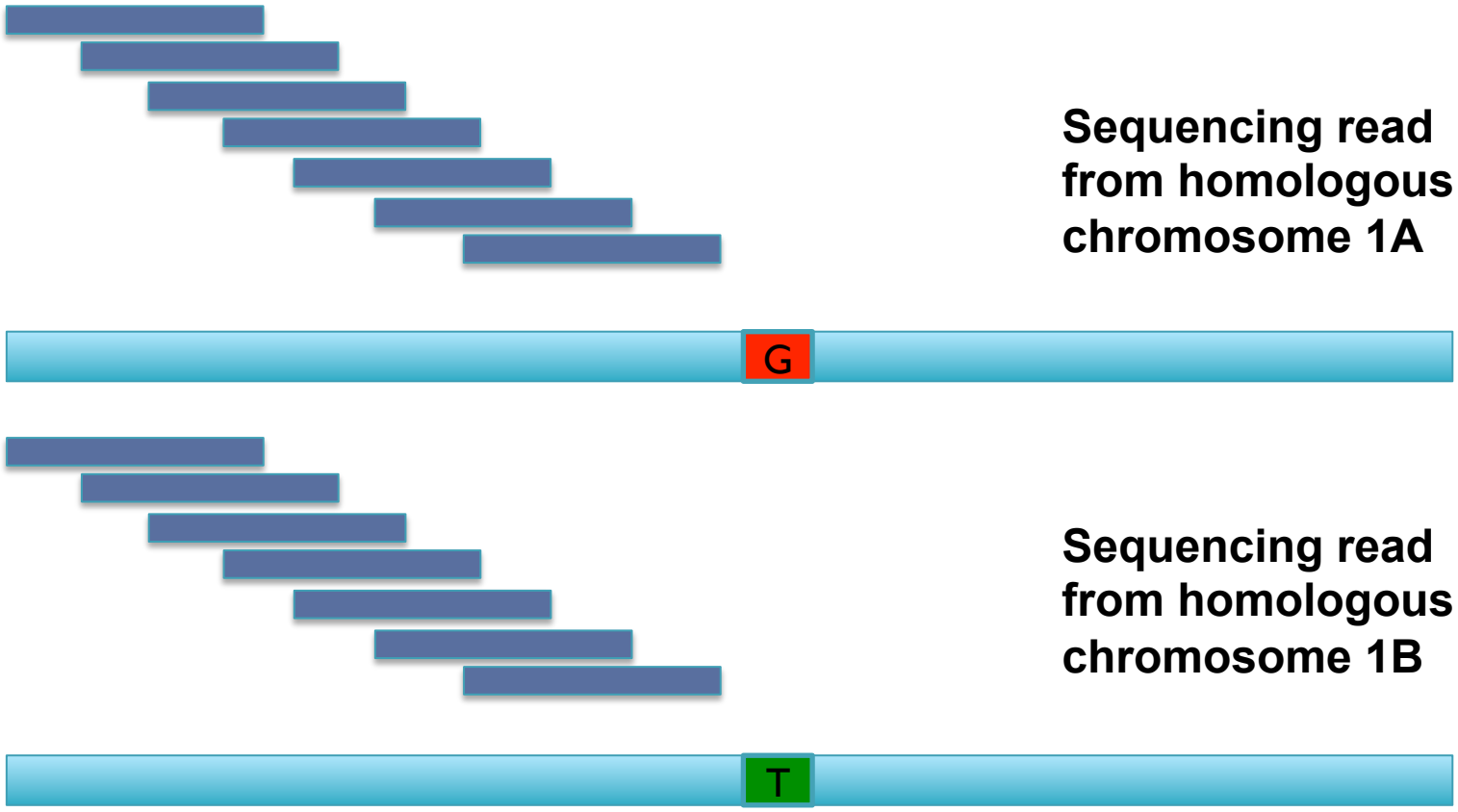
Sequencing read from homologous chromosome 1A

Sequencing read from homologous chromosome 1B

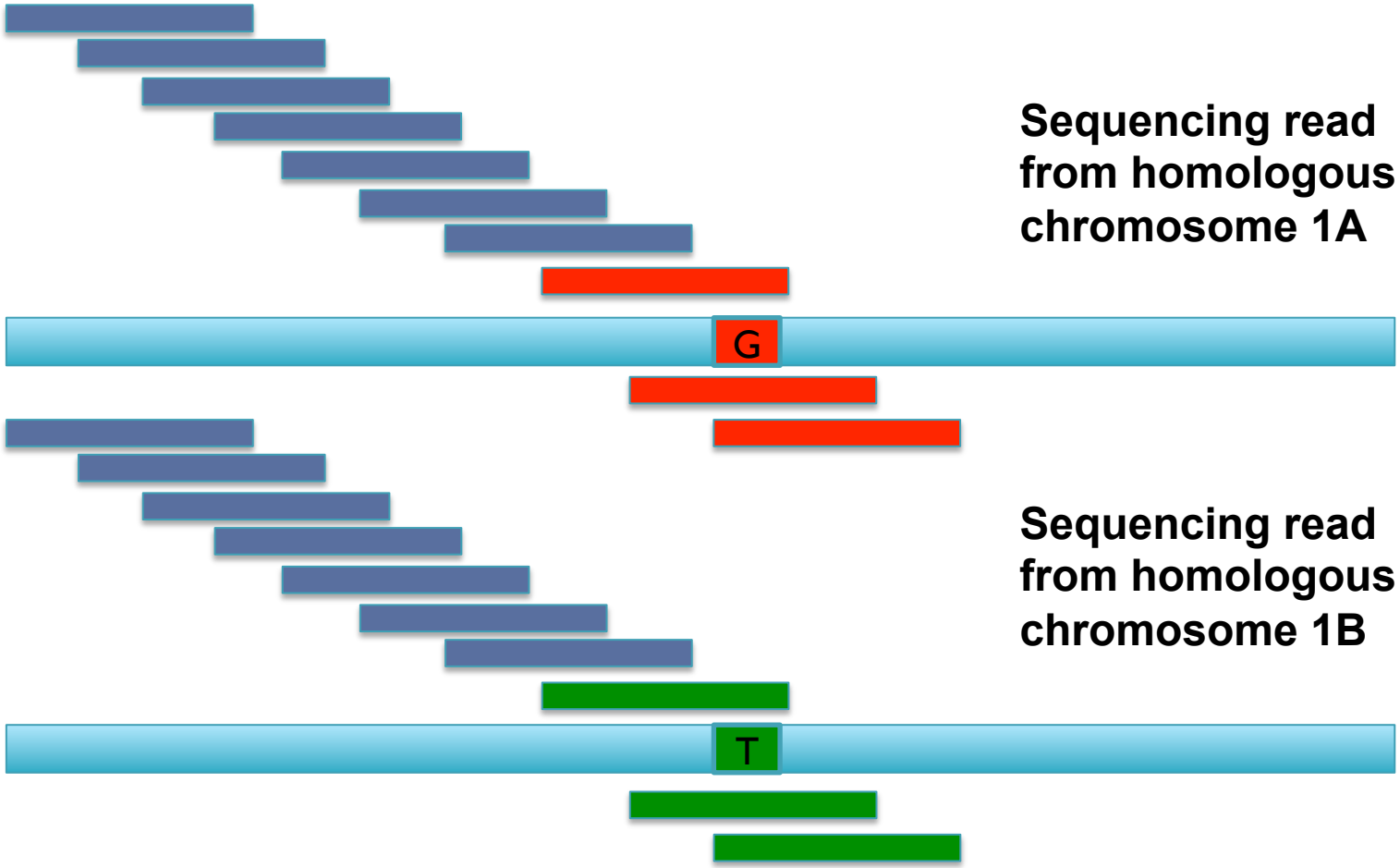
Heterozygous Kmer counting



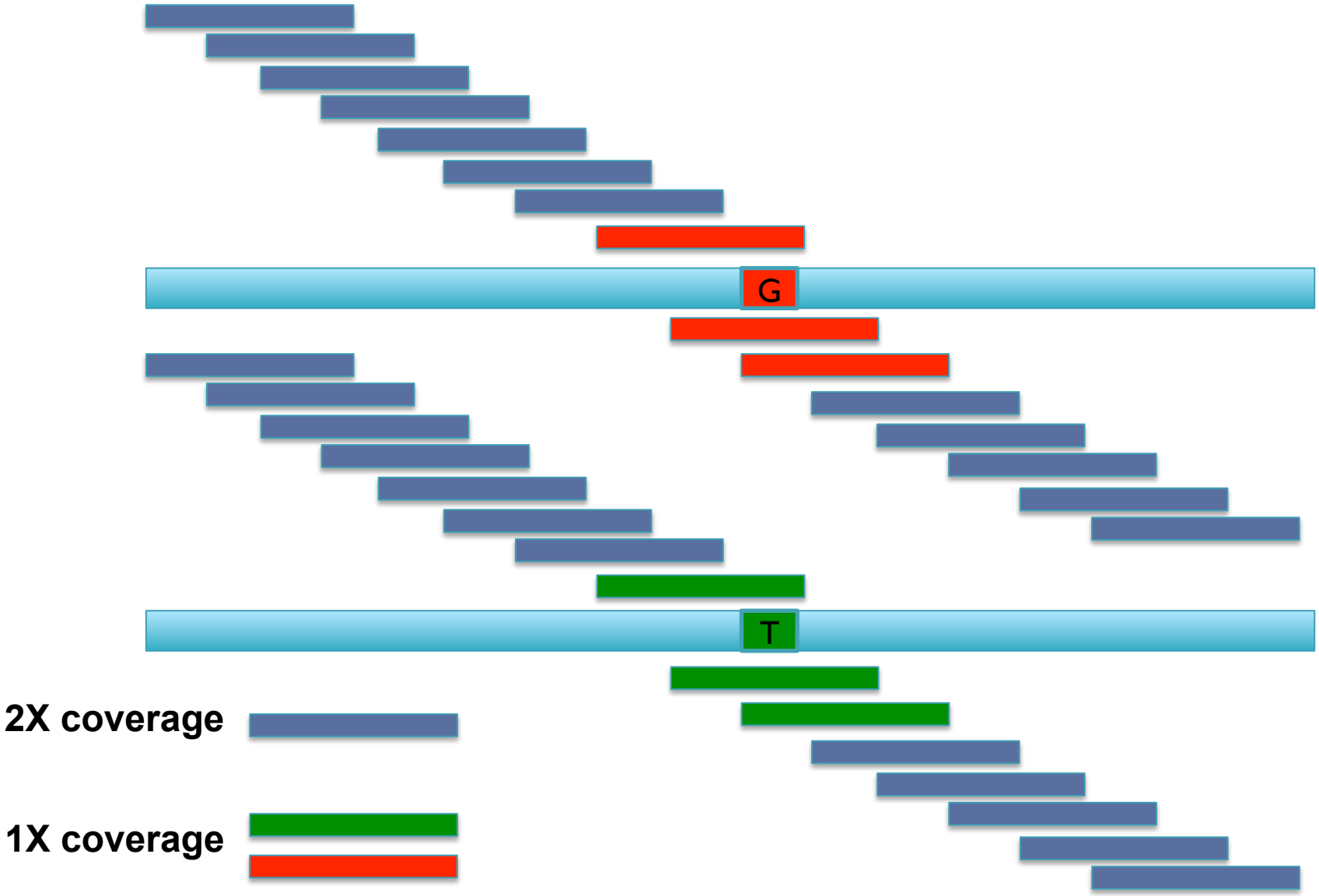
Heterozygous Kmer counting



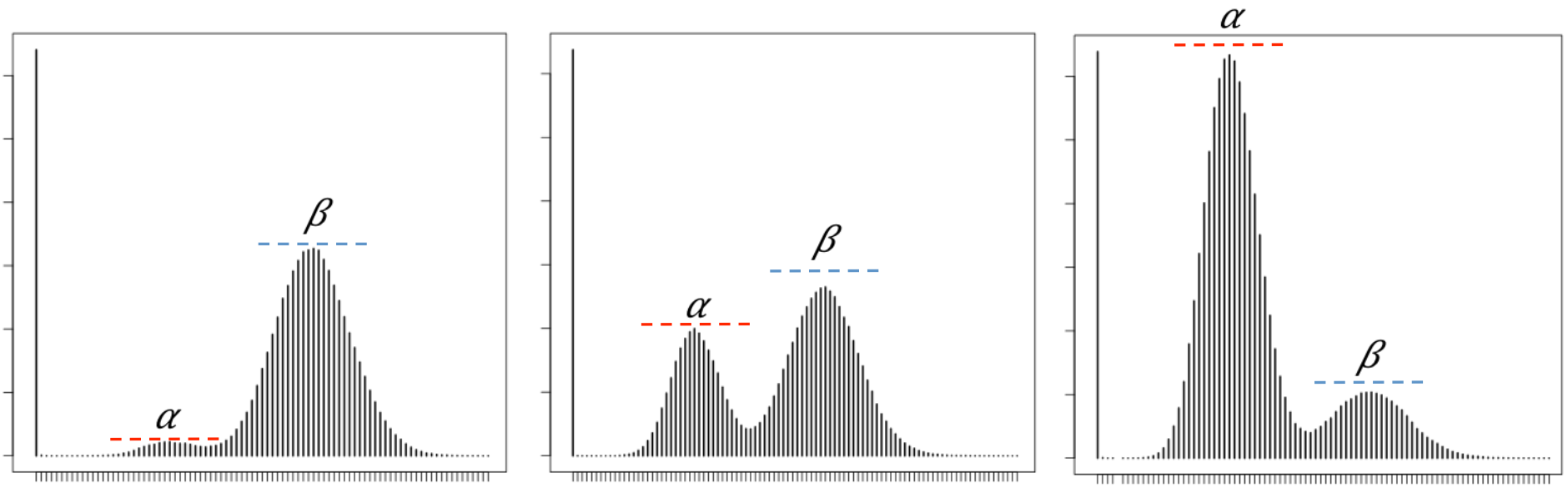
Heterozygous Kmer counting



Heterozygous Kmer counting



Heterozygous Kmer Profiles



0.1% heterozygosity **1% heterozygosity** **5% heterozygosity**

- ***Heterozygosity creates a characteristic "double-peak" in the Kmer profile***
  - Second peak at twice k-mer coverage as the first: heterozygous kmers average 50x coverage, homozygous kmers average 100x coverage

- ***Relative heights of the peaks is directly proportional to the heterozygosity rate***
  - The peaks are balanced at around 1.25% because each heterozygous SNP creates 2*k heterozygous kmers (typically k = 21)

GenomeScope Model

$$f(x) = G\Big\{\alpha NB(x, \lambda, \lambda/\rho) + \beta NB(x, 2\lambda, 2\lambda/\rho) + \gamma NB(x, 3\lambda, 3\lambda/\rho) + \delta NB(x, 4\lambda, 4\lambda/\rho)\Big\}$$

Analyze k-mer profiles using a mixture model of 4 negative binominal components

- Components centered at 1,2,3,4 * λ
- Four components capture heterozygous and homozygous unique (α,β) and 2 copy repeats (γ,δ). Higher order repeats do not contribute a significant number of kmers
- Negative binomial instead of Poisson to account for over dispersion observed in real data (especially PCR duplicates); variance modeled by ρ

$$\alpha = 2(1-d)(1-(1-r)^k) + 2d(1-(1-r)^k)^2 + 2d((1-r)^k)(1-(1-r)^k)$$

$$\beta = (1-d)((1-r)^k) + d(1-(1-r)^k)^2$$

$$\gamma = 2d((1-r)^k)(1-(1-r)^k)$$

$$\delta = d(1-r)^{2k}$$

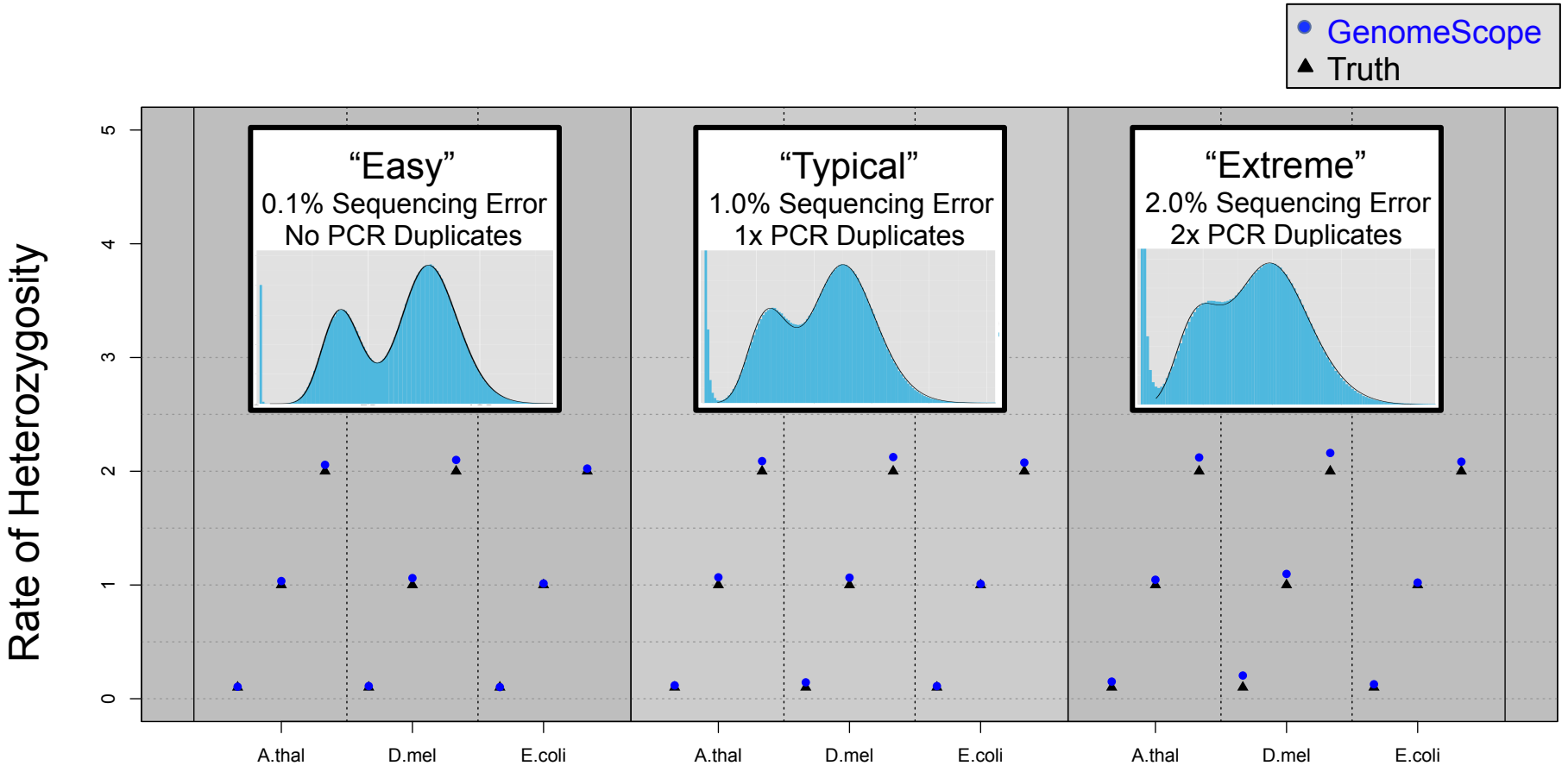
k is the *k-mer* length

r is the rate of heterozygosity

d represents the percentage of the genome that is two-copy repeat

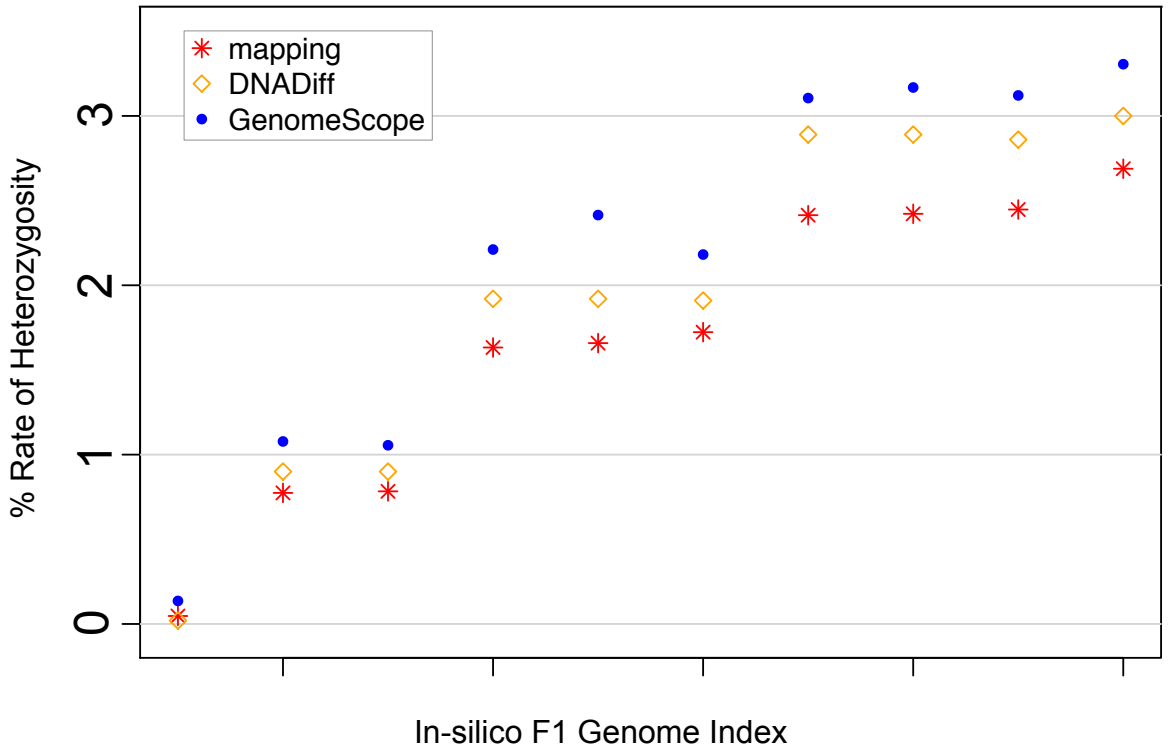
Fit model with nls, infer rate of heterozygosity, genome size, unique/repetitive content, sequencing error rate, rate of PCR duplicates

Simulated Results



Introduce SNPs into A. thaliana, D. melanogaster, or E. coli at known rates, simulate shotgun sequencing with specified rates of sequencing error and PCR duplications

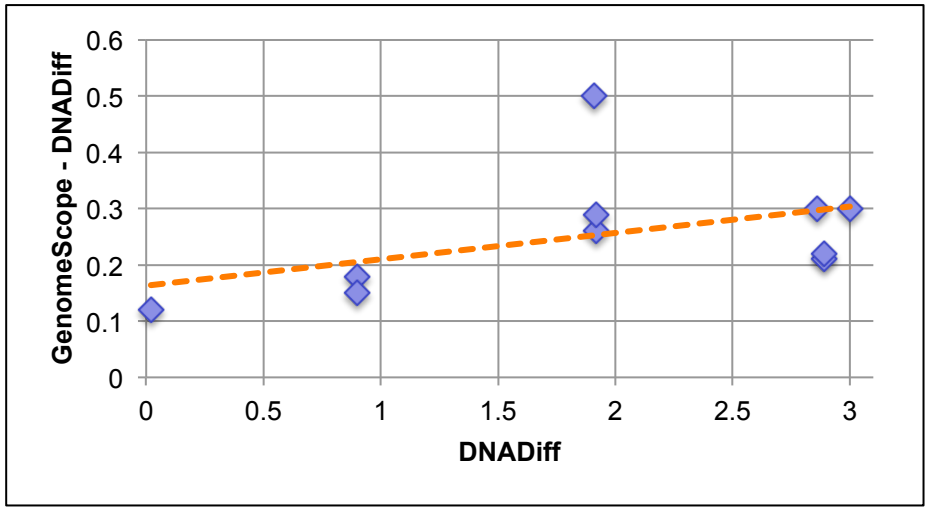
In silico E. coli population sequencing
"Synthetic F1 Genome"

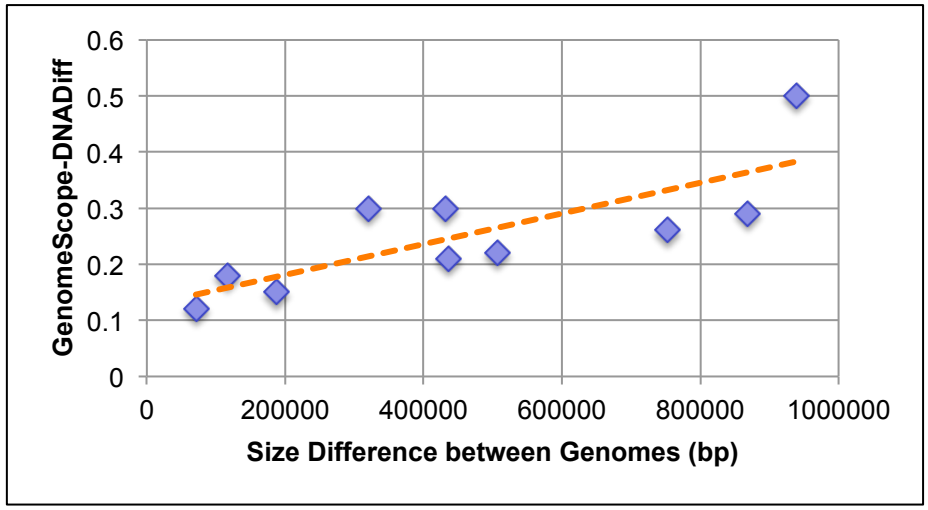


Mix equal numbers of real Illumina reads from pairs of 5 different E. coli isolates that have finished genomes with varying rates of similarity

Compare results to mapping pipeline (BWA+SAMTools) and MUMmer/DNADiff

Understanding DNADiff



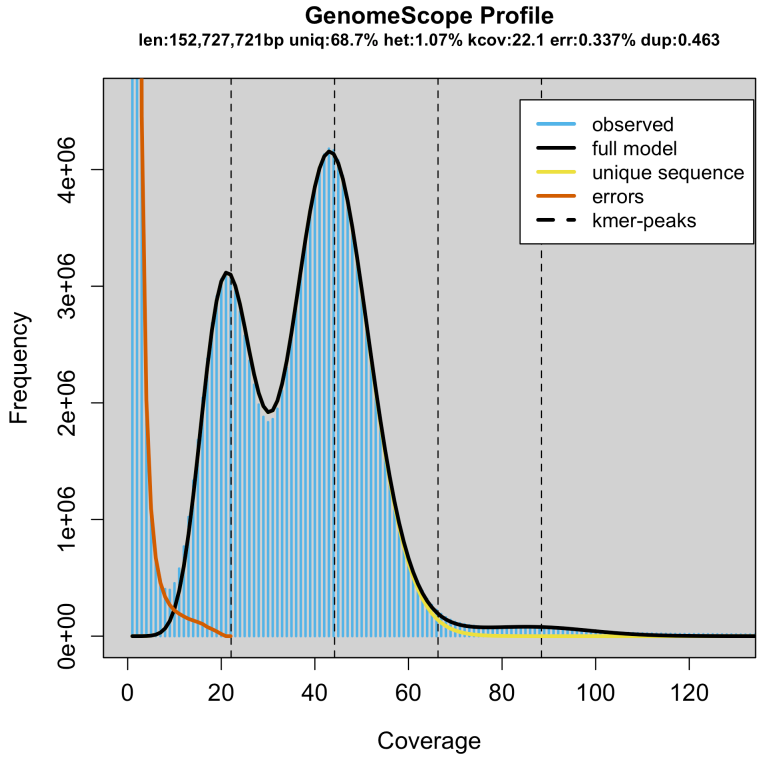


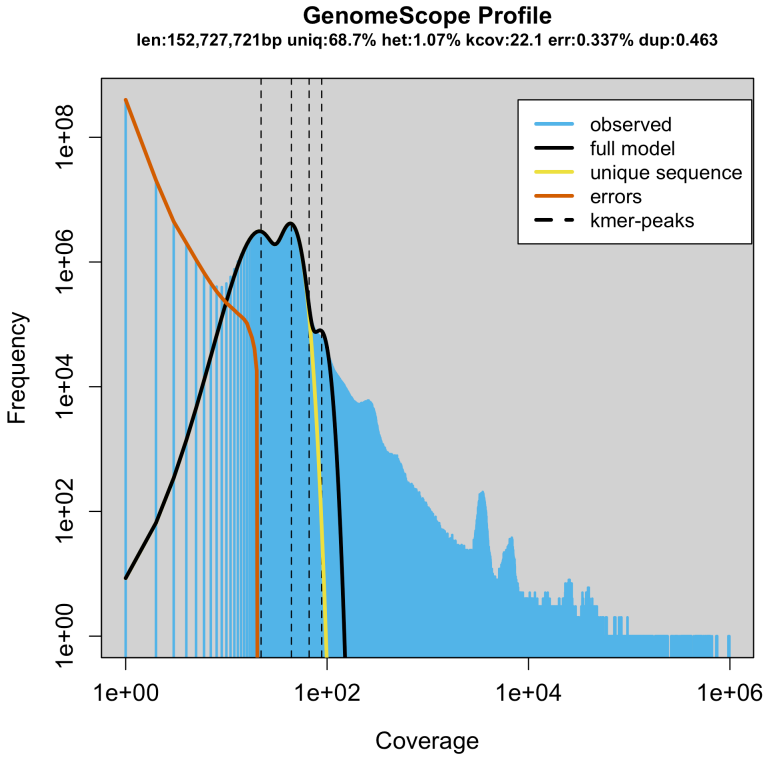
Observe that the difference between the rate of heterozygosity estimated by GenomeScope was generally higher than DNADiff, and that it was correlated with the rate of heterozygosity

The difference was strongly correlated with the size difference between the genomes

Conclude that DNADiff is underestimating the true rate because it doesn't include bases in regions that don't align!

GenomeScope: Fast genome analysis from short reads





Evaluated on several genomes with published rates of heterozygosity:

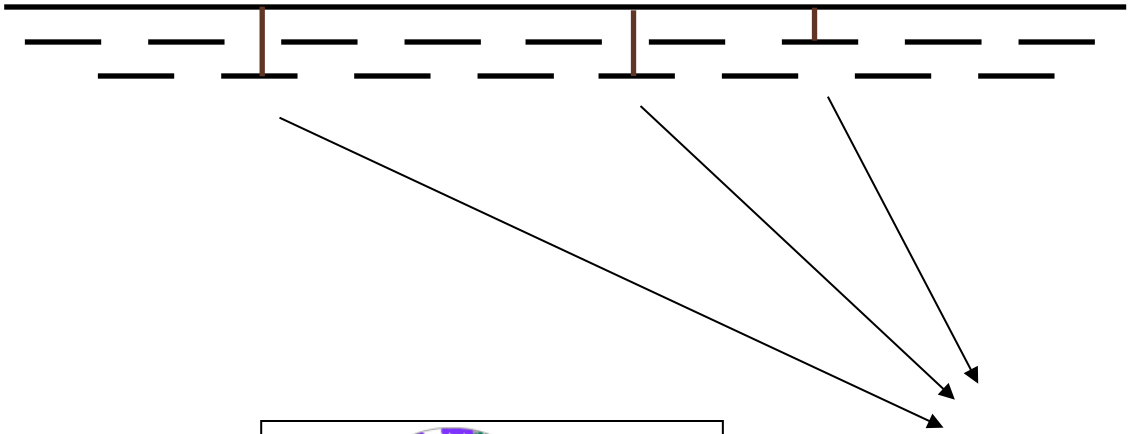
- *L. calcarifer* (Asian seabass), *D. melanogaster* (fruit fly), *M. undulates* (budgerigar), *A. thaliana* Col-Cvi F1 (thale cress), *P. bretschneideri* (pear), C. gigas (Pacific oyster)
- Agrees well with published results:
  - Rate of heterozygosity is typically higher but likely correct.
  - Genome size of plants inflated by organelle sequences (exclude very high freq. kmers)

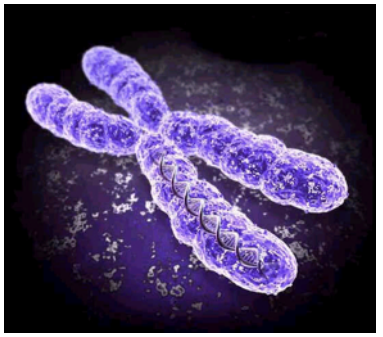
Vurture, GW*, Sedlazeck FJ*, Nattestad, M, Underwood, C, Fang, H, Gurtowski, J, Schatz, MC. *bioRxiv*

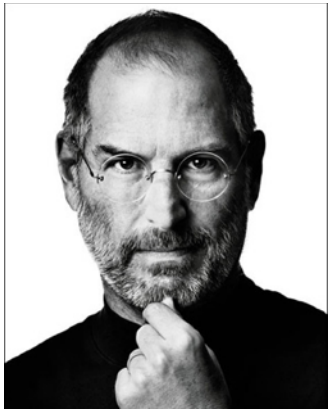
Part I:
Quick Review

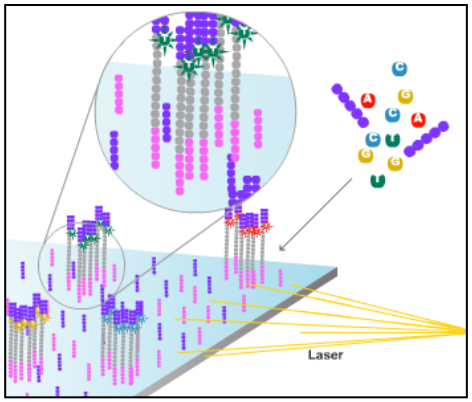
Personal Genomics

How does your genome compare to the reference?







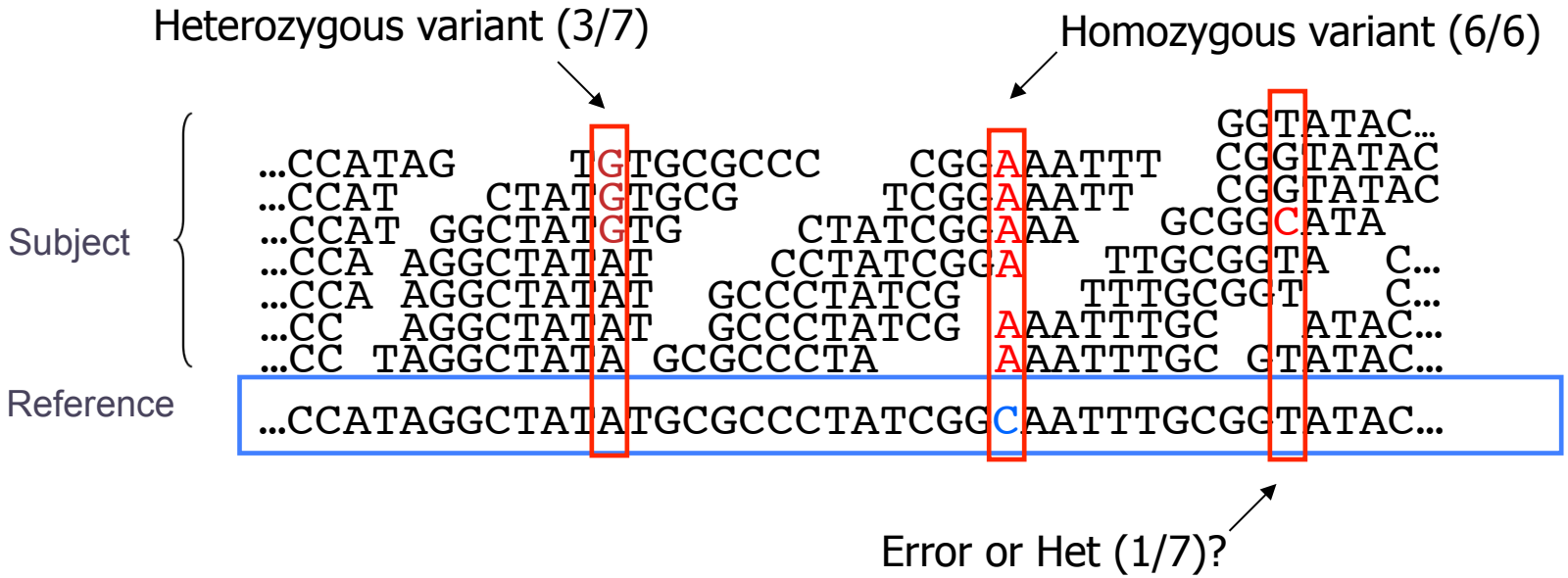


Heart Disease

Cancer

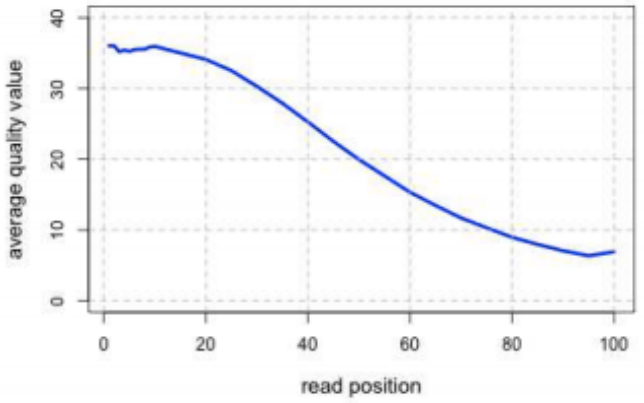
Creates magical technology

Genotyping Theory



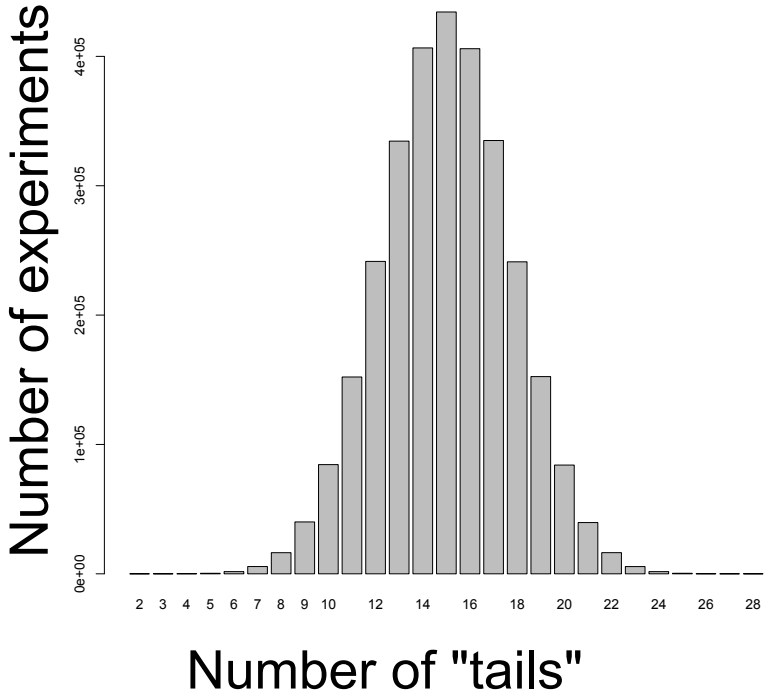
- If there were no sequencing errors, identifying SNPs would be very easy: any time a read disagrees with the reference, it must be a variant!

- Sequencing instruments make mistakes
  - Quality of read decreases over the read length



- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times

So, with 30 tosses (reads), we are much more likely to see an even mix of alternate and reference alleles at a heterozygous locus in a genome



This is why at least a "30X" (30 fold sequence coverage) genome is recommended: it confers sufficient power to distinguish heterozygous alleles and from mere sequencing errors

P(3/30 het) <?> P(3/30 err)

So, with 30 tosses (reads), we are much more likely to see an even mix o at a e

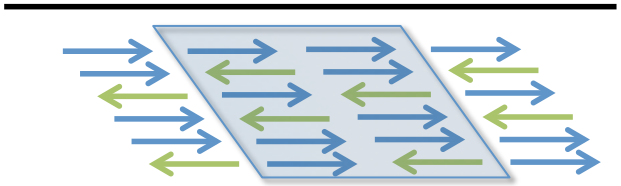
Number of experiments

ast a uence ne is onfers stinguish and from errors

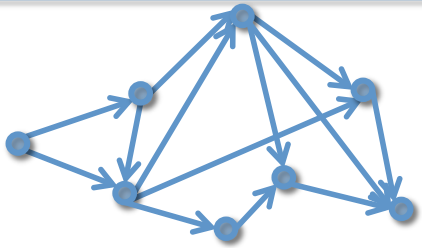
/30 err)

Scalpel Algorithm

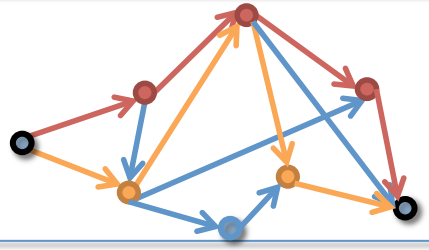
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



Decompose reads into overlapping *k*-mers and construct de Bruijn graph from the reads

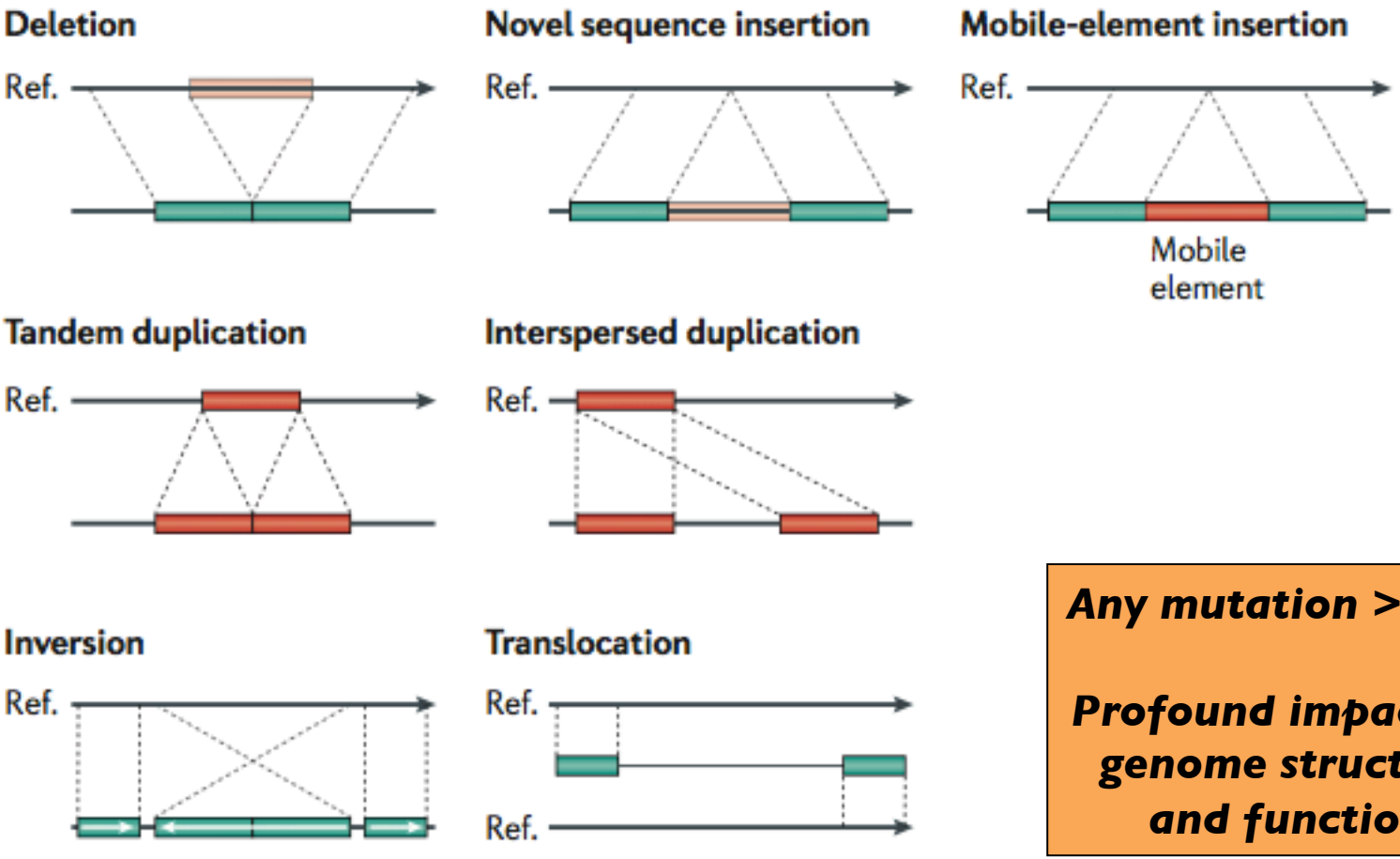


Find end-to-end haplotype paths spanning the region



Align assembled sequences to reference to detect mutations

Structural Variations



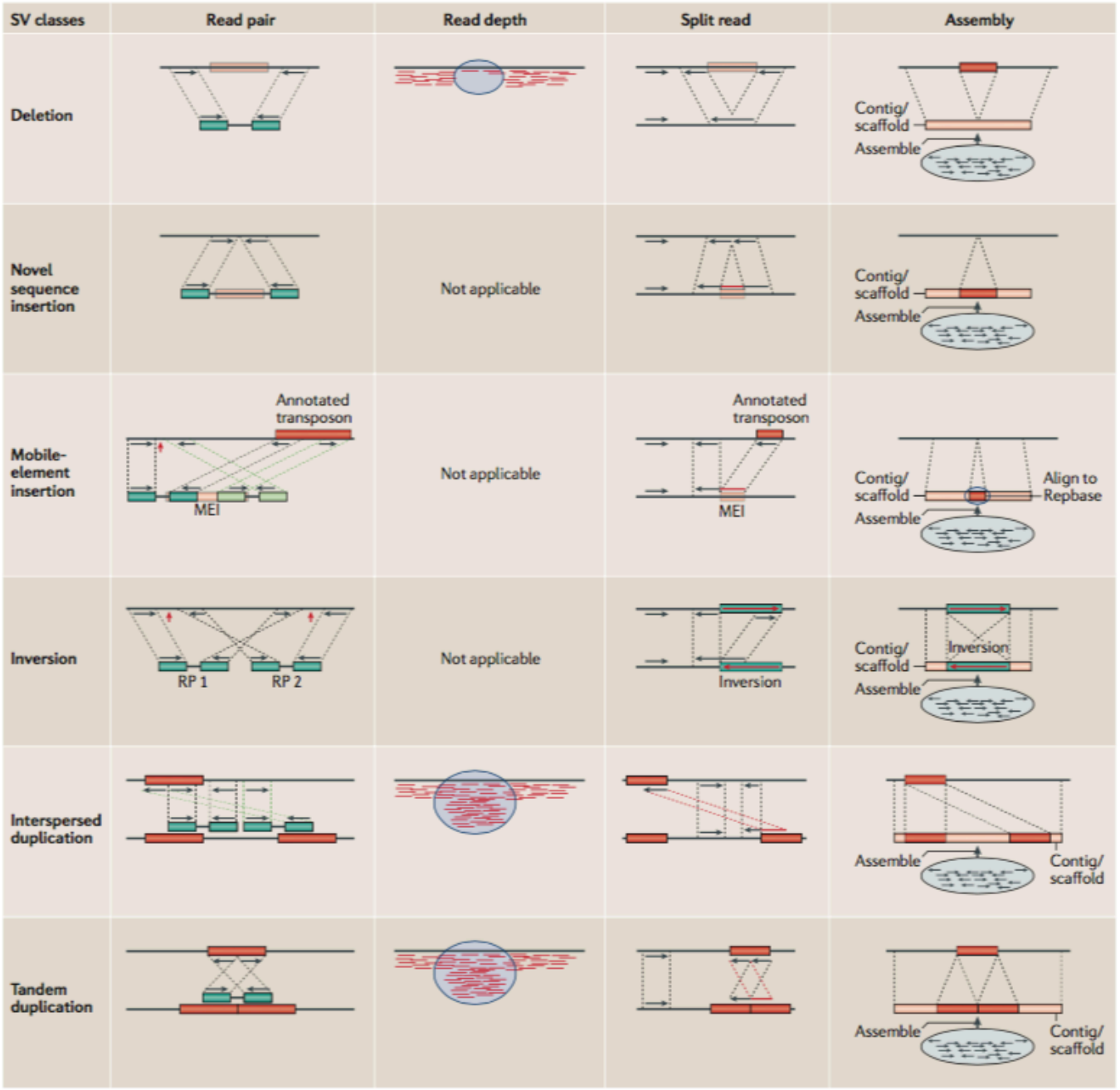
Any mutation >50bp

Profound impact on genome structure and function

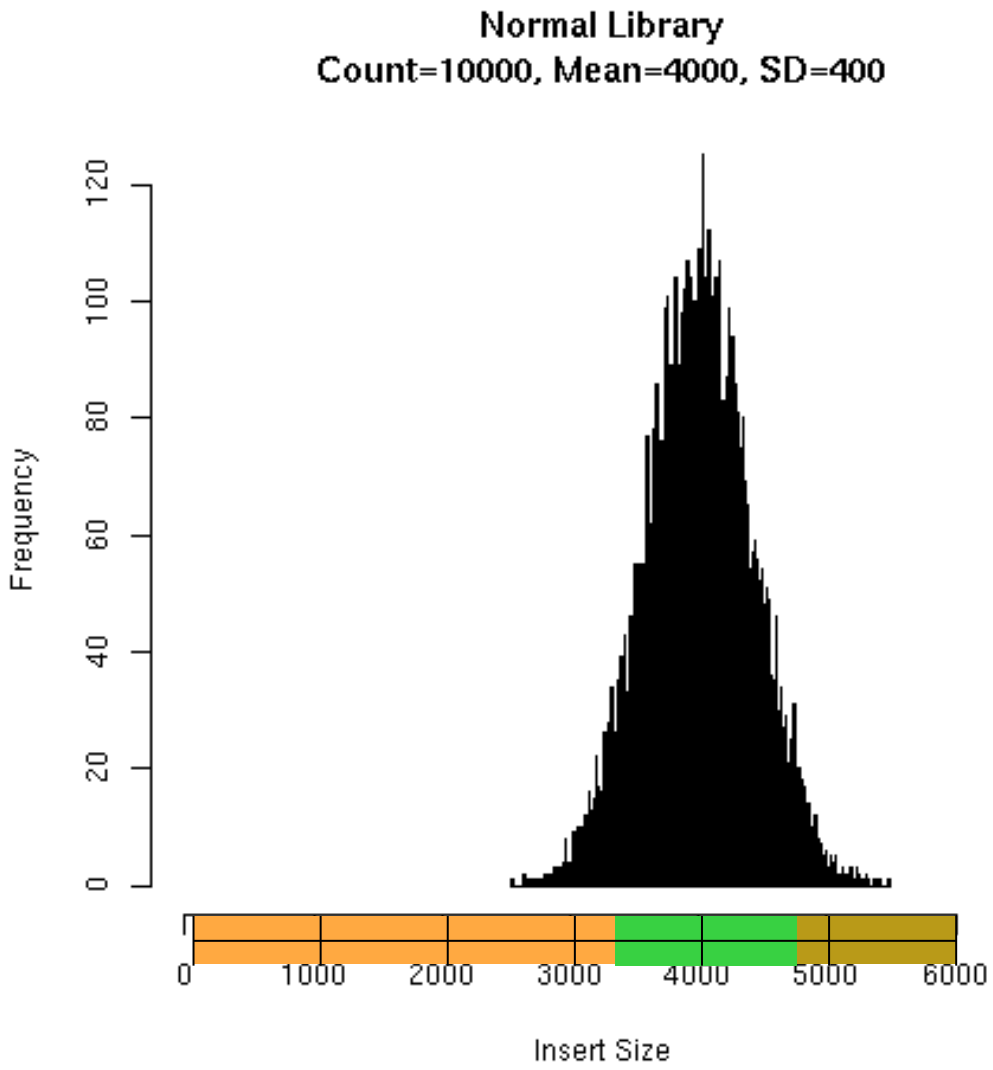
Genome structural variation discovery and genotyping

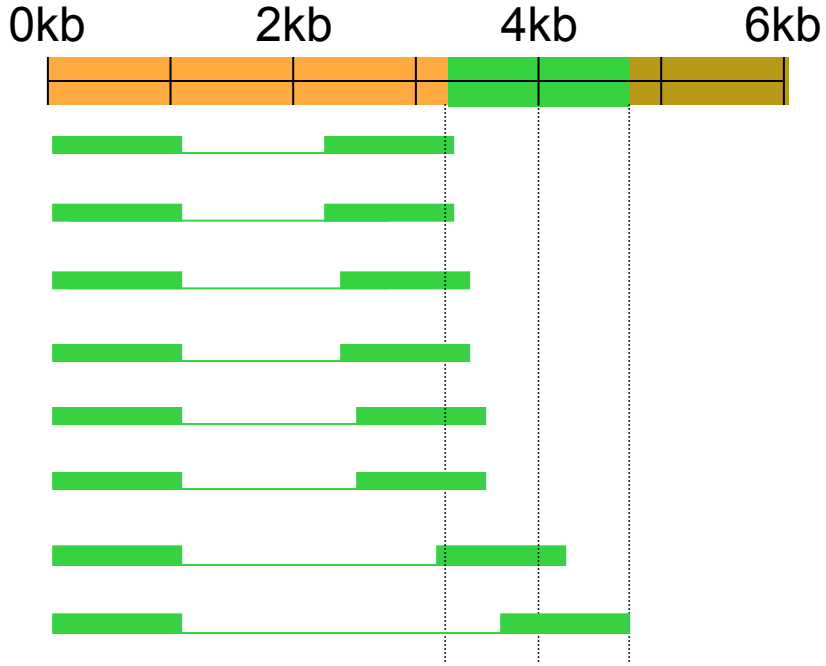
Alkan, C, Coe, BP, Eichler, EE (2011) *Nature Reviews Genetics*. May;12(5):363-76. doi: 10.1038/nrg2958.

Structural Variation Sequence Signatures



C/E-Statistic: Compression





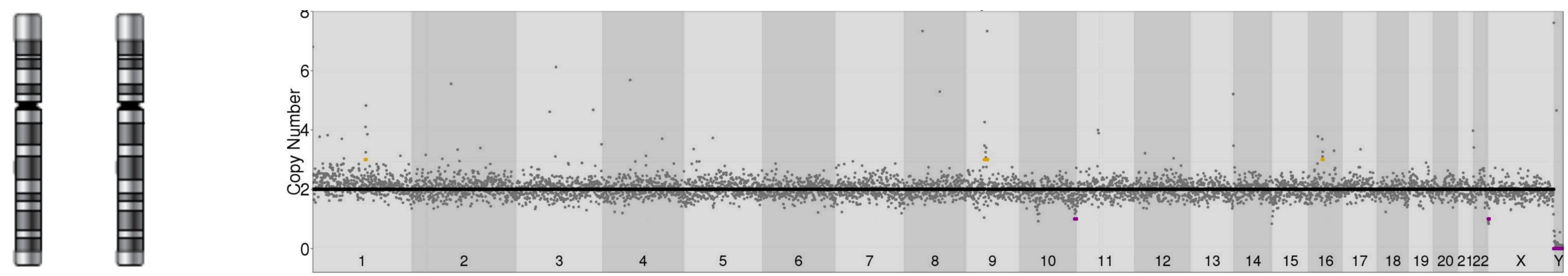
8 inserts: 3.2 kb-4.8kb

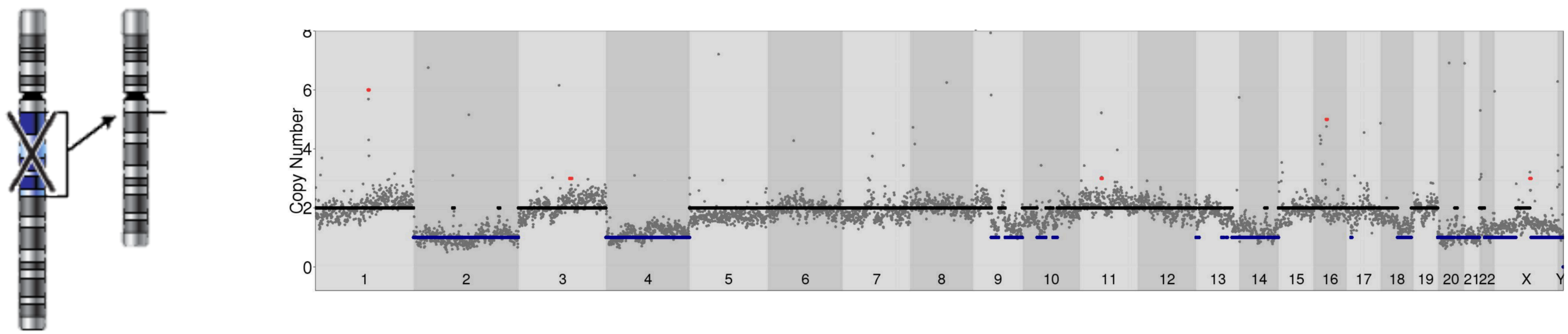
Local Mean: 3488

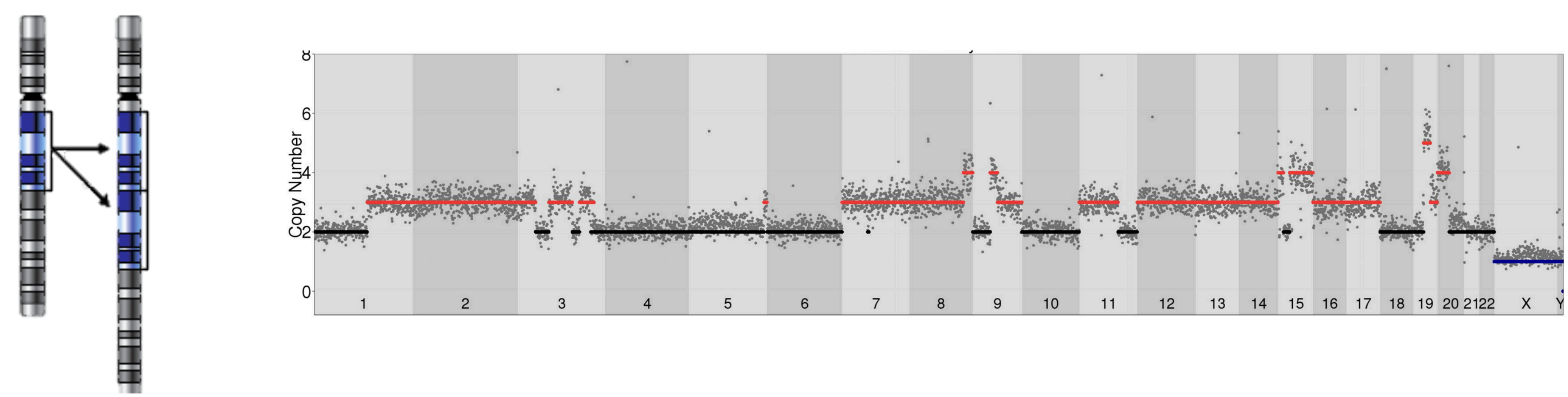
C/E Stat: $\frac{(3488-4000)}{(400 / \sqrt{8})} = -3.62$

C/E Stat ≤ -3.0 indicates Compression

What are Copy Number Variations?







A probabilistic framework for SV discovery

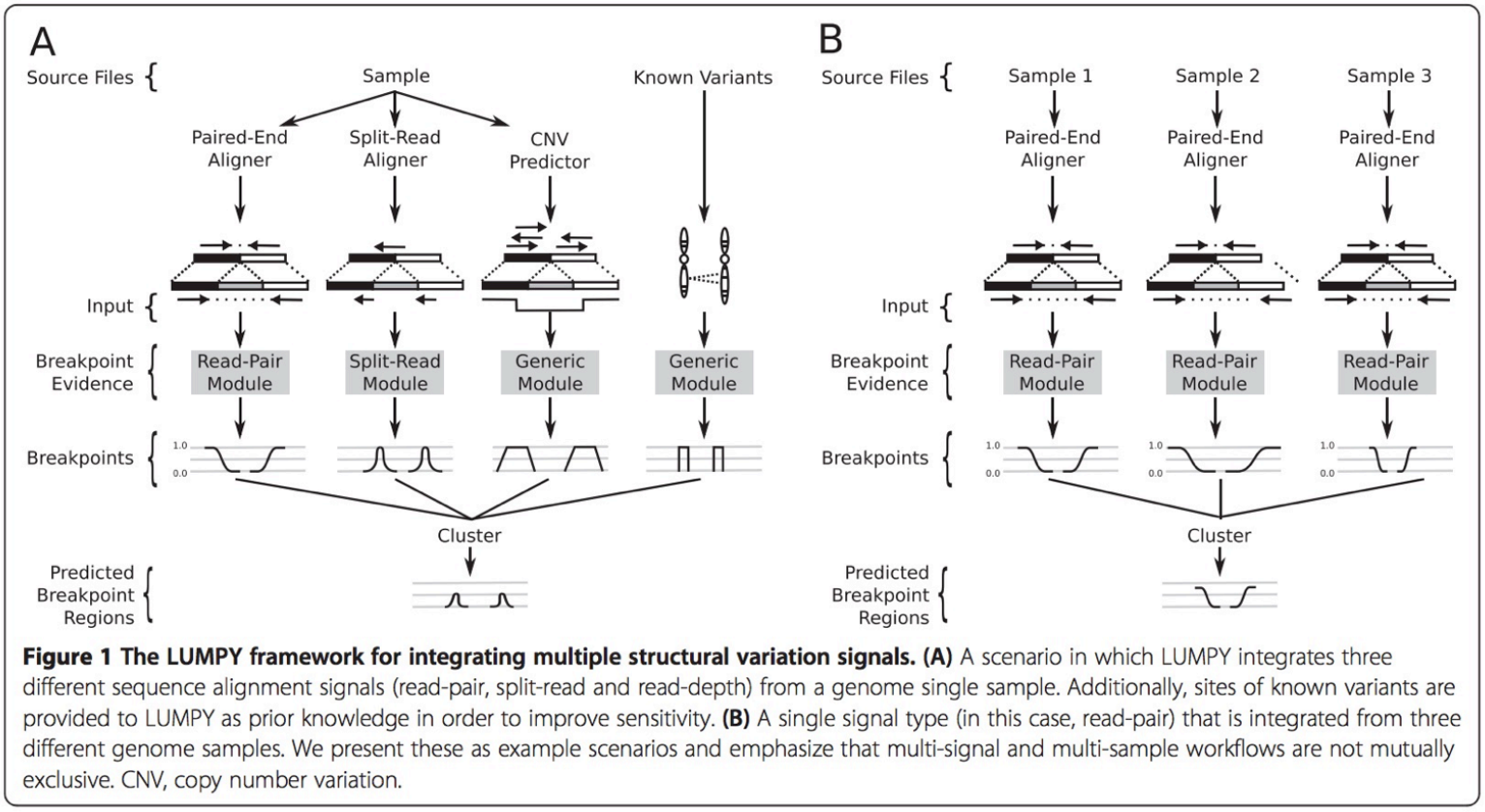


Figure 1 The LUMPY framework for integrating multiple structural variation signals. (A) A scenario in which LUMPY integrates three different sequence alignment signals (read-pair, split-read and read-depth) from a genome single sample. Additionally, sites of known variants are provided to LUMPY as prior knowledge in order to improve sensitivity. **(B)** A single signal type (in this case, read-pair) that is integrated from three different genome samples. We present these as example scenarios and emphasize that multi-signal and multi-sample workflows are not mutually exclusive. CNV, copy number variation.

Lumpy integrates paired-end mapping, split-read mapping, and depth of coverage for better SV discovery accuracy

Ryan Layer

The dirty secrets of SV discovery

Secret #1: Often many false positives

- Short reads + heuristic alignment + rep. genome
  = systematic alignment artifacts (false calls)
- Chimeras and duplicate molecules
- Ref. genome errors (e.g., gaps, mis-assemblies)
- **ALL** SV mapping studies use strict filters for above

Secret #2: The false negative rate is also typically high

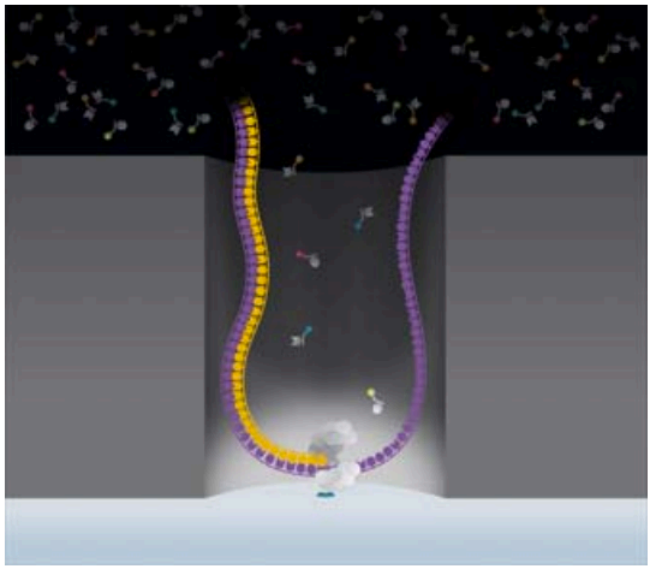
- Most current datasets have low to moderate ***physical*** coverage due to small insert size (~10-20X)
- Breakpoints are enriched in repetitive genomic regions that pose problems for sensitive read alignment
- FILTERING!
- The false negative rate is usually hard to measure, but is thought to be extremely high for most paired-end mapping studies (>30%)
- When searching for spontaneous mutations in a family or a tumor/normal comparison, a false negative call in one sample can be a false positive somatic or de novo call in another.

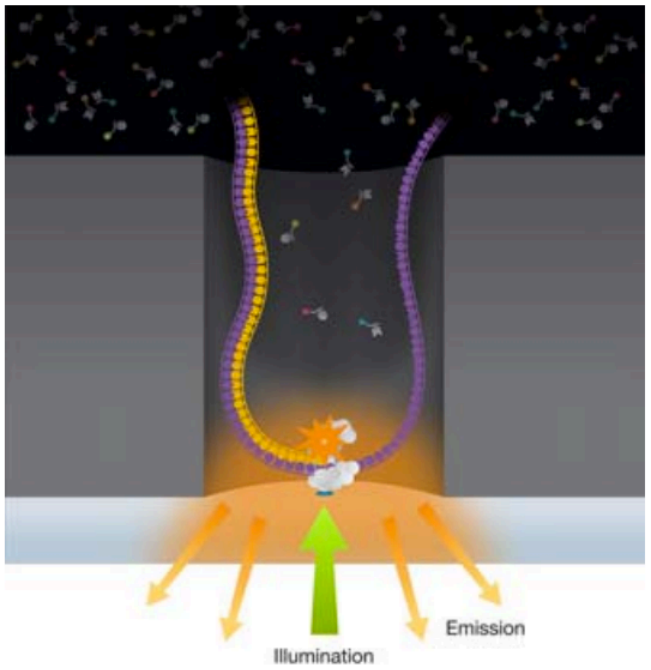
Part 2:
3rd Generation Sequencing

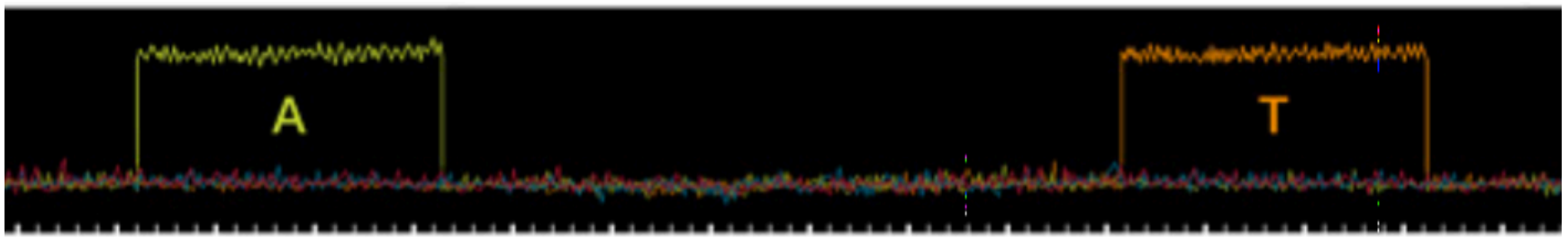
PacBio Single Molecule Real Time Sequencing (SMRT-sequencing)

PacBio: SMRT Sequencing

Imaging of florescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).

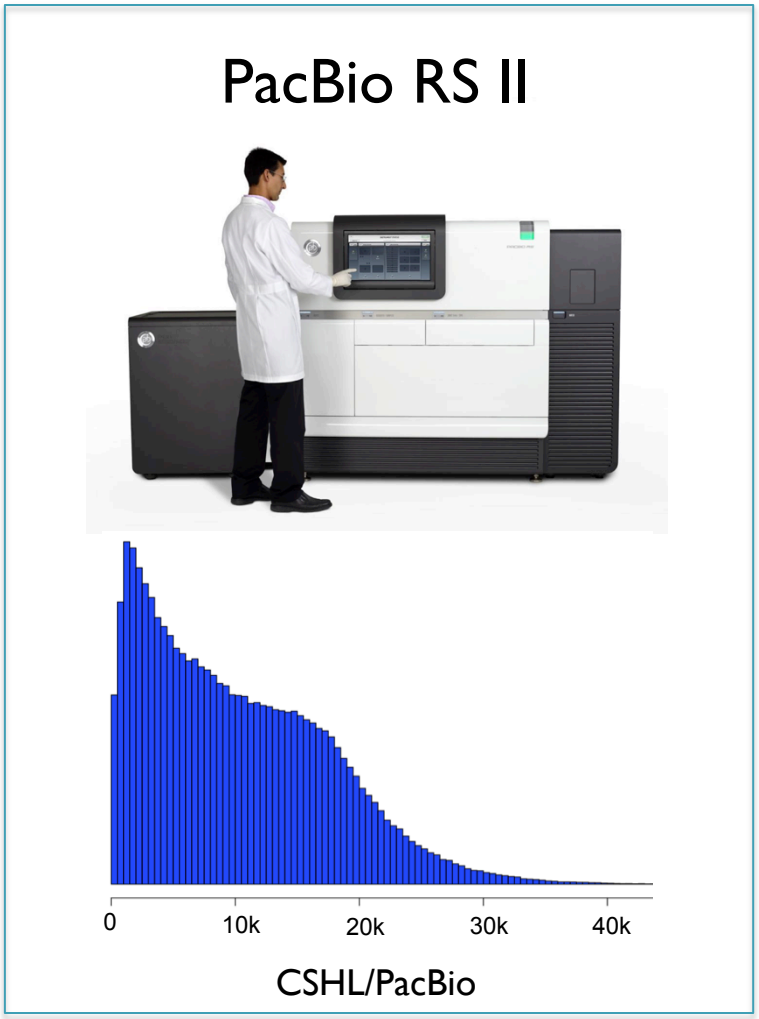






http://www.youtube.com/watch?v=v8p4ph2MAvI

SMRT Sequencing Data




```
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

ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACTTATTCACAATTAG
||||||| ||||||||  |||||| ||||| ||||||||||||||||||
ATCCAGT-GAAAATATA--TTATGC-ATCCA-GAACTTATTCACAATTAG
```

Sample of 100k reads aligned with BLASR requiring >100bp alignment
Average overall accuracy: 83.7%, 11.5% insertions, 3.4% deletions, 1.4% mismatch

LETTER

doi:10.1038/nature13907

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

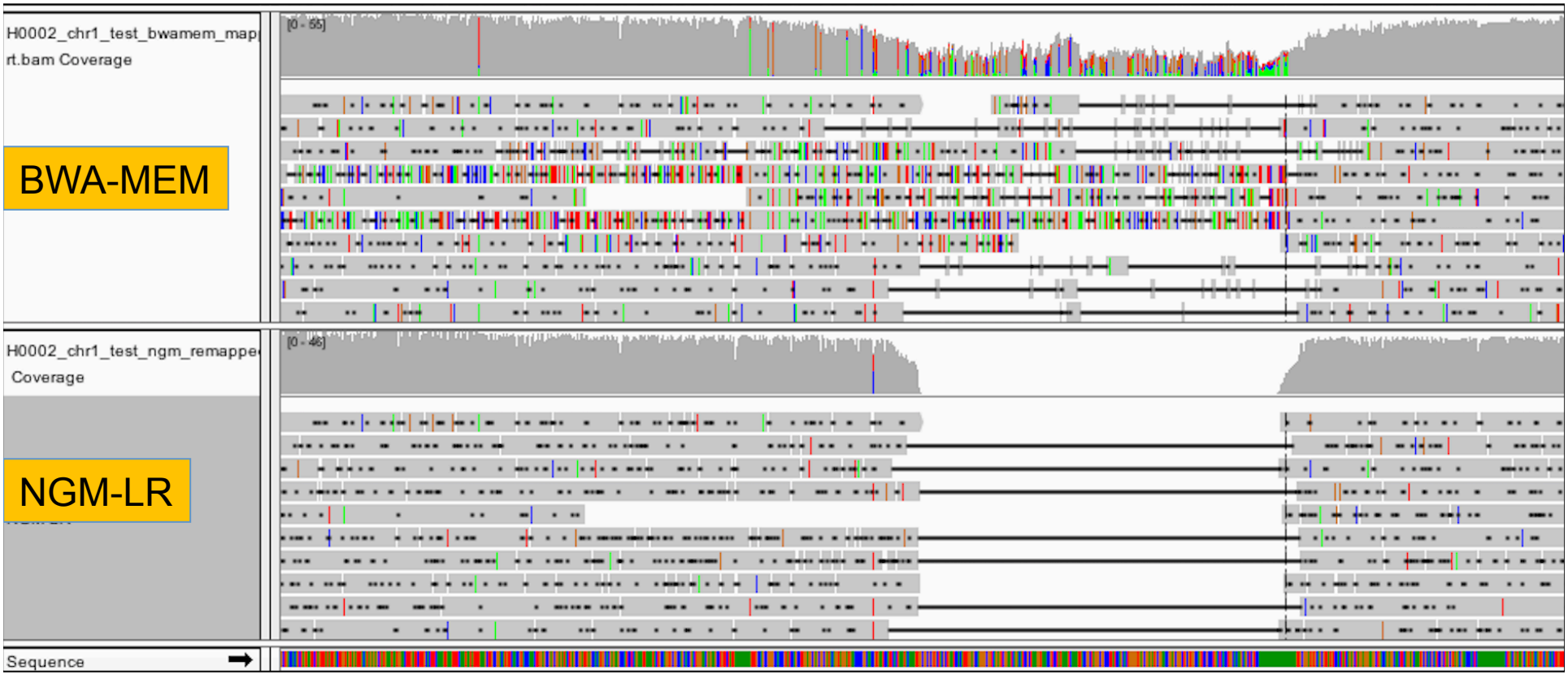
Mark J. P. Chaisson[1], John Huddleston[1,2], Megan Y. Dennis[1], Peter H. Sudmant[1], Maika Malig[1], Fereydoun Hormozdiari[1], Francesca Antonacci[3], Urvashi Surti[4], Richard Sandstrom[1], Matthew Boitano[5], Jane M. Landolin[5], John A. Stamatoyannopoulos[1], Michael W. Hunkapiller[5], Jonas Korlach[5]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

NGM-LR + Sniffles: PacBio SV Analysis Tools

https://github.com/philres/ngmlr & https://github.com/fritzsedlazeck/Sniffles



Improved SV Variant Detection with long reads

1. NGM-LR: Improve mapping of noisy long reads: improved seeding, convex gap scoring

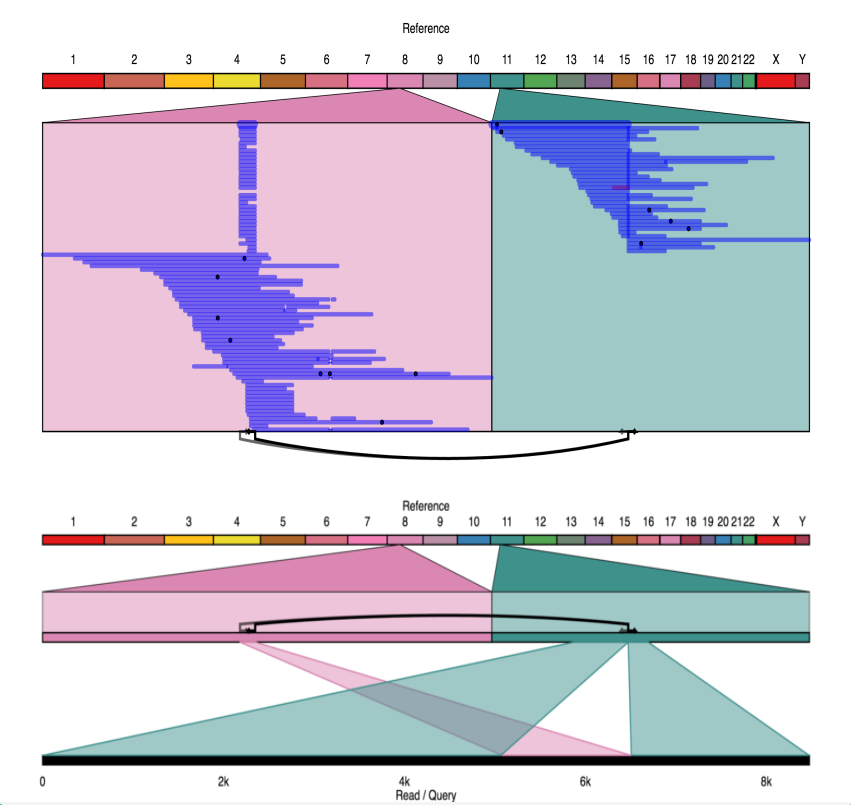
2. Sniffles: Integrates evidence from split-reads, alignment fidelity, breakpoint concordance

Sniffles PacBio Variant Calls

Sniffles calls

	All SVs (50bp+)	Large SVs (10kbp+)
Deletions	7,389	164
Duplications	1,284	139
Insertions	8,382	4
Inversions	229	116
Translocations	170	170
All	***17,454***	***593***

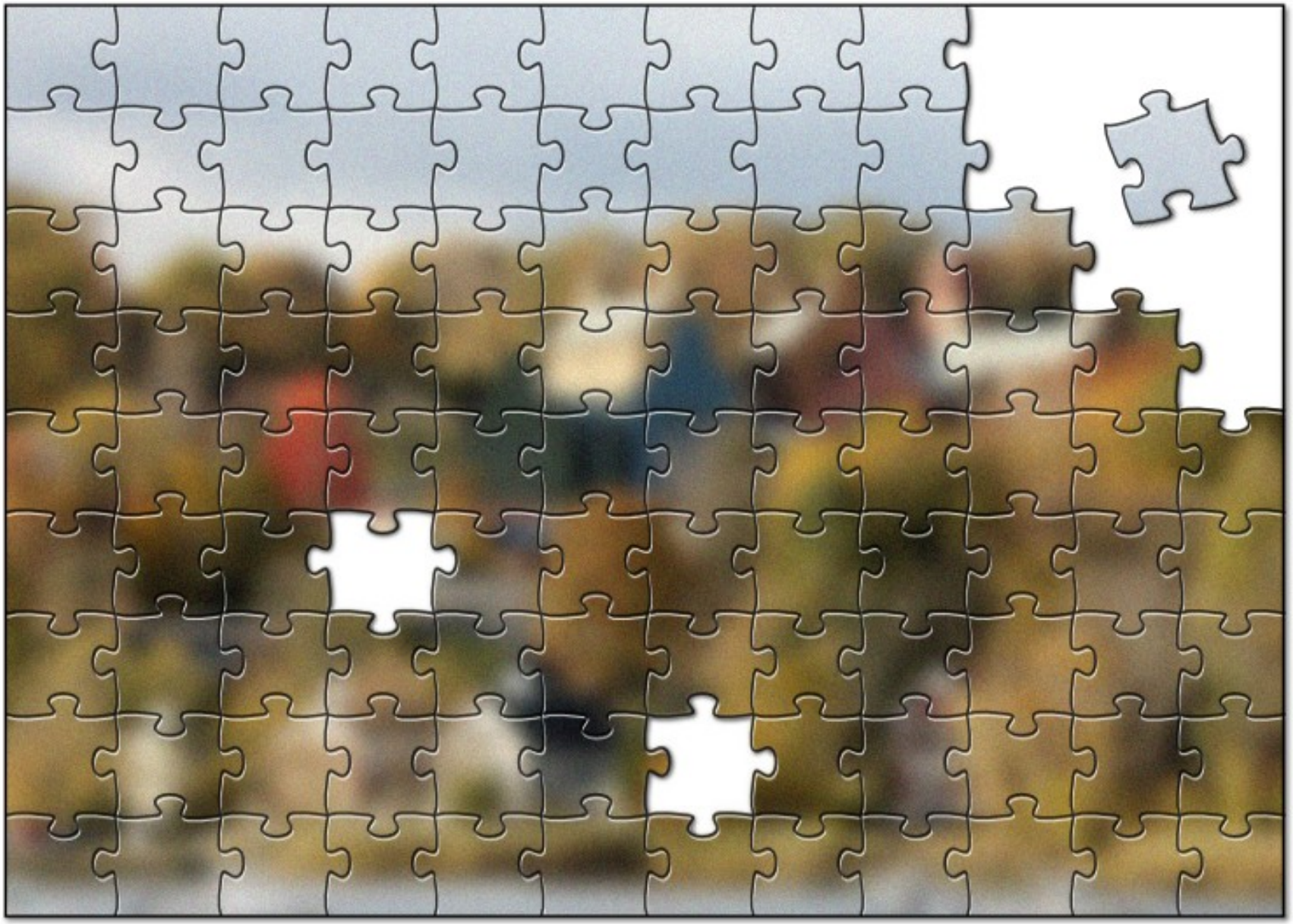
Translocation in Ribbon



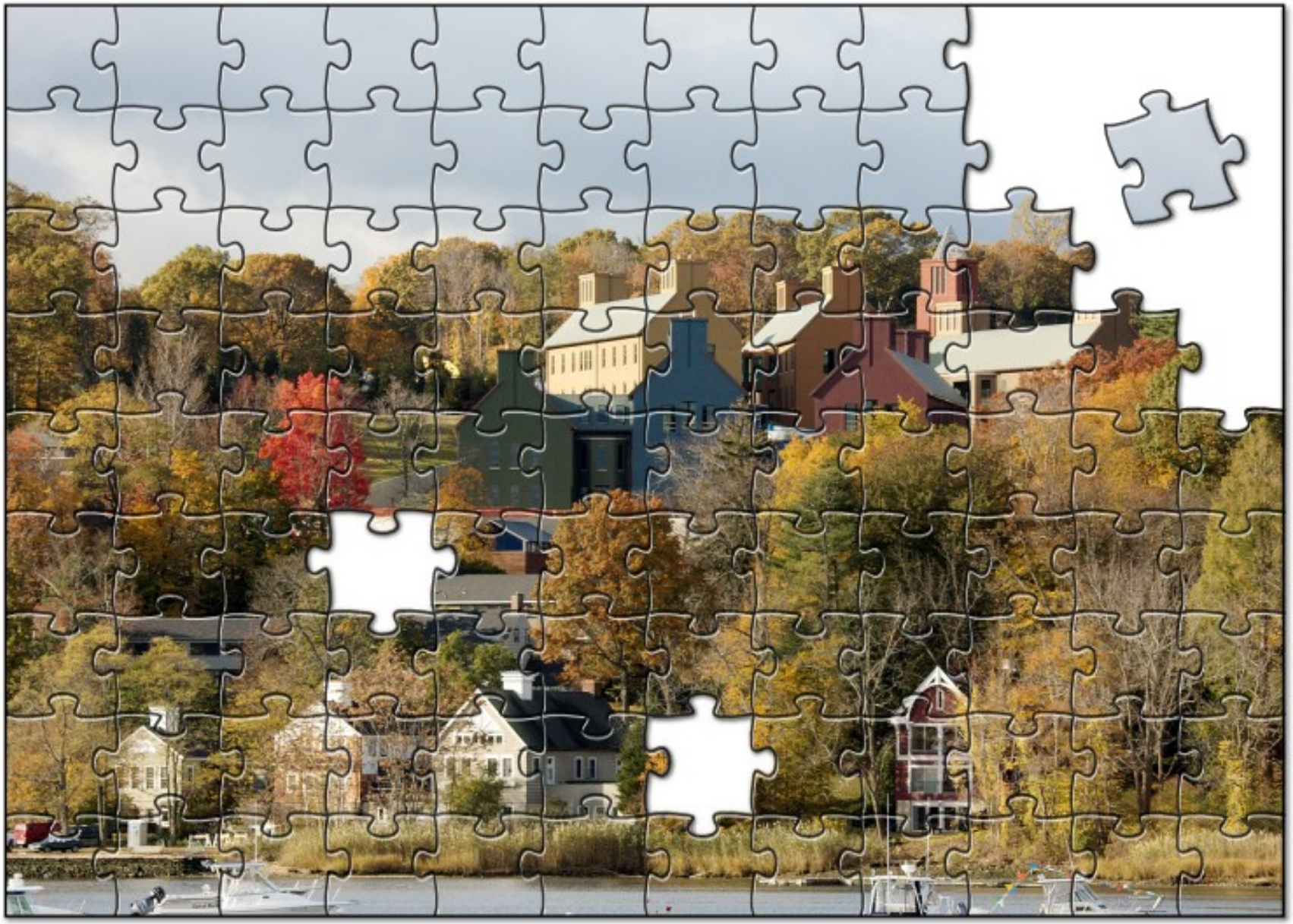
Ribbon: Visualizing complex genome alignments and structural variation
Nattestad et al. (2016) *bioRxiv* doi: http://dx.doi.org/10.1101/082123

Structural Variation Sequence Signatures

Single Molecule Sequences

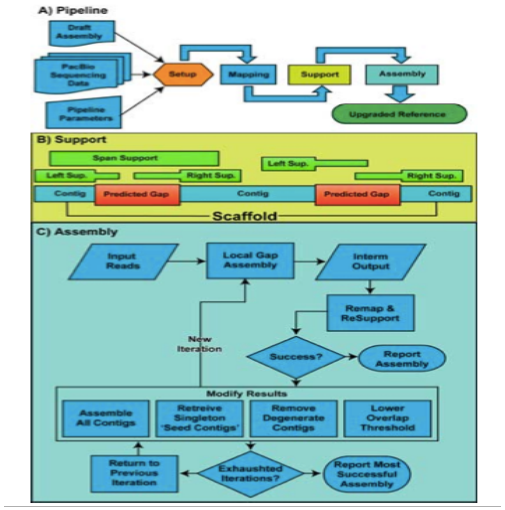


“Corrective Lens” for Sequencing



PacBio Assembly Algorithms

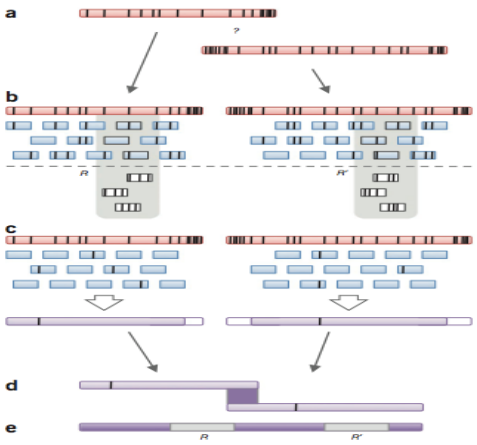
PBJelly



Gap Filling and Assembly Upgrade

English *et al* (2012)
PLOS One. 7(11): e47768

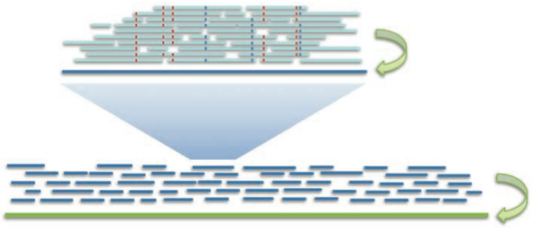
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

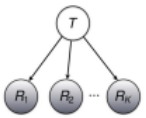
Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

MHAP/FALCON & Quiver



$$\Pr(\mathbf{R} \mid T)$$

$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$



Quiver Performance Results
Comparison to Reference Genome
(M. ruber ; 3.1 MB ; SMRT® Cells)

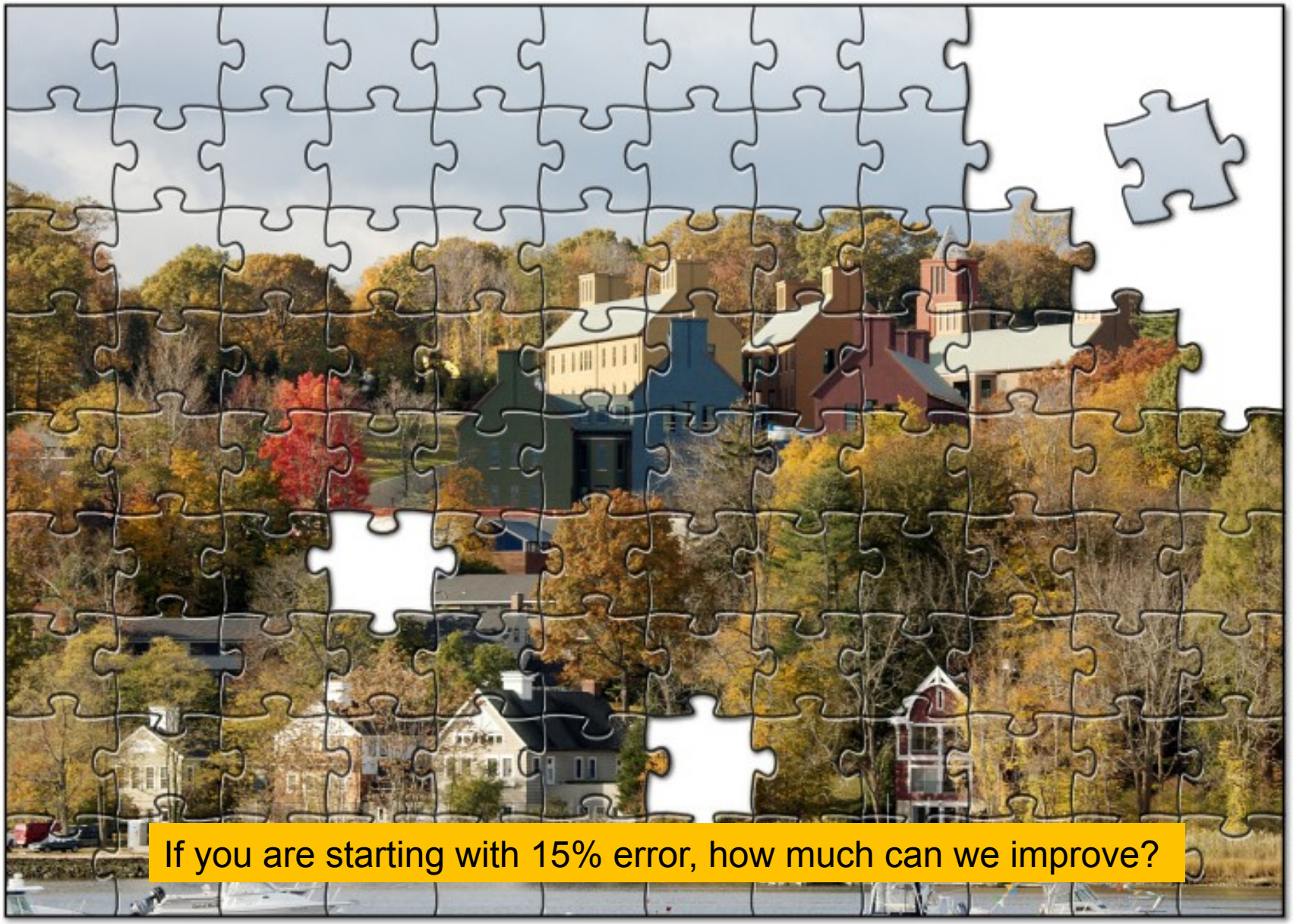
	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

PB-only Correction & Polishing

Chin *et al* (2013)
Nature Methods. 10:563–569

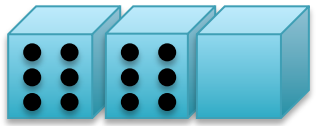
< 5x PacBio Coverage > 50x

“Corrective Lens” for Sequencing

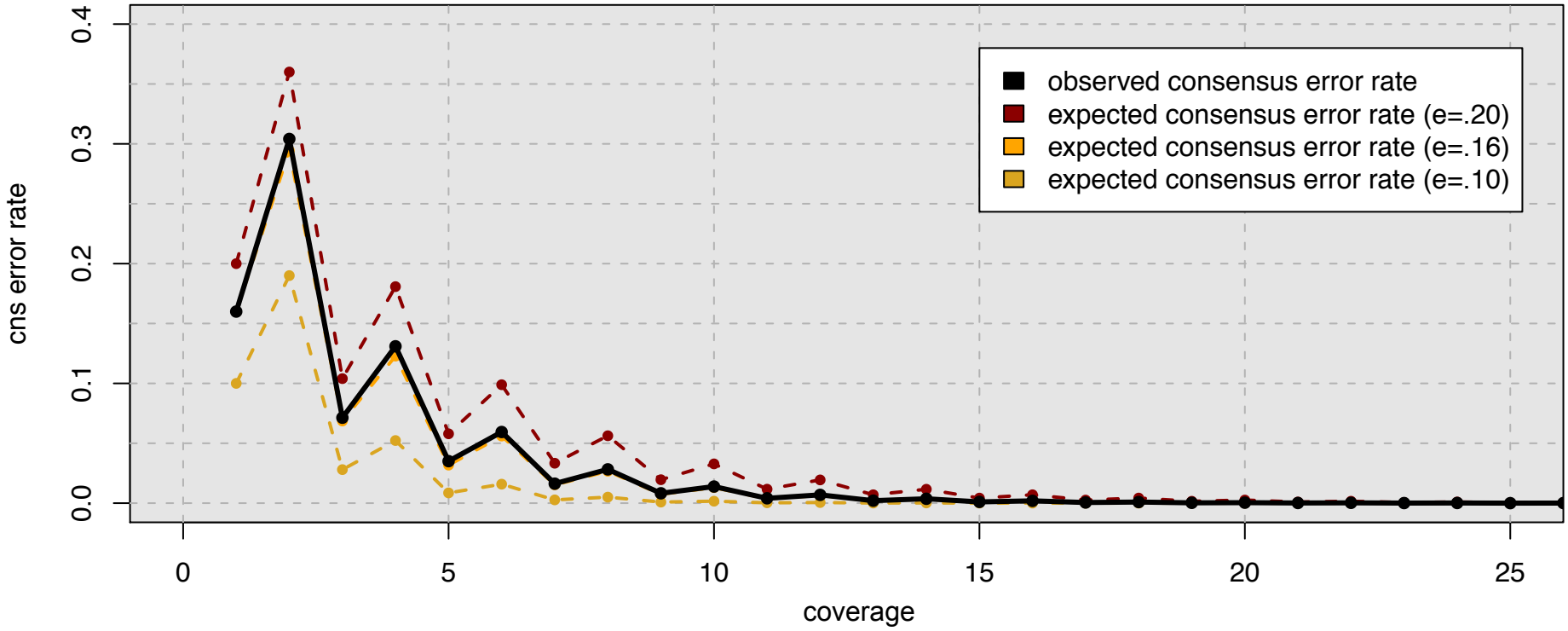


Consensus Quality: Probability Review

Roll *n* dice => What is the probability that at least half are 6's

n	*Min to Win*	*Winning Events*	*P(Win)*
1		*1/6*	*16.7%*
2		*P(1 of 2) + P(2 of 2)*	*30.5%*
3		*P(2 of 3) + P(3 of 3)*	*7.4%*
4		*P(2 of 4) + P(3 of 4) + P(4 of 4)*	*13.2%*
5		*P(3 of 5) + P(4 of 5) + P(5 of 5)*	*3.5%*
n	*ceil(n/2)*	$\sum_{i=\lceil n/2 \rceil}^{n} P(i \ of \ n) = \sum_{i=\lceil n/2 \rceil}^{n} \binom{n}{i} (p)^{i} (1-p)^{n-i}$	

Consensus Accuracy and Coverage



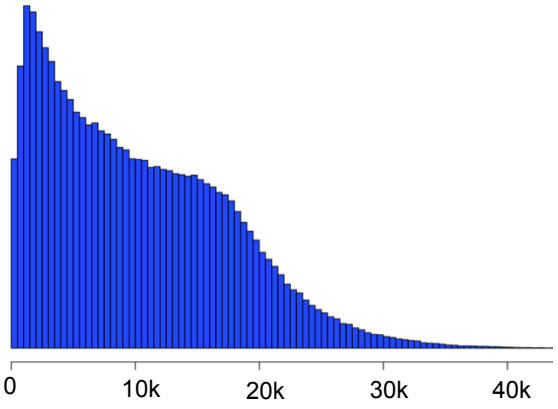
Coverage can overcome random errors

- Dashed: error model from binomial sampling; solid: observed accuracy
- For same reason, CCS is extremely accurate when using 5+ subreads

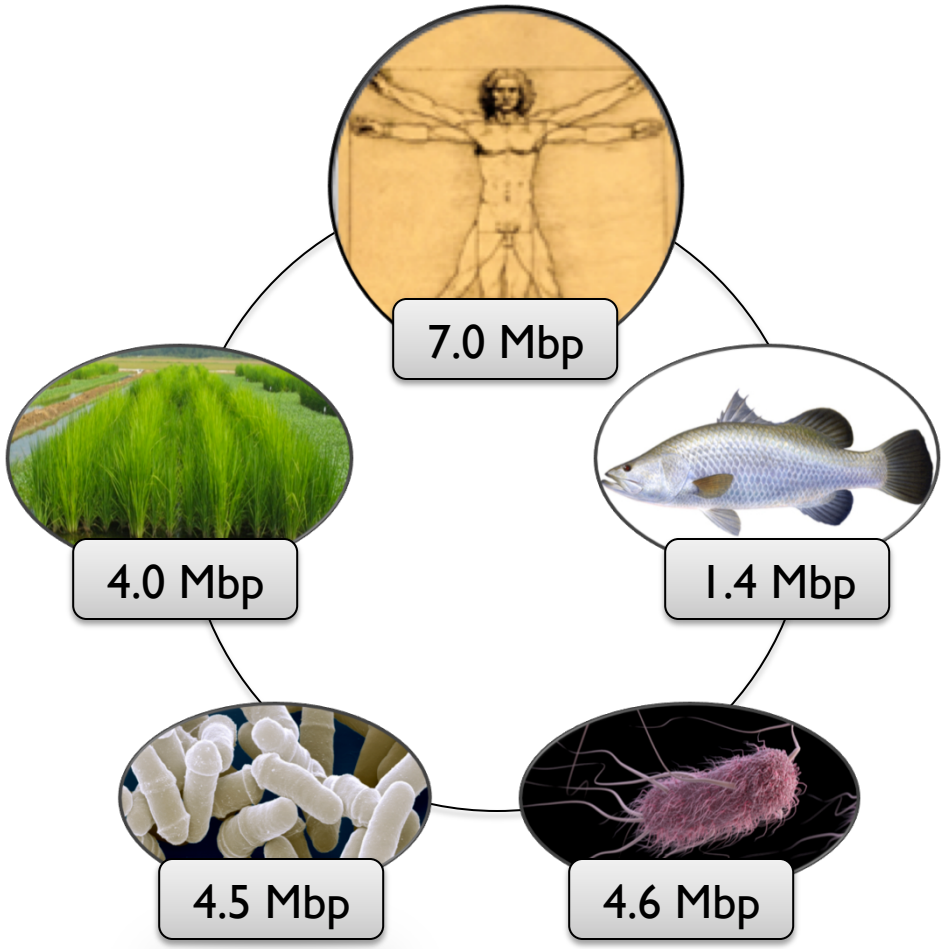
$$CNS\,Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

3rd Gen Long Read Sequencing

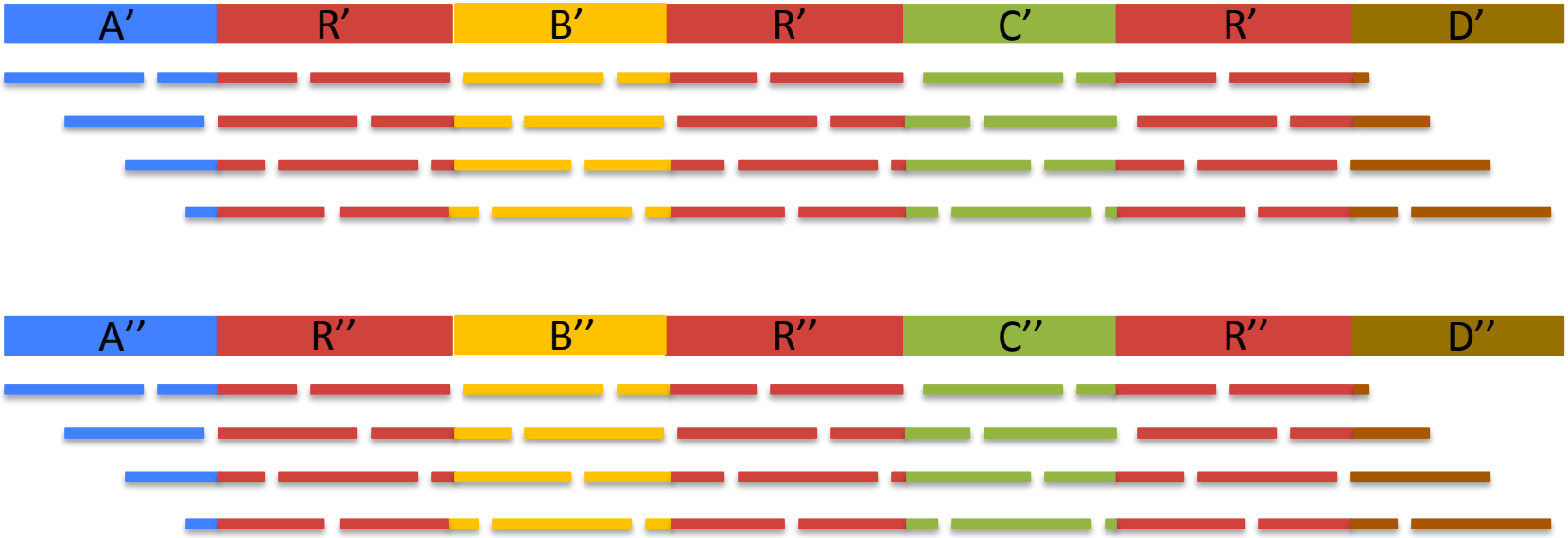
PacBio RS II



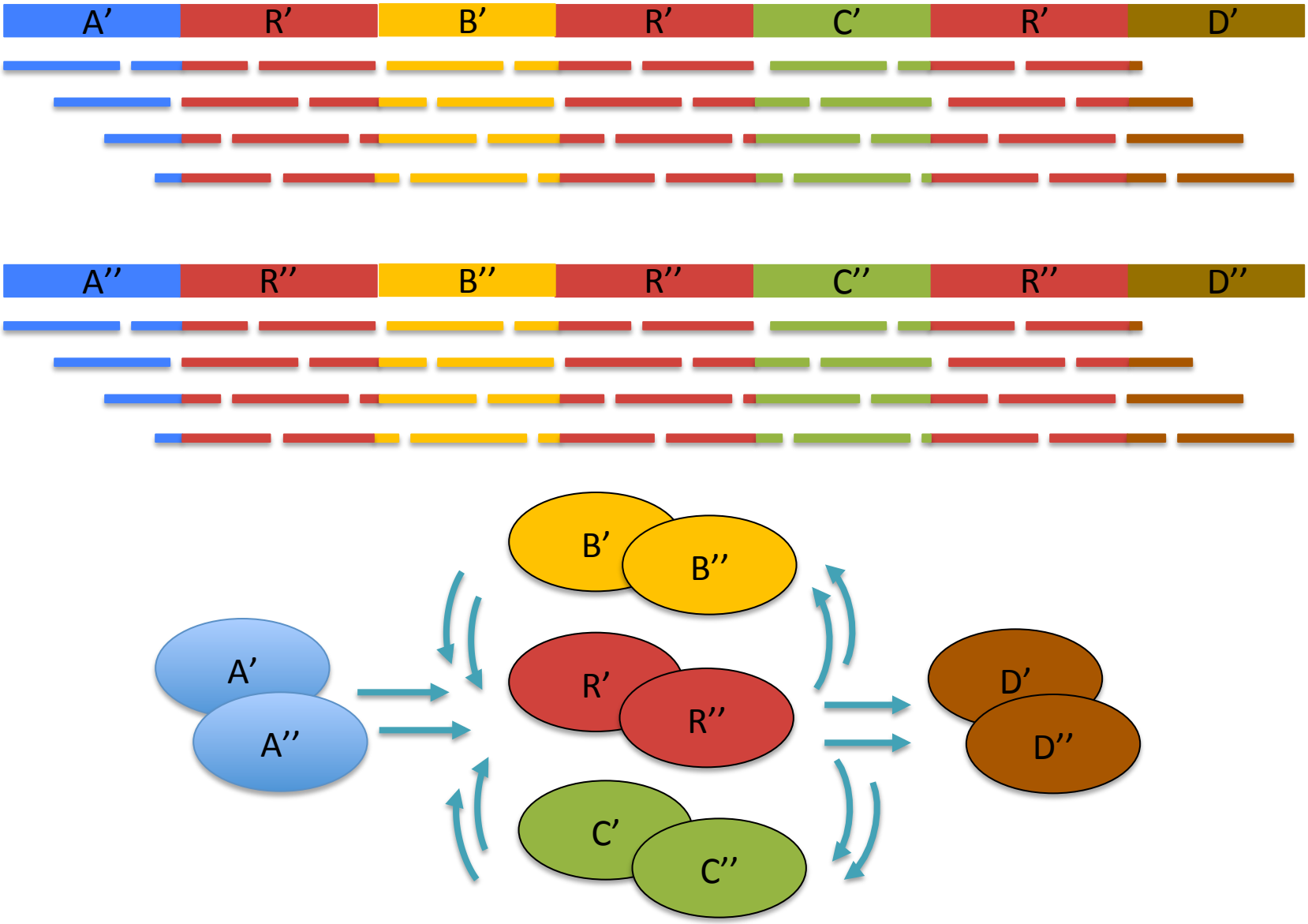
CSHL/PacBio



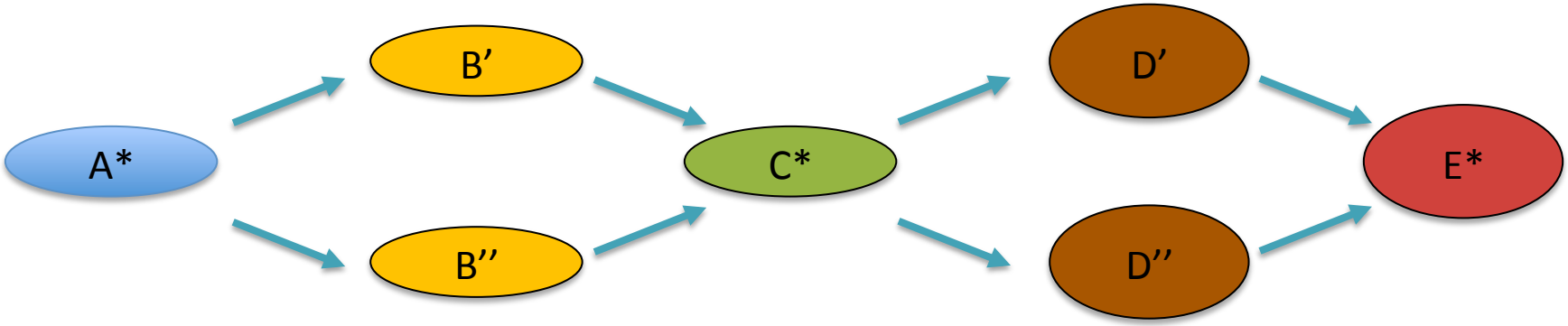
Diploid Assembly Complexity



Diploid Assembly Complexity



Diploid Assembly Problems



Assembly becomes more fragmented

- *A. thaliana* inbred with short reads: ~100kbp contig N50
- *A. thaliana* outbred with short reads: ~10kbp contig N50

Assembly sequence & size will be distorted

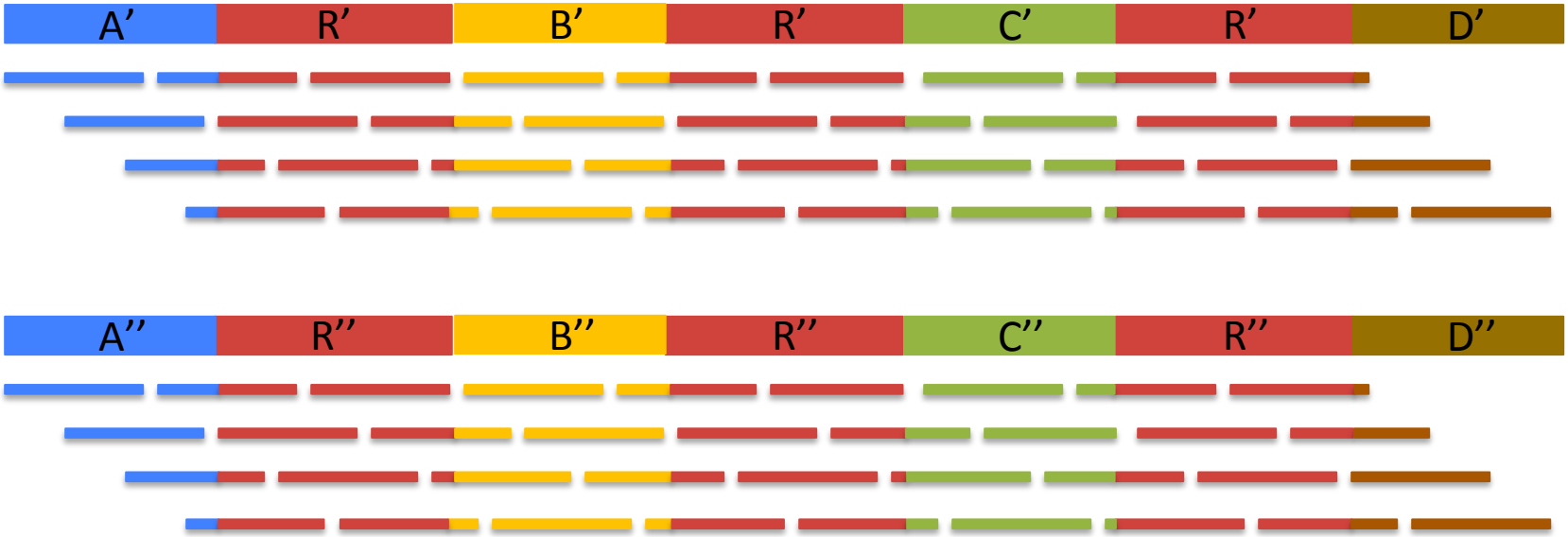
- Regions of low heterozygosity will be assembled together
  -> reduces assembly from true diploid size
- Regions of high heterozygosity will be split apart
  -> haplotypes may be next to each other in scaffolds or left out

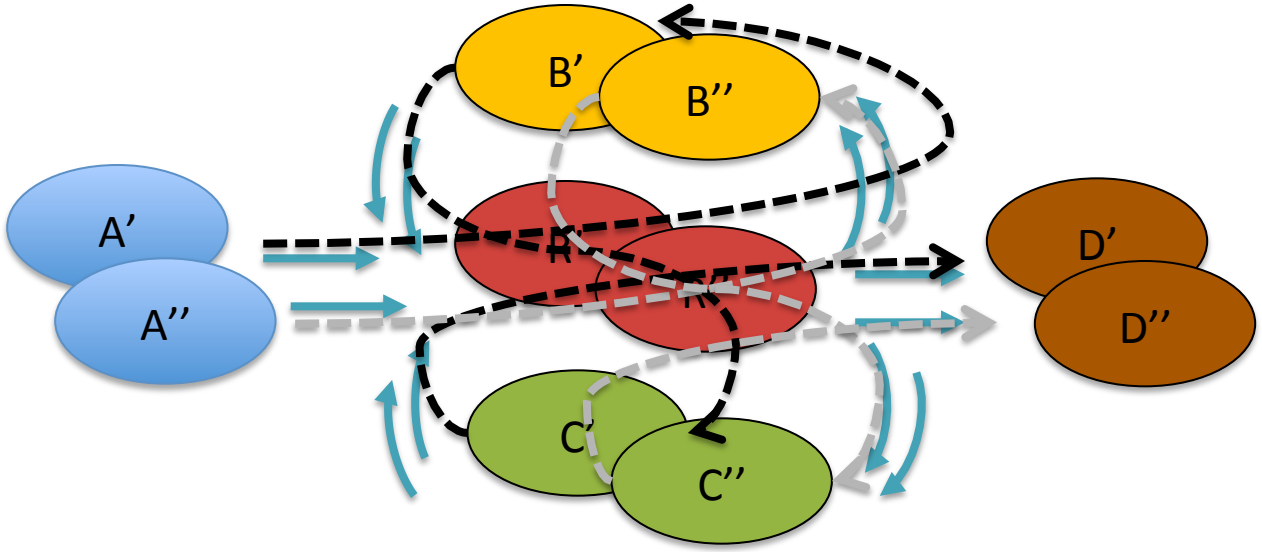
"Mosaic" consensus sequences*

- Sequence will arbitrarily switch from maternal to paternal alleles
- May be "read incoherent" and not supported by any sequencing reads

Critical genes may be assembled into 0, 1, or 2 copies (or more)!

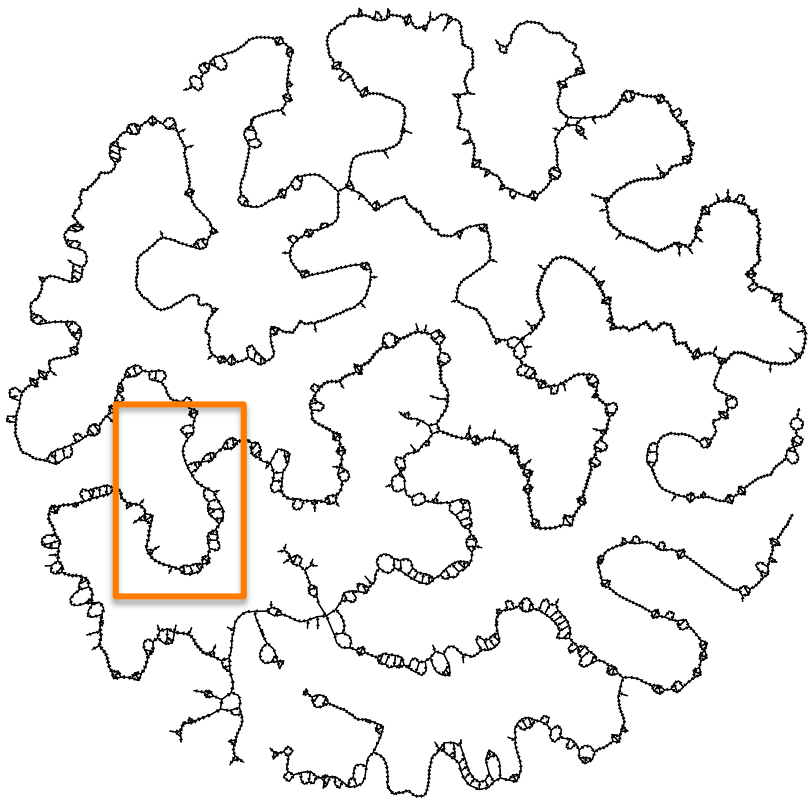
Assembly Complexity



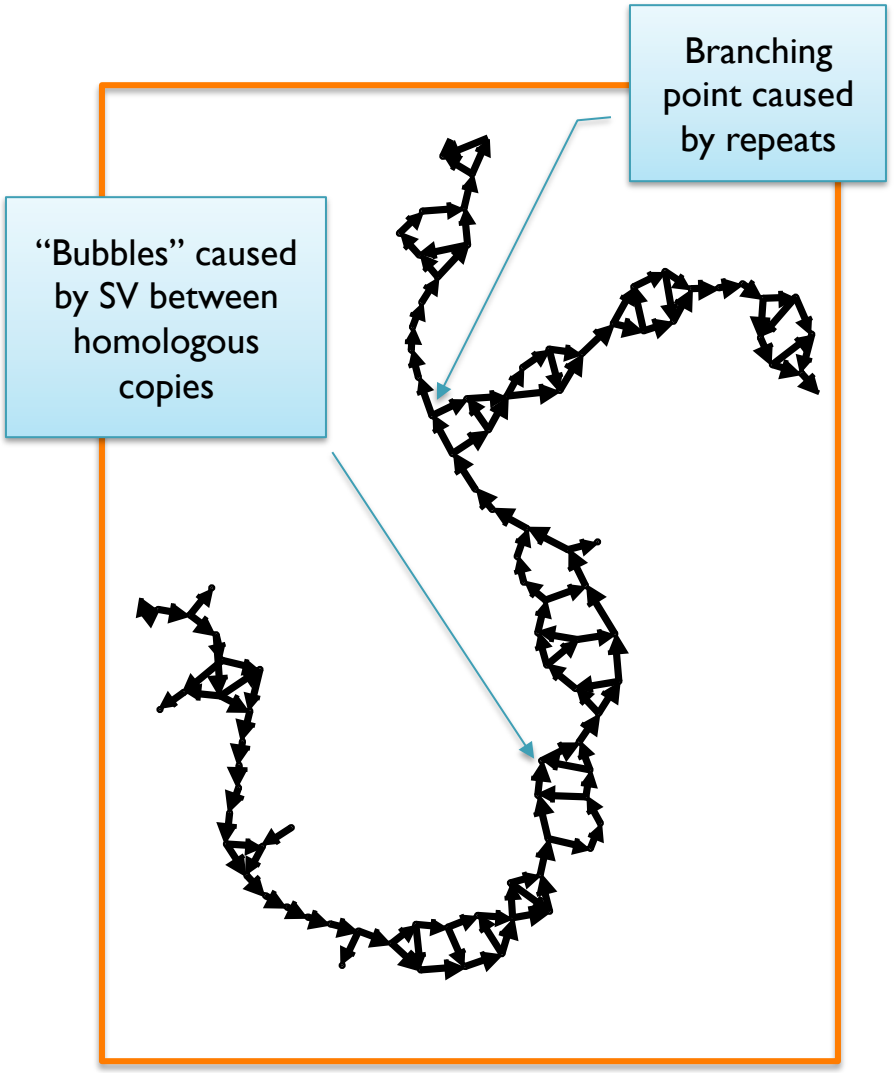


FALCON-unzip: Phased Diploid Genome Assembly

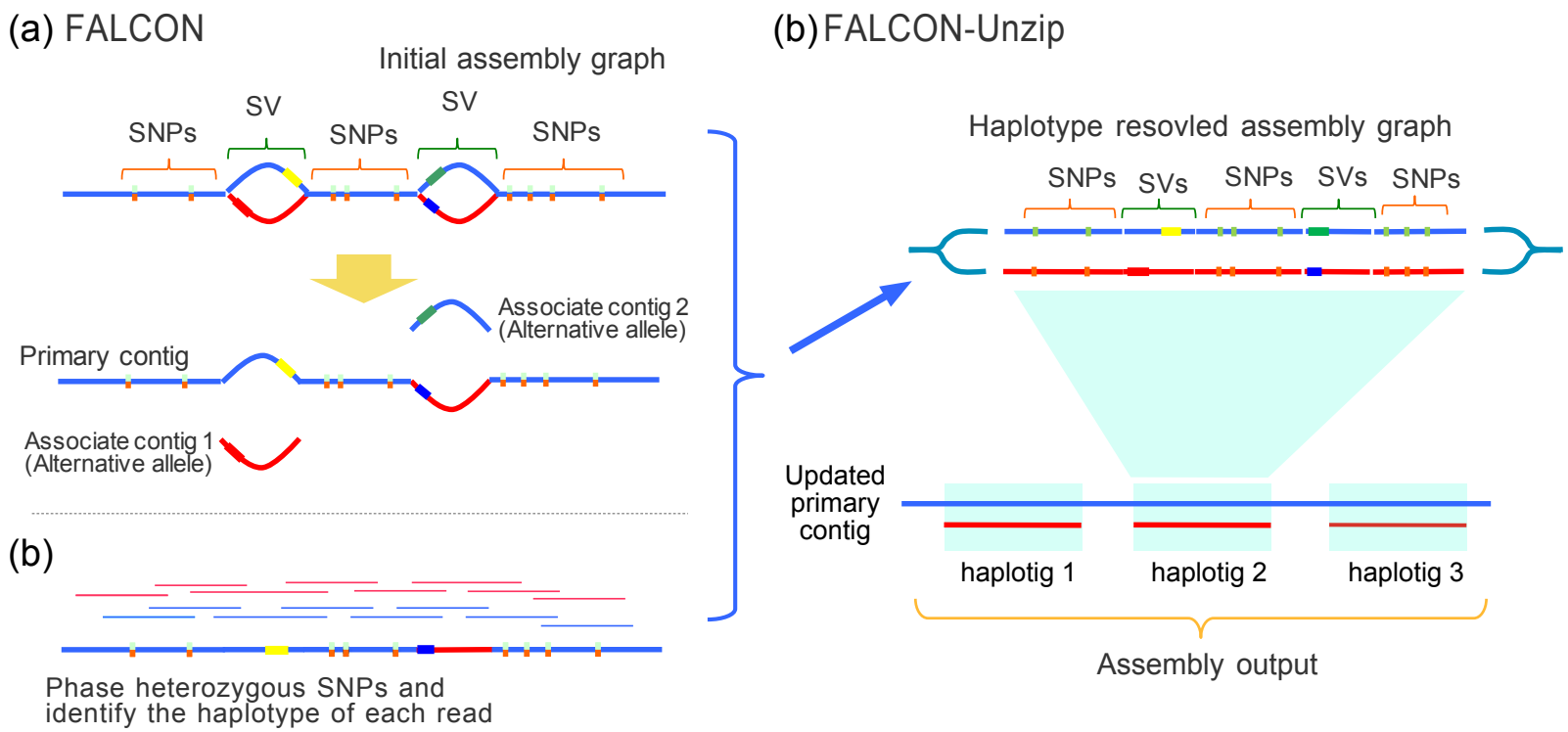
Assembly graph from A. thaliana Ler-0 + Col-0 data



The graph “diameter” ~ 12 M bp
Mean edge size=17.4 k bp



Algorithm overview



1. Assemble Genome with FALCON

- Consensus is a mosaic of the two alleles, except large SVs that form bubbles

2. Use bubbles to seed phasing in flanking regions

- Greedy analysis of heterozygous SNPs flanking SV regions

3. Update Assembly graph with phased sequences: Phased Haplotigs

A. thaliana Assemblies

Two inbred lines, CVI-0 and Col-0, were sequenced separately about 1.5 years ago with P5C3 chemistry

- Compare Col-0 assembly to TAIR reference
- Establish very high quality reference for CVI

Characterize the variations between the two strains with the per-strain haploid assemblies:

- High SV density: big SV every 80 kb
- High SNP density: SNP every 100 to 300 bp

In silico diploid dataset:

- Mixture of the two datasets to emulate a diploid genome at about 80x coverage.
- Useful for testing and development

Genuine diploid dataset:

- Sequencing of an F1 progeny to 120x coverage

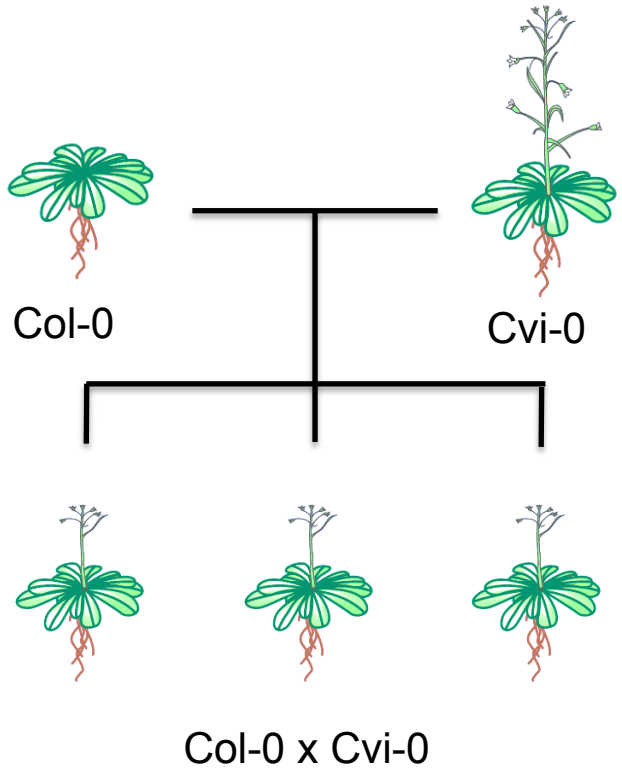
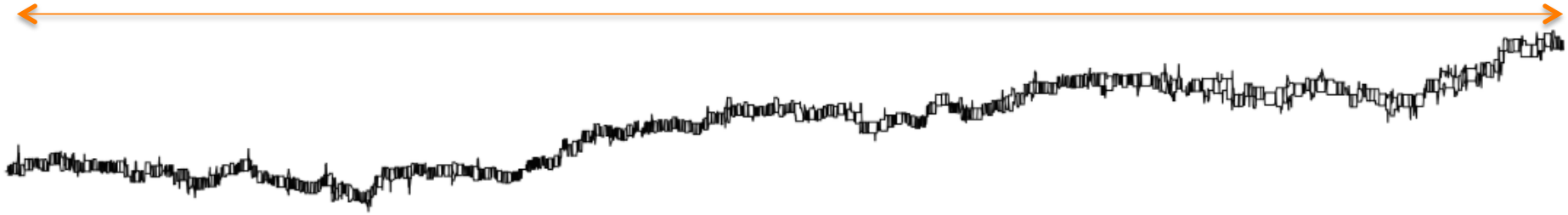
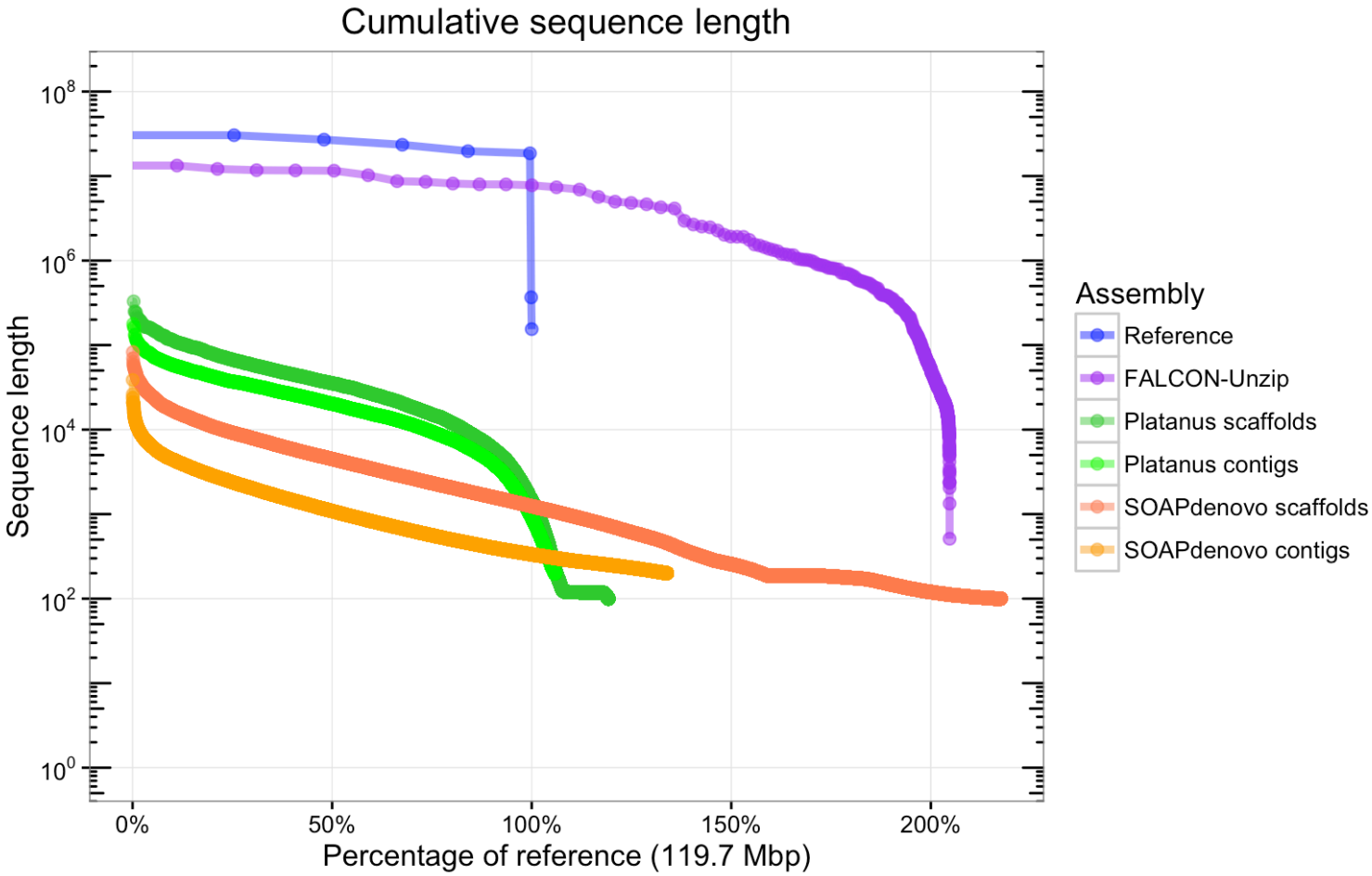


Image credits:
Pajoro, et al, Trends in plant science 21.1 (2016): 6-8.

9.49 Mb haplotype fused assembly graph



A. thaliana F1 Assembly Results



Cumulative sequence length of three *Arabidopsis* F1 assemblies created by FALCON-Unzip, Platanus, and SOAPdenovo compared to the TAIR10 reference.

FALCON-unzip:
Phased Diploid Genome Assembly with PacBio Long Reads




	C. pyxidata (Coral fungus)	Cabernet Sauvignon	*T. guttata* (Zebra finch)[$*]	Human*
Haploid Genome Size:	~ 44 Mb	~ 500 Mb	~1.2 Gb	~ 3 Gb
Sequencing Coverage	4.1 Gb / 95x	73.7 Gb / 147x	50 Gb / 42x	255 Gb / 85x
Primary contig size	41.9 Mb	591.0 Mb	1.07 Gb	2.76 Gb
Primary contig N50	1.5 Mb	2.2 Mb	3.23 Mb	22.9 Mb
Haplotig size	25.5 Mb	372.2 Mb	0.84 Gb	2.0 Gb
Haplotig N50	872 kb	767 kb	910 kb	330 kb

[$] Thanks to Erich Jarvis for permission to use preliminary data

* Preliminary results. Fast file system and efficient computational infrastructure are currently needed for large genomes.

Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing
Chin, CS, Peluso, P, Sedlazeck, FJ, Nattestad, M, Concepcion, GT, Clum, A, Dunn, C, O'Malley, R, Figueroa-Balderas, R, Morales-Cruz, A, Cramer, GR, Delledonne, M, Luo, C, Ecker, JR, Cantu, D, Rank, DR., Schatz, MC
(2016) Nature Methods doi:10.1038/nmeth.4035

PacBio Roadmap

PacBio RS II

$750k instrument cost
1895 lbs

~$75k / human @ 50x

SMRTcell

150k Zero Mode Waveguides
~10kb average read length
~1 GB / SMRTcell
~$500 / SMRTcell

PacBio Roadmap

PacBio Sequel

$350k instrument cost
841 lbs

~$30k / human @ 50x

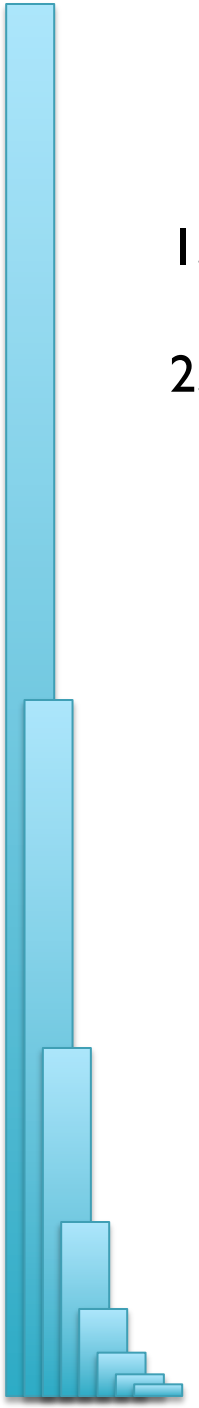
SMRTcell v2

1M Zero Mode Waveguides
~15kb average read length
~10 GB / SMRTcell
~$1000 / SMRTcell

Next Steps

1. Get ready for assignment 2

2. Check out the course webpage




Welcome to Applied Comparative Genomics
https://github.com/schatzlab/appliedgenomics

Questions?