Genomic Data Analysis on Spark+Hadoop

Ryan Williams

DataEngConf NYC

11/4/2016

Agenda

- Intro
- Genomics crash course
- Guacamole: somatic mutation calling on Spark
- Other applications / interesting algorithms

Hammer Lab

- Computational lab in the department of Genetics and Genomic Sciences at Mount Sinai
- Principal investigator: Jeff Hammerbacher
- Focus on informatics for cancer immunotherapy
- Software developed at github.com/hammerlab

Genomics / Sequencing Overview

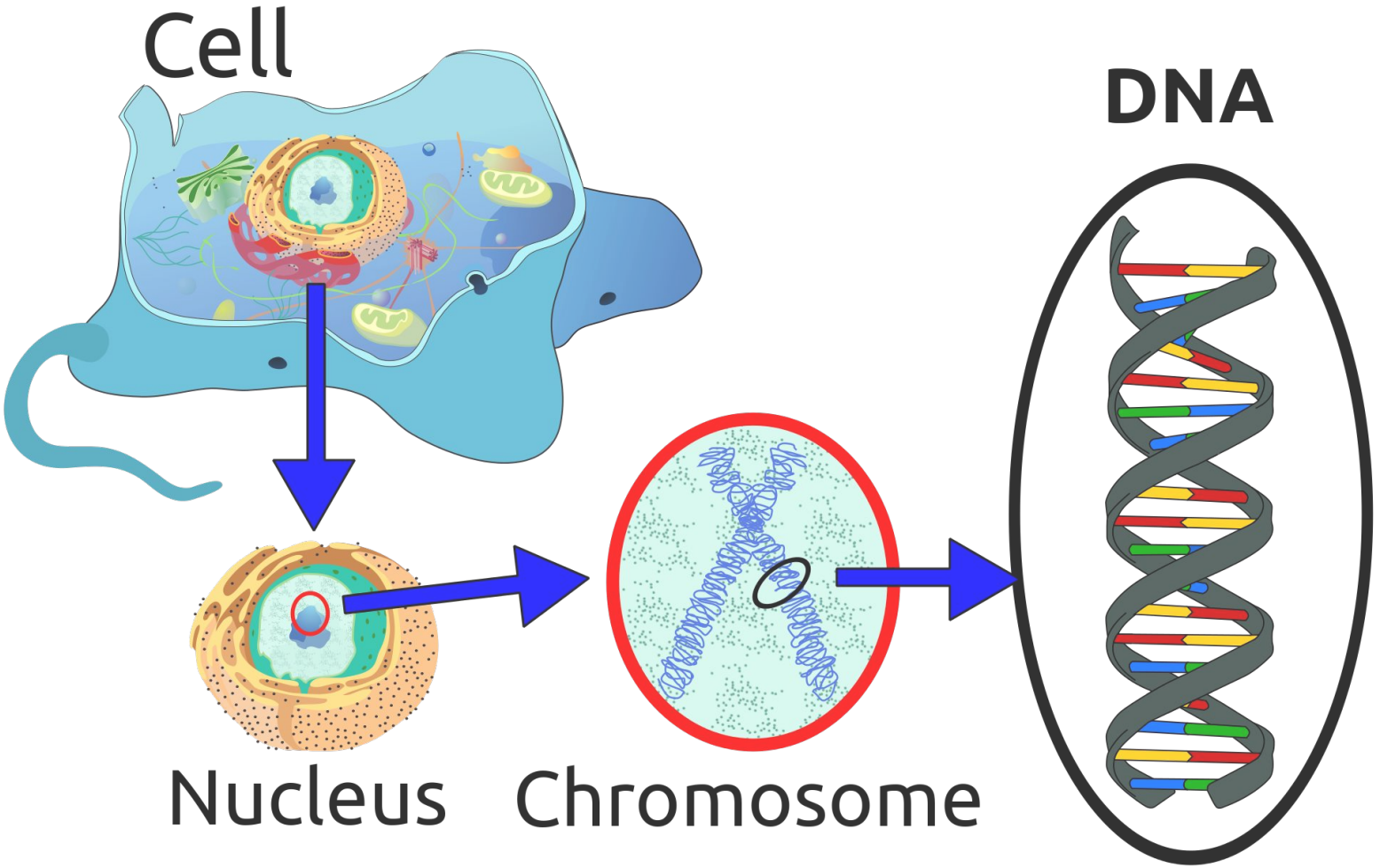
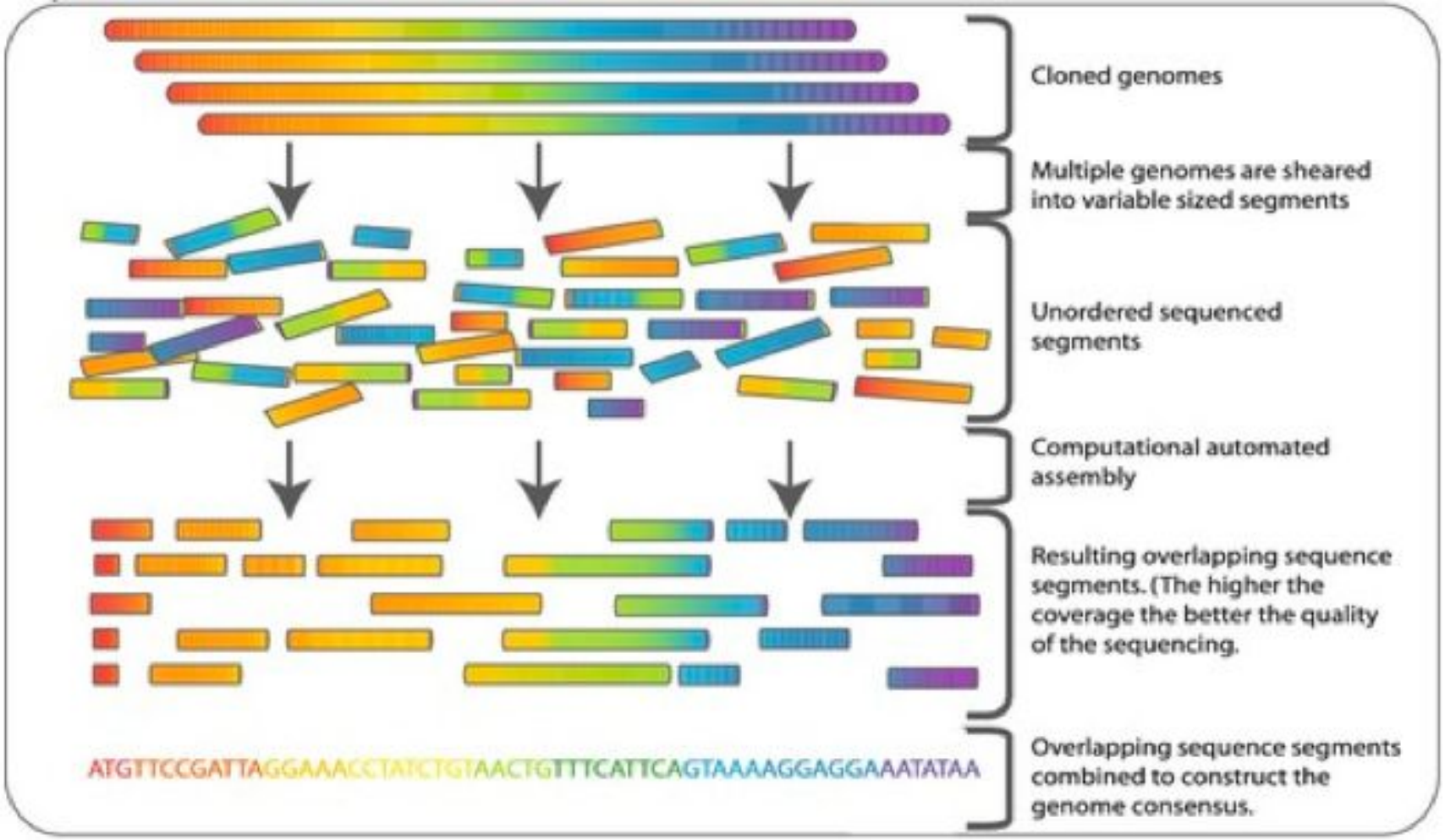


Image source: https://en.wikipedia.org/wiki/DNA#/media/File:DNA_Structure%2BKey%2BLabelled.pn_NoBB.png

http://blogs.plos.org/dnascience/2016/01/21/can-a-quirky-chromosome-create-a-second-human-species/



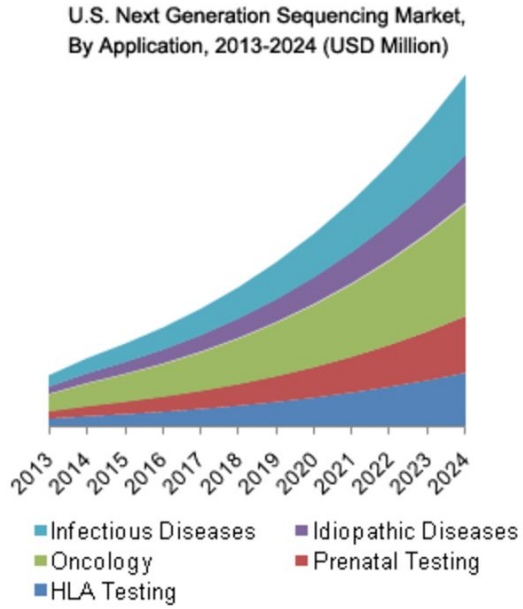
http://www.wikiwand.com/en/Shotgun_sequencing

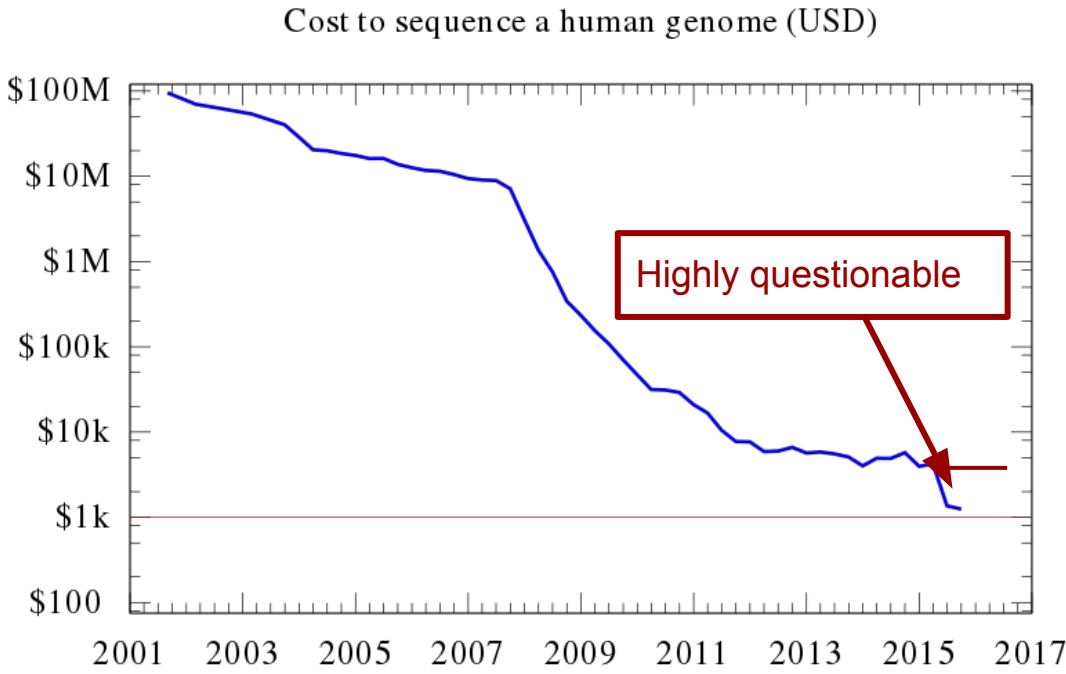
Storing Human Genomes

- 1 genome ≈ 3B base-pairs
- Theory:
  - “2 bits per base-pair” (A, C, G, T)
  - ⇒ 1 genome ≈ 750MB
  - <1% unique, person to person
  - 7BN genomes ≈ 50PB
- Reality:
  - 1BN 100bp “reads”
    - ⇒ 100BN sequenced base-pairs
    - Cover the genome at average depth 30 (“30x coverage”)
  - 2-bit base, 1-byte quality score ⇒ **100GB / genome**
  - **100-100k genomes ⇒ 10TB-10PB**

Sequencing Human Genomes

- Human Genome Project, 1990-2003
- 1000 Genomes Project, 2008-2012
- 100k Genomes Project, 2012-???

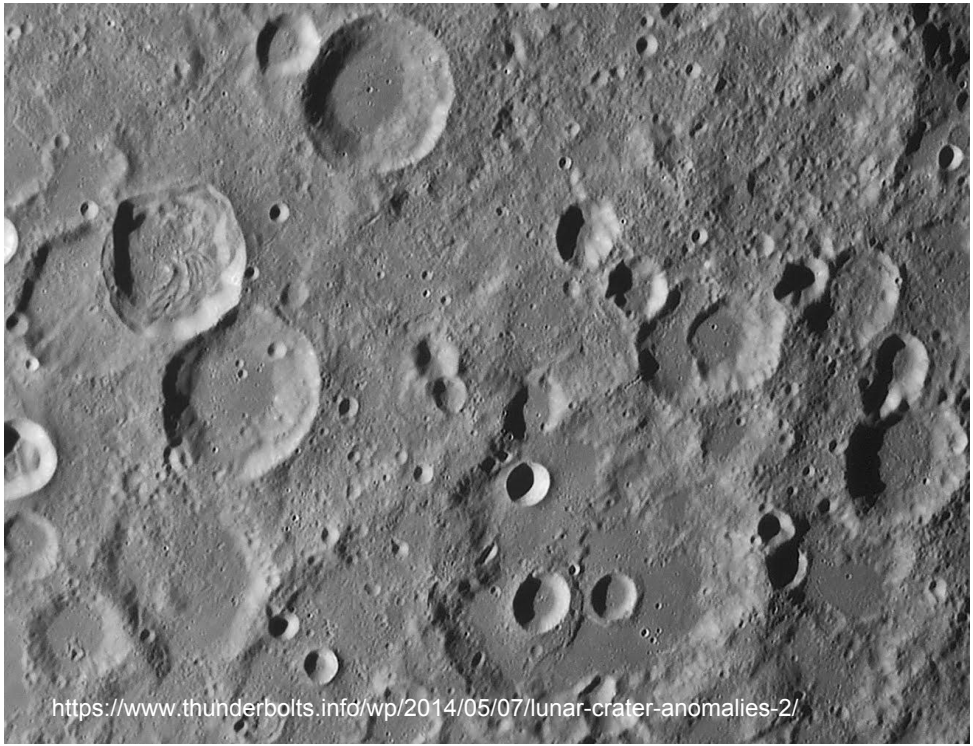


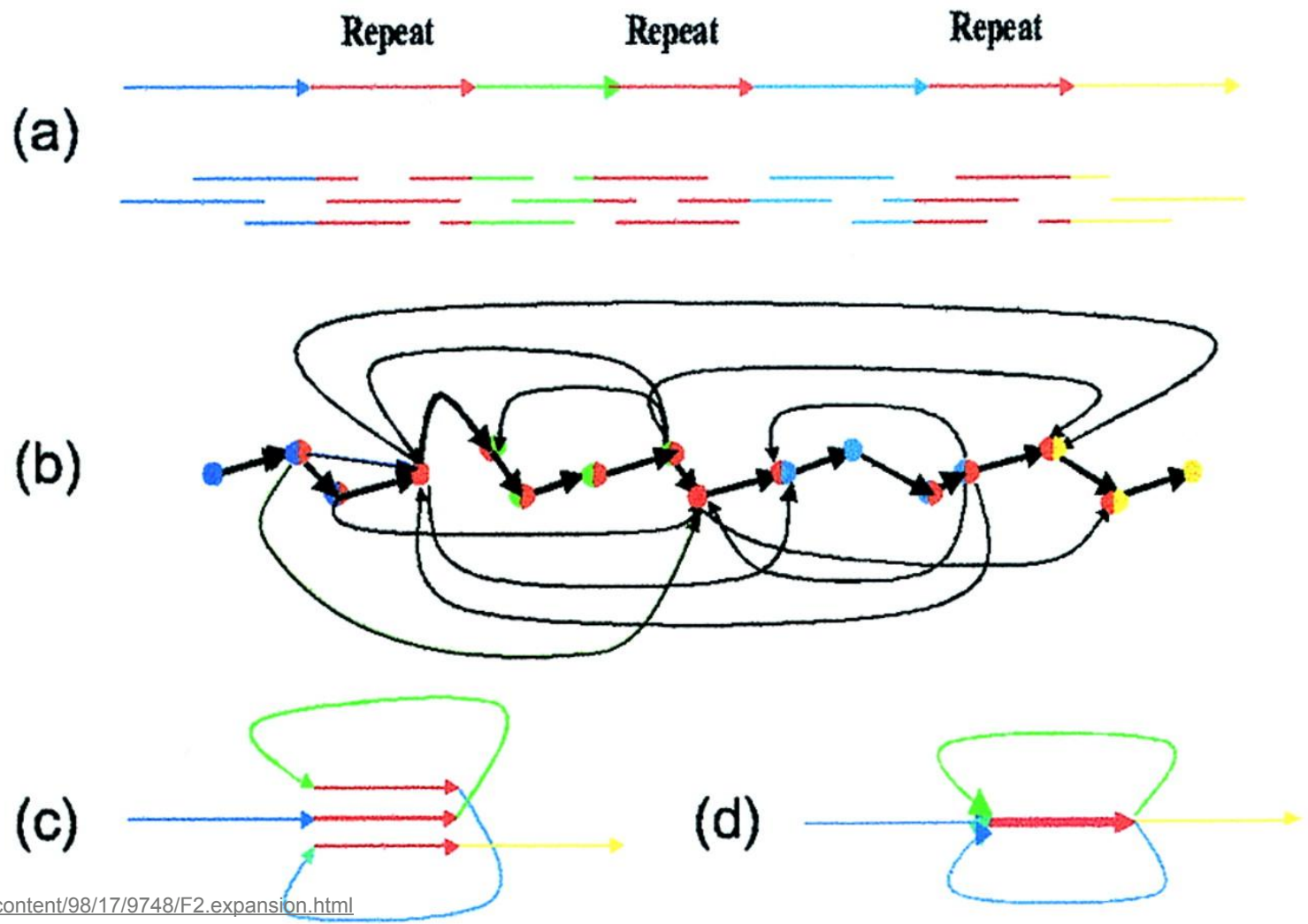


Genome Alignment / Assembly

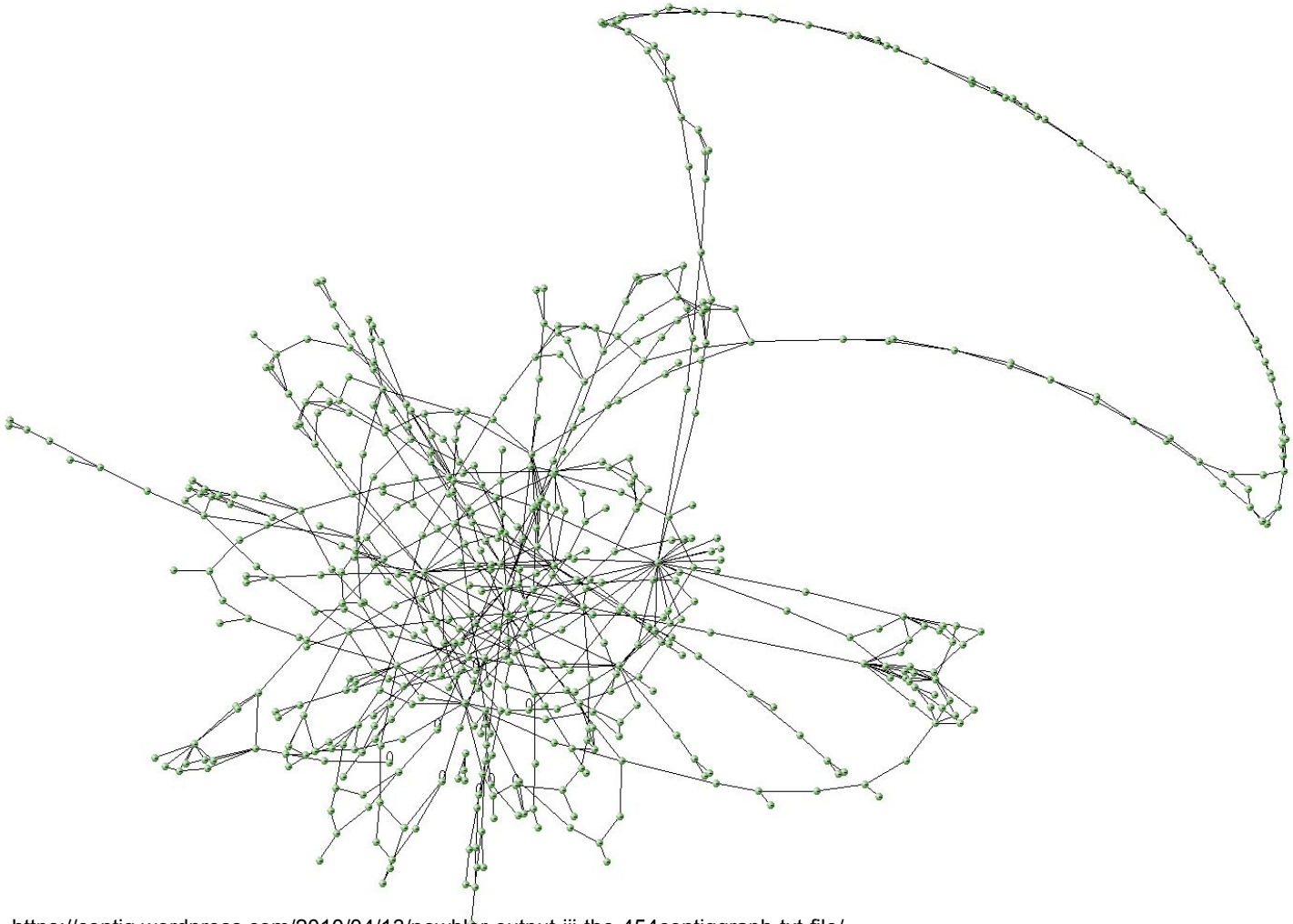
Genome structure makes things difficult

- Excessive repetitiveness
- 20% retrotransposons
  - L1: 7000bp, 100k copies
- Pseudogenes
- Impossible to resolve with “short reads”

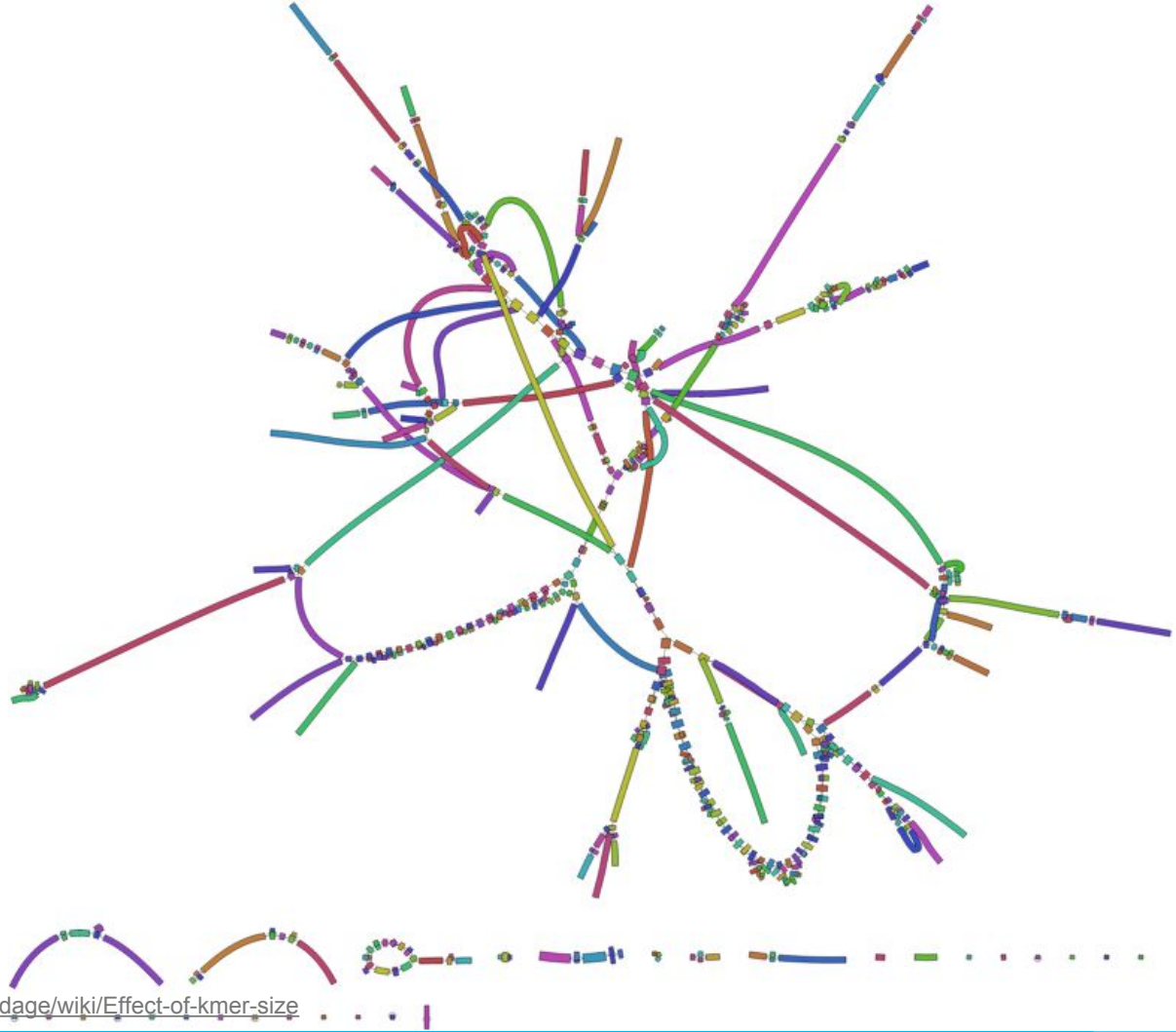




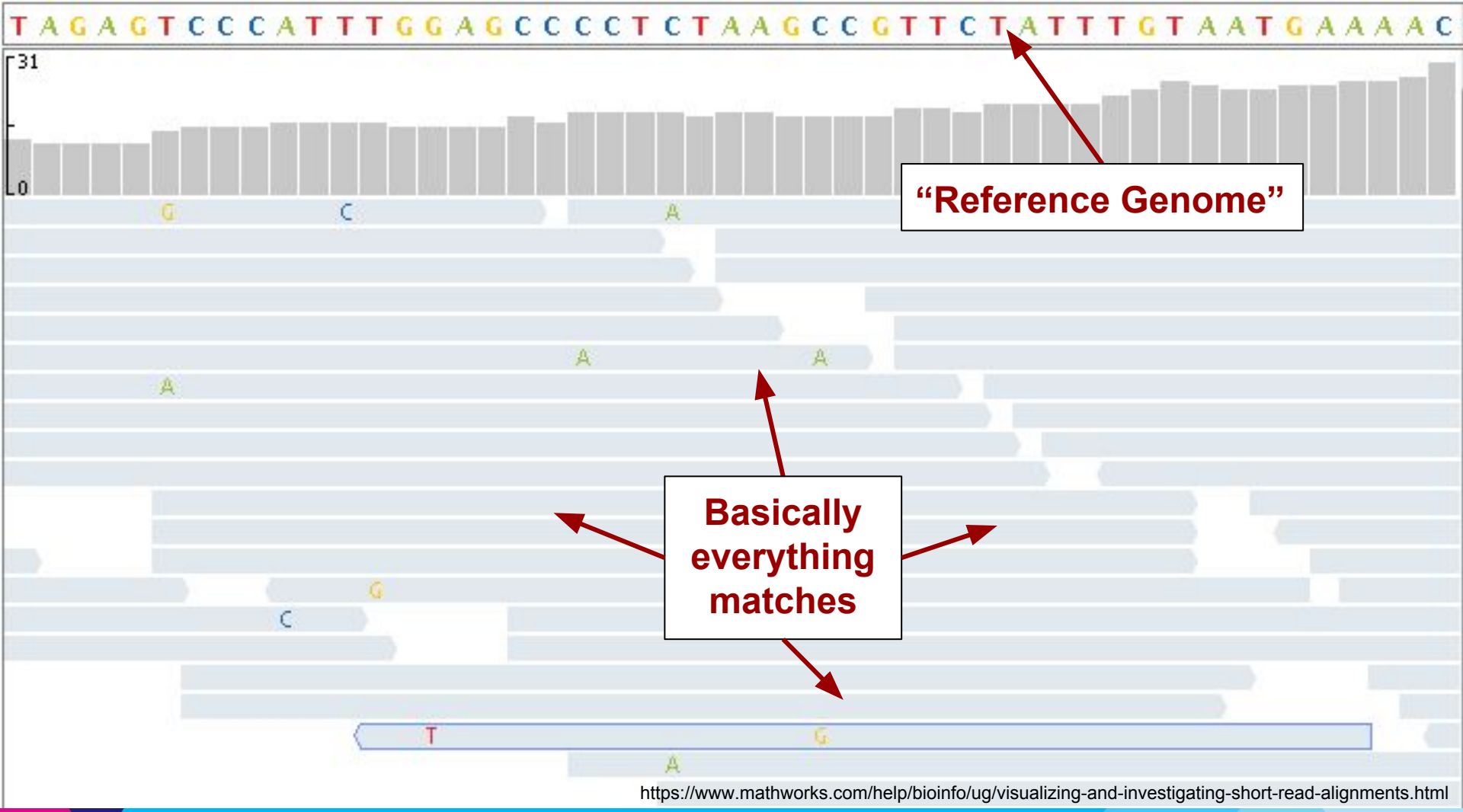
http://www.pnas.org/content/98/17/9748/F2.expansion.html



https://contig.wordpress.com/2010/04/13/newbler-output-iii-the-454contiggraph-txt-file/

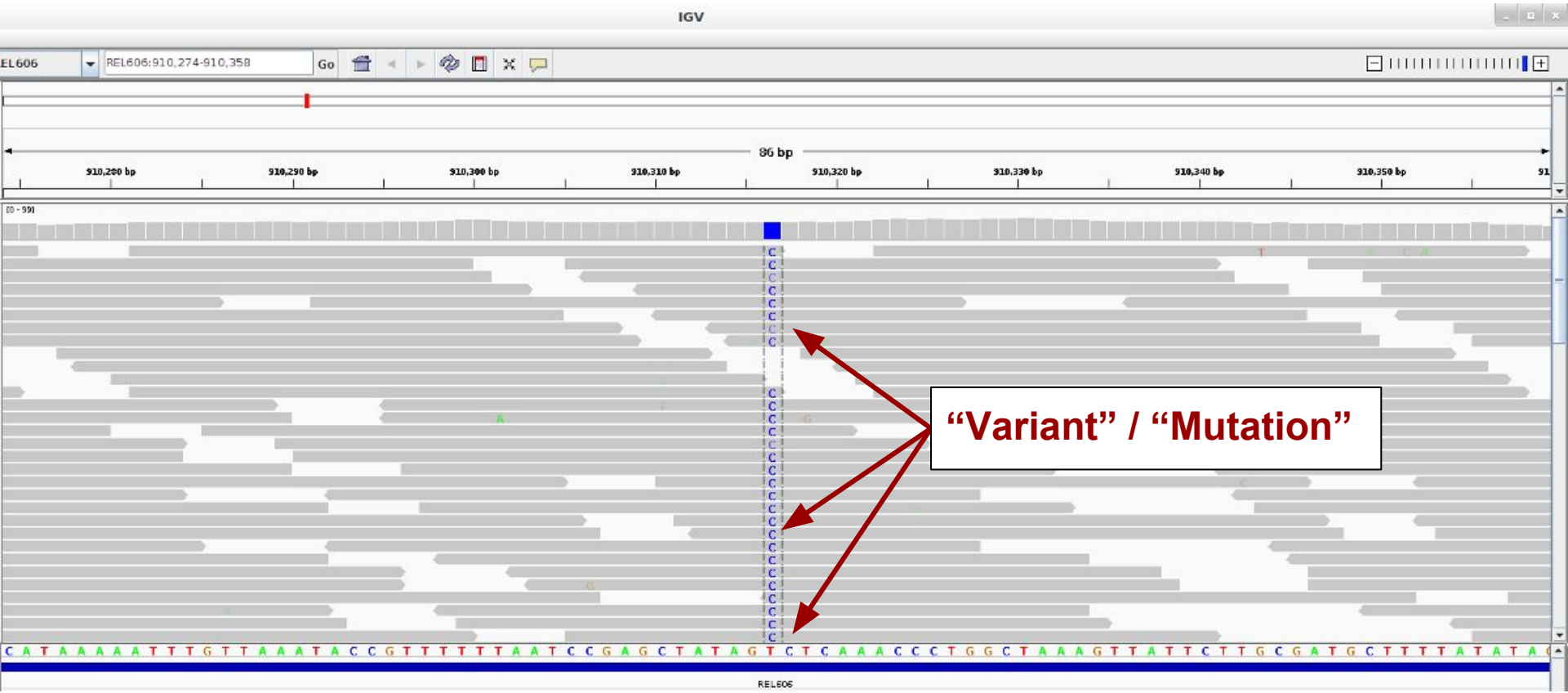


https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size



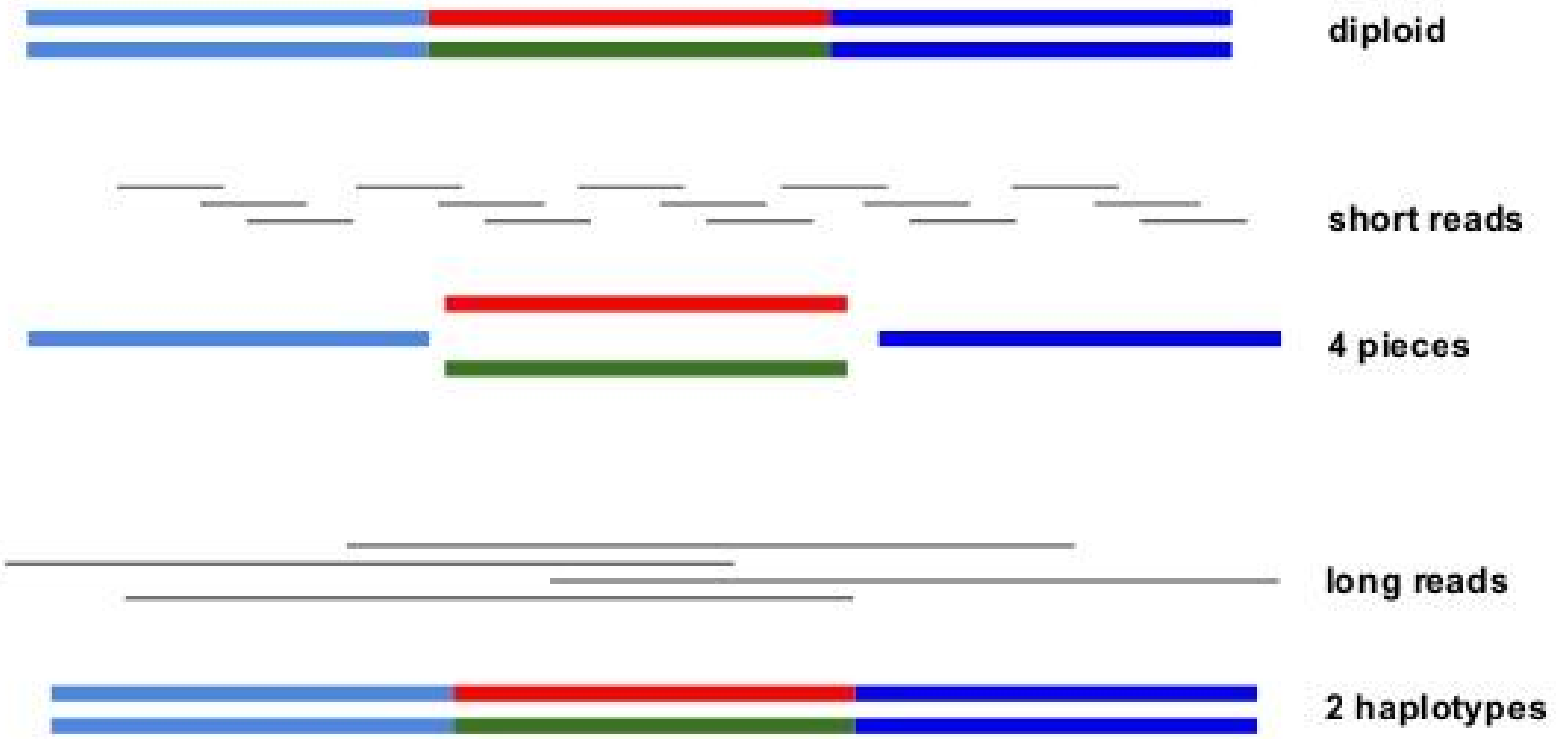
https://www.mathworks.com/help/bioinfo/ug/visualizing-and-investigating-short-read-alignments.html

Variant Calling

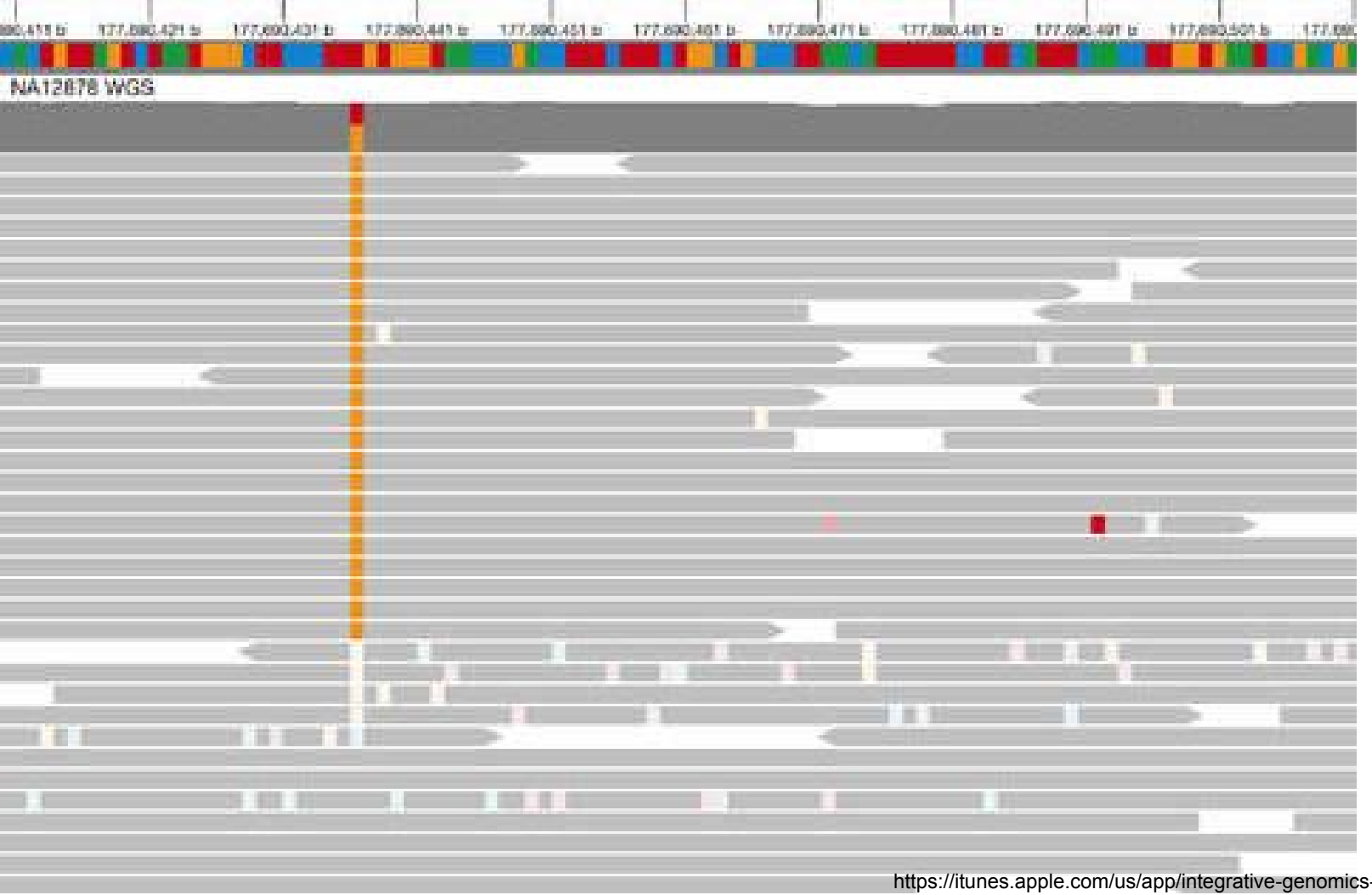


https://wikis.utexas.edu/display/bioiteam/IGV+Tutorial

Heterozygosity



http://www.slideshare.net/torstenseemann/long-read-sequencing-wehi-bioinformatics-seminar-tue-16-june-2015



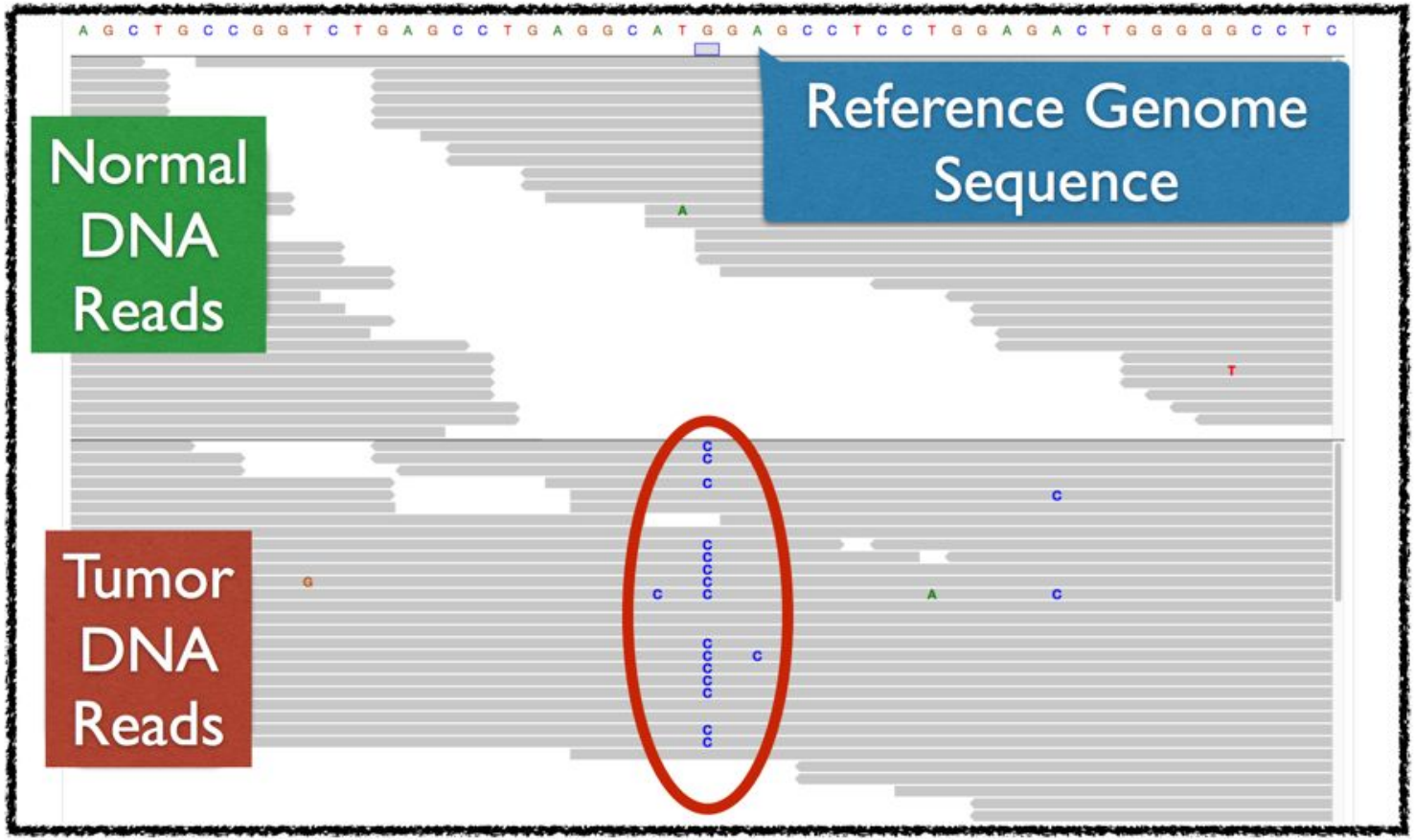
https://itunes.apple.com/us/app/integrative-genomics-viewer/id900183022?mt=8

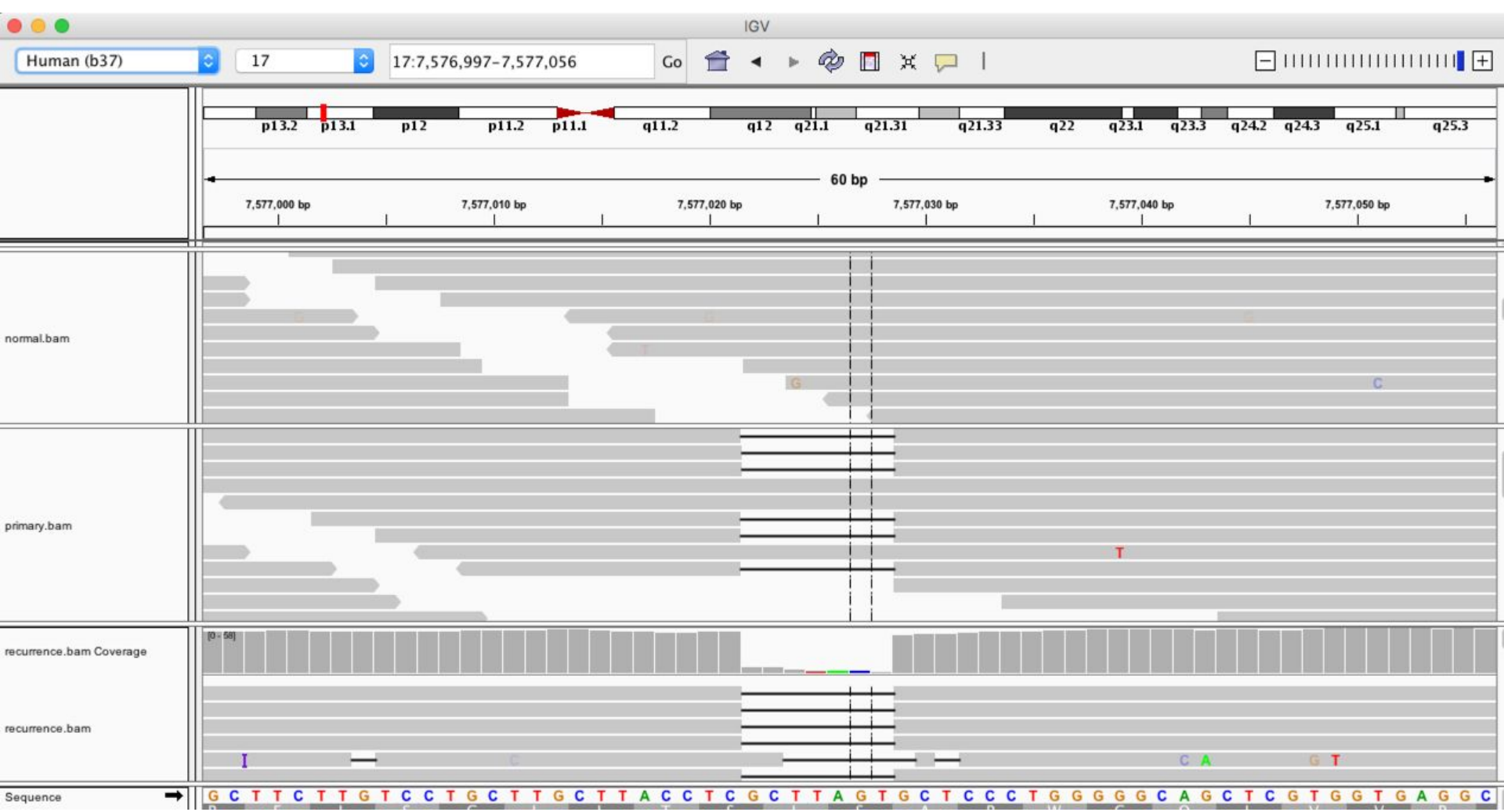
Somatic Mutation Calling

Somatic Mutation Calling

- Cells come from two populations
  - e.g. “normal” and tumor cells
- Find mutations specific to cancer cells

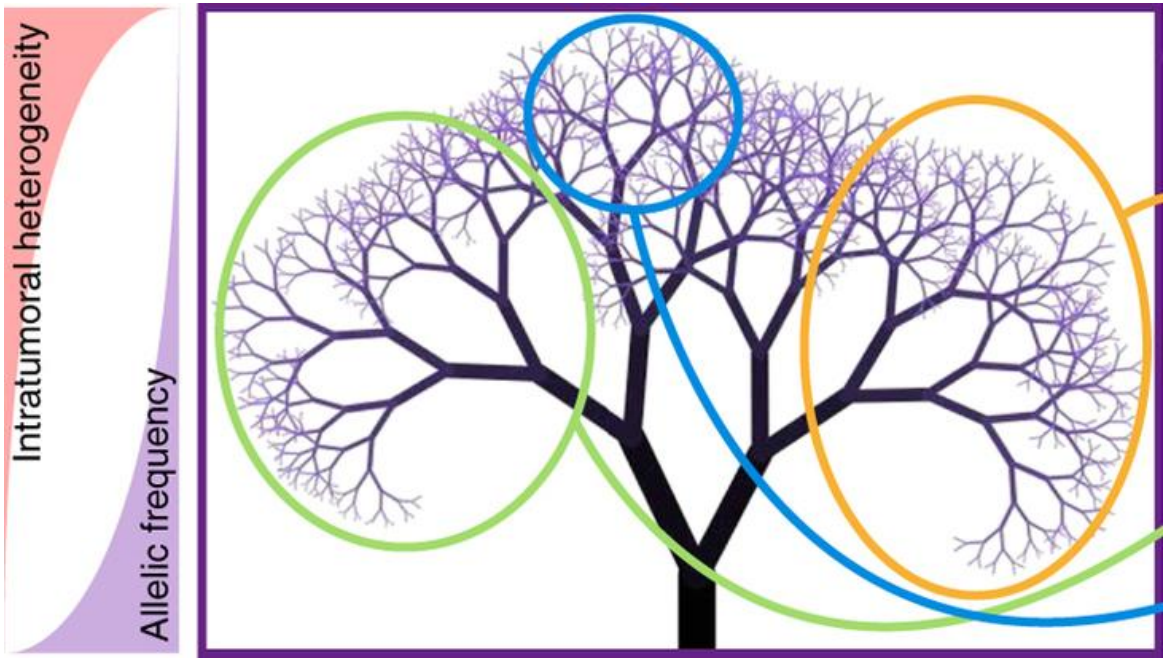
Somatic Mutation Calling




IGV
Human (b37)
17
17:7,576,997-7,577,056
Go
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.33
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
60 bp
7,577,000 bp
7,577,010 bp
7,577,020 bp
7,577,030 bp
7,577,040 bp
7,577,050 bp
normal.bam
primary.bam
recurrence.bam Coverage
recurrence.bam
Sequence
G C T T C T T G T C C T G C T T G C T T A C C T C G C T T A G T G C T C C C T G G G G G C A G C T C G T G G T G A G G C

Somatic Mutation Calling

- Underdetermined!
  - Sub-clonality
  - Tumor sample purity



Agreement on somatic variant calls across tools is surprisingly poor

Exome sequencing

SNVs

	EBCall	Mutect	Seurat	Shimmer	Somatic Sniper	Strelka	Varscan 2	Virmid
EBCall	1.00	0.51	0.60	0.36	0.50	0.43	0.57	0.48
Mutect	0.20	1.00	0.47	0.18	0.25	0.26	0.30	0.33
Seurat	0.08	0.16	1.00	0.10	0.26	0.09	0.19	0.12
Shimmer	0.32	0.41	0.66	1.00	0.39	0.39	0.46	0.40
Somatic Sniper	0.08	0.10	0.31	0.07	1.00	0.06	0.20	0.09
Strelka	0.52	0.80	0.79	0.53	0.52	1.00	0.69	0.68
Varscan	0.21	0.28	0.52	0.20	0.50	0.21	1.00	0.24
Virmid	0.37	0.64	0.72	0.35	0.47	0.43	0.51	1.00

Krøigård, A.B. et al., 2016. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. Plos One, 11(3), p.e0151664. Available at: http://dx.plos.org/10.1371/journal.pone.0151664.

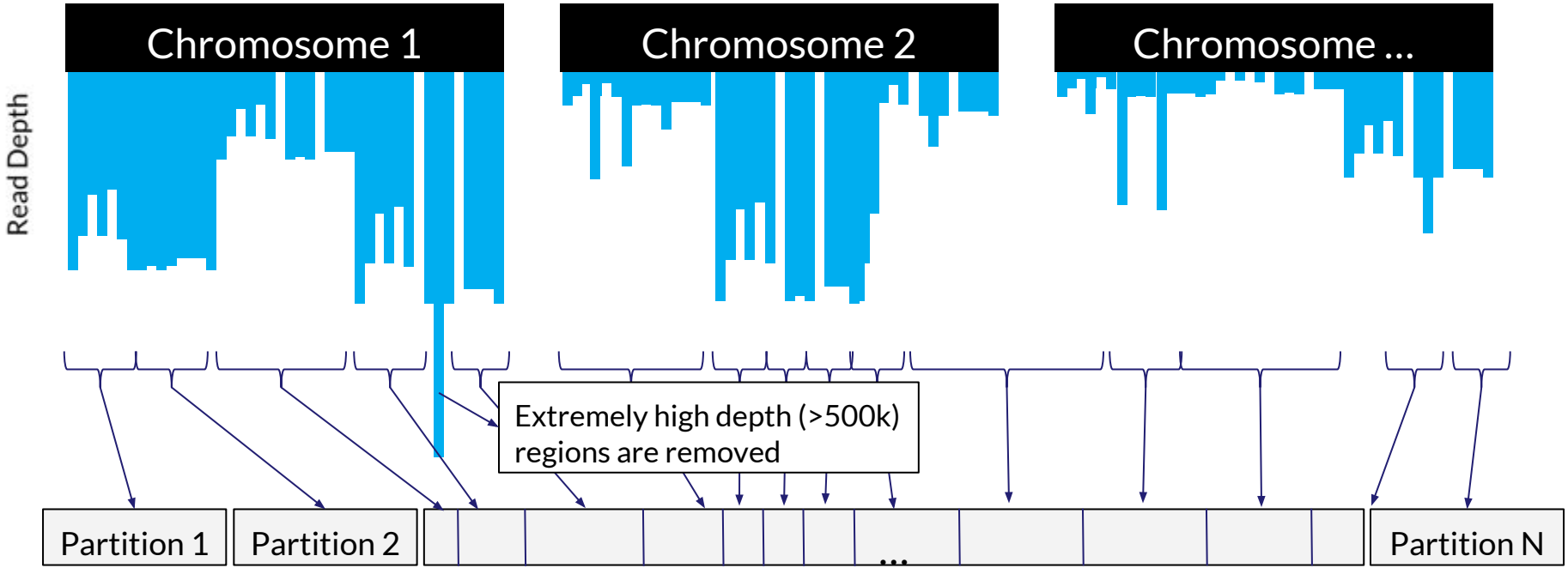
Guacamole: Somatic Mutation Calling with Apache Spark

A typical Guacamole analysis

1. Partition the genome
2. Partition reads according to (1)
3. Build pileups at each site
4. Apply user-supplied function at each pileup
5. Write output

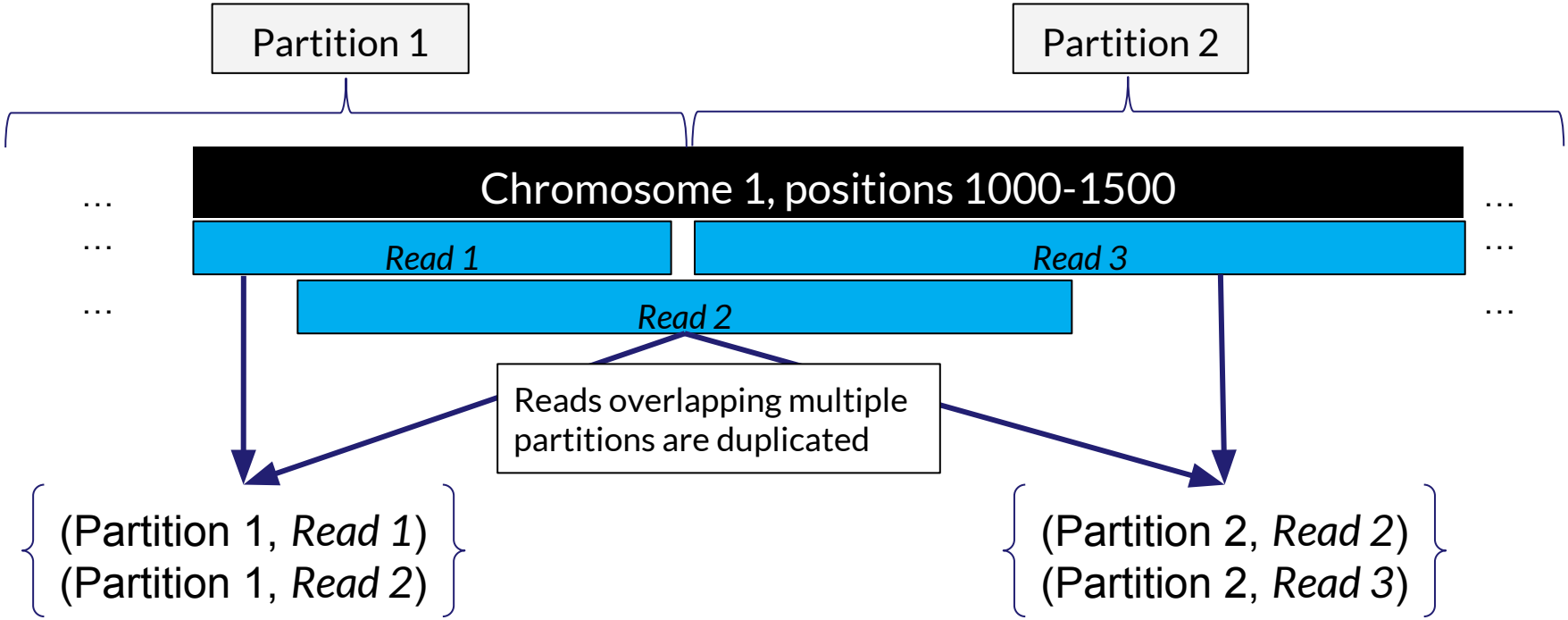
Step 1: Partition the genome

- Partition the genome into intervals, balancing the number of reads overlapping each partition
- Each interval will correspond to one Spark partition

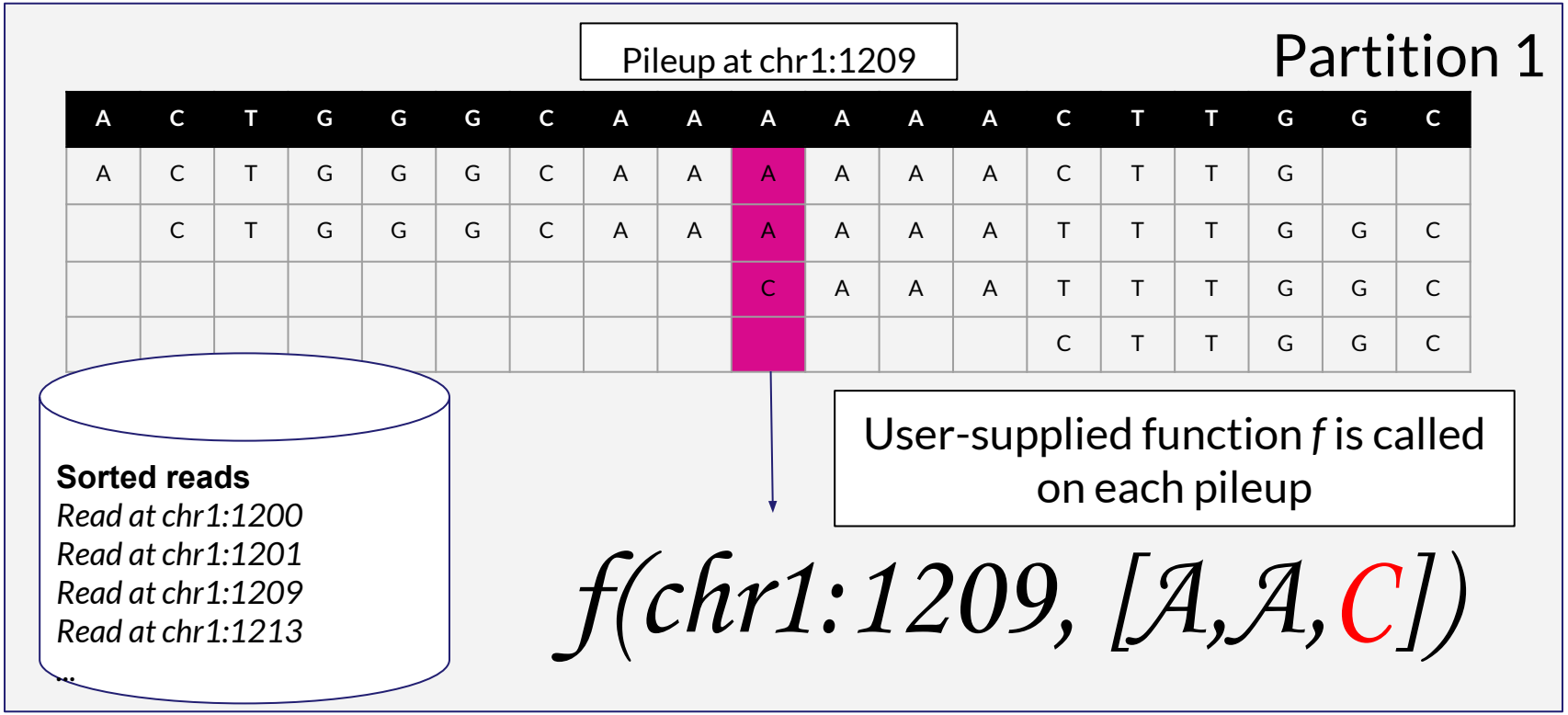


Step 2: Partition reads

- All-to-all shuffle of reads based on the genomic partition in Step 1
- A copy of each read goes to each partition it overlaps



Step 3: each partition streams through reads to generate pileups



Benchmarking

Testing cluster

Hardware	
Nodes	100
Cores	2400
Memory	12.5 TB
Storage	3.6 PB

Software	
Spark	1.6.1
Hadoop	2.6.0-cdh5.5.1
OS	CentOS 7.2.1511

Guacamole speed

Guacamole	
2 WGS samples (DREAM Synth4)	22 minutes
3 Whole Genome Samples (AOCS-034)	31 minutes
10 Whole Exome Samples (PT189)	52 minutes

Mutect	
2 WGS samples (DREAM Synth4) chromosome 1	158 minutes
We compare to chromosome 1 single-node runtime because Mutect is typically parallelized by chromosome, and chromosome 1 takes the longest	

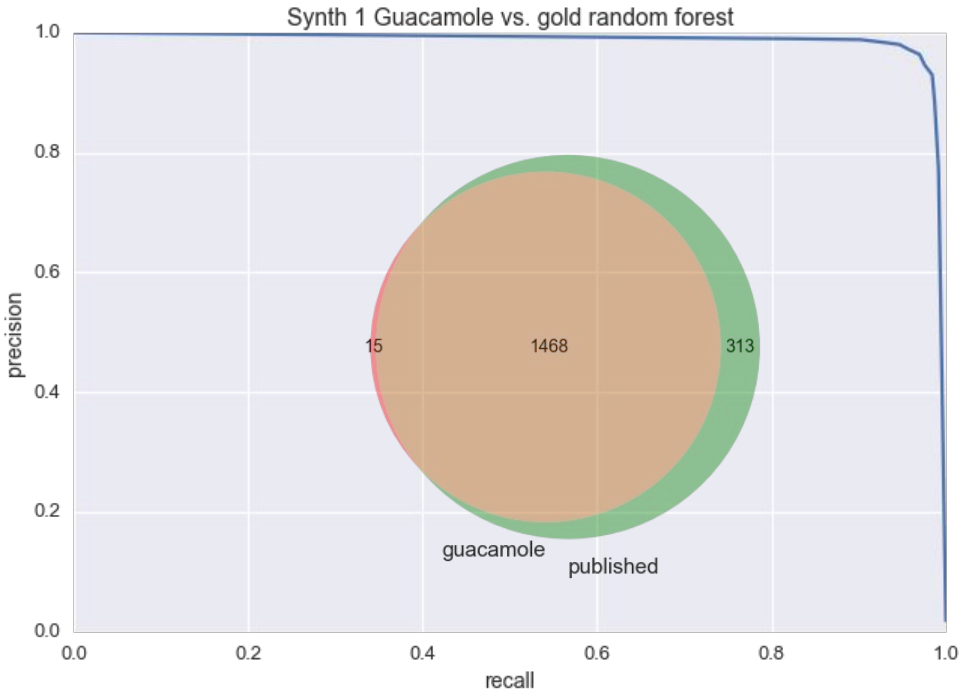
WIP: Model-based optimization

Features

- Raw likelihood
- Difference of ref and alt likelihood
- Variant allele fractions
- Allele depths
- Strand bias

Methodology

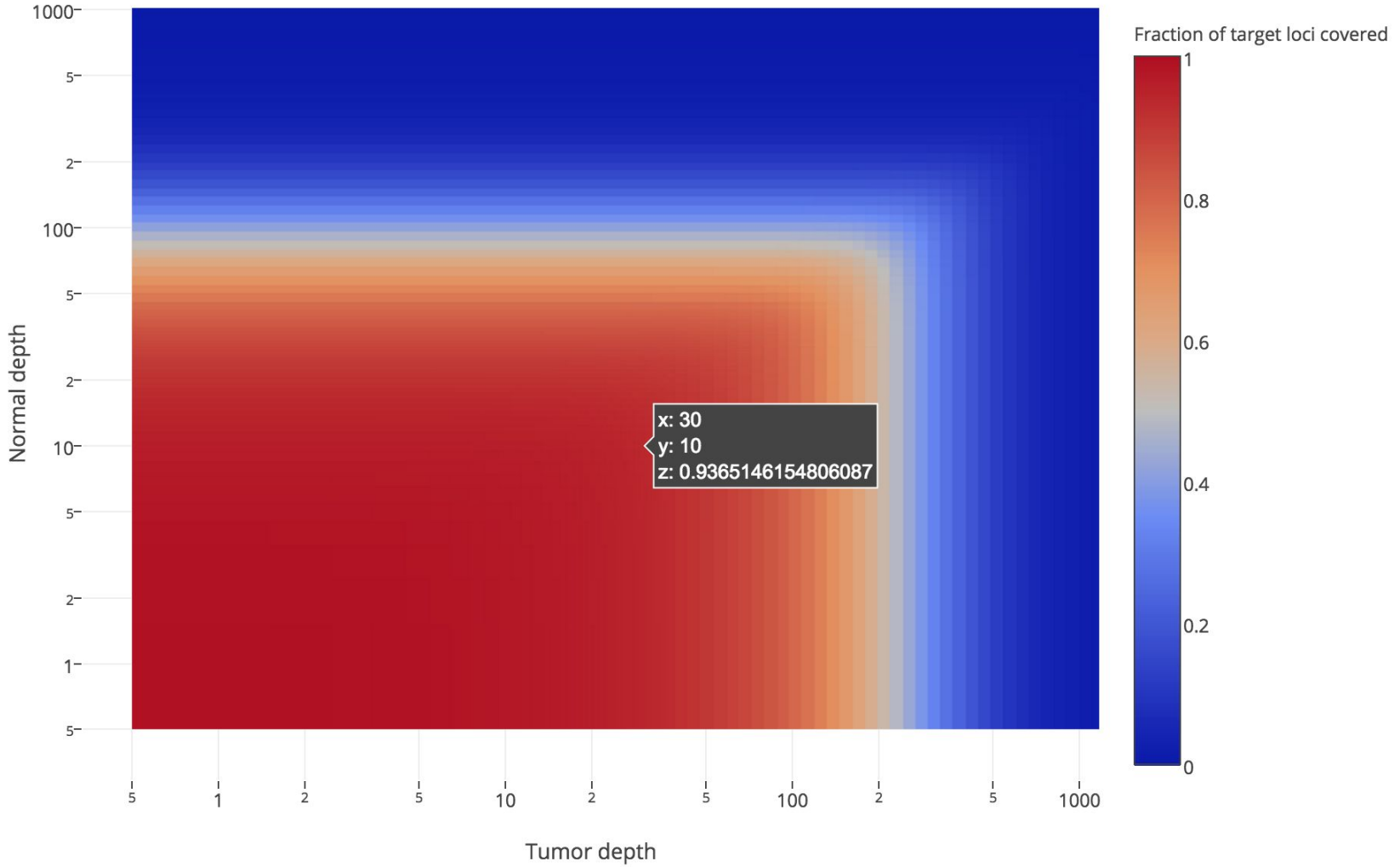
- Random forest
- 1:1 train/test split



Other applications: QC Analysis

Exome Sequencing

- “Exome”: just the genes. 1% of genome.
- Question: how much of the exome was covered with at least X reads in one sample and Y reads in the other.


Fraction of target loci covered
1
0.8
0.6
0.4
0.2
0
Normal depth
1000
5
2
100
5
2
10
5
2
1
5
x: 30
y: 10
z: 0.9365146154806087
5
1
2
5
10
2
5
100
2
5
1000
Tumor depth

Distributed 2D-Prefix-Sum

- Demo / viz
- Spark implementation: hammerlab/magic-rdds

Thanks!