Variant Calling

Michael Schatz

Feb 12, 2020

Lecture 6: Applied Comparative Genomics

Assignment 2: Genome Assembly
Due Wednesday Feb 12 @ 11:59pm

1. ***Setup Docker/Ubuntu***
2. ***Initialize Tools***
3. ***Download Reference Genome & Reads***
4. ***Decode the secret message***
    1. *Estimate coverage, check read quality*
    2. *Check kmer distribution*
    3. *Assemble the reads with spades*
    4. *Align to reference with MUMmer*
    5. *Extract foreign sequence*
    6. *dna-encode.pl -d*

https://github.com/schatzlab/appliedgenomics2020/blob/master/assignments/assignment2/README.md

Genomic Coordinates

What are coordinates of "TAC" in GATTACA?

1-based coordinates

- Base 4 through 6: [4,6] "closed"
- Base 4 through 7: [4,7) "half-open"
- 3 bases starting at base 4: [4, +3]

```
GATTACA
1234567
```

0-based coordinates

- Position 3 through 5: [3,5] "closed"
- Position 3 through 6: [3,6) "half-open"
- 3 bases starting at position 3: [3, +3]

```
GATTACA
0123456
```

Genomic Conventions

1-based coordinates

- BLAST/MUMmer alignments
- Ensembl Genome Browser
- SAM, VCF, GFF and Wiggle

```
GATTACA
1234567
```

0-based coordinates

- BAM, BCFv2, BED, and PSL
- UCSC Genome Browser
- C/C++, Perl, Python, Java

```
GATTACA
0123456
```

Always double check the manual!
You will get this wrong someday ☹

Assignment 3: Due Wed Feb 19

Assignment 3: Coverage, Genome Assembly, and Variant Calling

Assignment Date: Wednesday, Feb. 12, 2020
Due Date: Wednesday, Feb. 19, 2020 @ 11:59pm

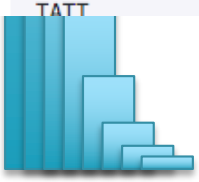
Question 1. Coverage simulator [10 pts]

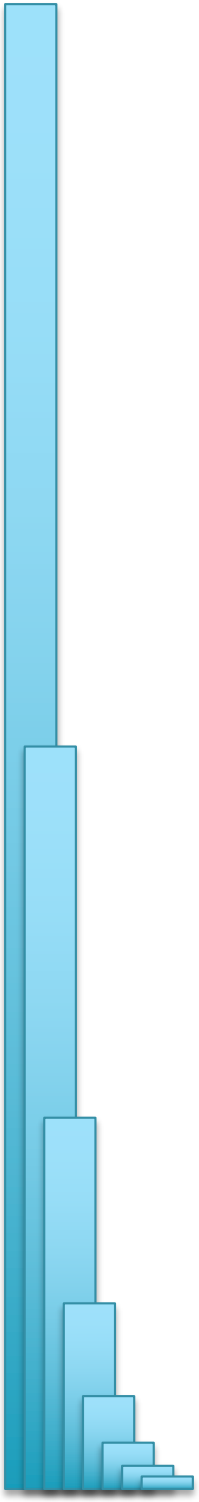
- Q1a. How many 100bp reads are needed to sequence a 1Mbp genome to 5x coverage?
- Q1b. In the language of your choice, simulate sequencing 5x coverage of a 1Mbp genome and plot the histogram of coverage. Note you do not need to actually output the sequences of the reads, you can just randomly sample positions in the genome and record the coverage. You do not need to consider the strand of each read. The start position of each read should have a uniform random probabilty at each possible starting position (1 through 999,900). You can record the coverage in an array of 1M positions. Overlay the histogram with a Poisson distribution with lambda=5
- Q1c. Using the histogram from 1b, how much of the genome has not been sequenced (has 0x coverage). How well does this match Poisson expectations?
- Q1d. Now repeat the analysis with 15x coverage: 1. simulate the appropriate number of reads, 2. make a histogram, 3. overlay a Poisson distribution with lambda=15, 4. compute the number of bases with 0x coverage, and 5. evaluated how well it matches the Poisson expectation.

Question 2. de Bruijn Graph construction [10 pts]

- Q2a. Draw (by hand or by code) the de Bruijn graph for the following reads using k=3 (assume all reads are from the forward strand, no sequencing errors, complete coverage of the genome)

```
ATTC
ATTG
CATT
CTTA
GATT
TATT
```



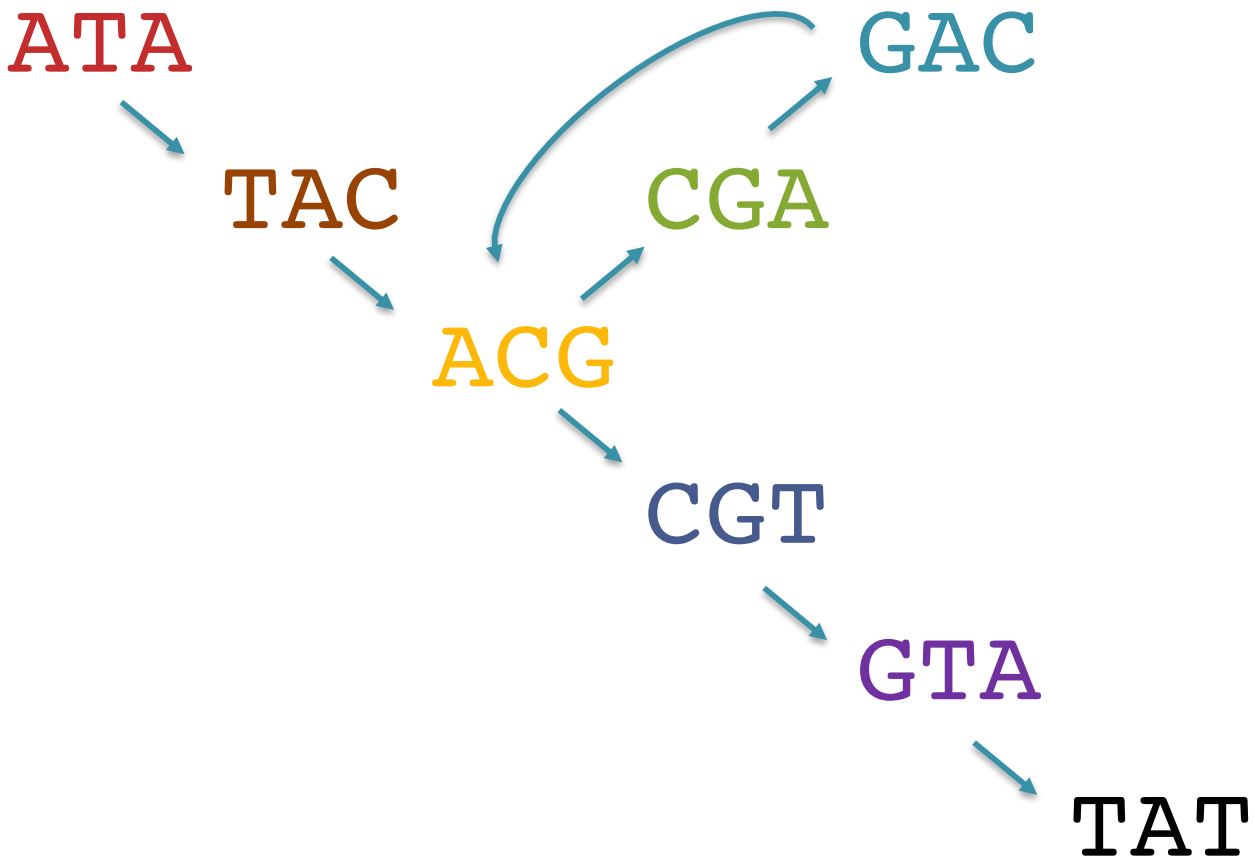


Part 1: Recap

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~
~~ACGT~~
~~ATAC~~
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
~~TACG~~

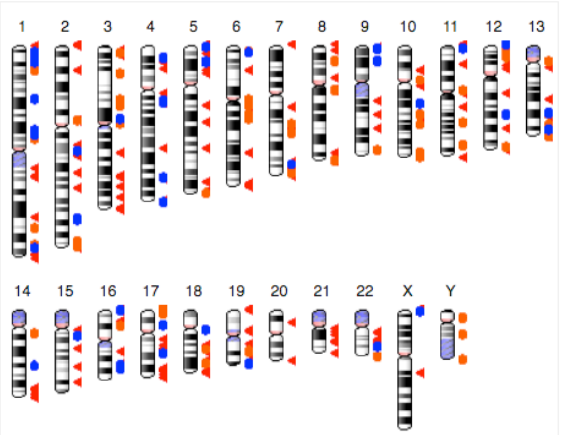


Note: there is no edge from ATA to TAT

ATACGACGTAT

Human Genome Overview

Information about the continuing improvement of the human genome



Ideogram of the latest human assembly, GRCh38.p11

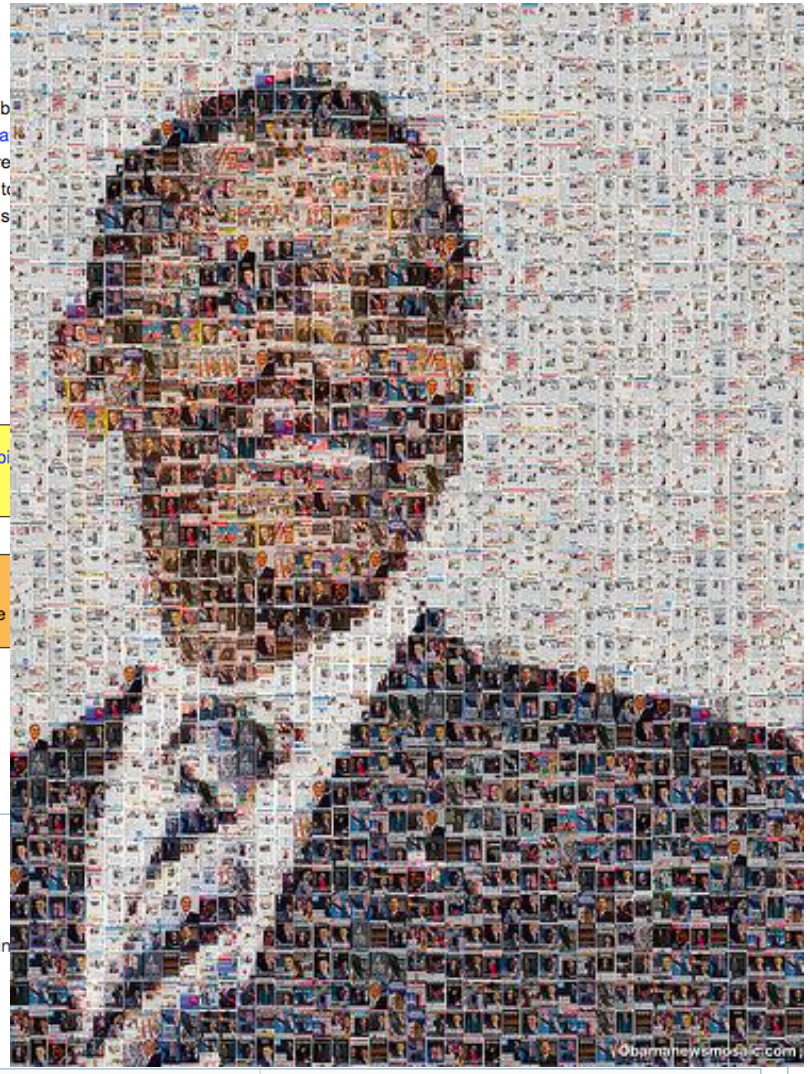
The GRC is working hard to provide the best possib
by both generating multiple representations (alterna
represented by a single path. Additionally, we are re
allows users who are interested in a specific locus t
affecting users who need chromosome coordinate s

Download data:

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

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

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



GRCh38.p11 | GRCh37.p13 | GRCh37

GRCh38.p11

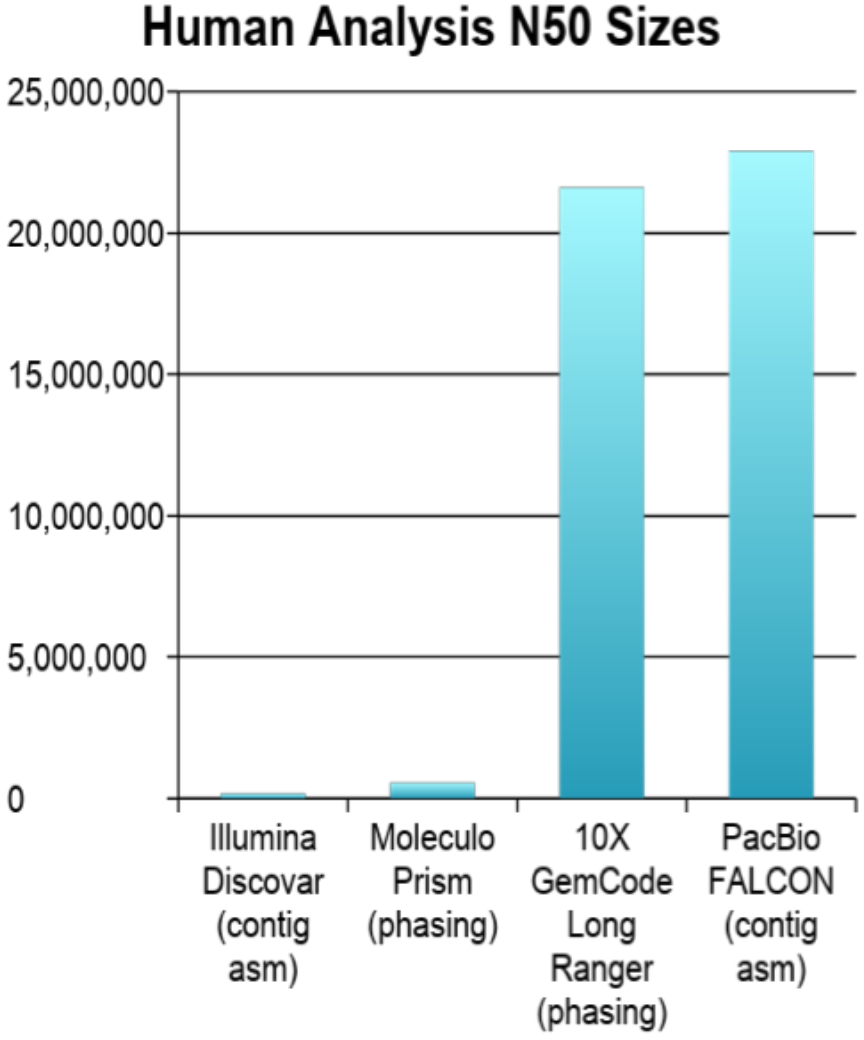
Release date: June 14, 2017
Release type: minor
Release notes: GRCh38.p11 is the eleventh patch release for the GRCh38 reference assembly. No chromosome coordin
of patch scaffolds is now: 64 FIX and 59 NOVEL.
Assembly accessions: GenBank: GCA_000001405.26 , RefSeq: GCF_000001405.37

Pseudoautosomal regions

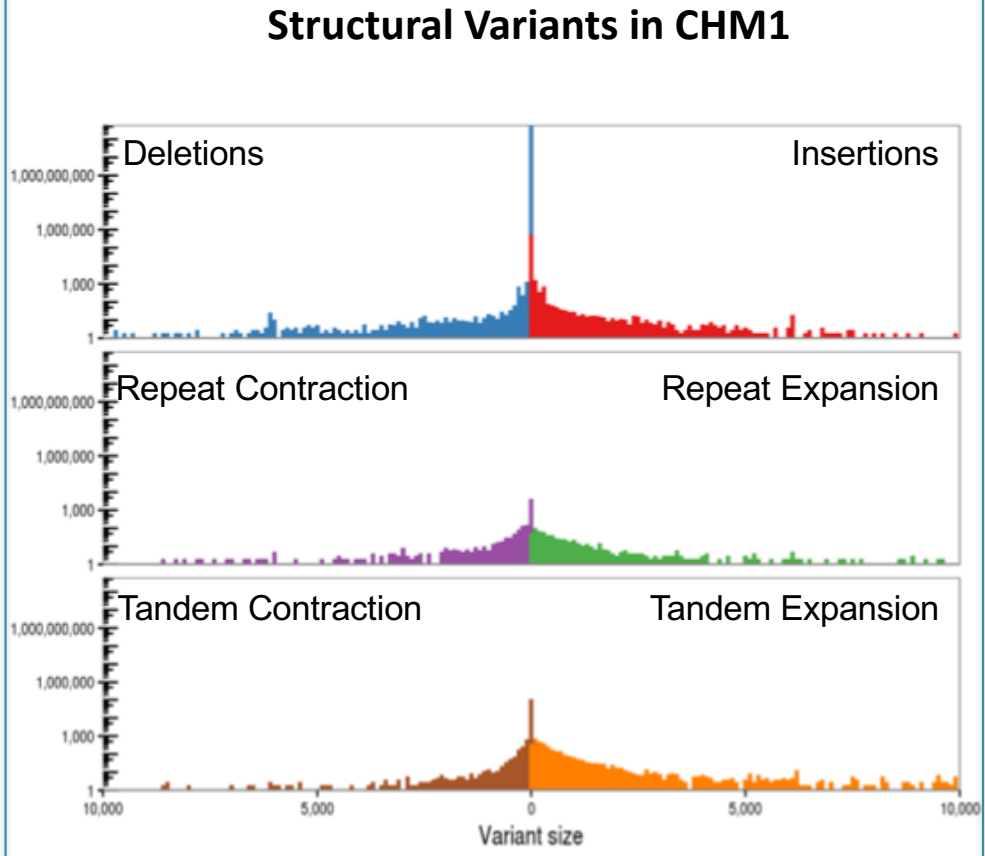
Name	Chr	Start	Stop
PAR#1	X	10,001	2,781,479
PAR#2	X	155,701,383	156,030,895
PAR#1	Y	10,001	2,781,479
PAR#2	Y	56,887,903	57,217,415

Genomics Arsenal in the Year 2020

Recent Long Read Assemblies

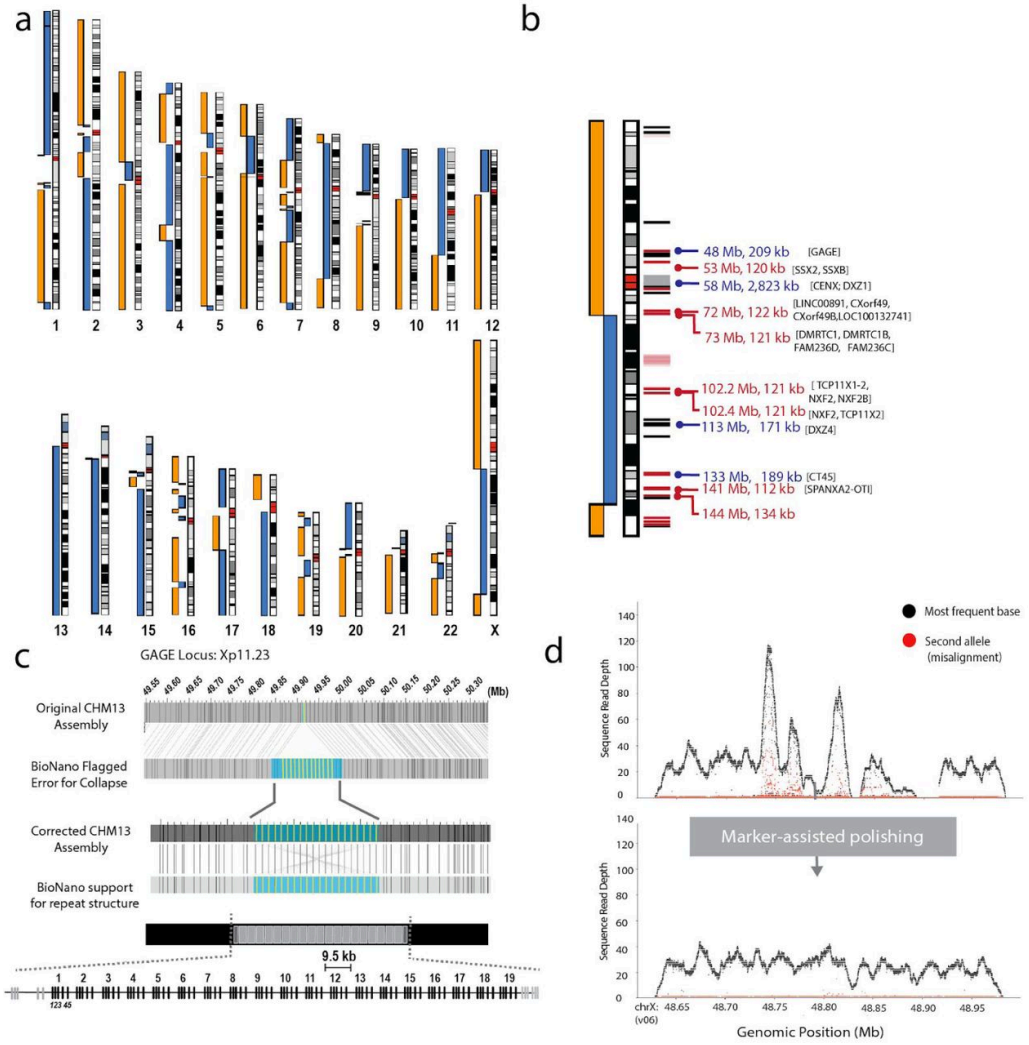


Third-generation sequencing and the future of genomics
Lee et al (2016) *bioRxiv*
doi: http://dx.doi.org/10.1101/048603

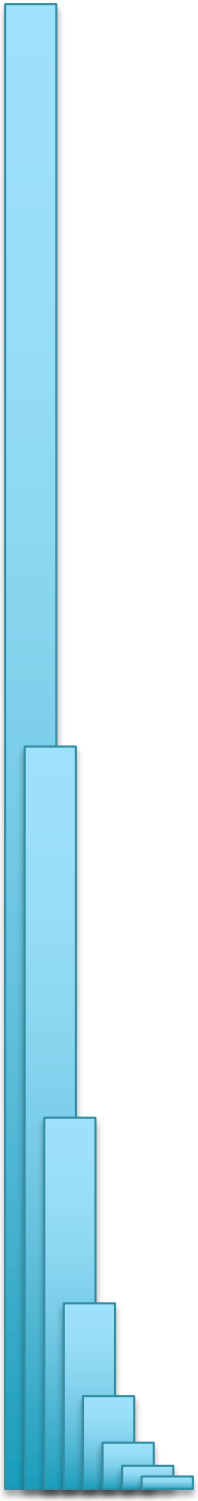


Assemblytics: a web analytics tool for the detection of variants from an assembly
Nattestad & Schatz (2016) *Bioinformatics*.
doi: 10.1093/bioinformatics/btw369

First Telomere-to-Telomere Human Chromosome



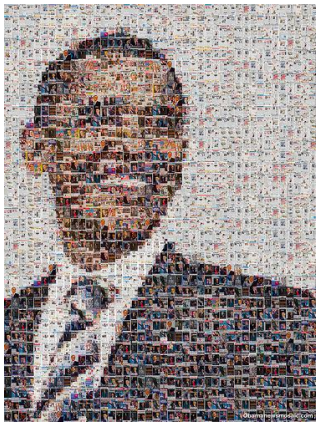
Telomere-to-telomere assembly of a complete human X chromosome
Miga et al. (2019) bioRxiv. https://doi.org/10.1101/735928

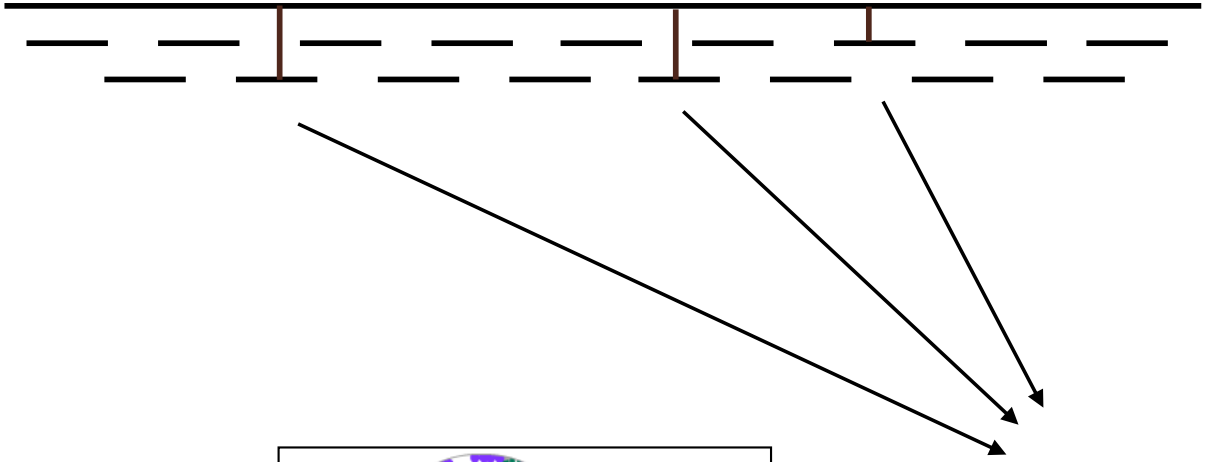


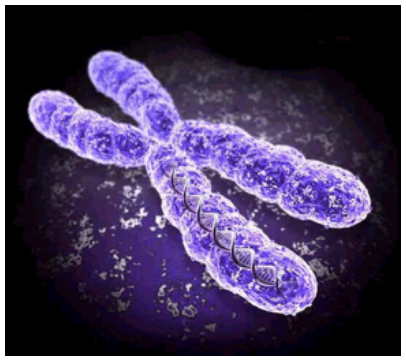
Part 2. Variant Calling

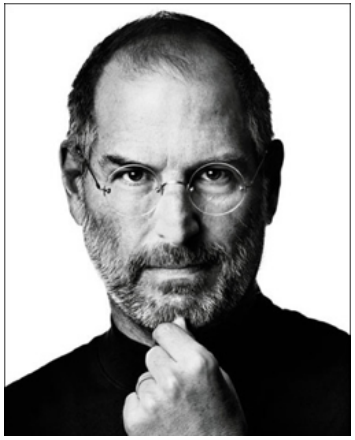
Personal Genomics

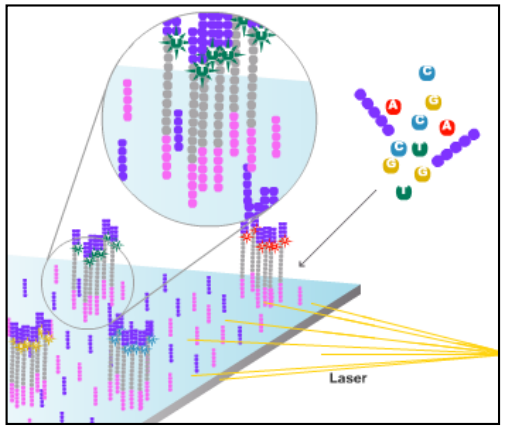
How does your genome compare to the reference?











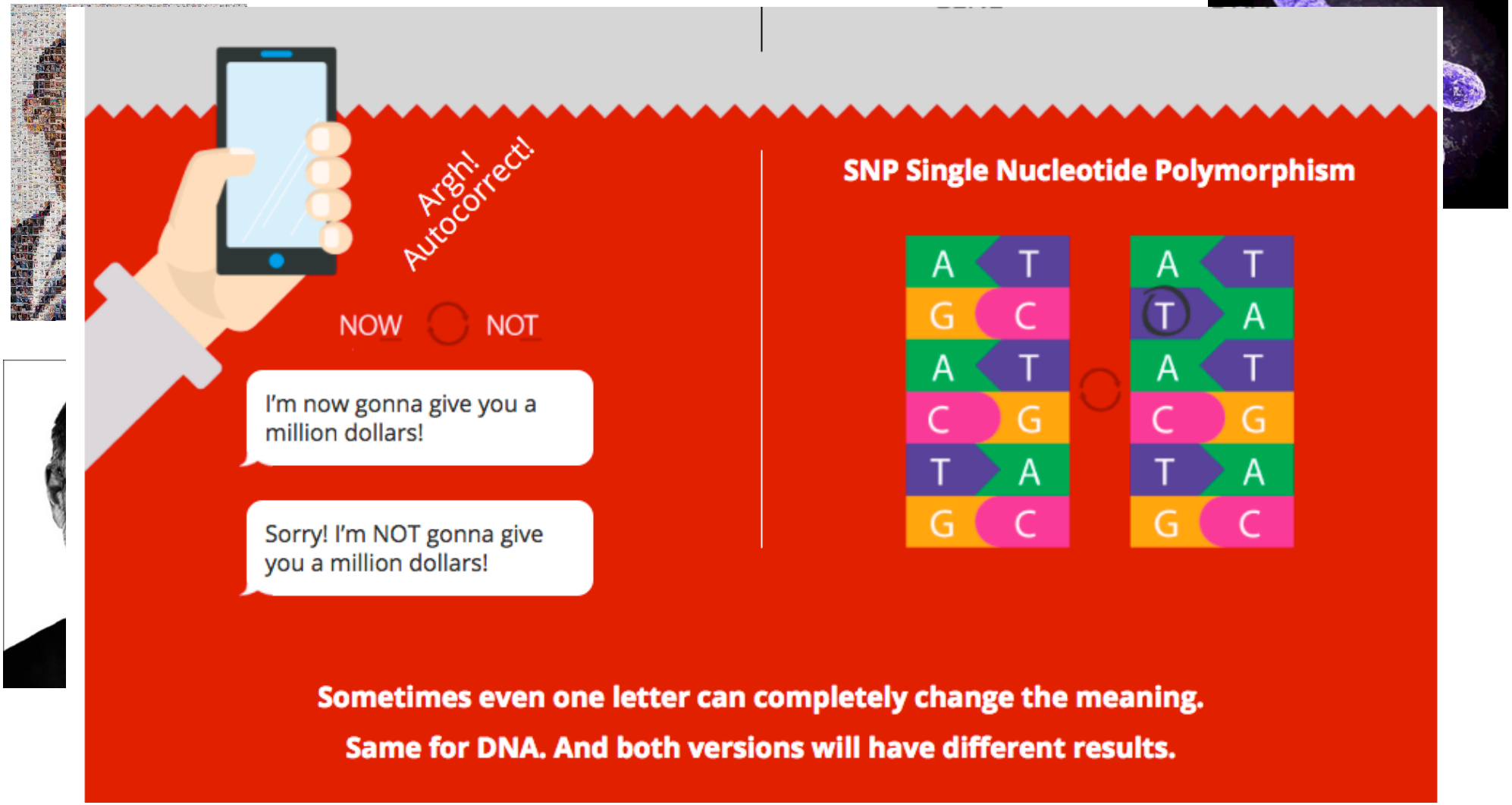
Heart Disease

Cancer

Creates magical technology

Personal Genomics

How does your genome compare to the reference?



Algorithm Overview

1. Split read into segments

Read: `CCAGTAGCTCTCAGCCTTATTTTACCCAGGCCTGTA`

Read (reverse complement): `TACAGGCCTGGGTAAAATAAGGCTGAGAGCTACTGG`

Policy: extract 16 nt seed every 10 nt

Seeds

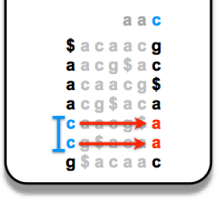
+, 0: `CCAGTAGCTCTCAGCC`
+, 10: `TCAGCCTTATTTTACC`
+, 20: `TTTACCCAGGCCTGTA`
-, 0: `TACAGGCCTGGGTAAA`
-, 10: `GGTAAAATAAGGCTGA`
-, 20: `GGCTGAGAGCTACTGG`

2. Lookup each segment and prioritize

Seeds

+, 0: `CCAGTAGCTCTCAGCC`
+, 10: `TCAGCCTTATTTTACC`
+, 20: `TTTACCCAGGCCTGTA`
-, 0: `TACAGGCCTGGGTAAA`
-, 10: `GGTAAAATAAGGCTGA`
-, 20: `GGCTGAGAGCTACTGG`

→ Ungapped alignment with FM Index



→ Seed alignments (as B ranges)

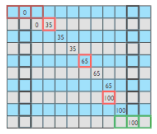
```
{ [211, 212], [212, 214] }
{ [653, 654], [651, 653] }
{ [684, 685] }
{ }
{ }
{ [624, 625] }
```

3. Evaluate end-to-end match

Extension candidates

```
SA:684, chr12:1955
SA:624, chr2:462
SA:211: chr4:762
SA:213: chr12:1935
SA:652: chr12:1945
```

→ SIMD dynamic programming aligner

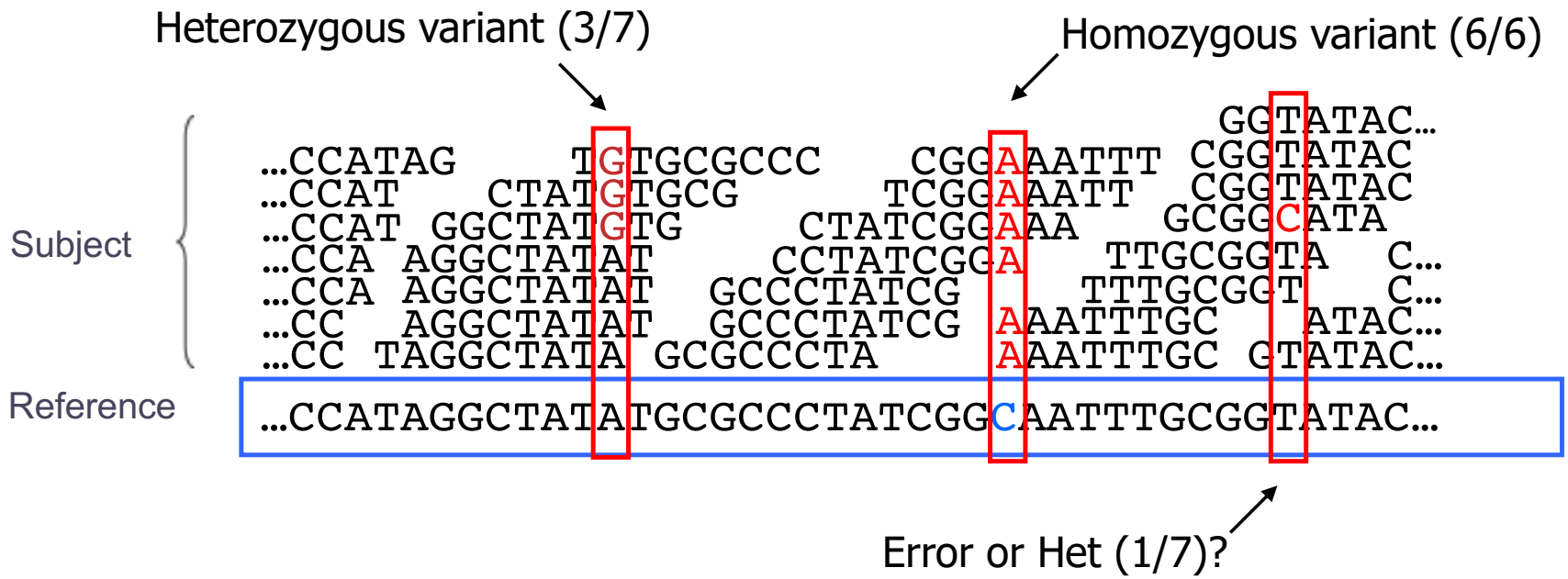


→ SAM alignments

```
r1   0    chr12    1936     0
     36M  *    0    0
     CCAGTAGCTCTCAGCCTTATTTTACCCAGGCCTGTA
     IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
     AS:i:0    XS:i:-2   XN:i:0
     XM:i:0    XO:i:0    XG:i:0
     NM:i:0    MD:Z:36   YT:Z:UU
     YM:i:0
...
```

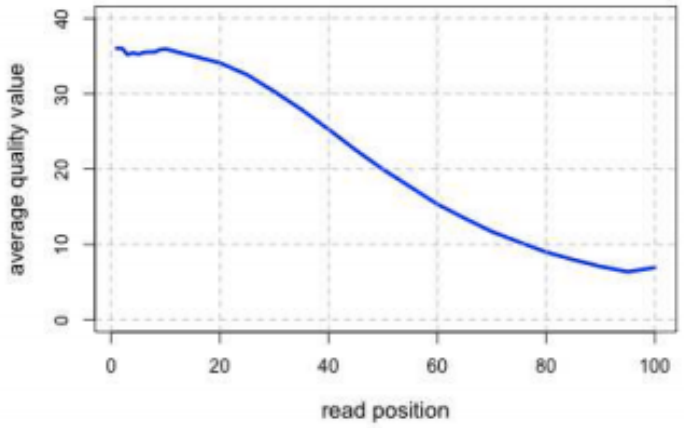
(Langmead & Salzberg, 2012)

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

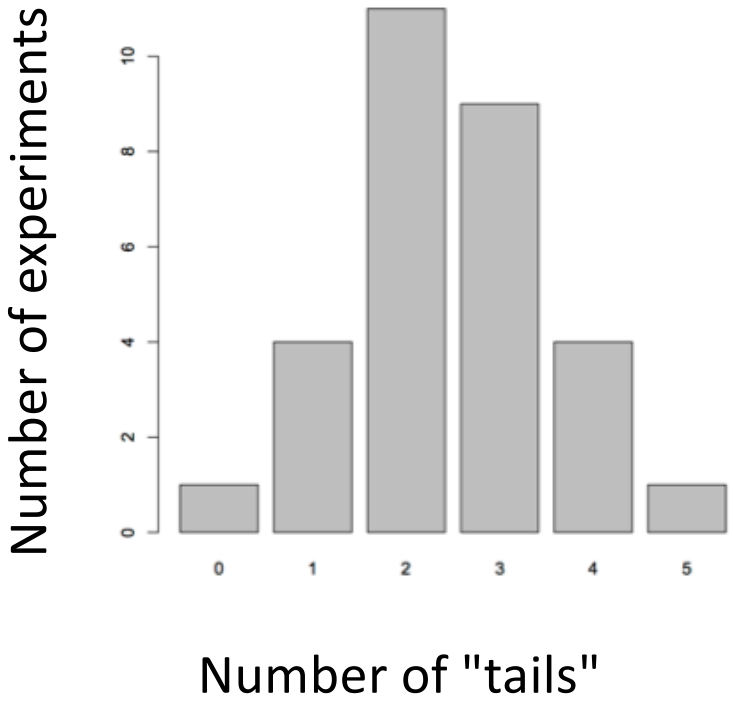
The Binomial Distribution: Adventures in Coin Flipping

P(heads) = 0.5

P(tails) = 0.5

Aaron Quinlan

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



R code:

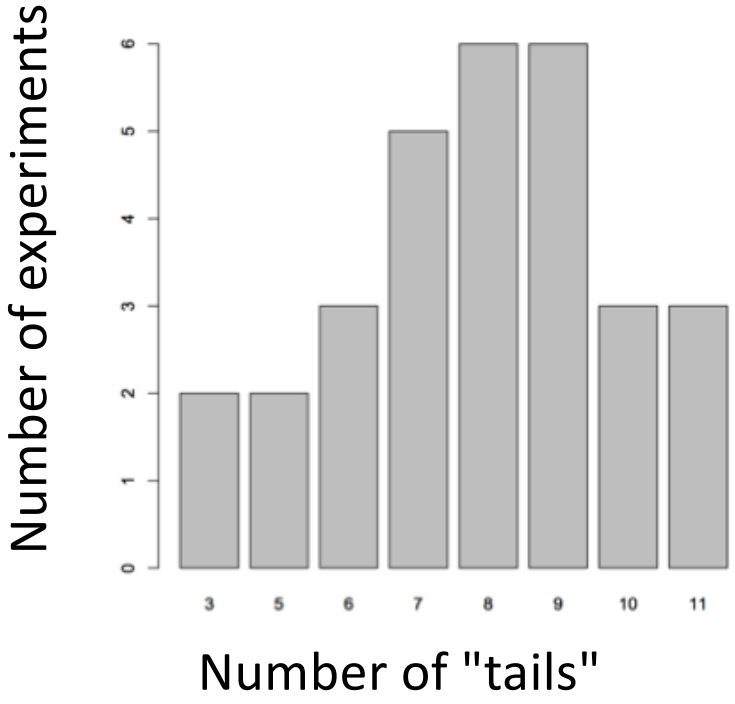
```
barplot(table(rbinom(30, 5, 0.5)))
```

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:

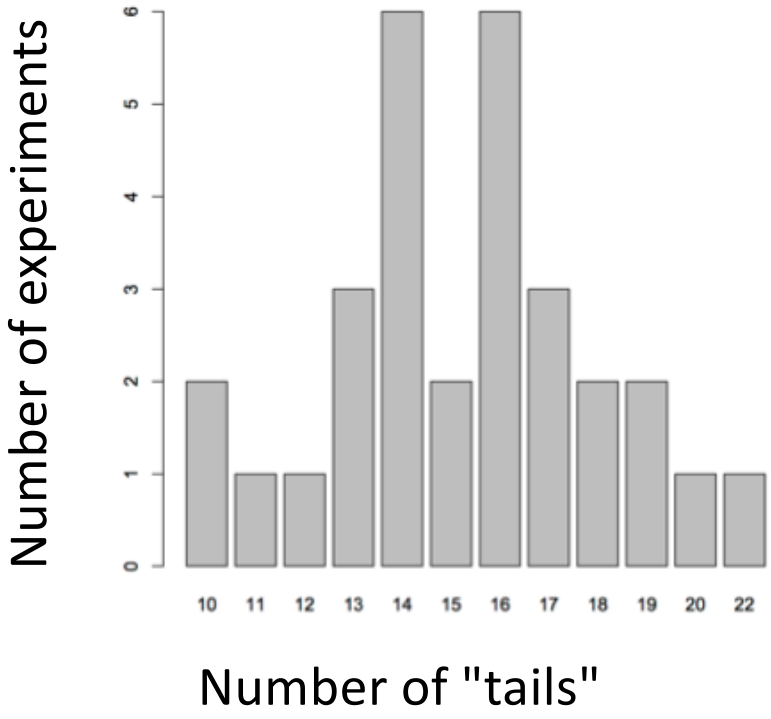
```
barplot(table(rbinom(30, 15, 0.5)))
```

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:

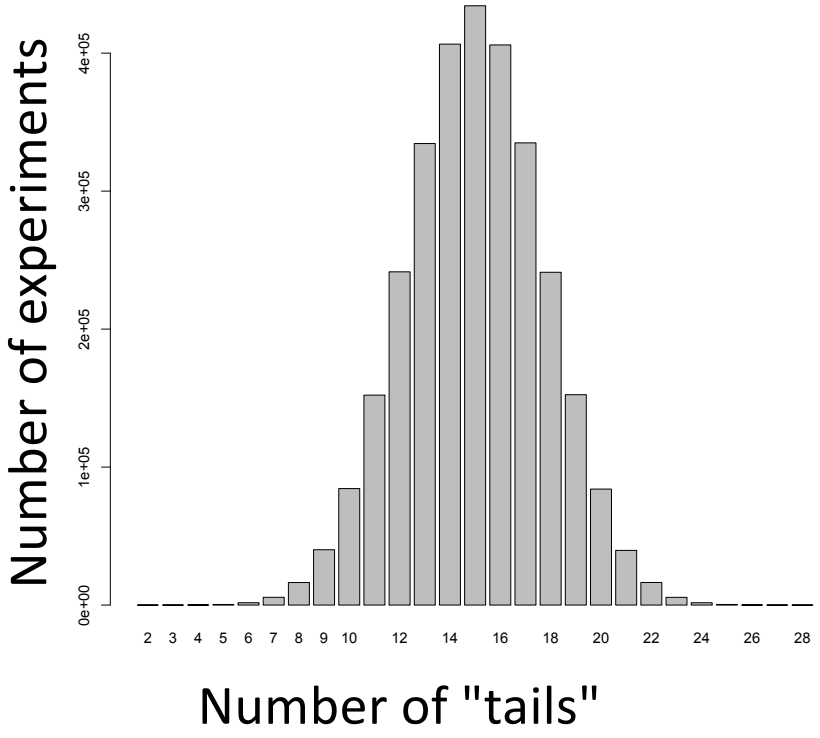
```
barplot(table(rbinom(30, 30, 0.5)))
```

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:

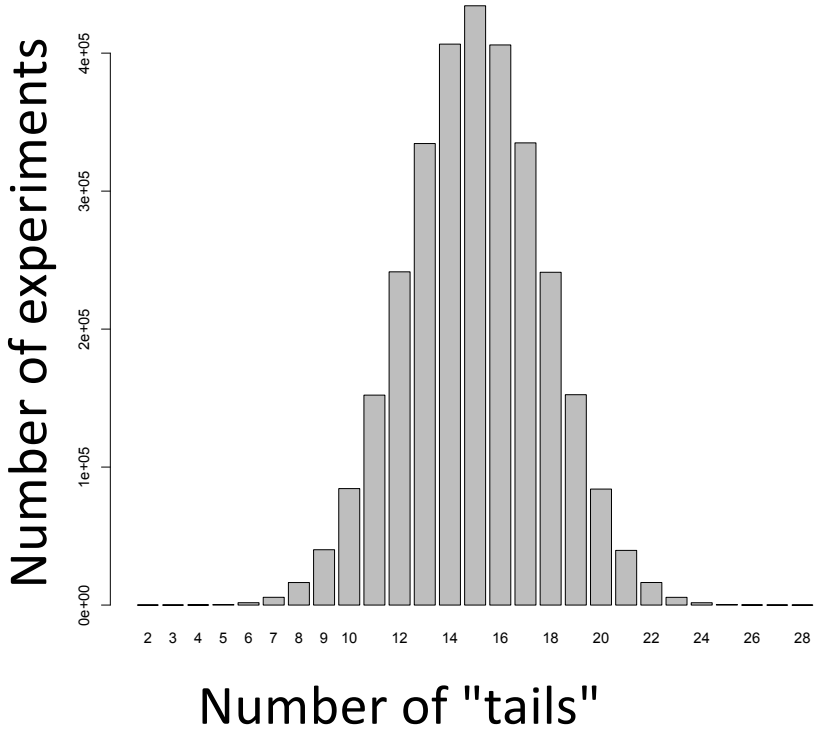
```
barplot(table(rbinom(3e6, 30, 0.5)))
```

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

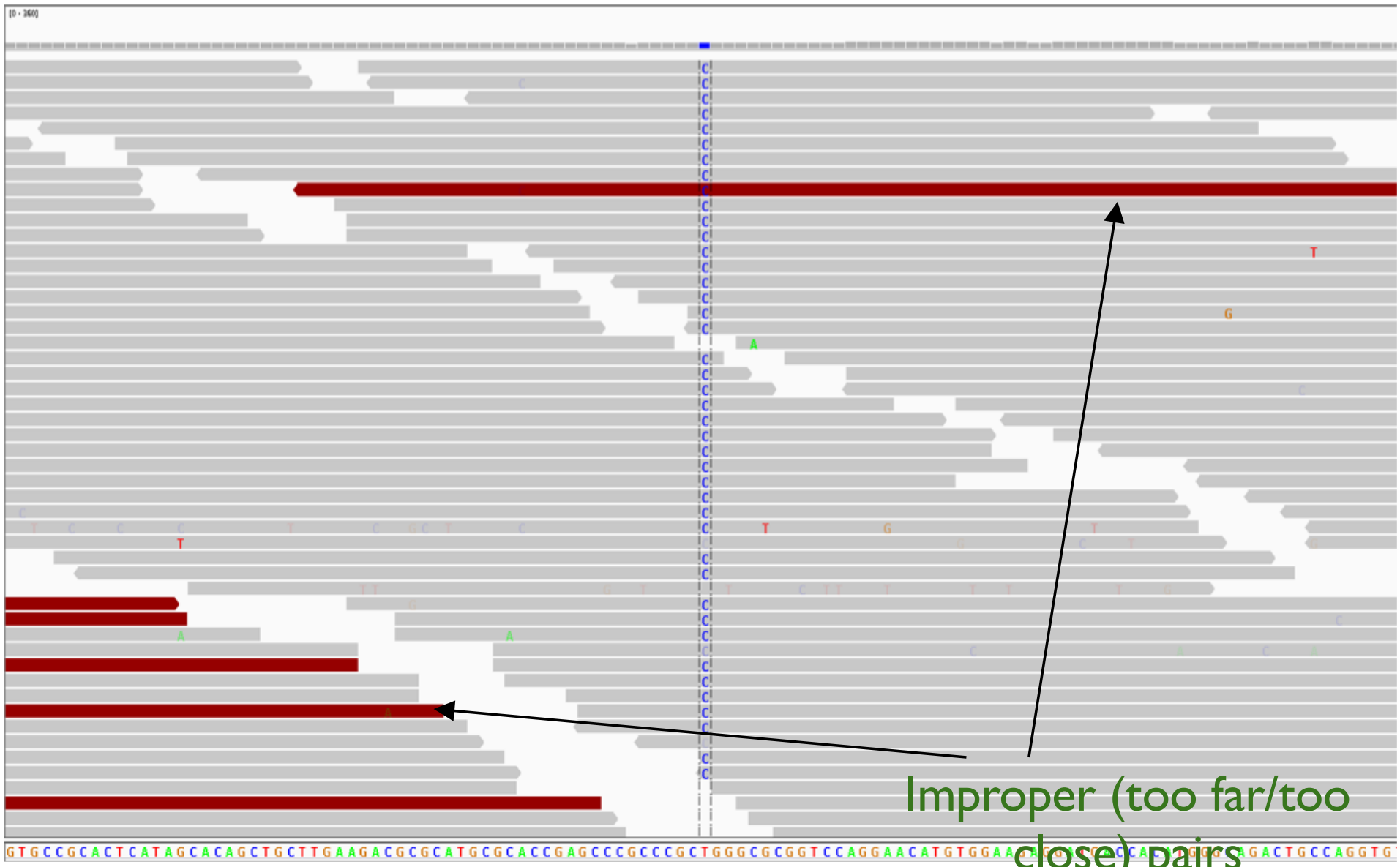


This is why at least a "30X" (30 fold sequence coverage) genome is recommended: it confers sufficient power to distinguish heterozygous alleles and from mere sequencing errors

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

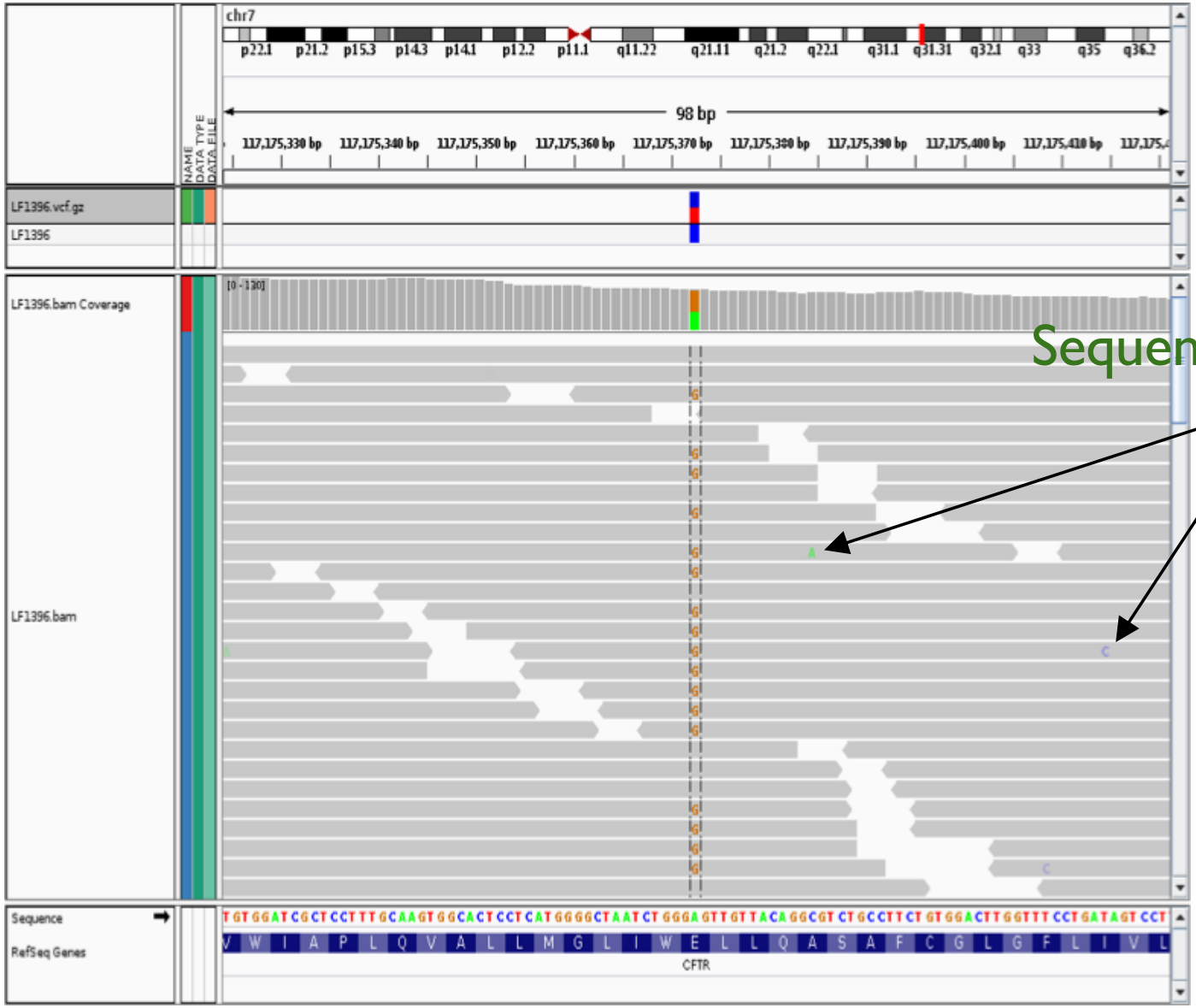
Some real examples of SNPs in IGV

Homozygous for the "C" allele



What else do you notice?

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

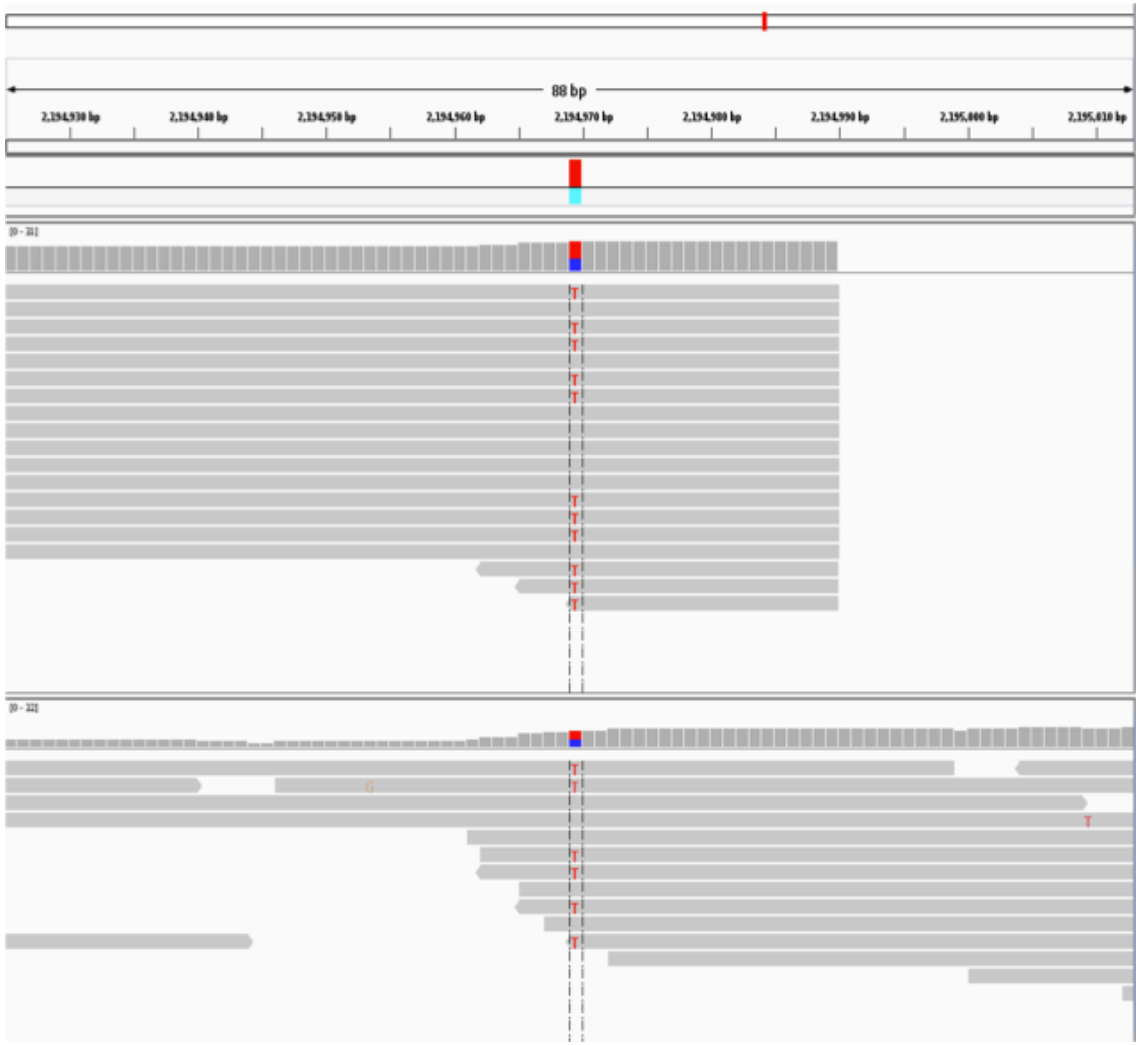


https://jchoigt.files.wordpress.com/2012/07/igv_e217g_snapshot.png

Heterozygous for the alternate allele

Individual 1

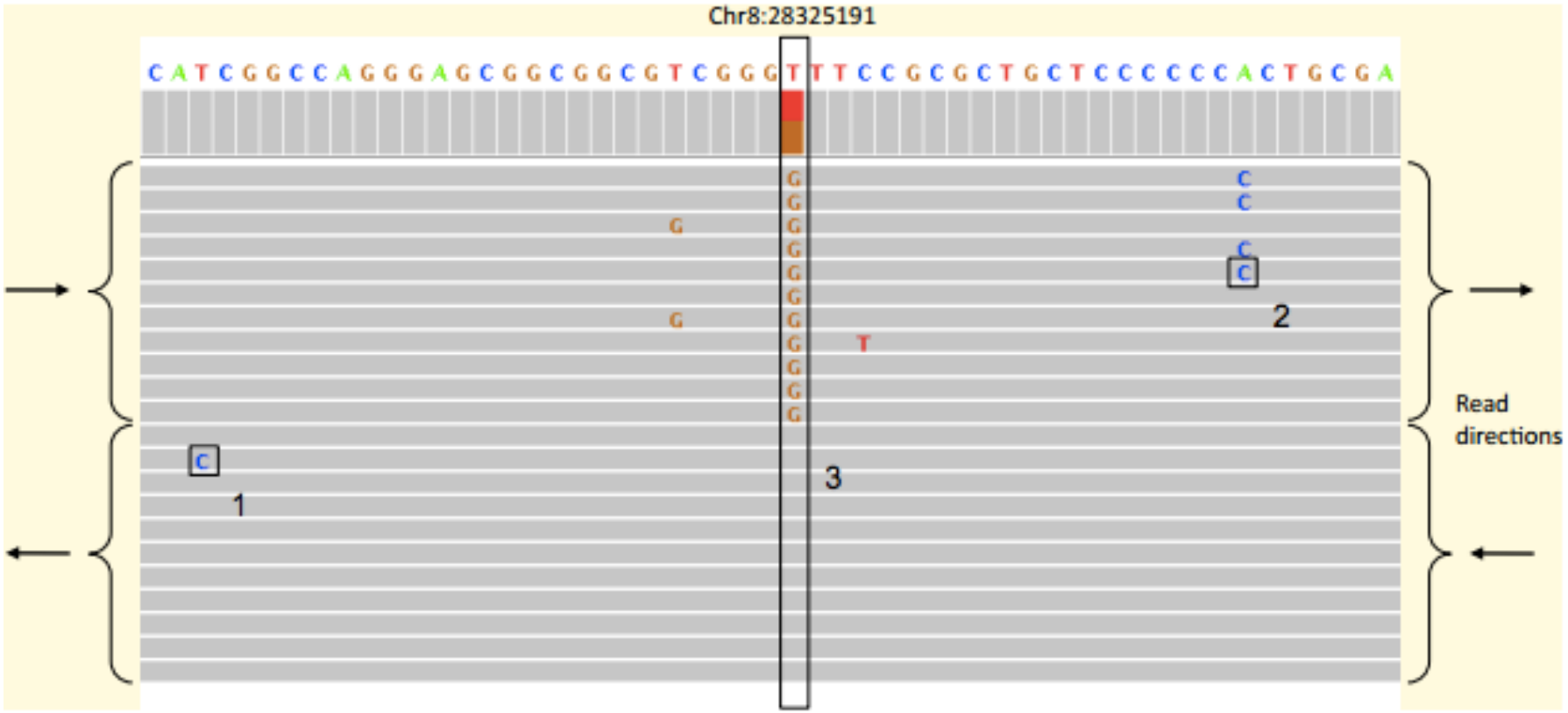
Individual 2



Which genotype prediction do you have more confidence in?

It is not always so easy ☹

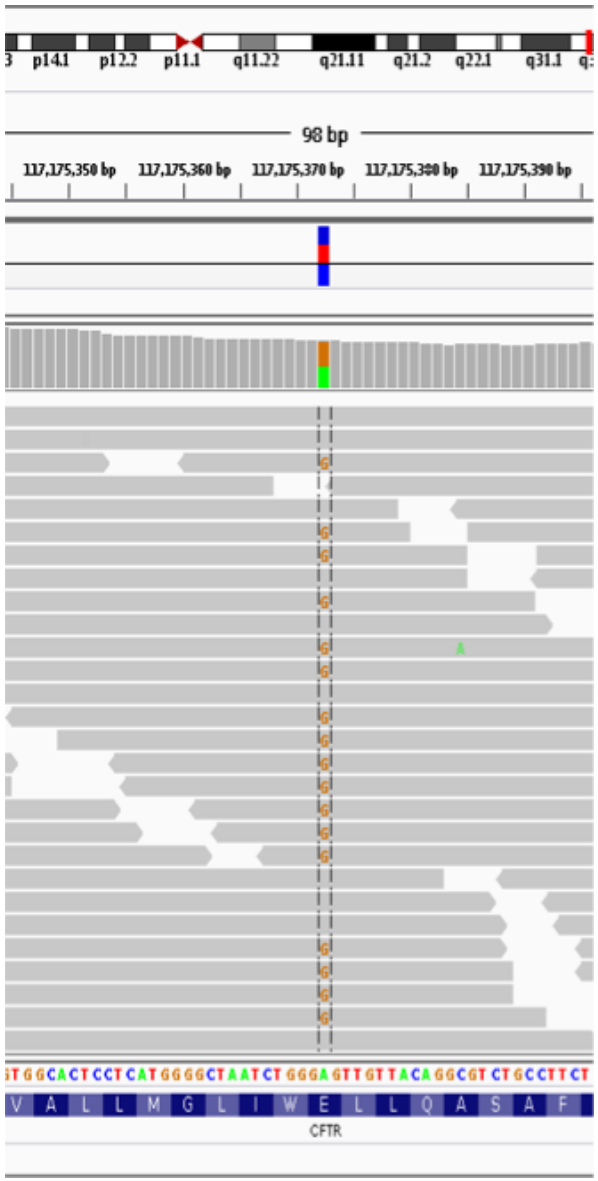
Beware of Systematic Errors



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

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

$$\Pr(\text{spam}|\text{words}) = \frac{\Pr(\text{words}|\text{spam})\Pr(\text{spam})}{\Pr(\text{words})}$$

Thomas Bayes

Portrait used of Bayes in a 1936 book,[1] but it is doubtful whether the portrait is actually of him.[2] No earlier portrait or claimed portrait survives.

Born	c. 1701 London, England
Died	7 April 1761 (aged 59) Tunbridge Wells, Kent, England
Residence	Tunbridge Wells, Kent, England
Nationality	English
Known for	Bayes' theorem

Signature

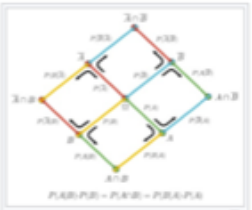
Statement of theorem [edit]

Bayes' theorem is stated mathematically as the following equation:[2]

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)},$$

where A and B are events and $P(B) \neq 0$.

- $P(A)$ and $P(B)$ are the probabilities of observing A and B without regard to each other.
- $P(A \mid B)$, a conditional probability, is the probability of observing event A given that B is true.
- $P(B \mid A)$ is the probability of observing event B given that A is true.



Visualization of Bayes' theorem by the superposition of two decision trees

History [edit]

Bayes' theorem was named after the Reverend Thomas Bayes (1701–1761), who studied how to compute a distribution for the probability parameter of a binomial distribution (in modern terminology). Bayes' unpublished manuscript was significantly edited by Richard Price before it was posthumously read at the Royal Society. Price edited[3] Bayes' major work "An Essay towards solving a Problem in the Doctrine of Chances" (1763), which appeared in "Philosophical Transactions,"[4] and contains Bayes' Theorem. Price wrote an introduction to the paper which provides some of the philosophical basis of Bayesian statistics. In 1765 he was elected a Fellow of the Royal Society in recognition of his work on the legacy of Bayes.[5][6]

The French mathematician Pierre-Simon Laplace reproduced and extended Bayes' results in 1774, apparently quite unaware of Bayes' work.[7][8] The Bayesian interpretation of probability was developed mainly by Laplace.[9]

Stephen Stigler suggested in 1983 that Bayes' theorem was discovered by Nicholas Saunderson, a blind English mathematician, some time before Bayes;[10][11] that interpretation, however, has been disputed.[12] Martyn Hooper[13] and Sharon McGrayne[14] have argued that Richard Price's contribution was substantial:

> By modern standards, we should refer to the Bayes–Price rule. Price discovered Bayes' work, recognized its importance, corrected it, contributed to the article, and found a use for it. The modern convention of employing Bayes' name alone is unfair but so entrenched that anything else makes little sense.[14]

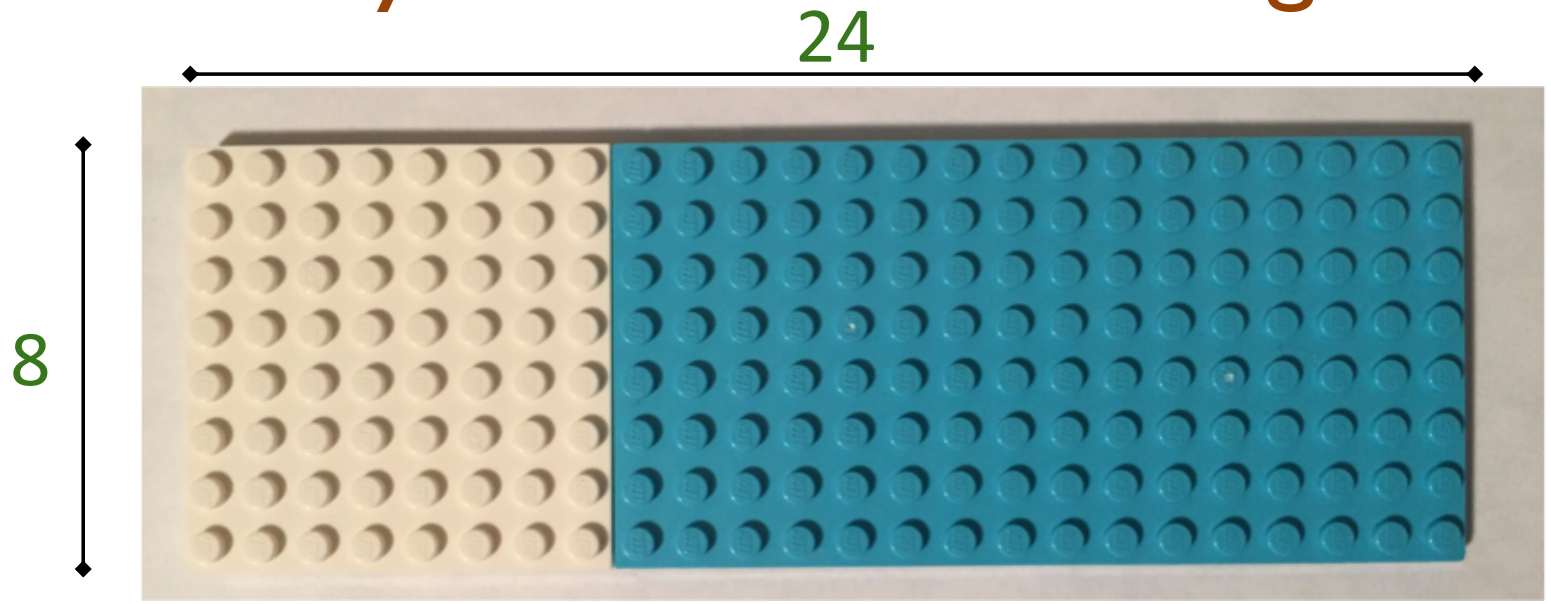
https://en.wikipedia.org/wiki/Thomas_Bayes

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.

Bayes' theorem with legos



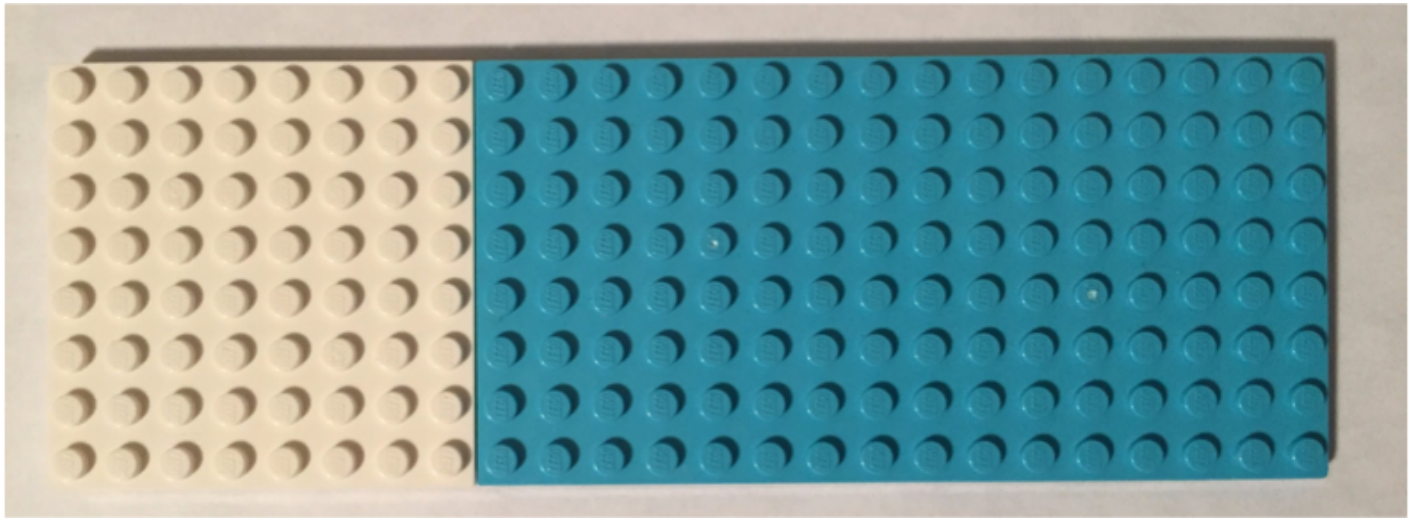
8x24 = 192 pegs, 64 are white, 128 are blue.

P(White) = 64 / 192 = **0.33**

P(Blue) = 128 / 192 = **0.67**

Inspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

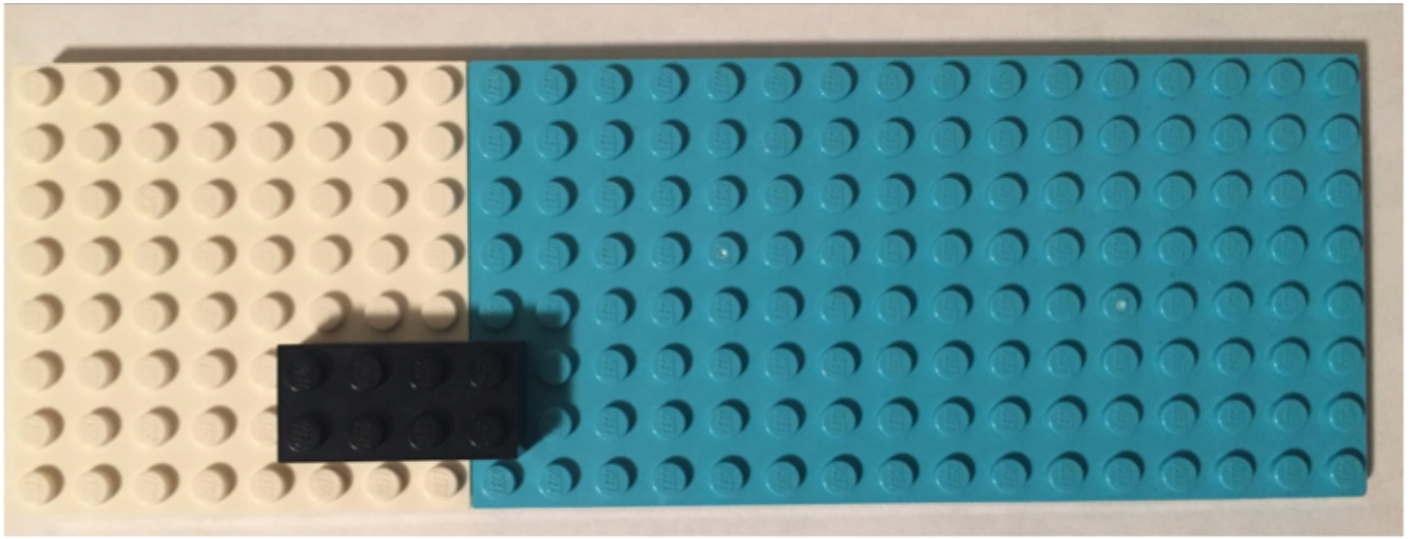
Our entire probability "space" must add up to 1.



P(White) + P(Blue) = 1

Inspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

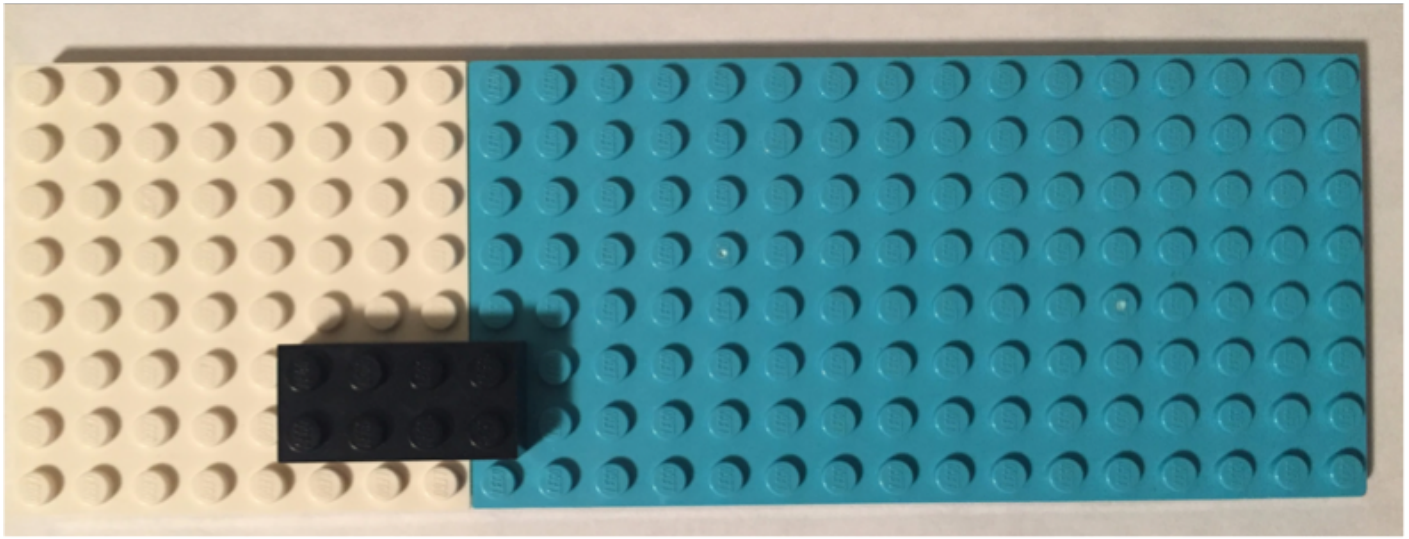
What is the probability of black?



P(Black) = 8 / 192 = 0.042

Inspired by
https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

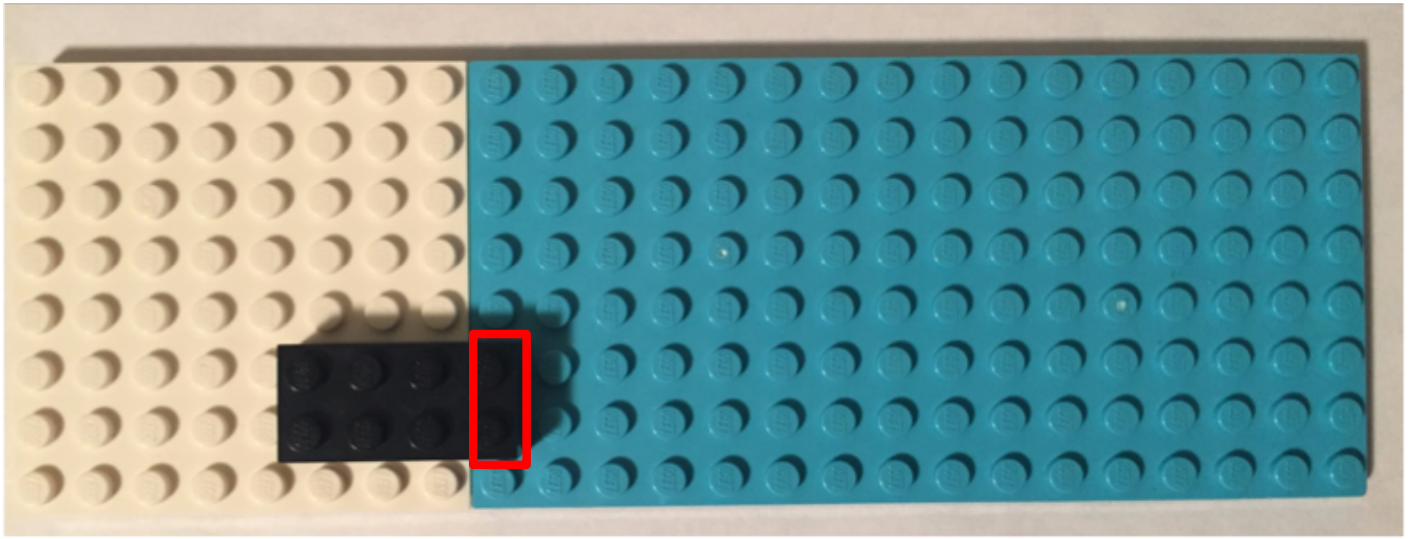
No, probability space is >1.
P(Black) is conditional on P(White) and P(Blue).



P(White) + P(Blue) + P(Black) = 1.042

Inspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

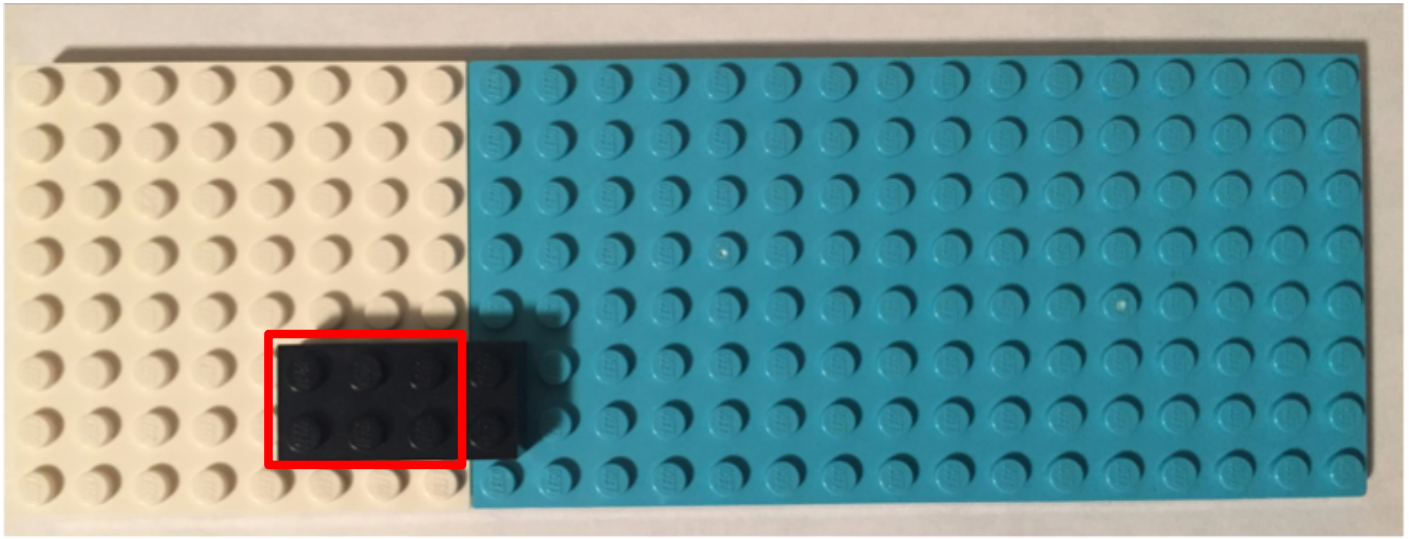
P(black | blue): "probability of black given that we are on a blue peg"



P(black | blue) = 2 / 128 = 0.015625

Inspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

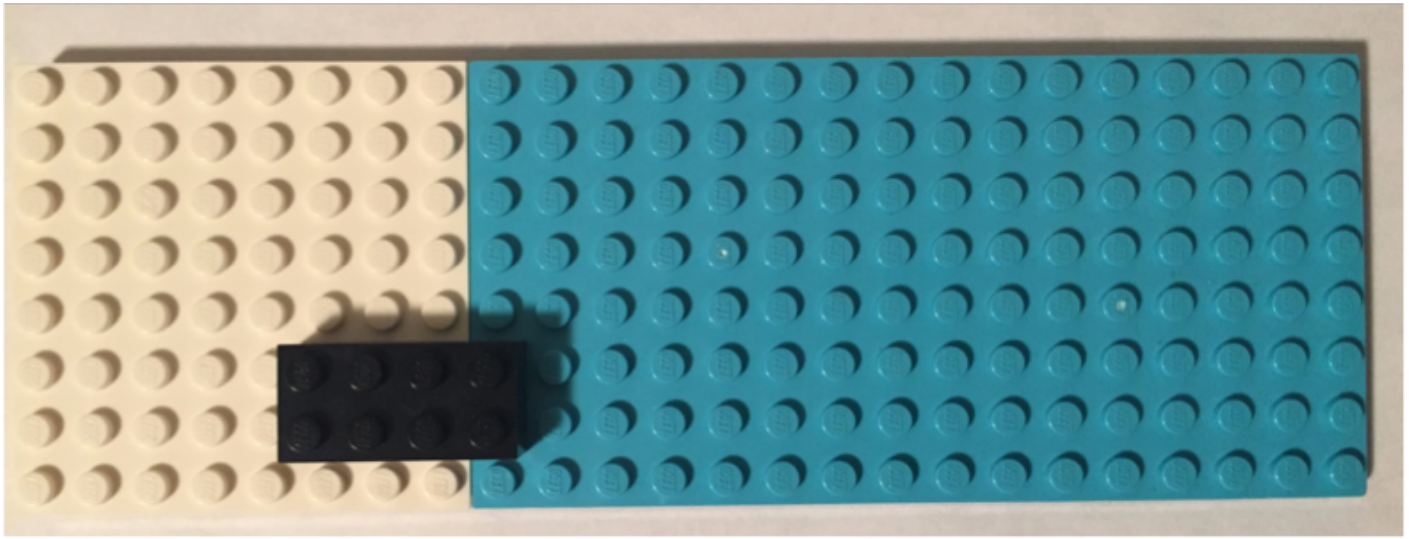
P(black | white): "probability of black given that we are on a white peg"



P(black | white) = 6 / 64 = 0.09375

Inspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

But what about the P(blue | black)?



P(blue | black) = 2 / 8 = 0.25

This intuition is formalized with Bayes' theorem.

Inspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

Bayes theorem

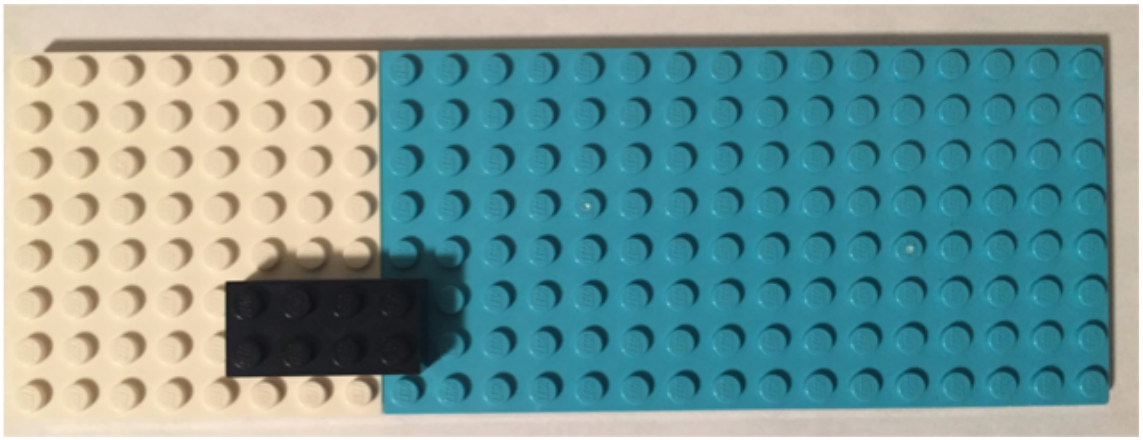
Prior Probability Of A

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

Posterior probability

Bayes theorem

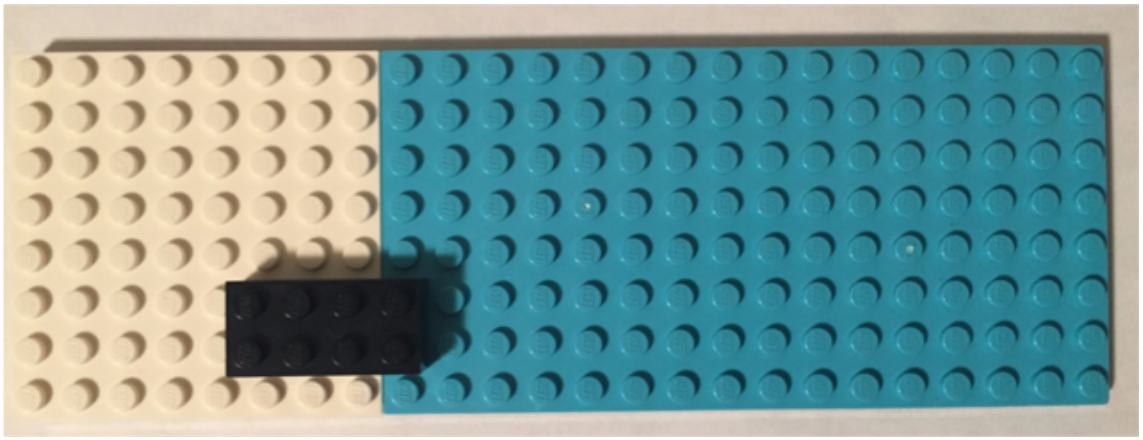
$$P(\text{black}|\text{white}) = \frac{P(\text{white}|\text{black}) * P(\text{black})}{P(\text{white})}$$



Inspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

Bayes theorem

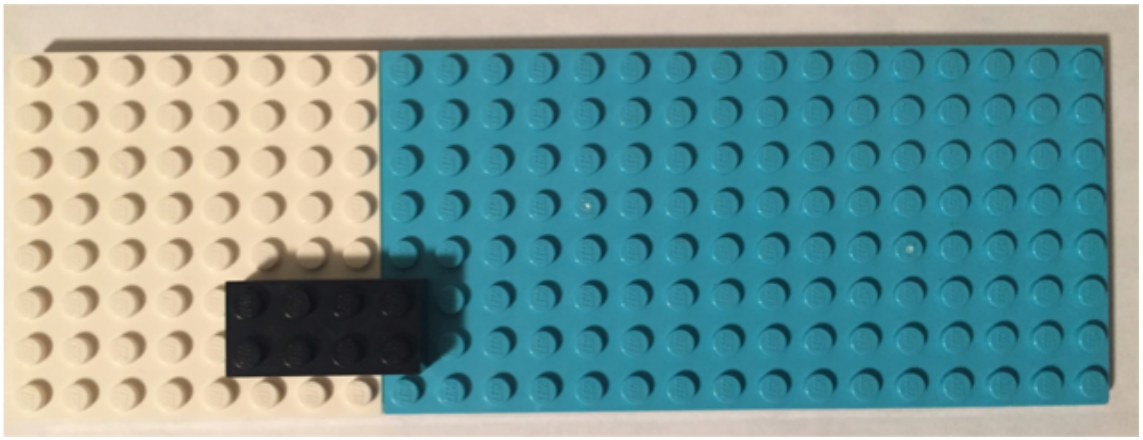
$$P(\text{black}|\text{white}) = \frac{0.75 * 0.0408}{0.33}$$



nspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

Bayes theorem

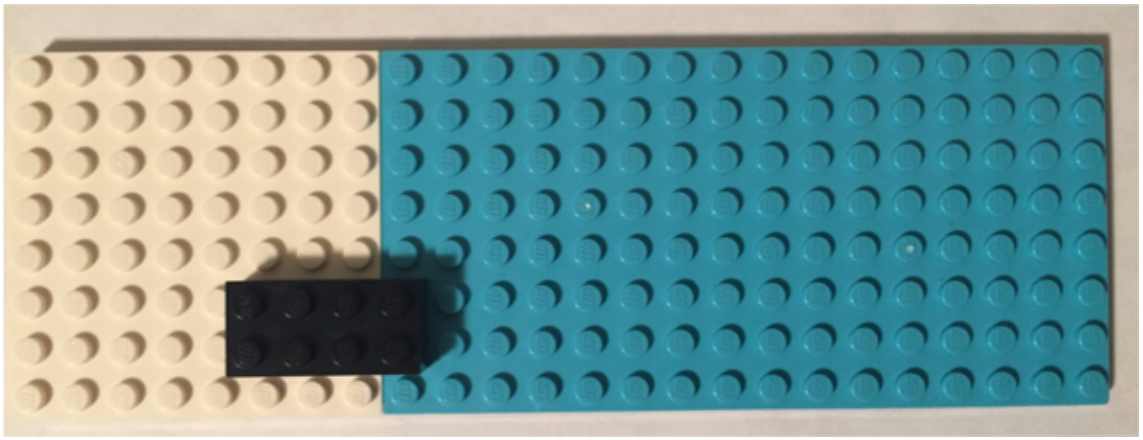
$$P(\text{black}|\text{white}) = 0.09375$$



Inspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

Bayes theorem

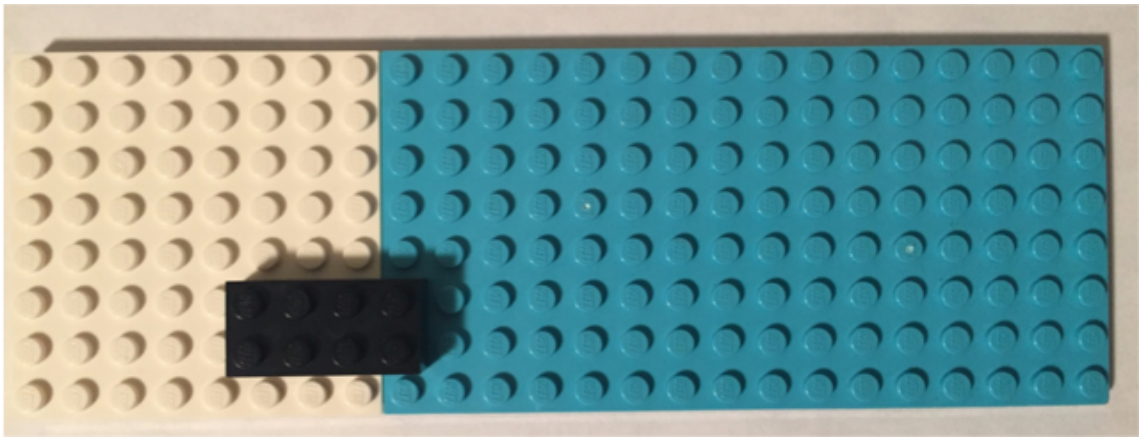
$$P(\text{white}|\text{black}) = \frac{P(\text{black}|\text{white}) * P(\text{white})}{P(\text{black})}$$



Inspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

Bayes theorem

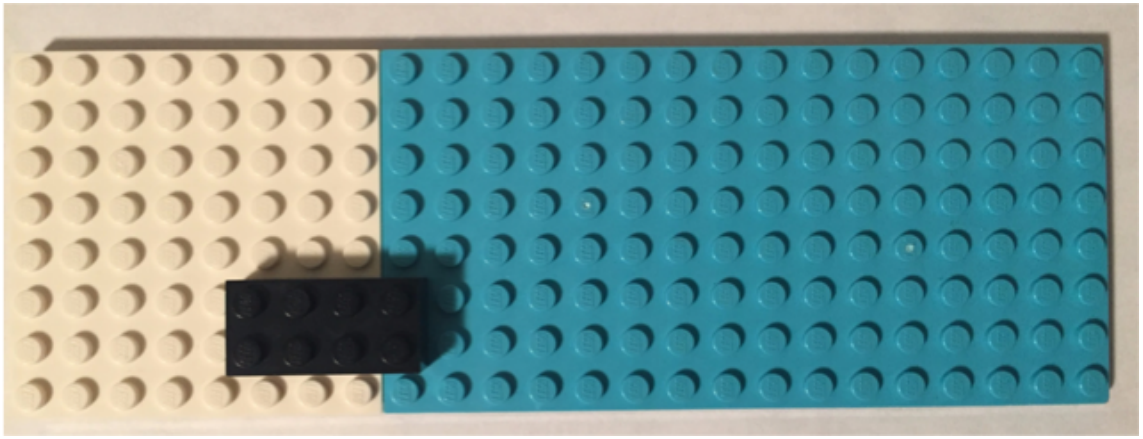
$$P(\text{white}|\text{black}) = \frac{0.09375 * 0.33}{0.0408}$$



Inspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

Bayes theorem

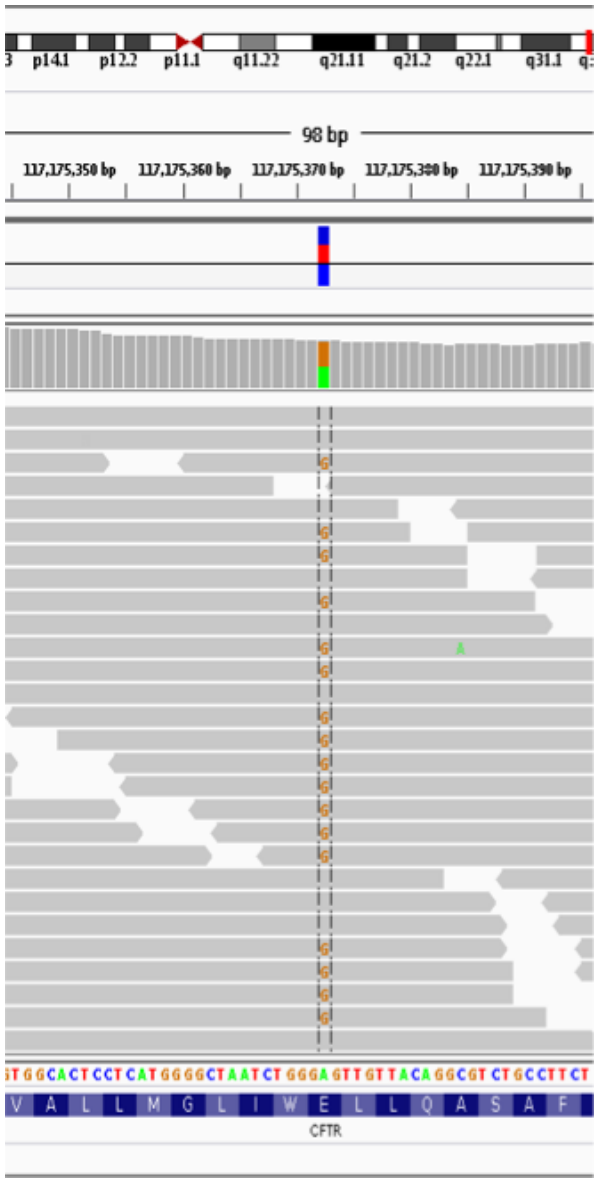
$$P(\text{white} \mid \text{black}) = 0.75$$



Inspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

While we can intuit these probabilities spatially with legos, the beauty of Bayes' theorem is that it can be generalized to situations that we cannot easily intuit.

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

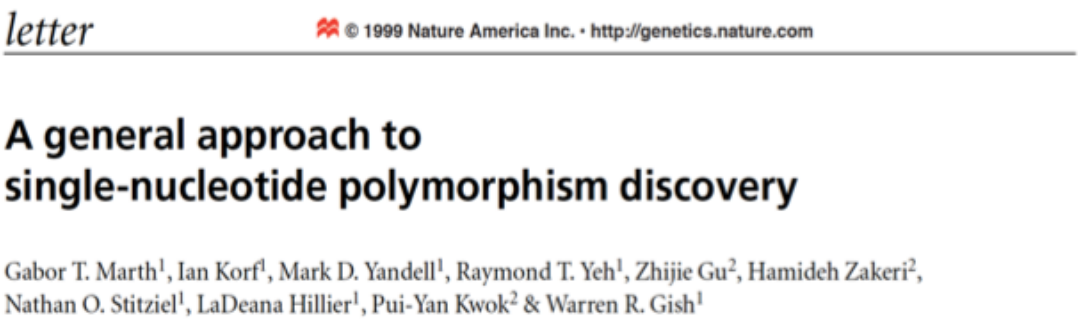
Expected (prior) polymorphism rate

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

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 © 1999 Nature America Inc. • http://genetics.nature.com

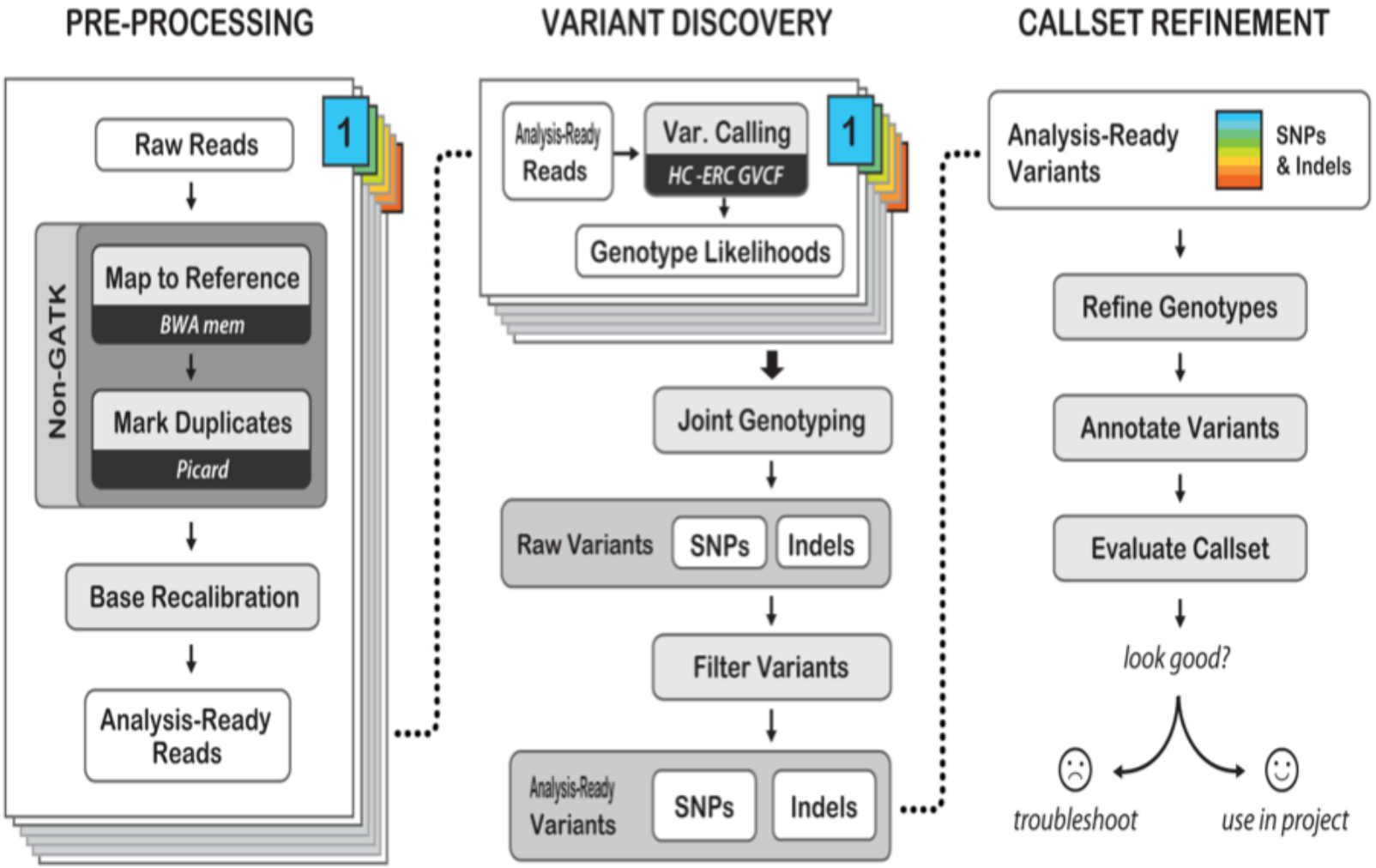
A general approach to single-nucleotide polymorphism discovery

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

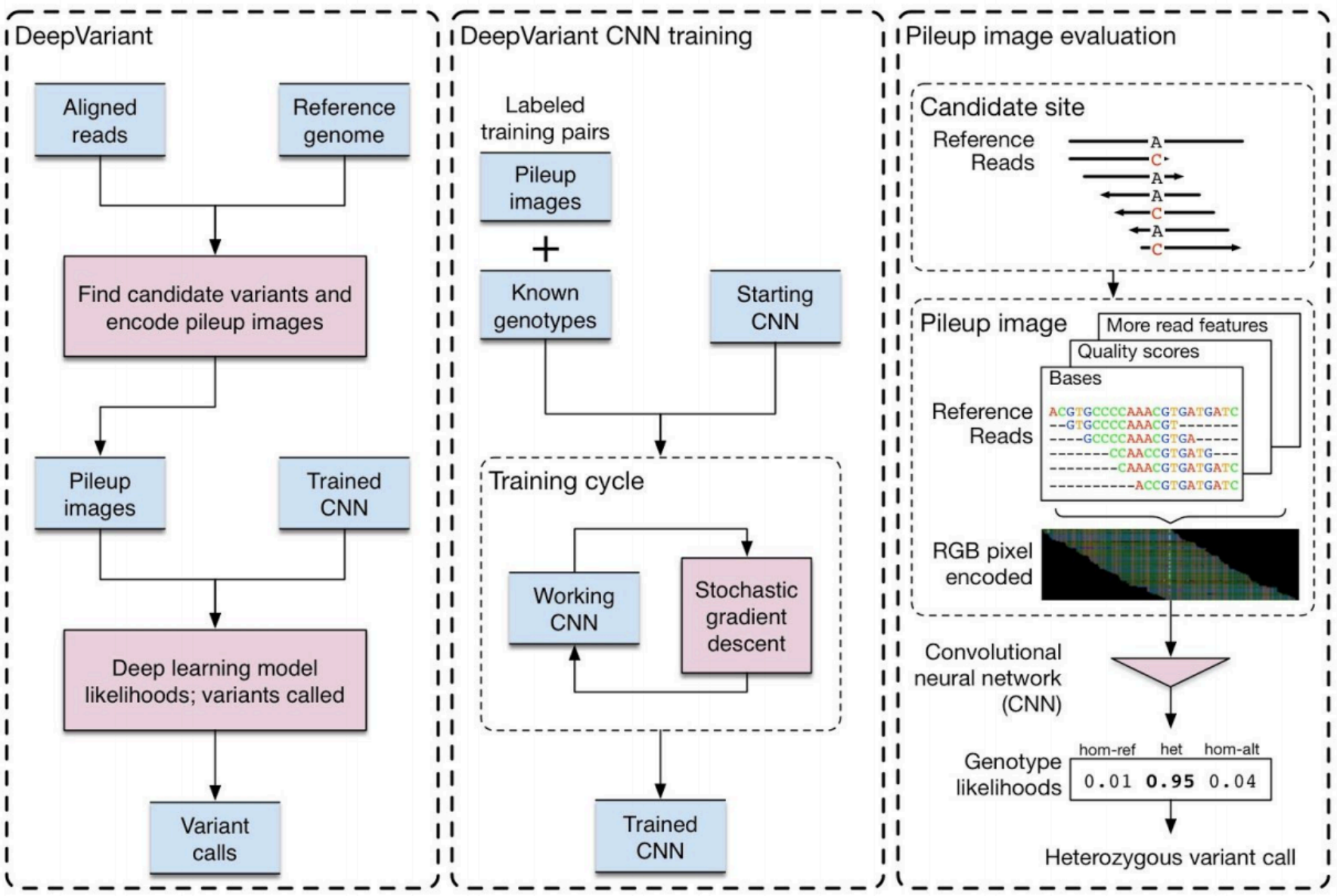
GATK workflow



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

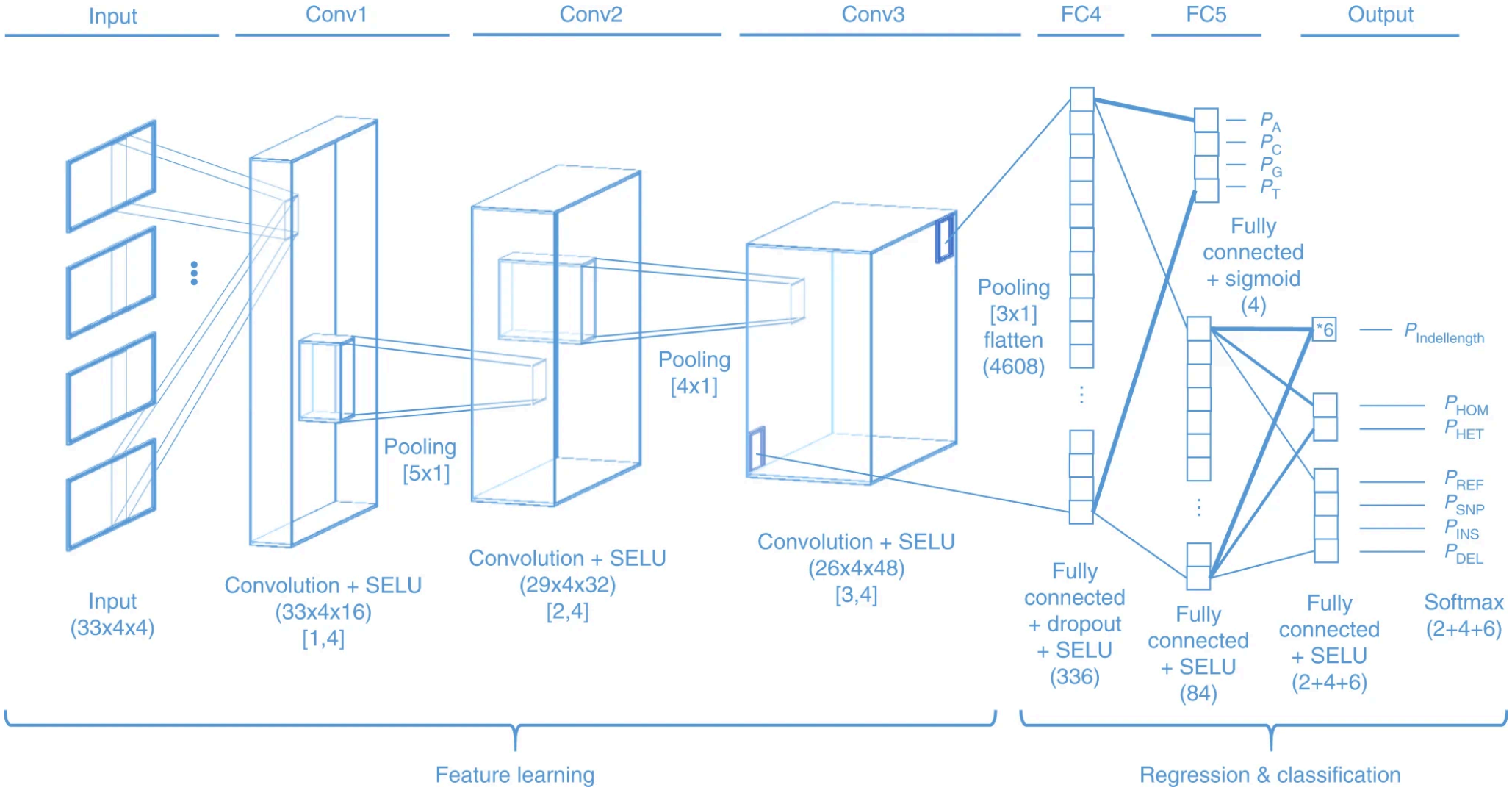
https://software.broadinstitute.org/gatk/img/BP_workflow_3.6.png

Deep Variant



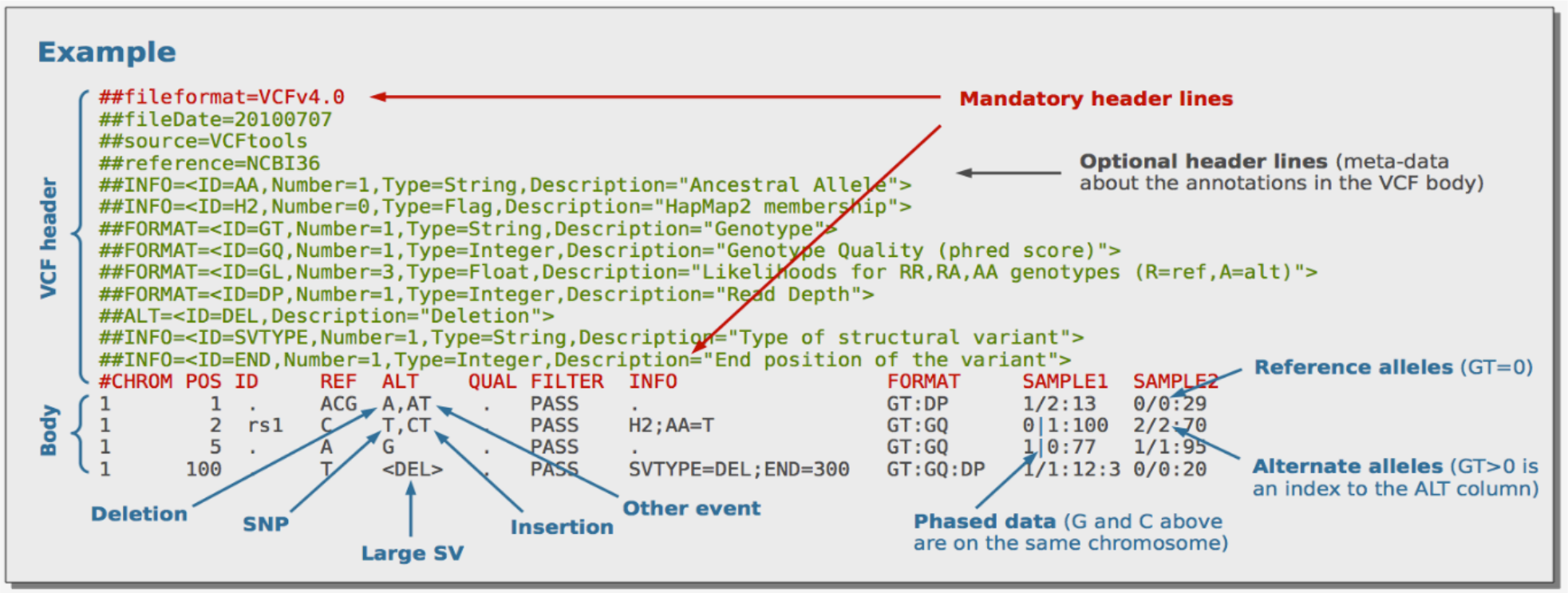
Creating a universal SNP and small indel variant caller with deep neural networks
Poplin et al. (2018) Nature Biotechnology. https://www.nature.com/articles/nbt.4235

Clairvoyant

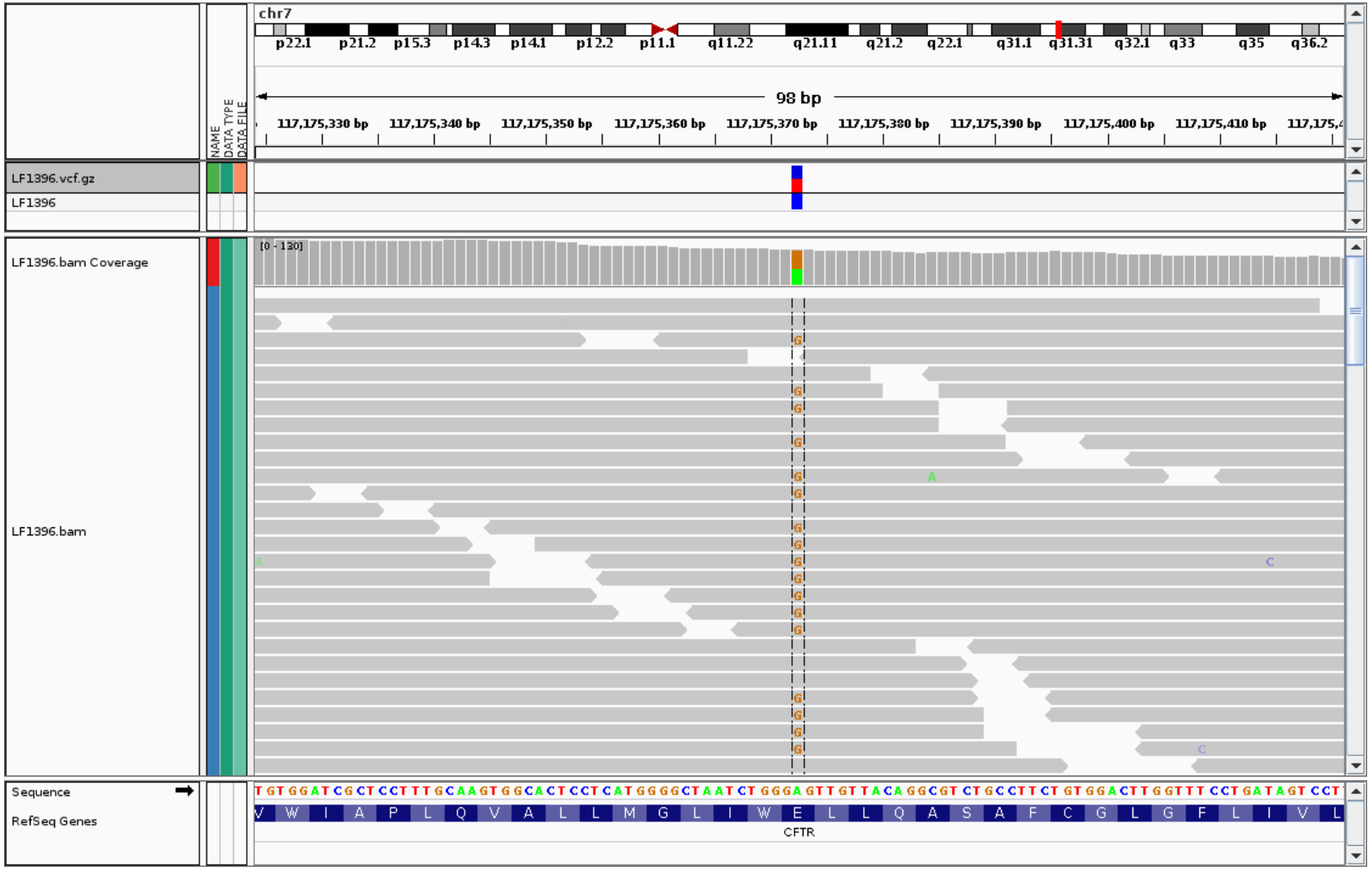


A multi-task convolutional deep neural network for variant calling in single molecule sequencing
Luo *et al.* (2019) Nature Communication. https://www.nature.com/articles/s41467-019-09025-z

VCF Format


Example
VCF header
Body
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 . T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

VCF Format




```
#CHROM POS       ID  REF ALT QUAL   FILTER INFO    FORMAT LF1396
chr7   117175373 .   A   G   90     PASS   AF=0.5  GT     0/1
```