Lecture 7. Variant Identification

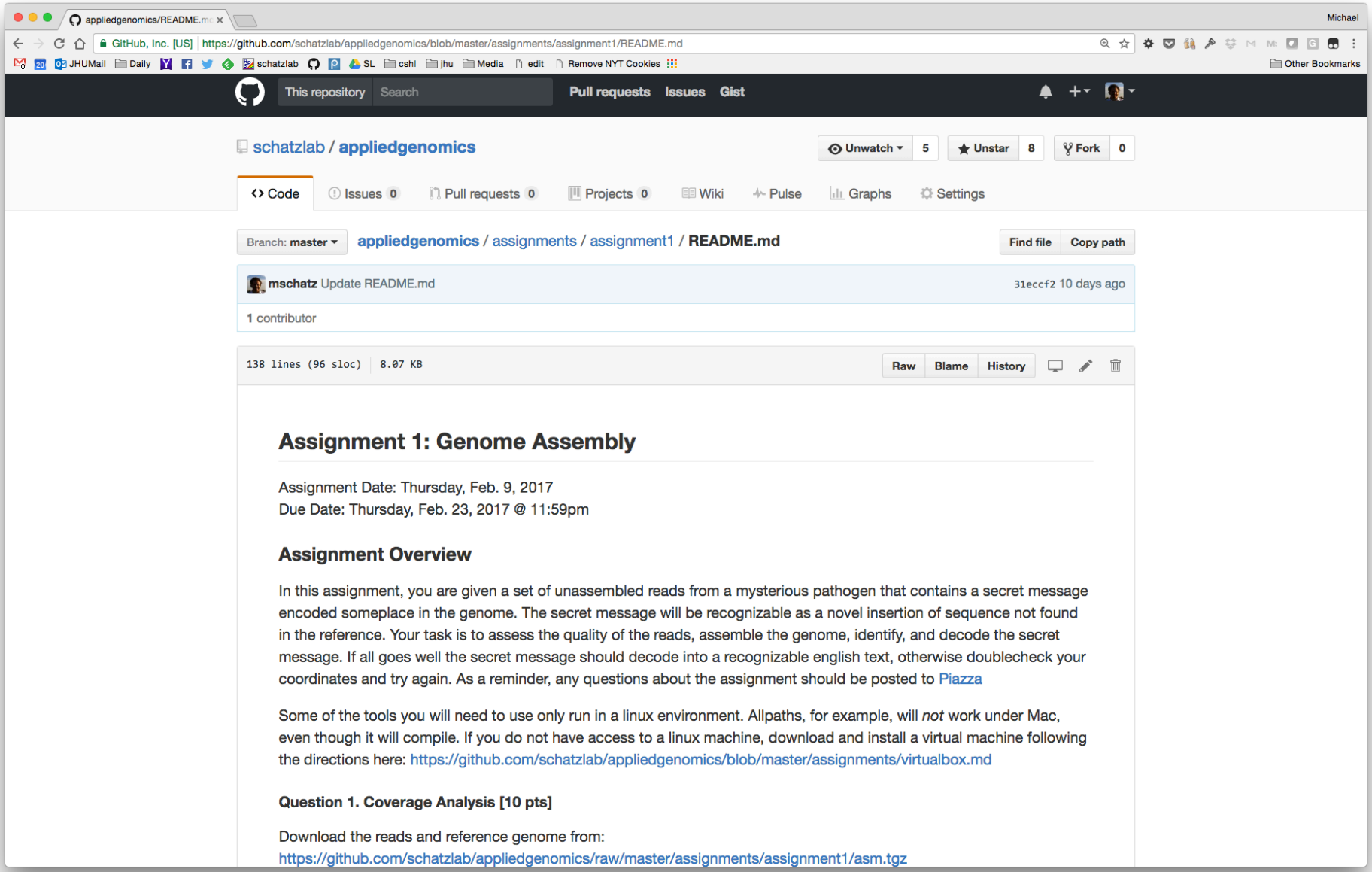
Michael Schatz

Feb 21, 2017

JHU 600.649: Applied Comparative Genomics

Assignment 1: Due Thursday @ 11:59pm

Email PDF to: jhuappliedgenomics@gmail.com



appliedgenomics/README.md

https://github.com/schatzlab/appliedgenomics/blob/master/assignments/assignment1/README.md

This repository Search Pull requests Issues Gist

schatzlab / appliedgenomics

Unwatch 5 Unstar 8 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Pulse Graphs Settings

Branch: master appliedgenomics / assignments / assignment1 / README.md Find file Copy path

mschatz Update README.md 31eccf2 10 days ago

1 contributor

138 lines (96 sloc) 8.07 KB Raw Blame History

Assignment 1: Genome Assembly

Assignment Date: Thursday, Feb. 9, 2017
Due Date: Thursday, Feb. 23, 2017 @ 11:59pm

Assignment Overview

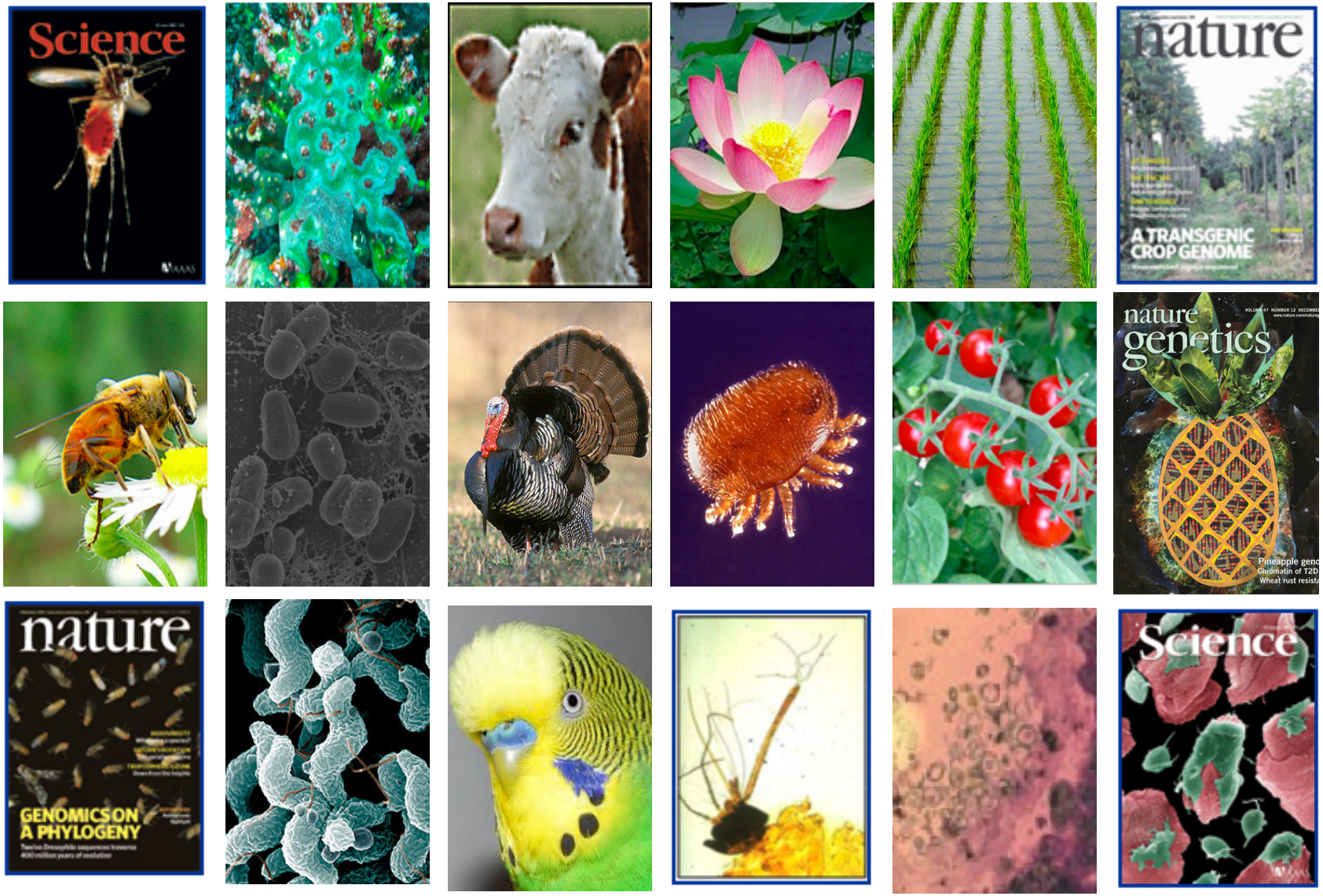
In this assignment, you are given a set of unassembled reads from a mysterious pathogen that contains a secret message encoded someplace in the genome. The secret message will be recognizable as a novel insertion of sequence not found in the reference. Your task is to assess the quality of the reads, assemble the genome, identify, and decode the secret message. If all goes well the secret message should decode into a recognizable english text, otherwise doublecheck your coordinates and try again. As a reminder, any questions about the assignment should be posted to Piazza

Some of the tools you will need to use only run in a linux environment. Allpaths, for example, will *not* work under Mac, even though it will compile. If you do not have access to a linux machine, download and install a virtual machine following the directions here: https://github.com/schatzlab/appliedgenomics/blob/master/assignments/virtualbox.md

Question 1. Coverage Analysis [10 pts]

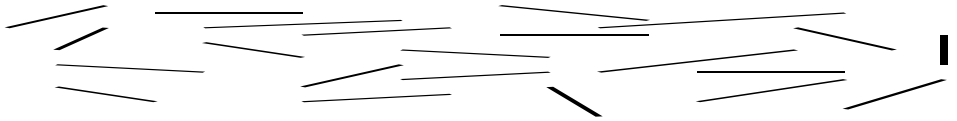
Download the reads and reference genome from:
https://github.com/schatzlab/appliedgenomics/raw/master/assignments/assignment1/asm.tgz

Genome Assembly



Assembling a Genome

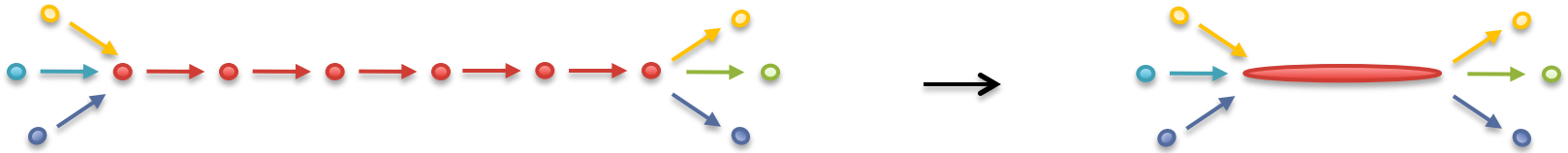
1. Shear & Sequence DNA



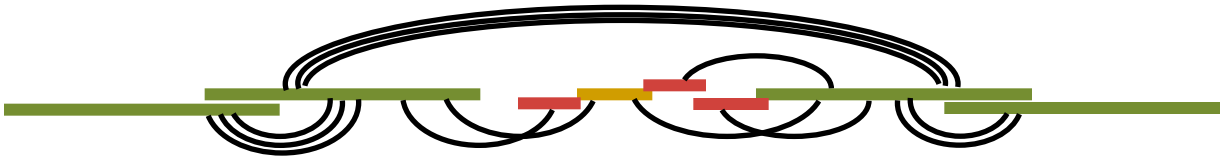
2. Construct assembly graph from reads (de Bruijn / overlap graph)

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

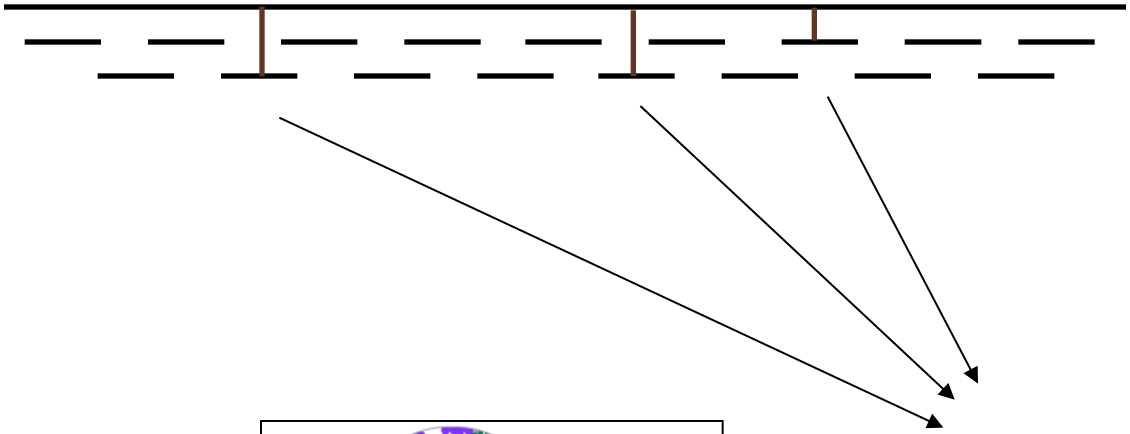


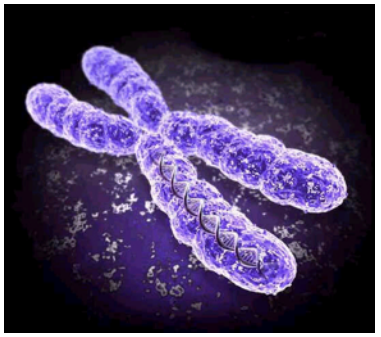
4. Detangle graph with long reads, mates, and other links

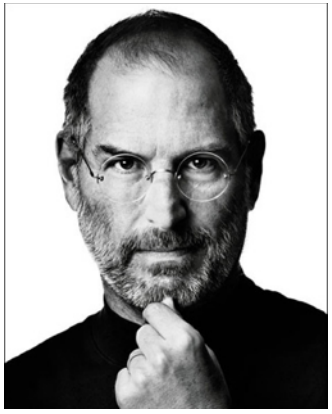


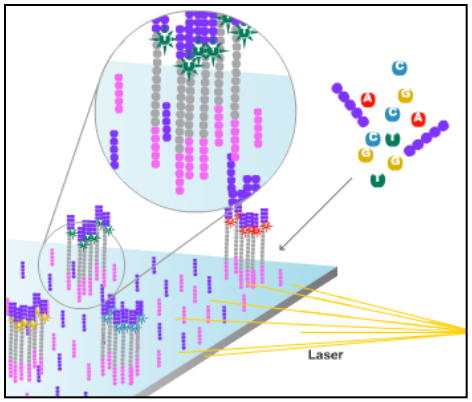
Personal Genomics

How does your genome compare to the reference?







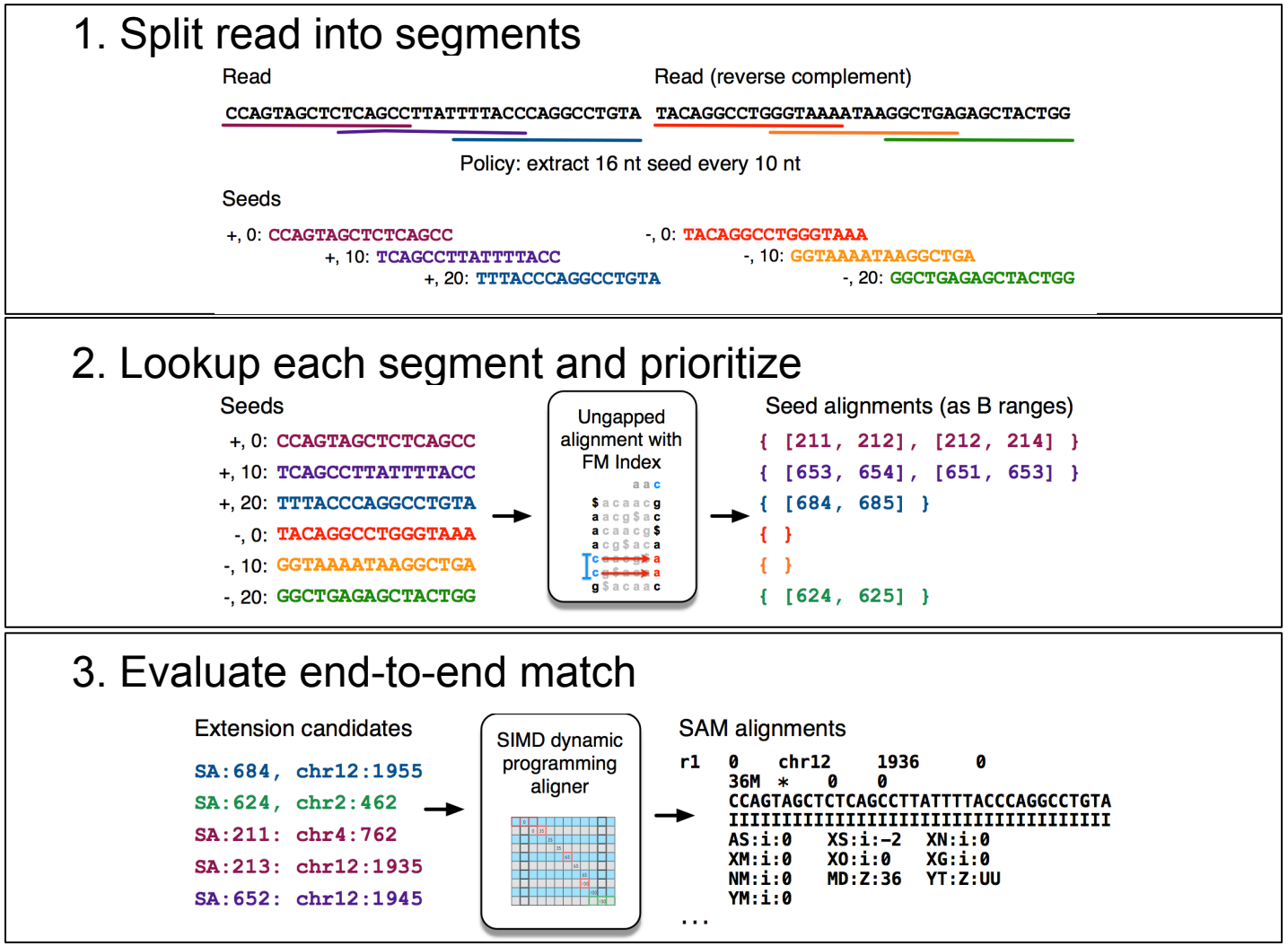


Heart Disease

Cancer

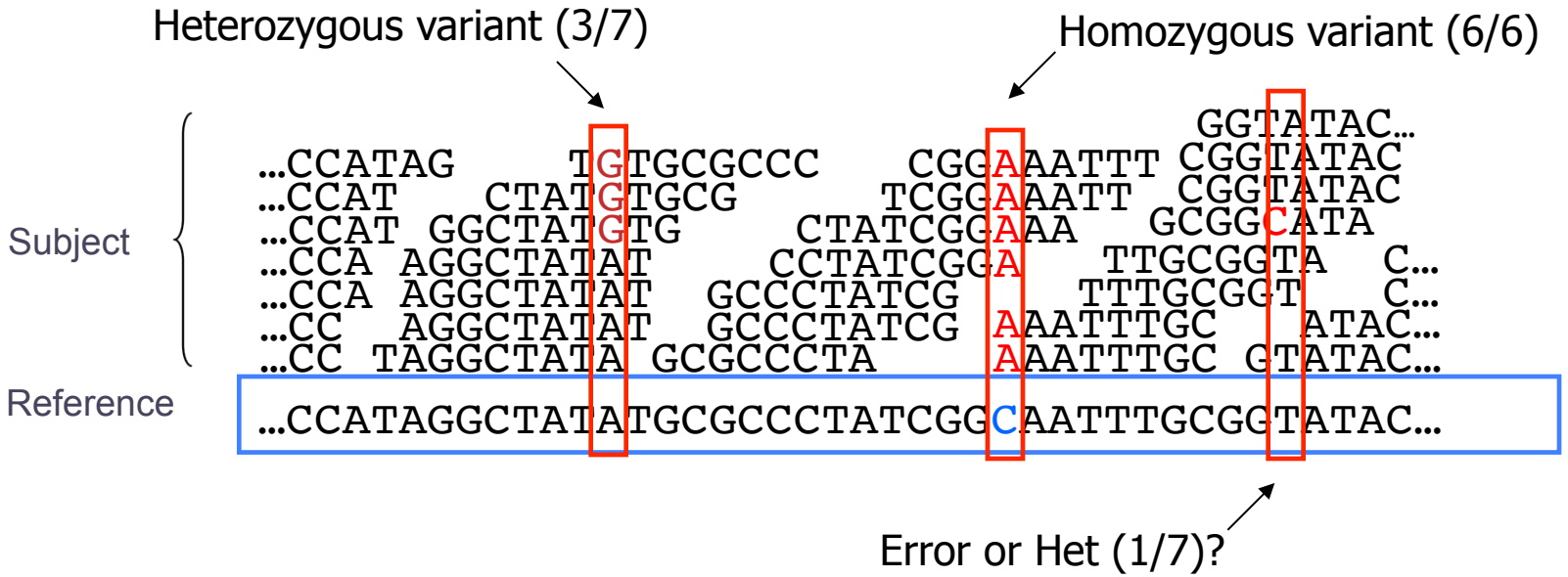
Creates magical technology

Read Mapping Overview



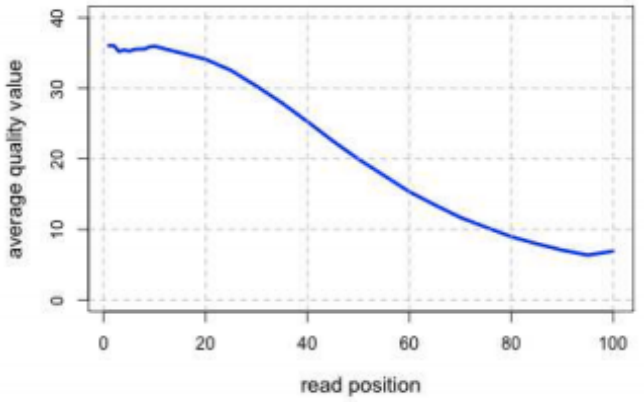
Fast gapped-read alignment with Bowtie 2
Langmead & Salzberg (2012) Nature Methods. doi:10.1038/nmeth.1923

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

The Binomial Distribution: Adventures in Coin Flipping

P(heads) = 0.5

P(tails) = 0.5

Aaron Quinlan

Thinking about allele sampling with the binomial distribution

The **binomial distribution** with parameters *n* and *p* is the discrete probability distribution of the number of successes in a sequence of *n* independent yes (e.g., "heads" or "reference allele") or no (e.g., "tails", or "alternate allele") experiments, each of which yields success with probability *p*.

The probability of getting exactly k successes in n trials is given by the probability mass function:

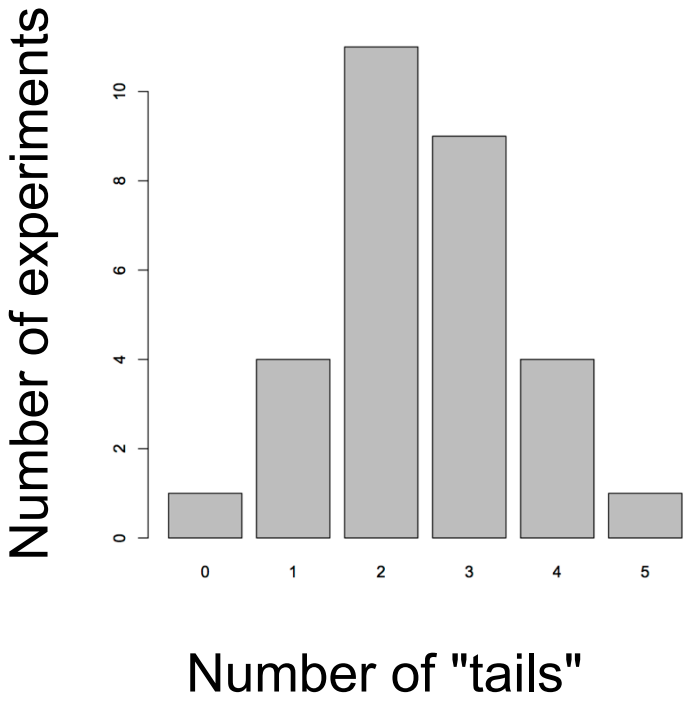
$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

What is the probability of seeing k=1 tails in n=3 flips of a fair coin with the probability of a tail (p) = 0.5?

3 choose 1 = 3; 0.5^1 = 0.5; $(1-0.5)^{(3-1)}$ = 0.25. So…. 3*0.5*0.25 = **0.375**

In R, the function would be: `dbinom(1, size=3, prob=0.5)`

What is the distribution of tails (alternate alleles) do we expect to see after 5 tosses (sequence reads)?

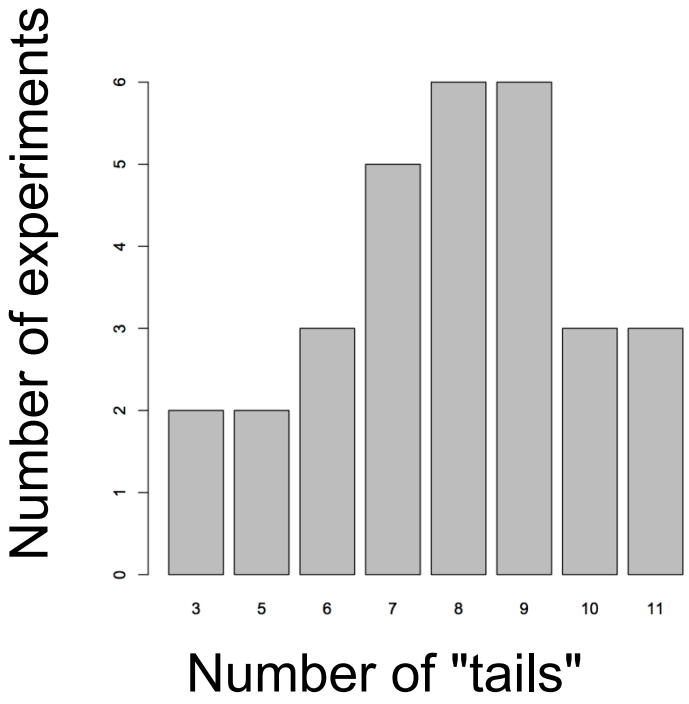


R code:

```
barplot(table(rbinom(30, 5, 0.5)))
```

30 experiments (students tossing coins)
5 tosses each
Probability of Tails

What is the distribution of tails (alternate alleles) do we expect to see after 15 tosses (sequence reads)?

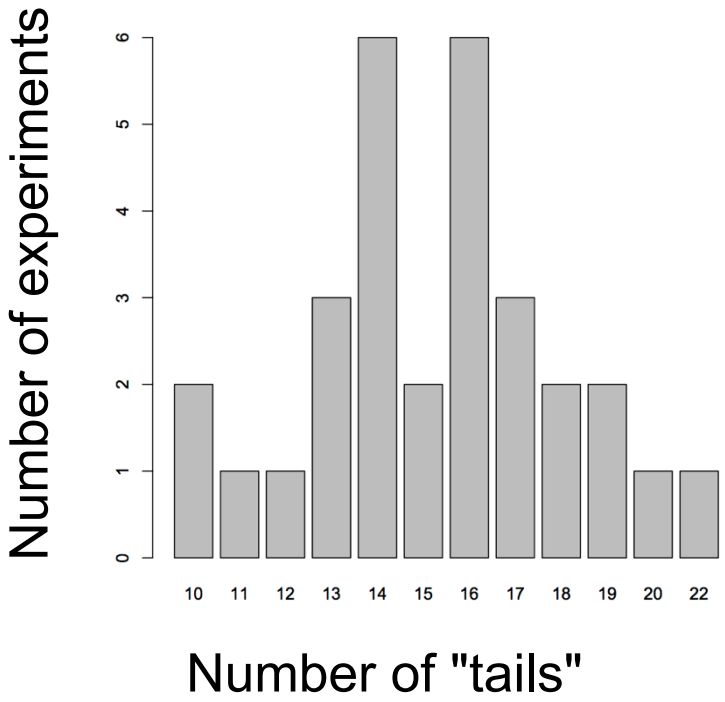


R code:

```
barplot(table(rbinom(30, 15, 0.5)))
```

30 experiments (students tossing coins)
15 tosses each
Probability of Tails

What is the distribution of tails (alternate alleles) do we expect to see after 30 tosses (sequence reads)?

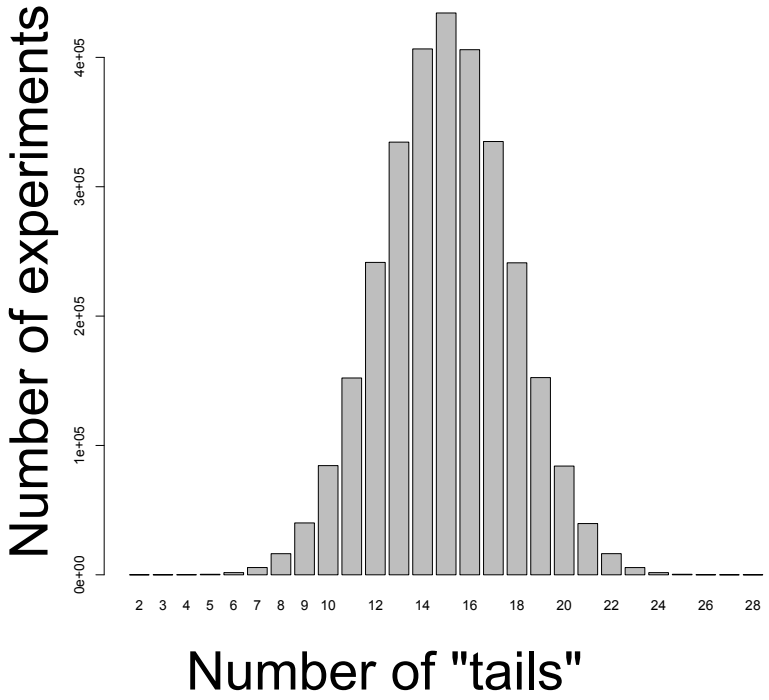


R code:

```
barplot(table(rbinom(30, 30, 0.5)))
```

30 experiments (students tossing coins)
30 tosses each
Probability of Tails

What is the distribution of tails (alternate alleles) do we expect to see after 30 tosses (sequence reads)?

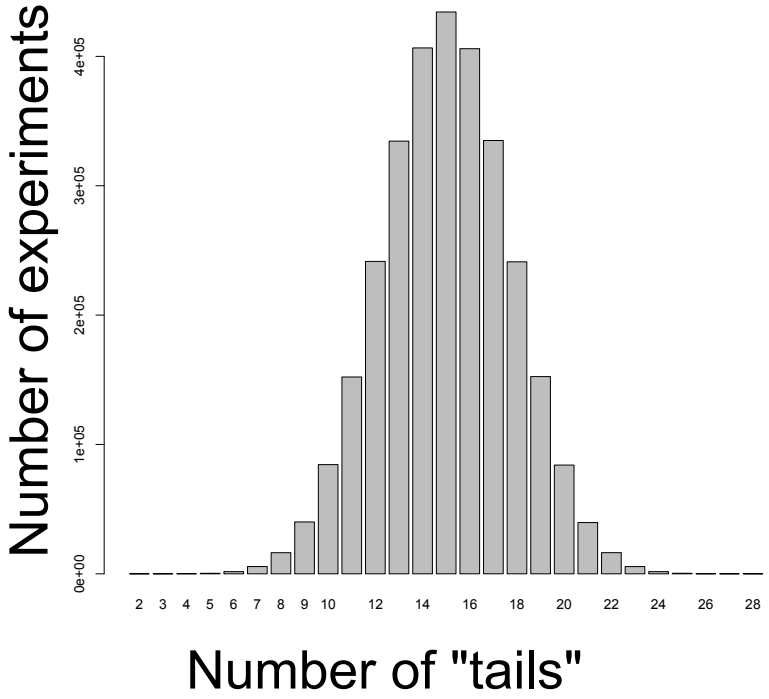


R code:

```
barplot(table(rbinom(3e6, 30, 0.5)))
```

3M experiments (students tossing coins)
30 tosses each
Probability of Tails

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

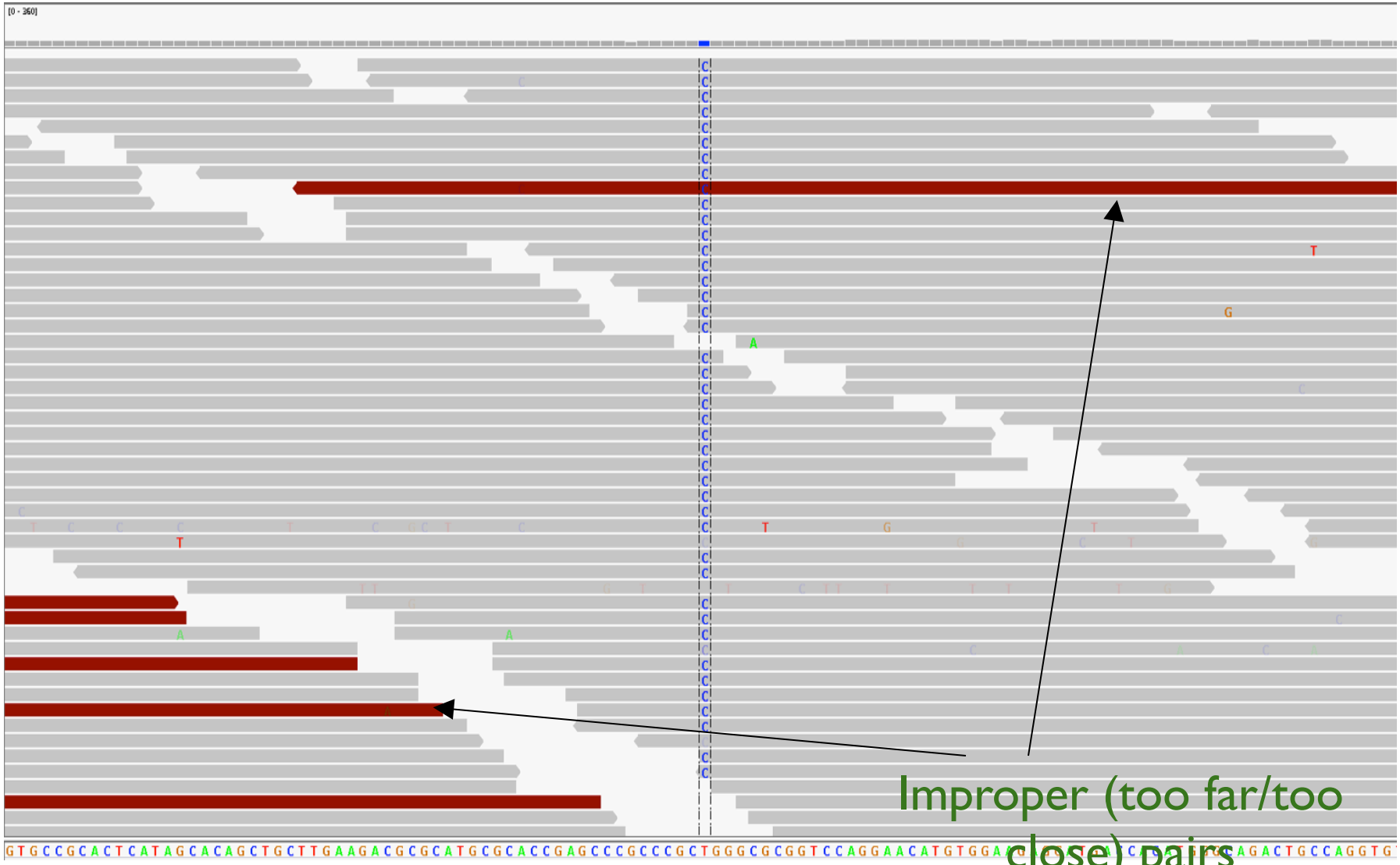


This is why at least a "30X" (30 fold sequence coverage) genome is recommended: it confers sufficient power to distinguish heterozygous alleles and from mere sequencing errors

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

Some real examples of SNPs in IGV

Homozygous for the "C" allele

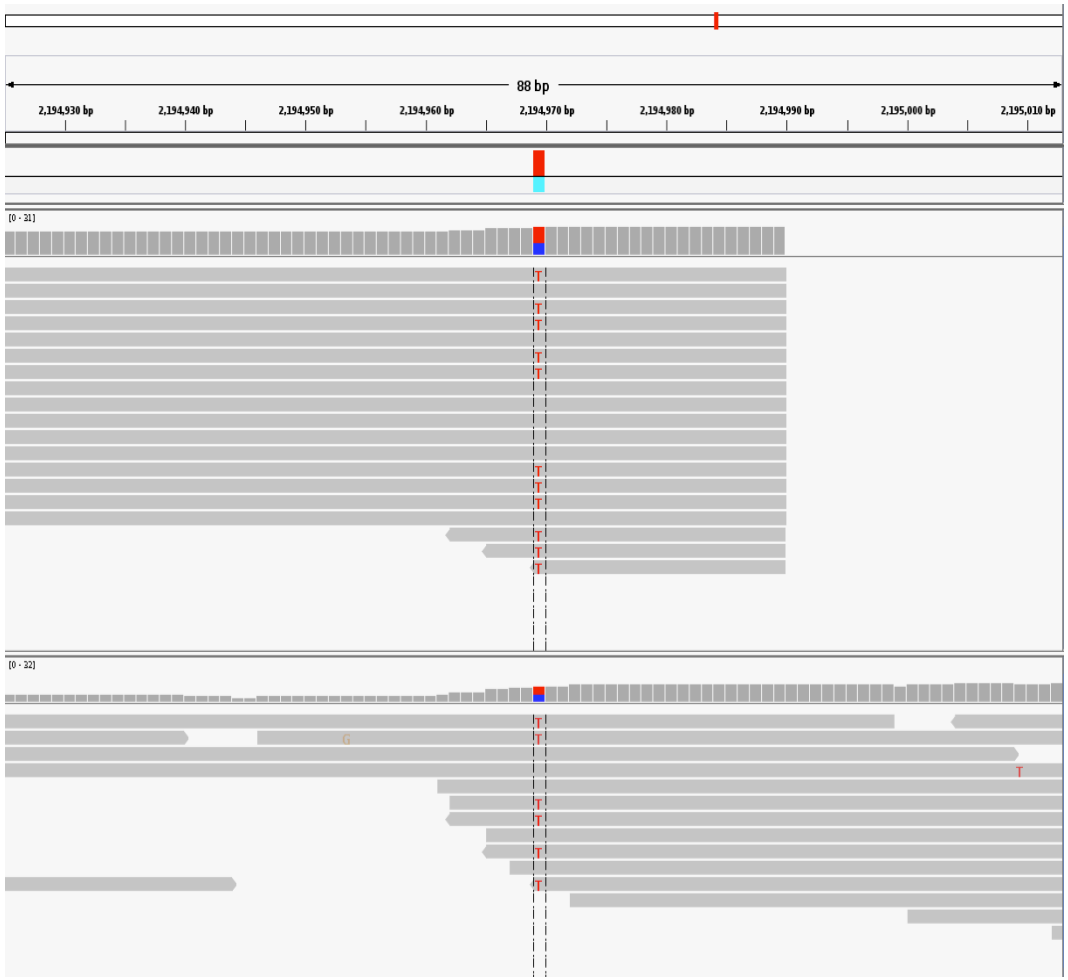


What else do you notice?

Heterozygous for the alternate allele

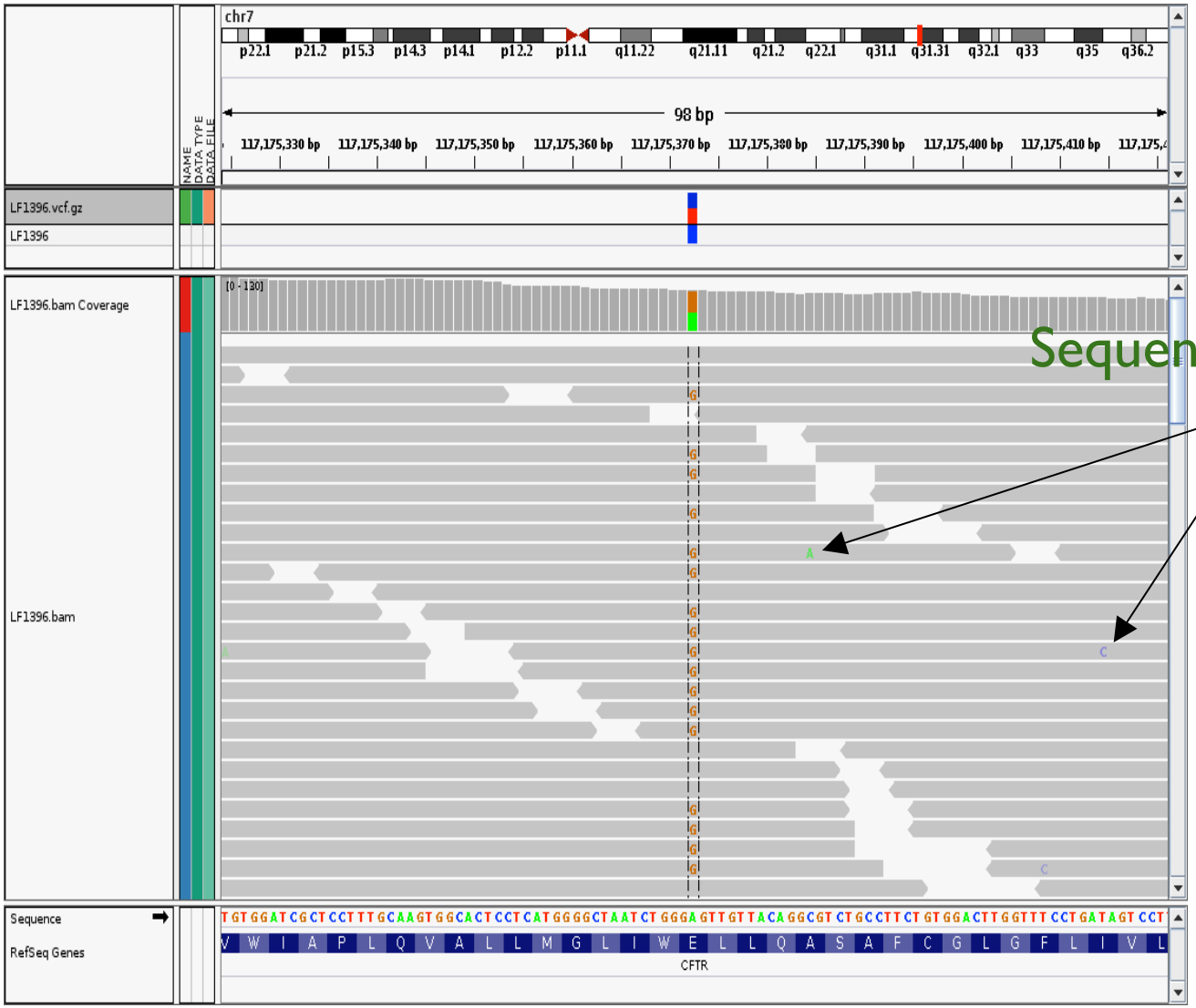
Individual 1

Individual 2



Which genotype prediction do you have more confidence in?

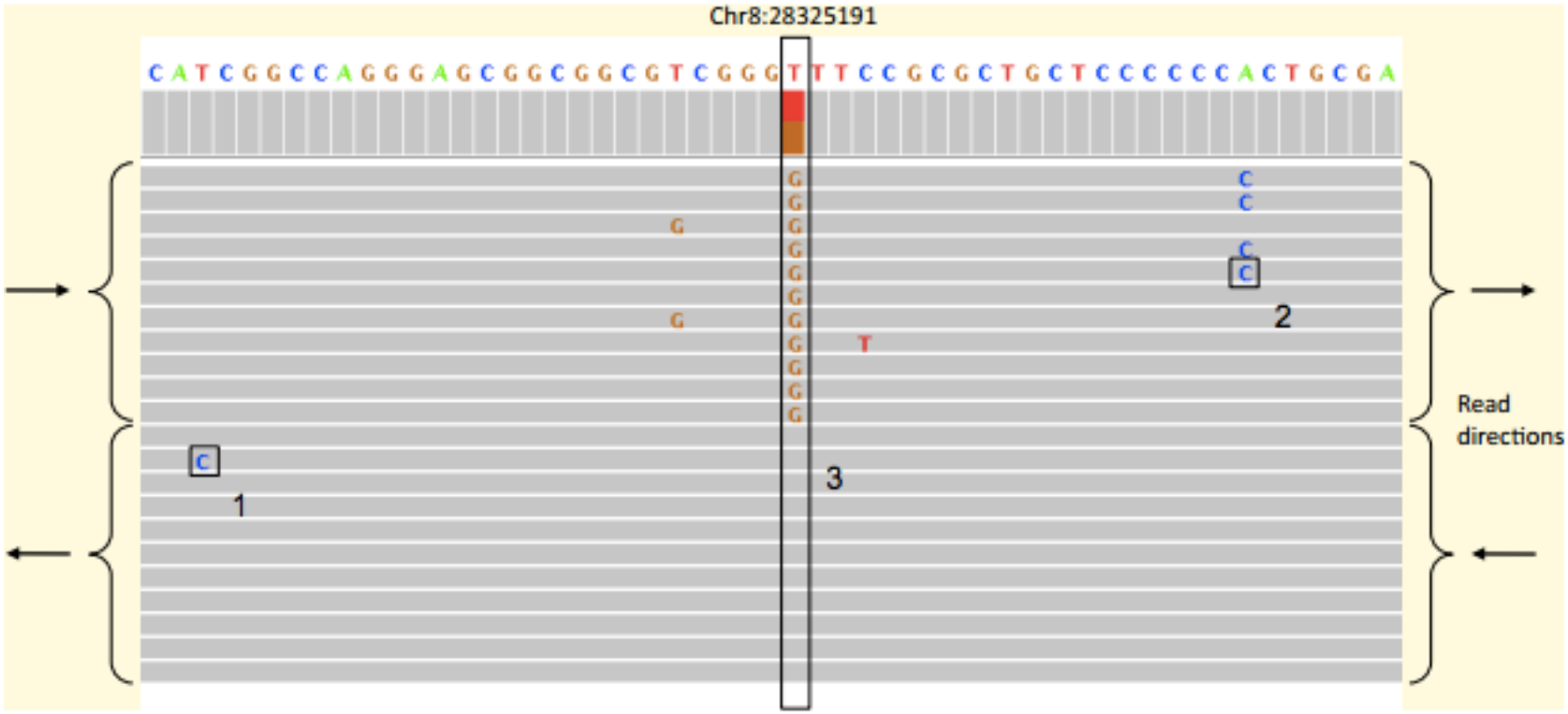
Sequencing errors fall out as noise (most of the time)



https://jchoigt.files.wordpress.com/2012/07/igv_e217g_snapshot.png

It is not always so easy ☹

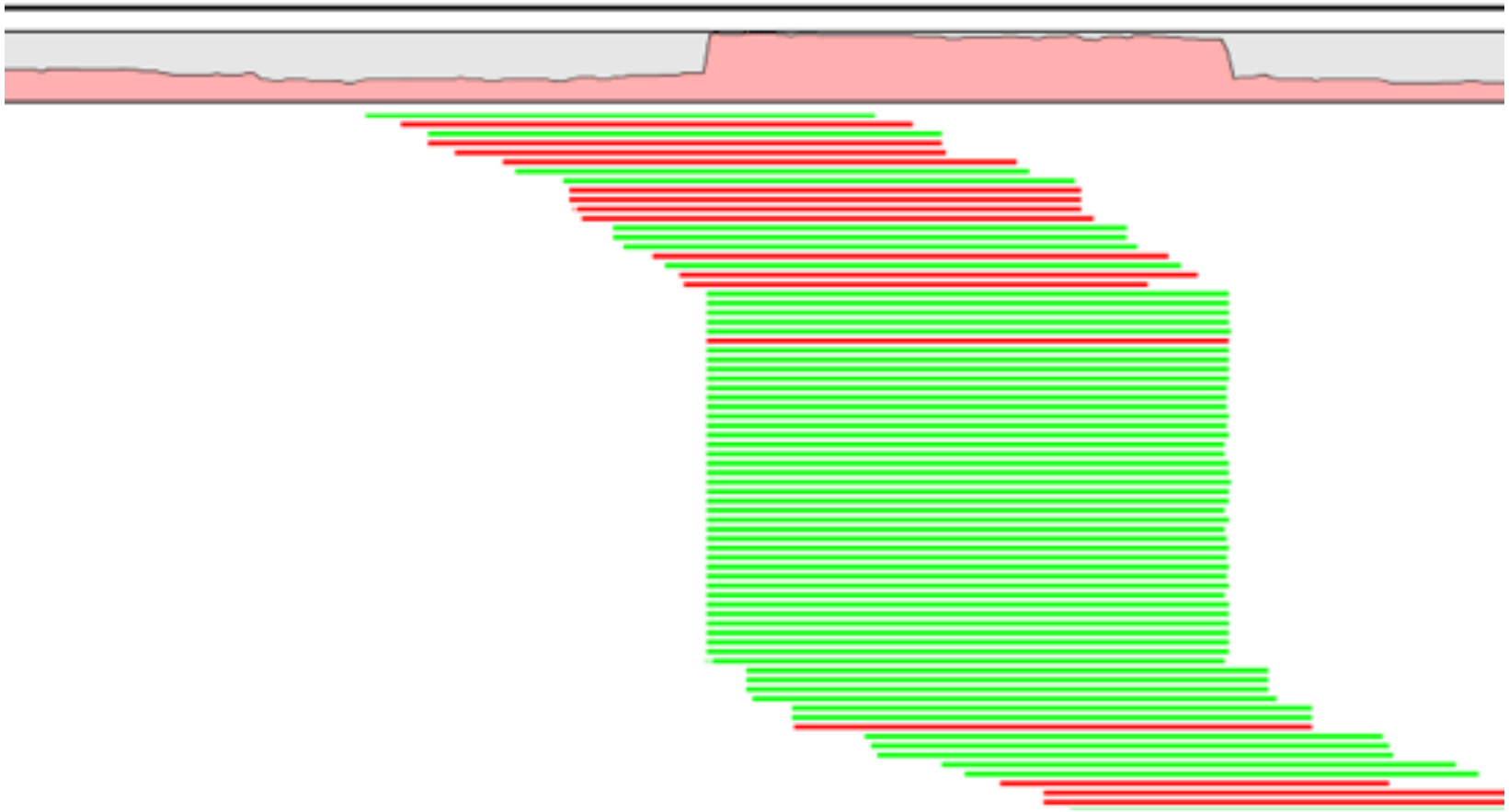
Beware of Systematic Errors



Identification and correction of systematic error in high-throughput sequence data
Meacham et al. (2011) *BMC Bioinformatics.* 12:451

A closer look at RNA editing.
Lior Pachter (2012) *Nature Biotechnology.* 30:246-247

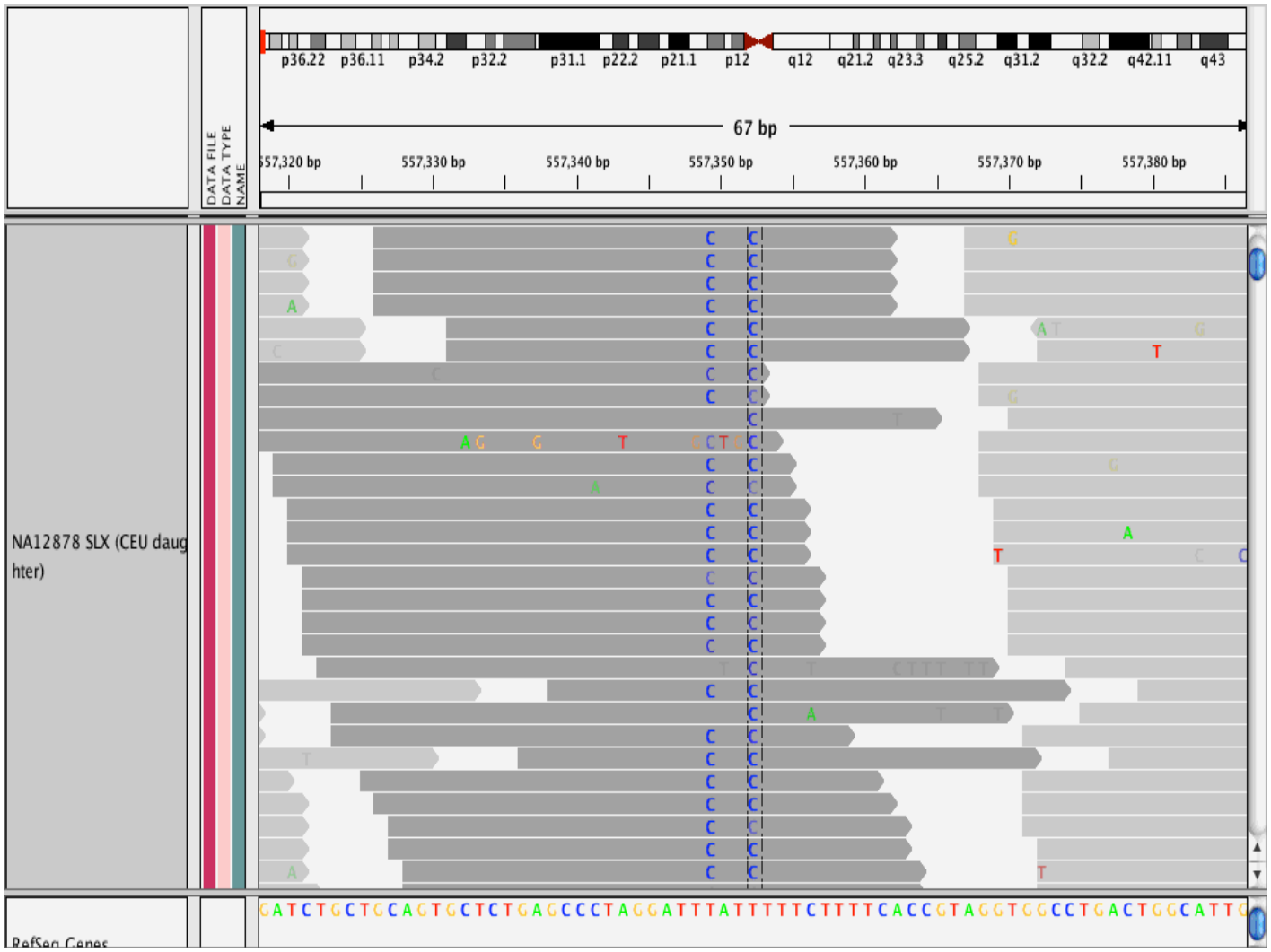
Beware of Duplicate Reads



The Sequence alignment/map (SAM) format and SAMtools.
Li et al. (2009) *Bioinformatics.* 25:2078-9

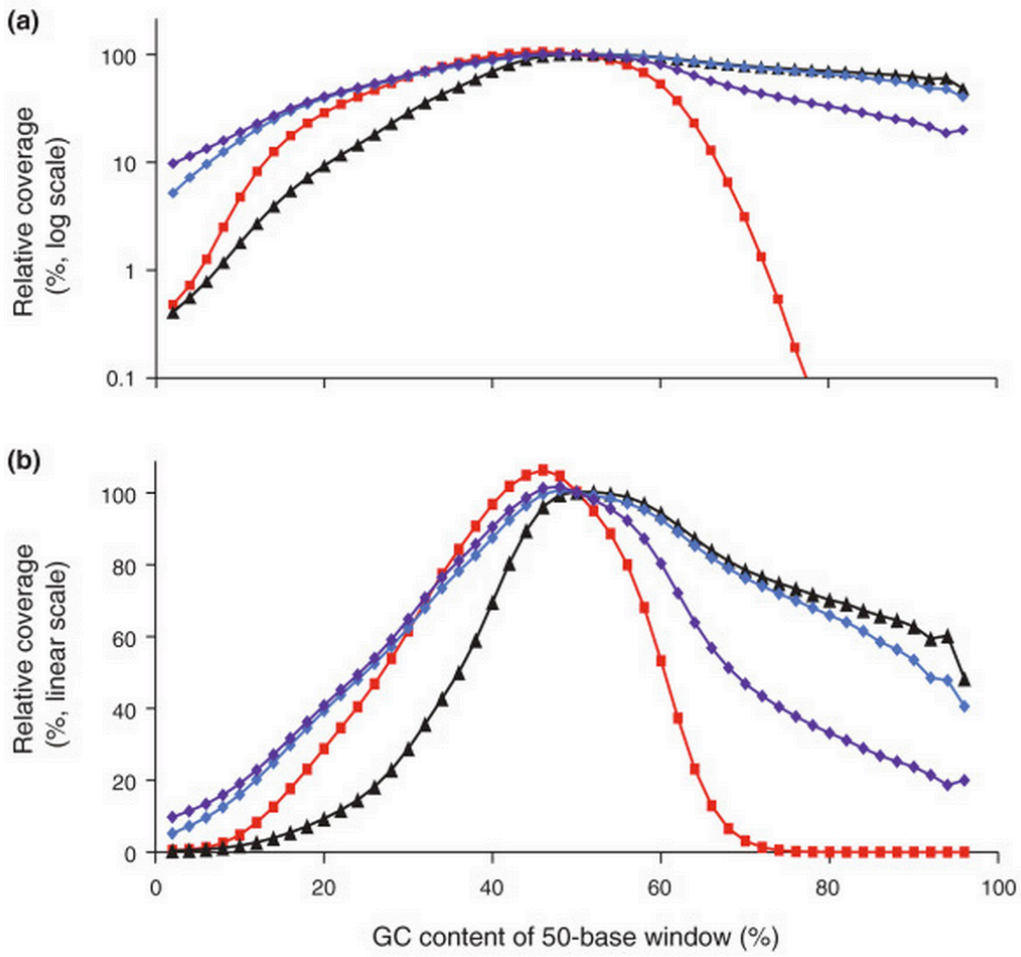
Picard: http://picard.sourceforge.net

Beware of strand bias from PCR



http://gatkforums.broadinstitute.org/gatk/discussion/49/using-variant-annotator

Beware of GC Biases

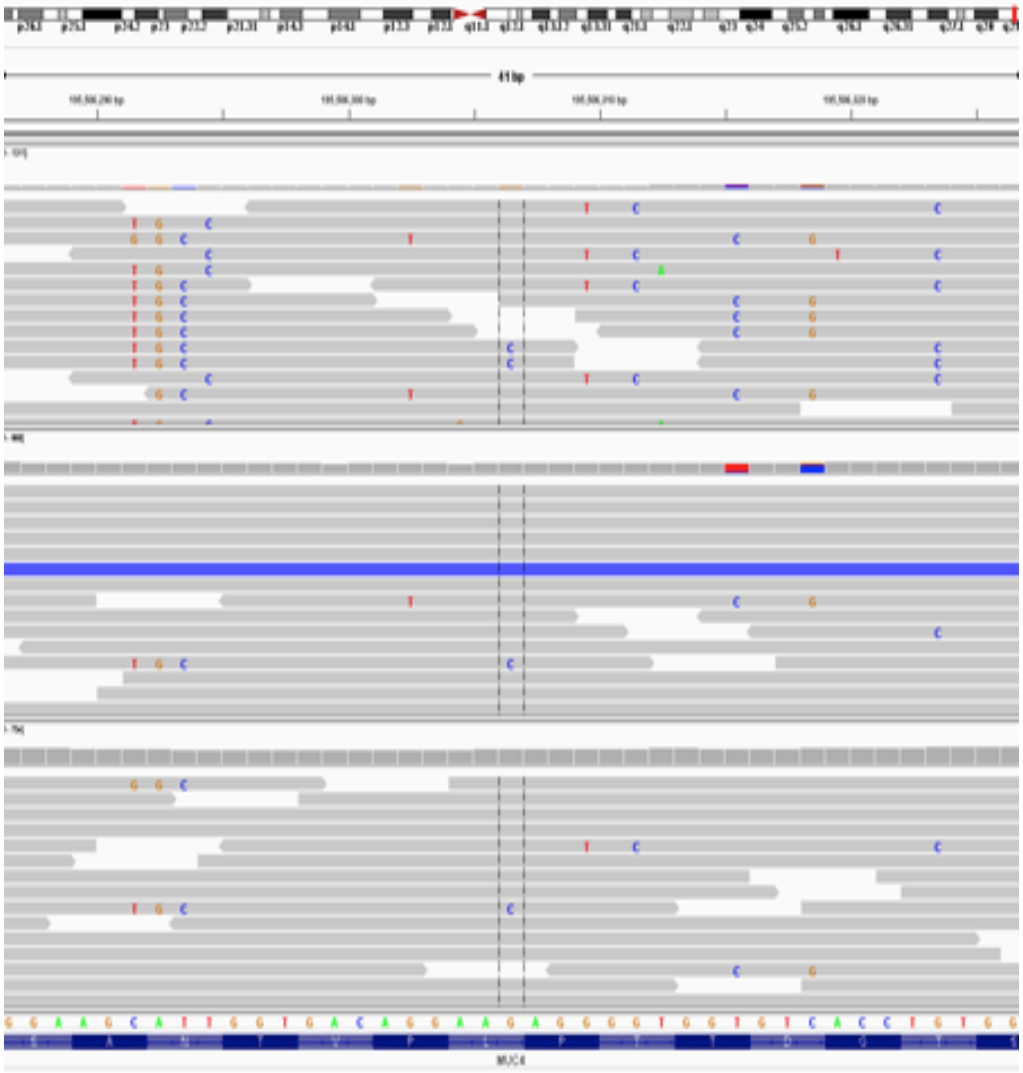


Illumina sequencing does not produce uniform coverage over the genome

- Coverage of extremely high or extremely low GC content will have reduced coverage in Illumina sequencing
- Biases primarily introduced during PCR; lower temperatures, slower heating, and fewer rounds minimize biases
- This makes it very difficult to identify variants (SNPs, CNVs, etc) in certain regions of the genome

Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.
Aird et al. (2011) *Genome Biology.* 12:R18.

Pileups of many differences from paralogy



RESEARCH ARTICLE | OPEN ACCESS

FLAGS, frequently mutated genes in public exomes

Casper Shyr, Maja Tarailo-Graovac, Michael Gottlieb, Jessica JY Lee, Clara van Karnebeek and Wyeth W Wasserman

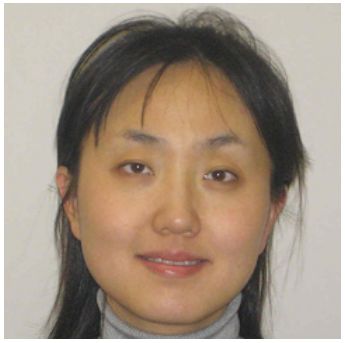
BMC Medical Genomics 2014 7:64 | DOI: 10.1186/s12920-014-0064-y | © Shyr et al.; licensee BioMed Central Ltd. 2014

Received: 16 June 2014 | Accepted: 24 October 2014 | Published: 3 December 2014

Open Peer Review reports

http://massgenomics.org/2013/06/ngs-false-positives.html

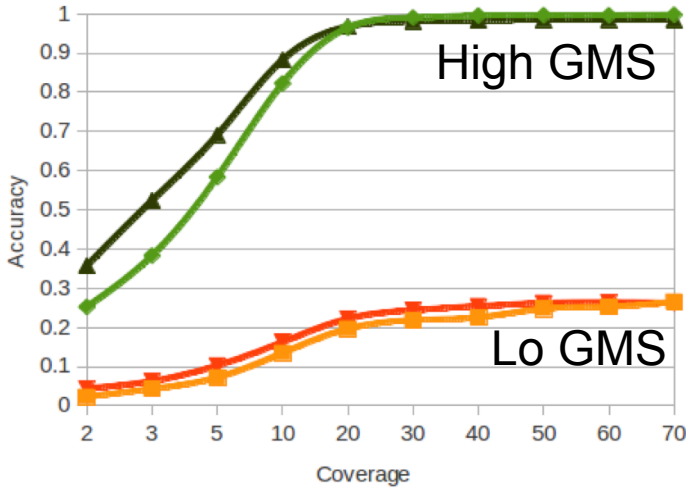
Beware of Mapping Errors



- Short read mapping is a essential for identifying mutations in the genome
  - Not every base of the genome can mapped equally well, especially because of repeats

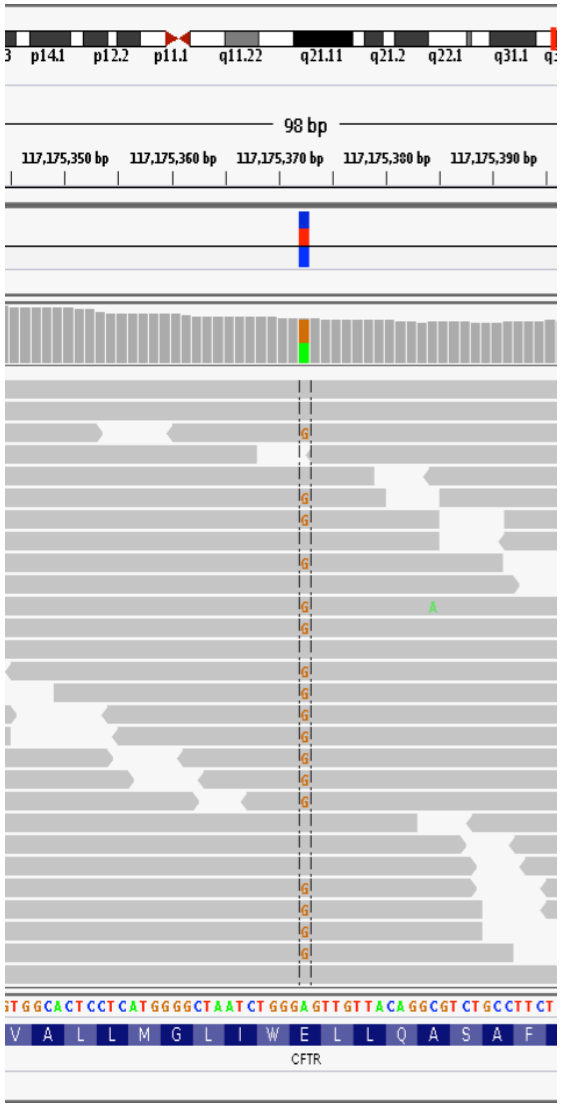
- Introduced a new probabilistic metric - the Genome Mappability Score - that quantifies how reliably reads can be mapped to every position in the genome
  - We have little power to measure 11-13% of the human genome, including of known clinically relevant variations
  - Errors in variation discovery are dominated by errors in low GMS regions

Species (build)	size	paired/single	whole (%)	transcription (%)
yeast (sc2)	12 Mbp	paired	94.85	95.04
		single	94.25	94.62
fly (dm3)	130 Mbp	paired	90.52	96.14
		single	89.70	95.94
mouse (mm9)	2.7 Gbp	paired	89.39	96.03
		single	87.47	94.75
human (hg19)	3.0 Gbp	paired	89.02	97.40
		single	87.79	96.38



Genomic Dark Matter: The reliability of short read mapping illustrated by the GMS.
Lee and Schatz (2012) *Bioinformatics*. doi: 10.1093/bioinformatics/bts330

What information is needed to decide if a variant exists?



- Depth of coverage at the locus
- Bases observed at the locus
- The base qualities of each allele
- The strand composition
- Mapping qualities
- Proper pairs?
- Expected polymorphism rate

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r99

PolyBayes: The first statistically rigorous variant detection tool.

letter

A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth[1], Ian Korf[1], Mark D. Yandell[1], Raymond T. Yeh[1], Zhijie Gu[2], Hamideh Zakeri[2], Nathan O. Stitziel[1], LaDeana Hillier[1], Pui-Yan Kwok[2] & Warren R. Gish[1]

Its main innovation was the use of Bayes's theorem

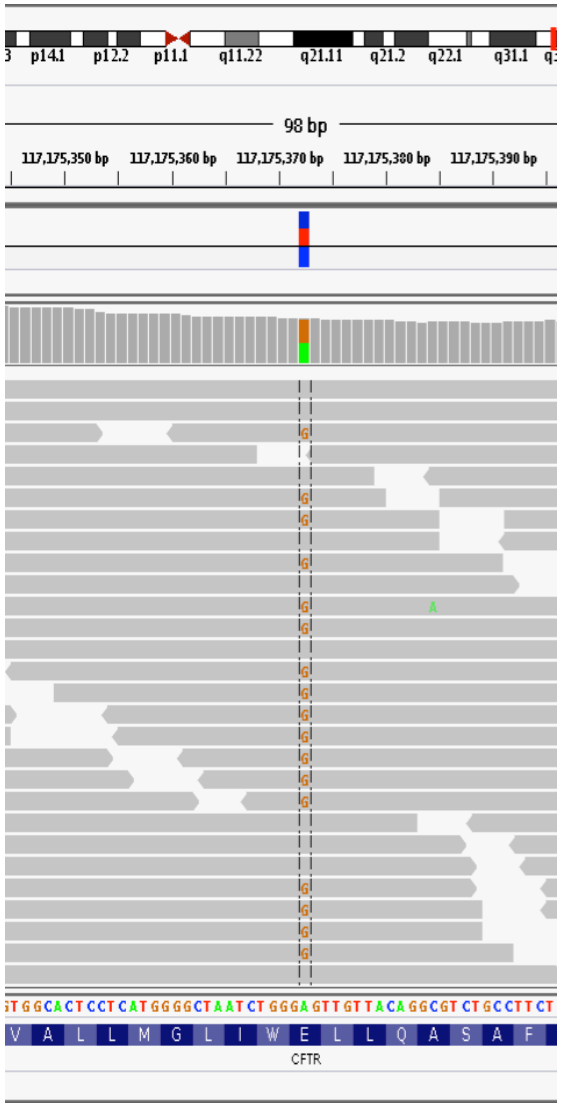
http://www.nature.com/ng/journal/v23/n4/full/ng1299_452.html

Bayes theorem

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Conditional probability. That is, the probability of A occurring, given that B has occurred.

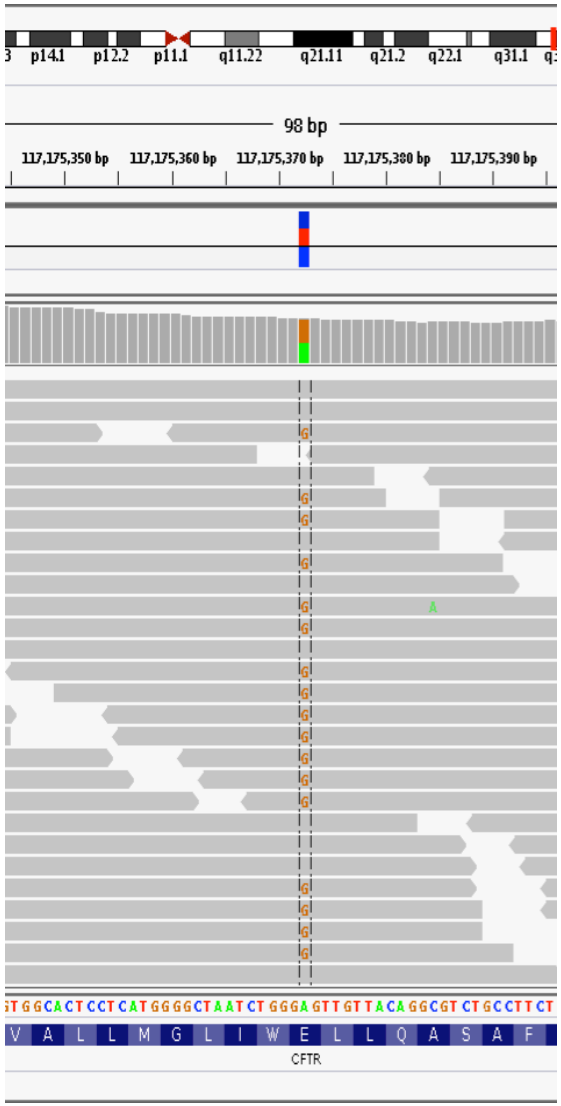
Bayesian SNP calling



$$P(SNP|Data) = \frac{P(Data|SNP) * P(SNP)}{P(Data)}$$

http://www.nature.com/ng/journal/v23/n4/full/ng1299_452.html

Bayesian SNP calling



$$P(SNP|Data) = \frac{P(Data|SNP) * P(SNP)}{P(Data)}$$

- Depth of coverage at the locus
- Bases observed at the locus
- The base qualities of each allele
- Transition or Transversion? Which type?
- The strand composition
- Mapping qualities
- Proper pairs?
- Expected polymorphism rate

http://www.nature.com/ng/journal/v23/n4/full/ng1299_452.html

PolyBayes: The first statistically rigorous variant detection tool.

letter

A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth[1], Ian Korf[1], Mark D. Yandell[1], Raymond T. Yeh[1], Zhijie Gu[2], Hamideh Zakeri[2], Nathan O. Stitziel[1], LaDeana Hillier[1], Pui-Yan Kwok[2] & Warren R. Gish[1]

Bayesian posterior probability

Base call + Base quality

Expected (prior) polymorphism rate

$$P(SNP) = \sum_{all\ variable\ S} \frac{\frac{P(S_1 \mid R_1)}{P_{Prior}(S_1)} \cdot \ldots \cdot \frac{P(S_N \mid R_N)}{P_{Prior}(S_N)} \cdot P_{Prior}(S_1,\ldots,S_N)}{\sum_{S_{i_1} \in [A,C,G,T]} \ldots \sum_{S_{i_N} \in [A,C,G,T]} \frac{P(S_{i_1} \mid R_1)}{P_{Prior}(S_{i_1})} \cdot \ldots \cdot \frac{P(S_{i_N} \mid R_1)}{P_{Prior}(S_{i_N})} \cdot P_{Prior}(S_{i_1},\ldots,S_{i_N})}$$

Probability of observed base composition (should model sequencing error rate)

http://www.nature.com/ng/journal/v23/n4/full/ng1299_452.html

PolyBayes: The first statistically rigorous variant detection tool.

letter

A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth[1], Ian Korf[1], Mark D. Yandell[1], Raymond T. Yeh[1], Zhijie Gu[2], Hamideh Zakeri[2], Nathan O. Stitziel[1], LaDeana Hillier[1], Pui-Yan Kwok[2] & Warren R. Gish[1]

This Bayesian statistical framework has been adopted by other modern SNP/INDEL callers such as FreeBayes, GATK, and samtools

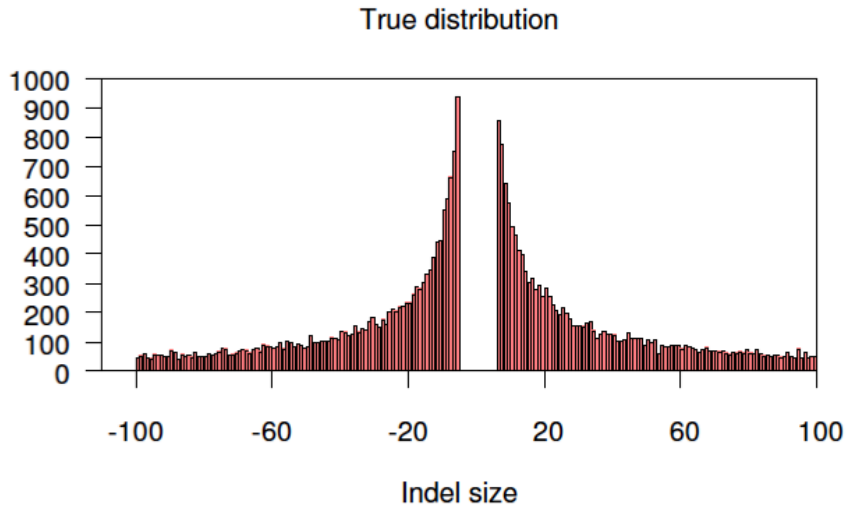
http://www.nature.com/ng/journal/v23/n4/full/ng1299_452.html

Variation Detection Complexity

SNPs + Short Indels

High precision and sensitivity

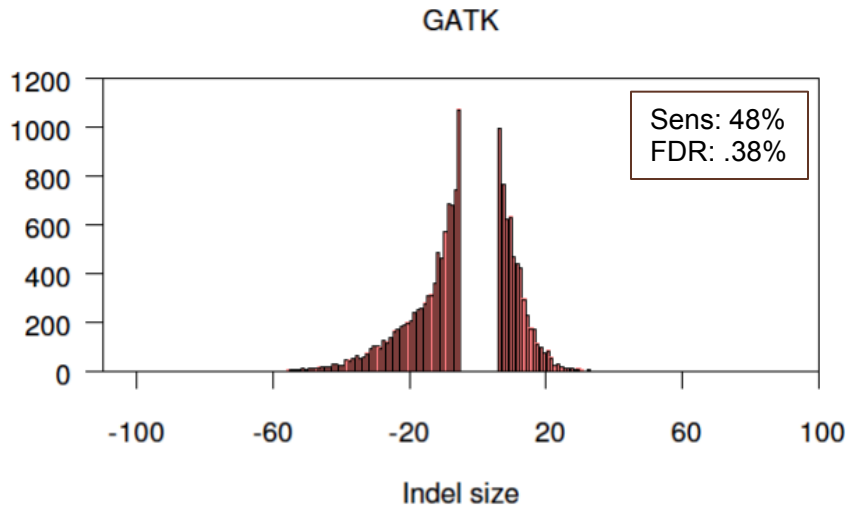
```
..TTTAGAATAG-CGAGTGC...
      ||||||| ||||
      AGAATAGGCGAG
```



"Long" Indels (>5bp)

Reduced precision and sensitivity

```
..TTTAG--------AGTGC...
  |||||         |||||
  TTTAGAATAGGC  |||||
        ATAGGCGAGTGC
```



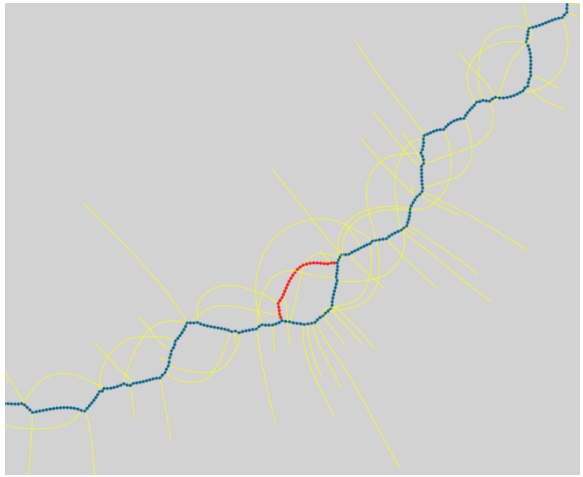
Analysis confounded by sequencing errors, localized repeats, allele biases, and mismapped reads

Scalpel: Haplotype Microassembly

DNA sequence **micro-assembly** pipeline for accurate detection and validation of *de novo* mutations (SNPs, indels) within exome-capture data.

Features

1. Combine mapping and assembly
2. Exhaustive search of haplotypes
3. De novo mutations

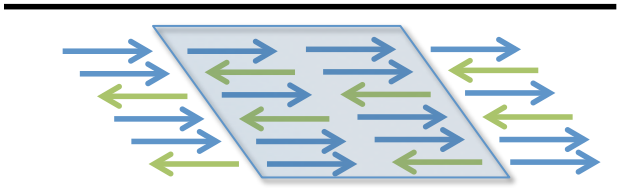


NRXN1 *de novo* SNP
(auSSC12501 chr2:50724605)

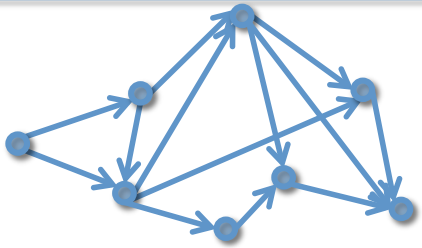
Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi et al. (2014) *Nature Methods*. doi:10.1038/nmeth.3069

Scalpel Algorithm

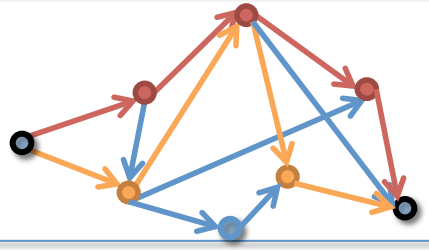
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



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

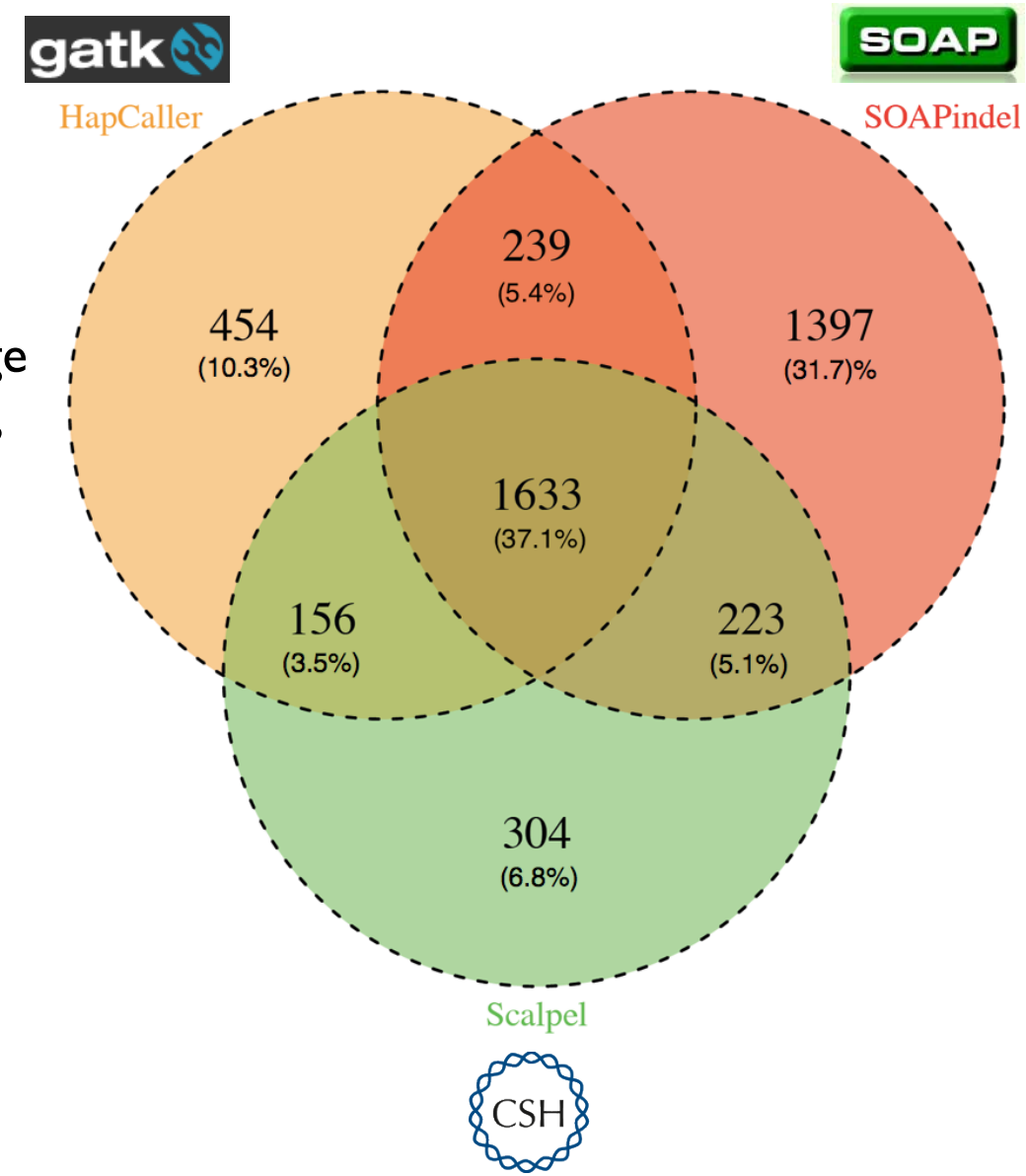
Experimental Analysis & Validation

Selected one deep coverage exome for deep analysis

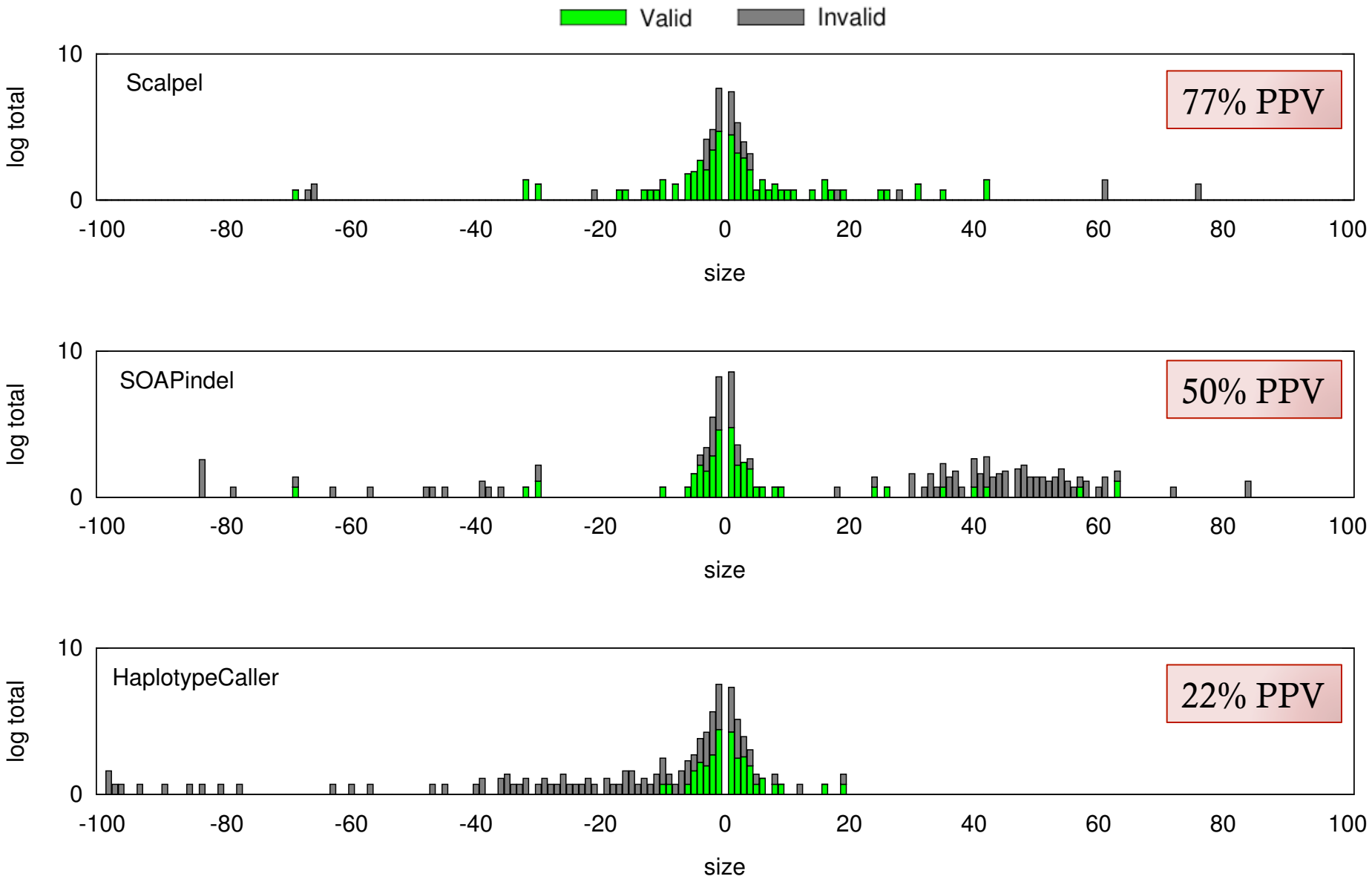
- Individual was diagnosed with ADHD and turrets syndrome
- 80% of the target at >20x coverage
- Evaluated with Scalpel, SOAPindel, and GATK Haplotype Caller

1000 indels selected for validation

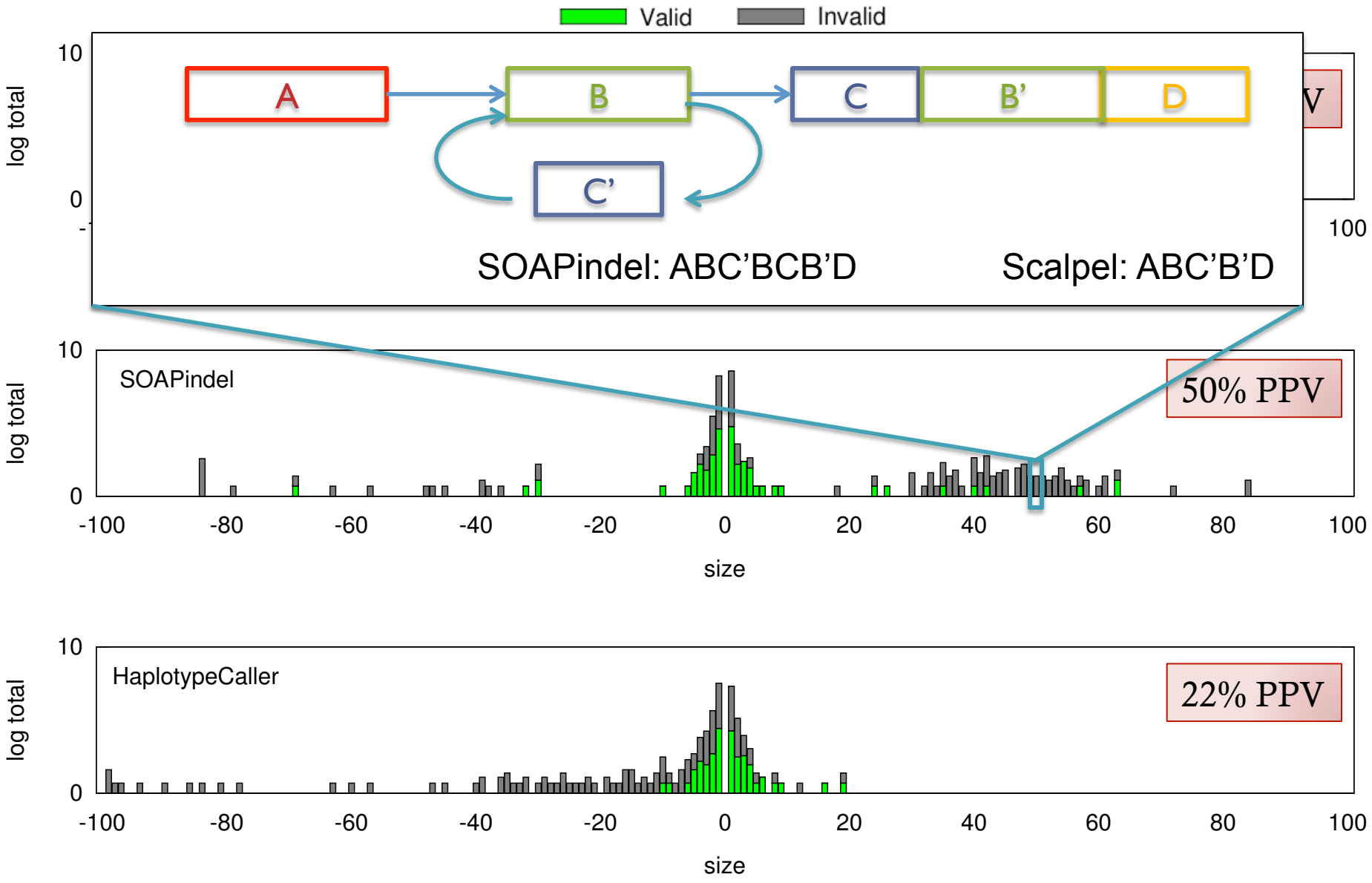
- 200 Scalpel
- 200 GATK Haplotype Caller
- 200 SOAPindel
- 200 within the intersection
- 200 long indels (>30bp)



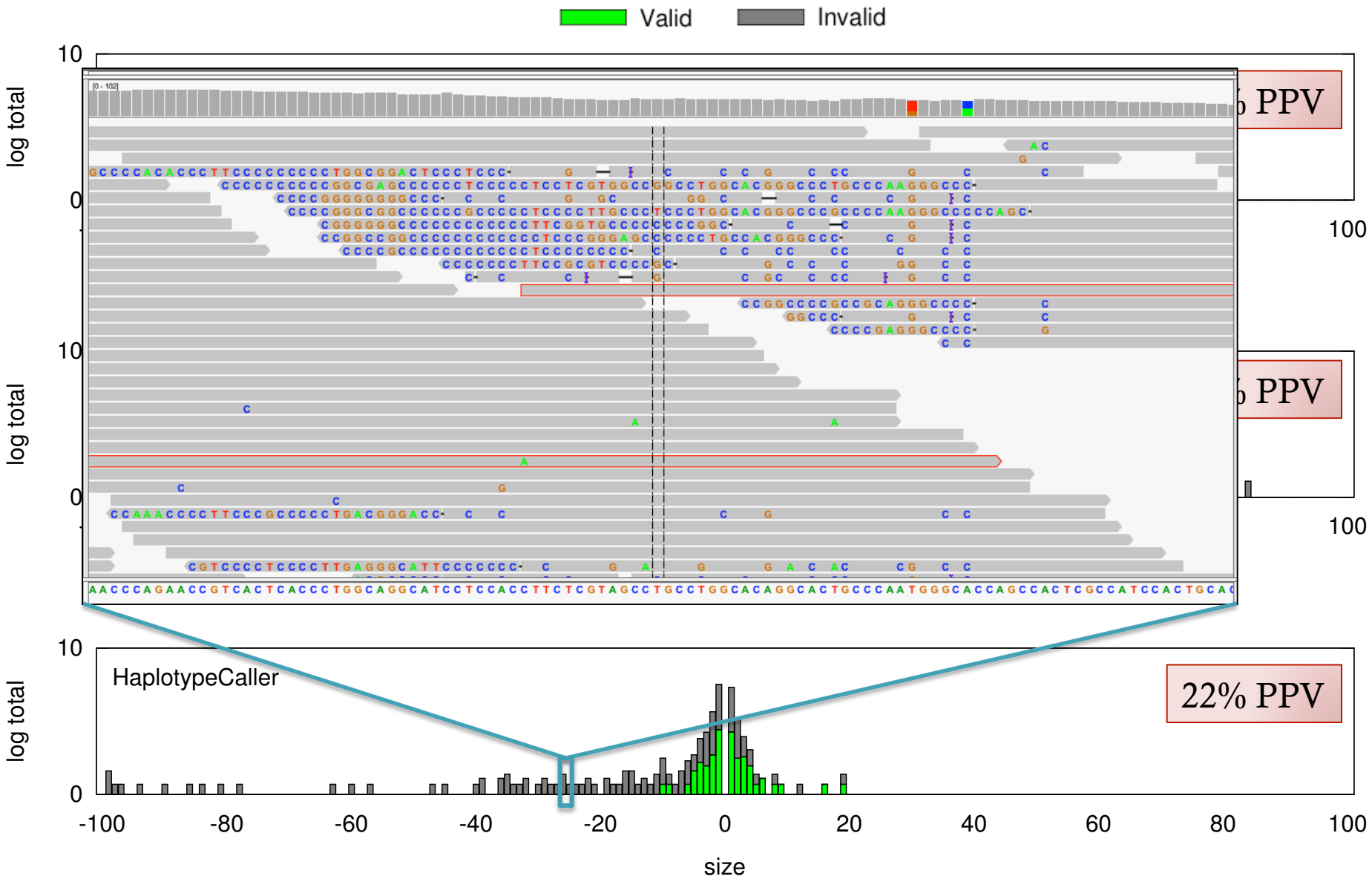
Scalpel Indel Validation



Scalpel Indel Validation



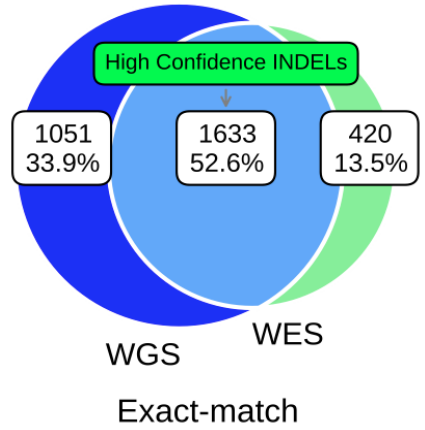
Scalpel Indel Validation



Refined indel analysis

Examine sources of indel errors

- Experimental validation of indels called from 30x whole genome vs. 110x whole exome of the same sample
- Most of the errors due to short microsatellite errors introduced during exome capture, also misses most long indels
- Recommend WGS for indel analysis instead



	All INDELs	Valid	PPV	INDELs >5bp	Valid (>5bp)	PPV (>5bp)
Intersection	160	152	95.0%	18	18	100%
WGS	145	122	84.1%	33	25	75.8%
WES	161	91	56.5%	1	1	100%

Reducing INDEL calling errors in whole-genome and exome sequencing data
Fang, H, Wu, Y, Narzisi, G, O'Rawe, JA, Jimenez Barrón LT, Rosenbaum, J, Ronemus, M, Iossifov I, Schatz, MC§, Lyon, GL§
Genome Medicine (2014) 6:89. doi:10.1186/s13073-014-0089-z

GATK: Genome Analysis Toolkit

NATURE GENETICS | TECHNICAL REPORT

日本語要約

A framework for variation discovery and genotyping using next-generation DNA sequencing data

Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler & Mark J Daly

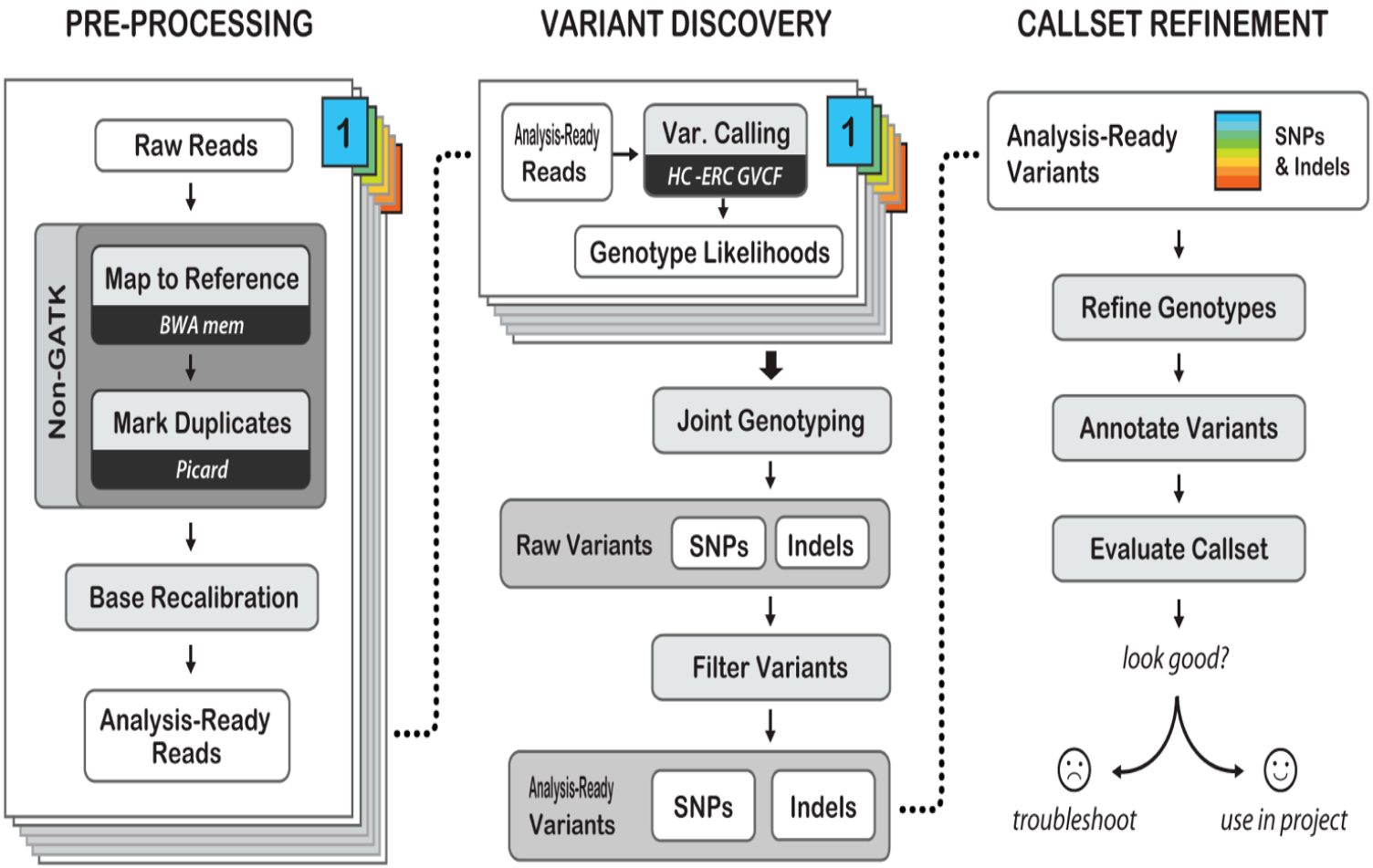
Affiliations | Contributions | Corresponding author

Nature Genetics **43**, 491–498 (2011) | doi:10.1038/ng.806

Received 27 August 2010 | Accepted 17 March 2011 | Published online 10 April 2011

http://www.nature.com/ng/journal/v43/n5/full/ng.806.html

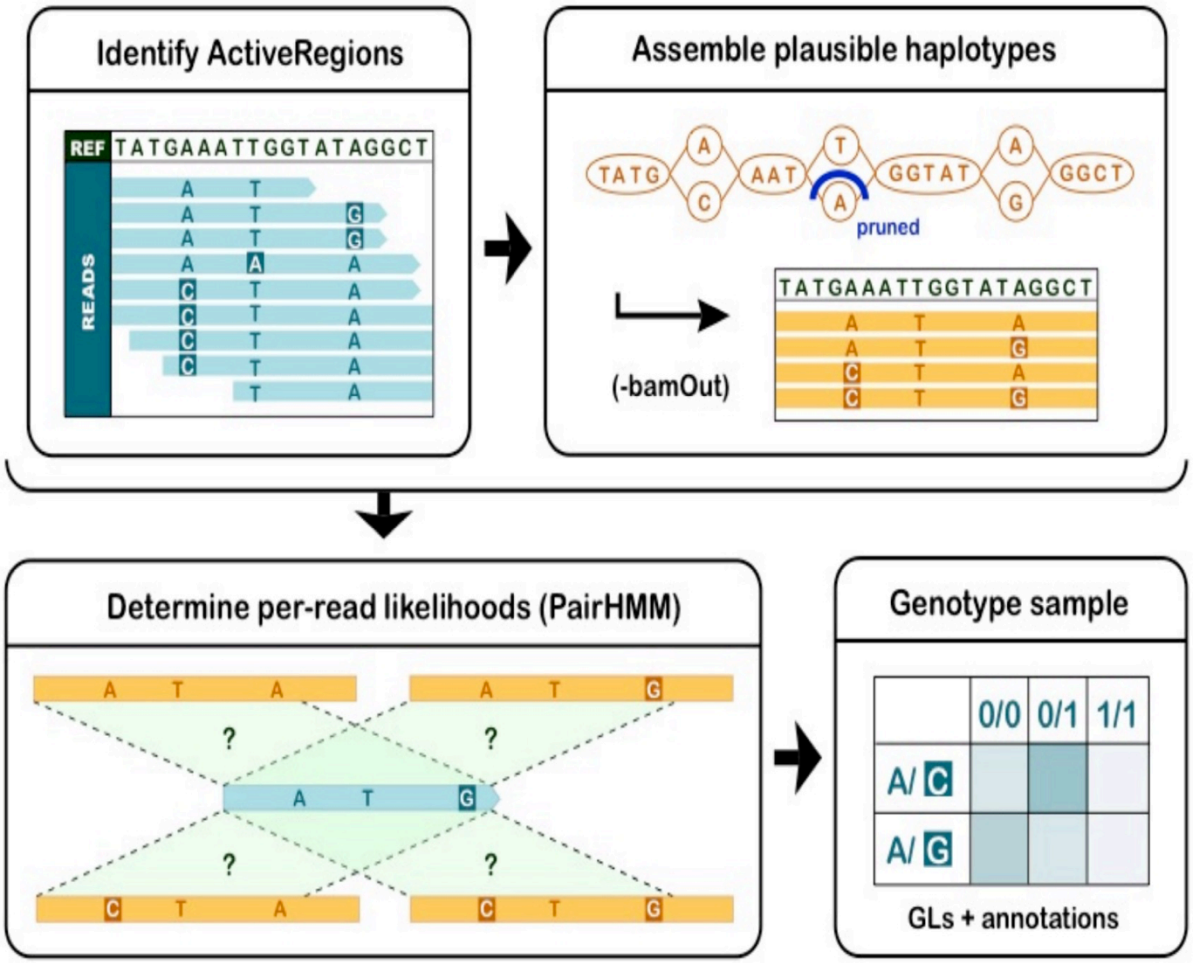
GATK workflow



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

https://software.broadinstitute.org/gatk/img/BP_workflow_3.6.png

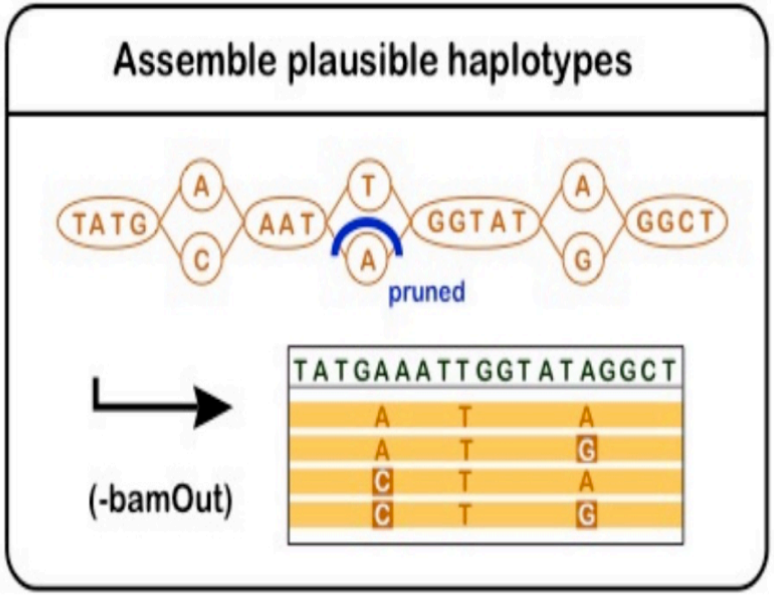
GATK Haplotype Caller



https://software.broadinstitute.org/gatk/events/slides/1506/GATKwr8-D-2-Variant_calling_and_joint_genotyping.pdf

Haplotype Caller "assembles" haplotypes

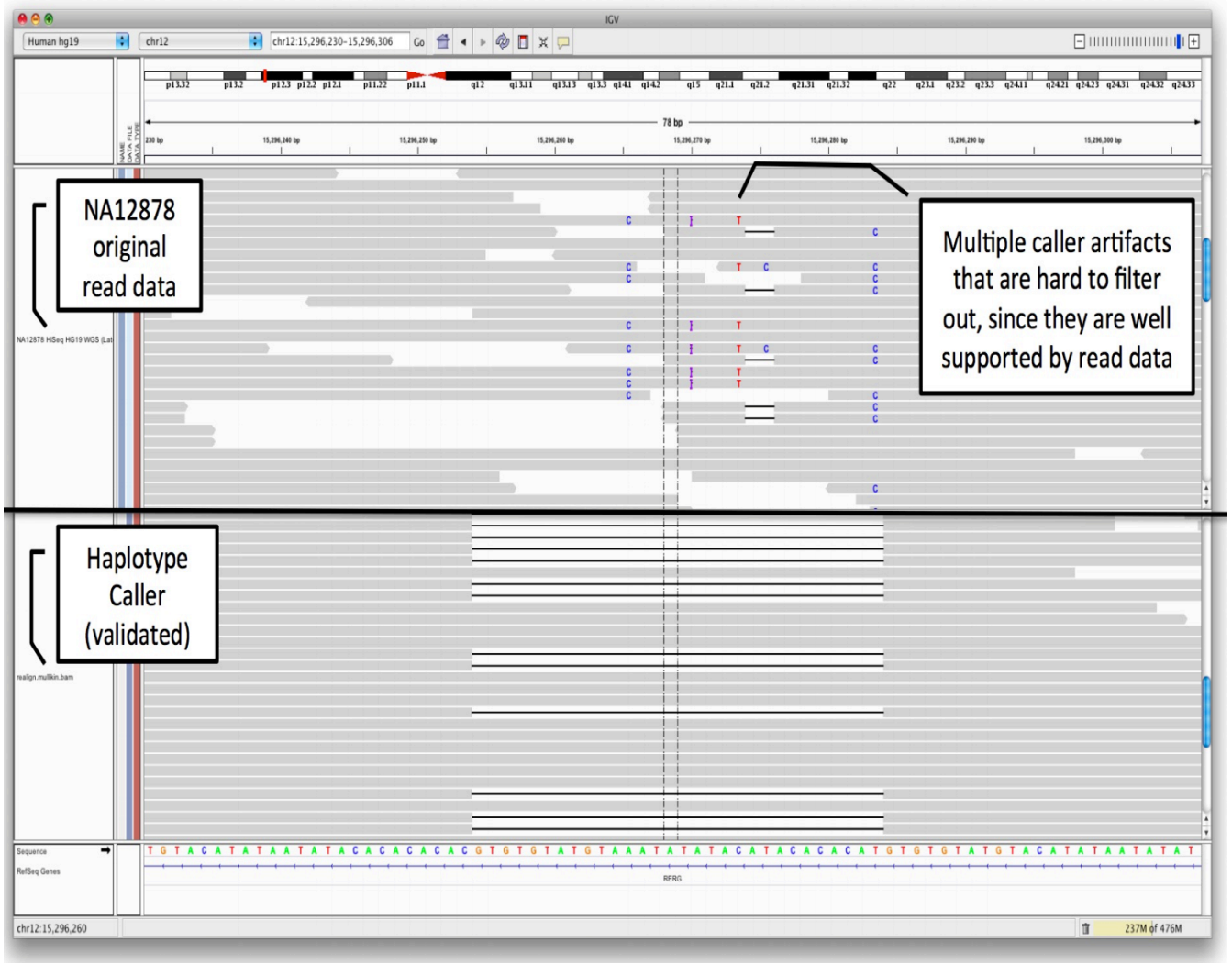
- Local re-assembly
- Traverse graph to collect most likely haplotypes
- Align haplotypes to ref using Smith-Waterman



Likely haplotypes + candidate variant sites

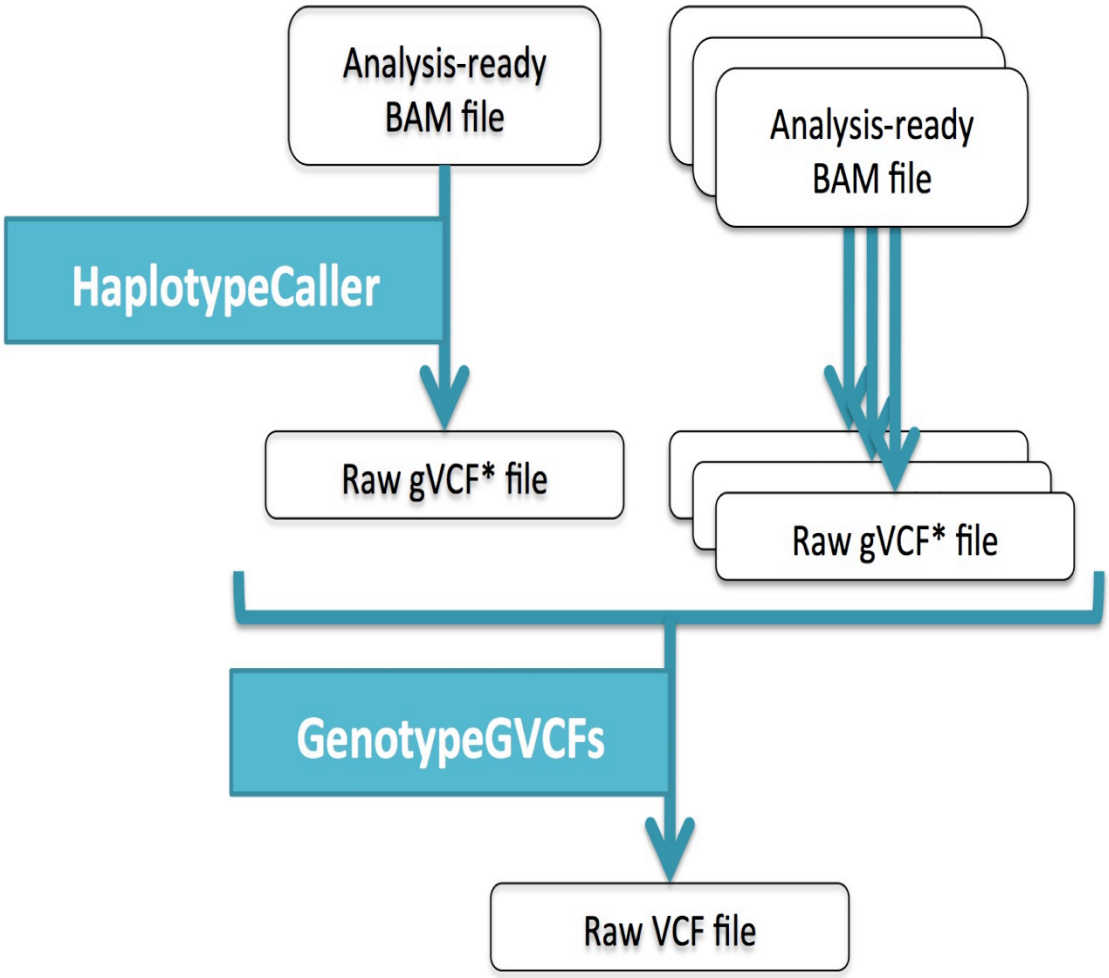
https://software.broadinstitute.org/gatk/events/slides/1506/GATKwr8-D-2-Variant_calling_and_joint_genotyping.pdf

Haplotype Caller "assembles" haplotypes



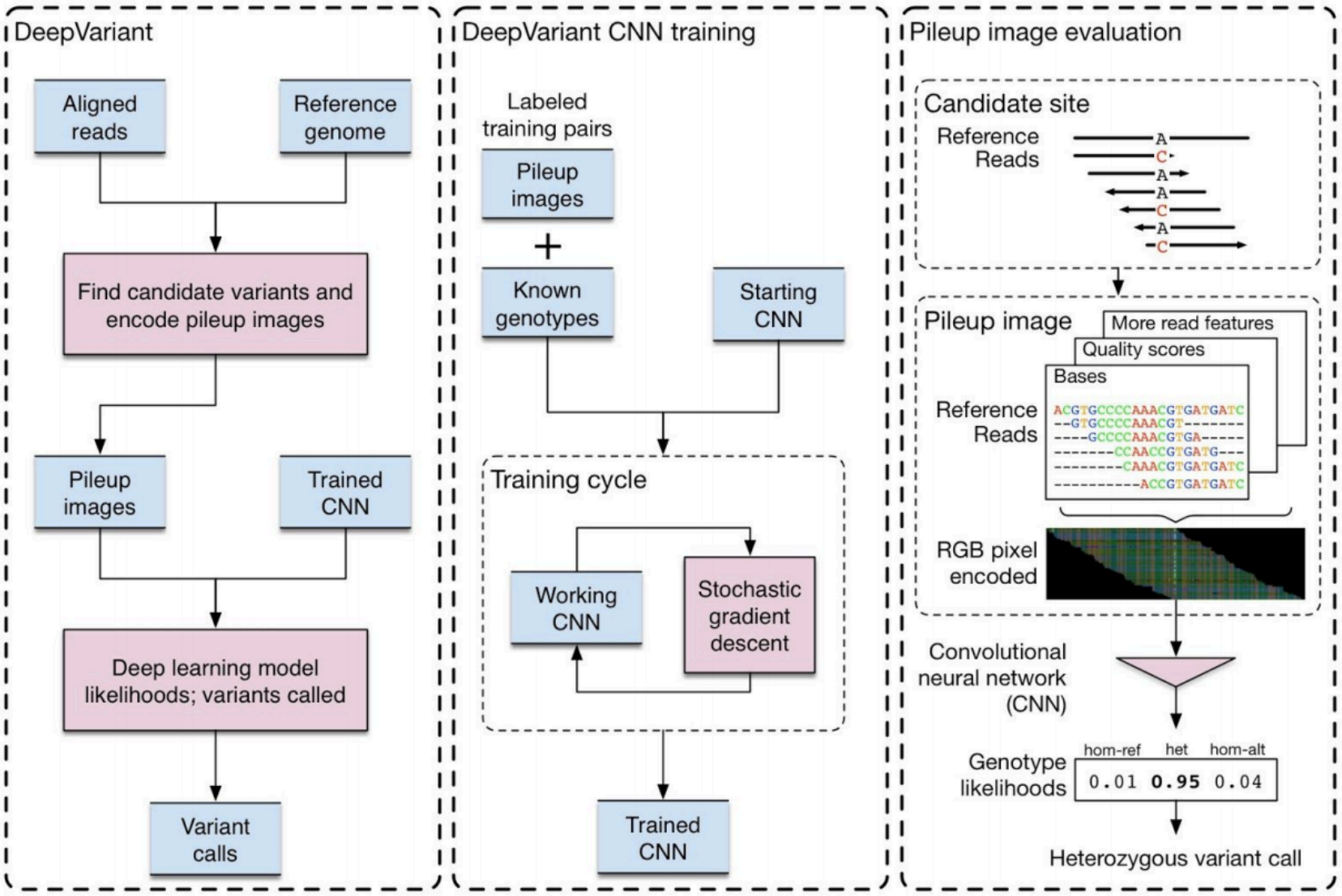
https://software.broadinstitute.org/gatk/events/slides/1506/GATKwr8-D-2-Variant_calling_and_joint_genotyping.pdf

GATK Produces a VCF file after genotyping across samples



https://software.broadinstitute.org/gatk/events/slides/1506/GATKwr8-D-2-Variant_calling_and_joint_genotyping.pdf

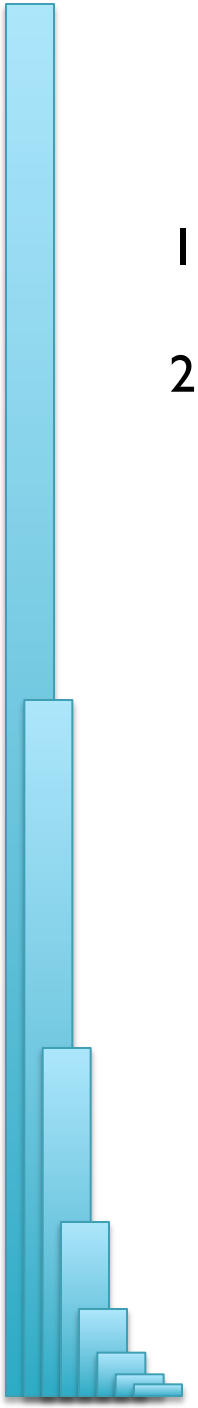
Deep Variant



Creating a universal SNP and small indel variant caller with deep neural networks
Poplin et al. (2016) bioRxiv. doi: https://doi.org/10.1101/092890

Next Steps

1. Finish Assignment 1
2. Check out the course webpage




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

Questions?