Variant Calling

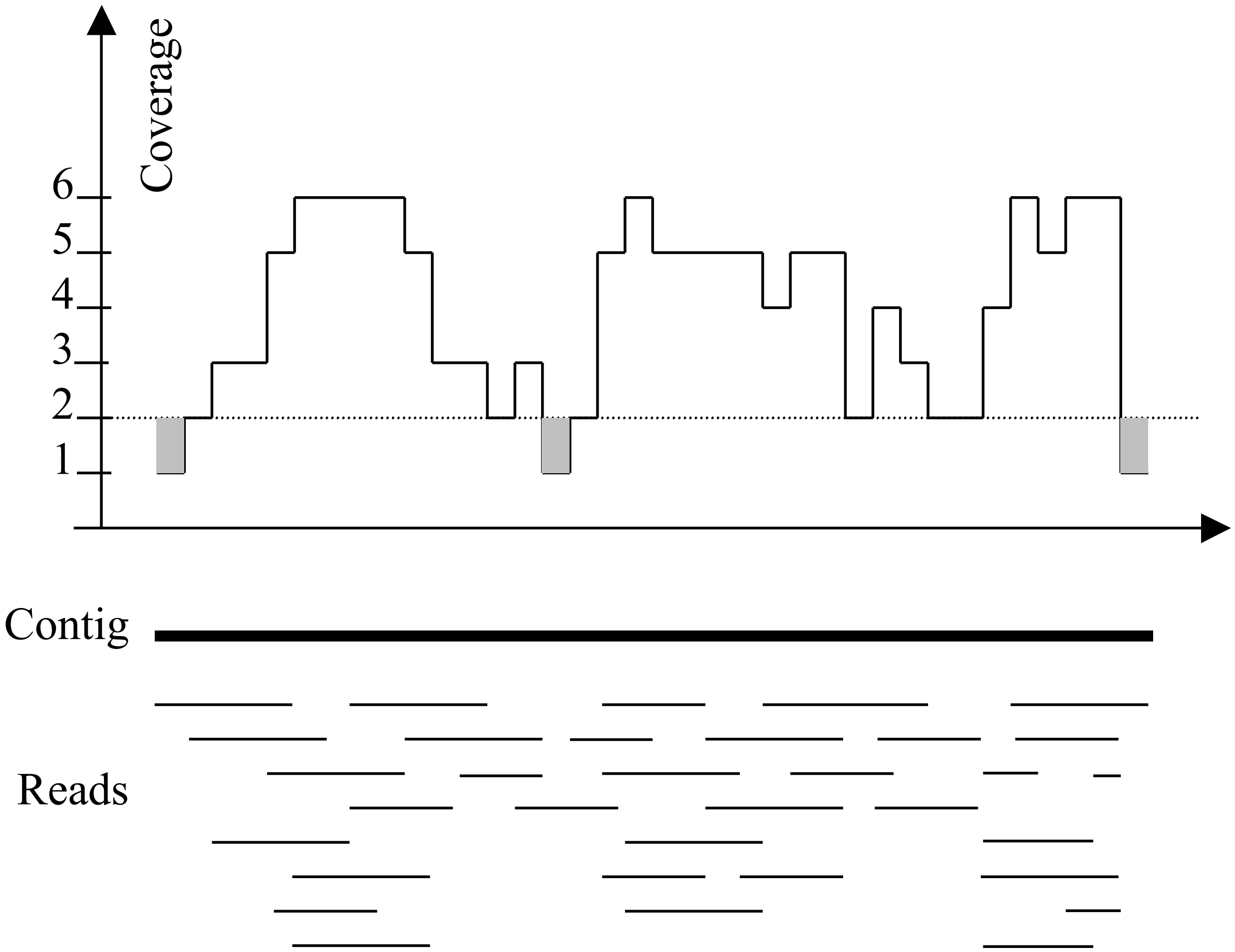
Michael Schatz

October 23 – Lecture 15

EN.601.452 Computational Biomedical Research

AS.020.415 Advanced Biomedical Research

Typical sequencing coverage

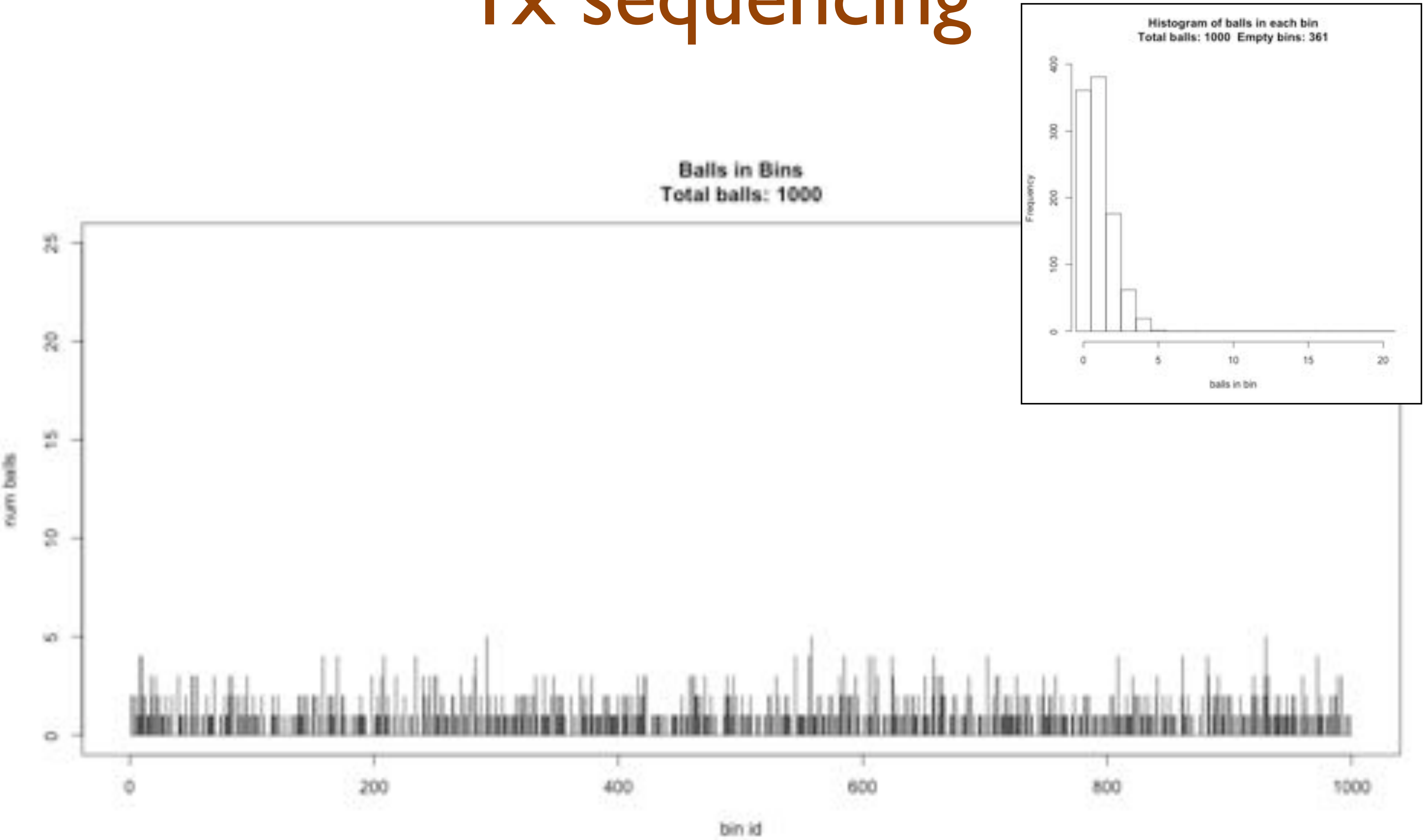


Imagine raindrops on a sidewalk

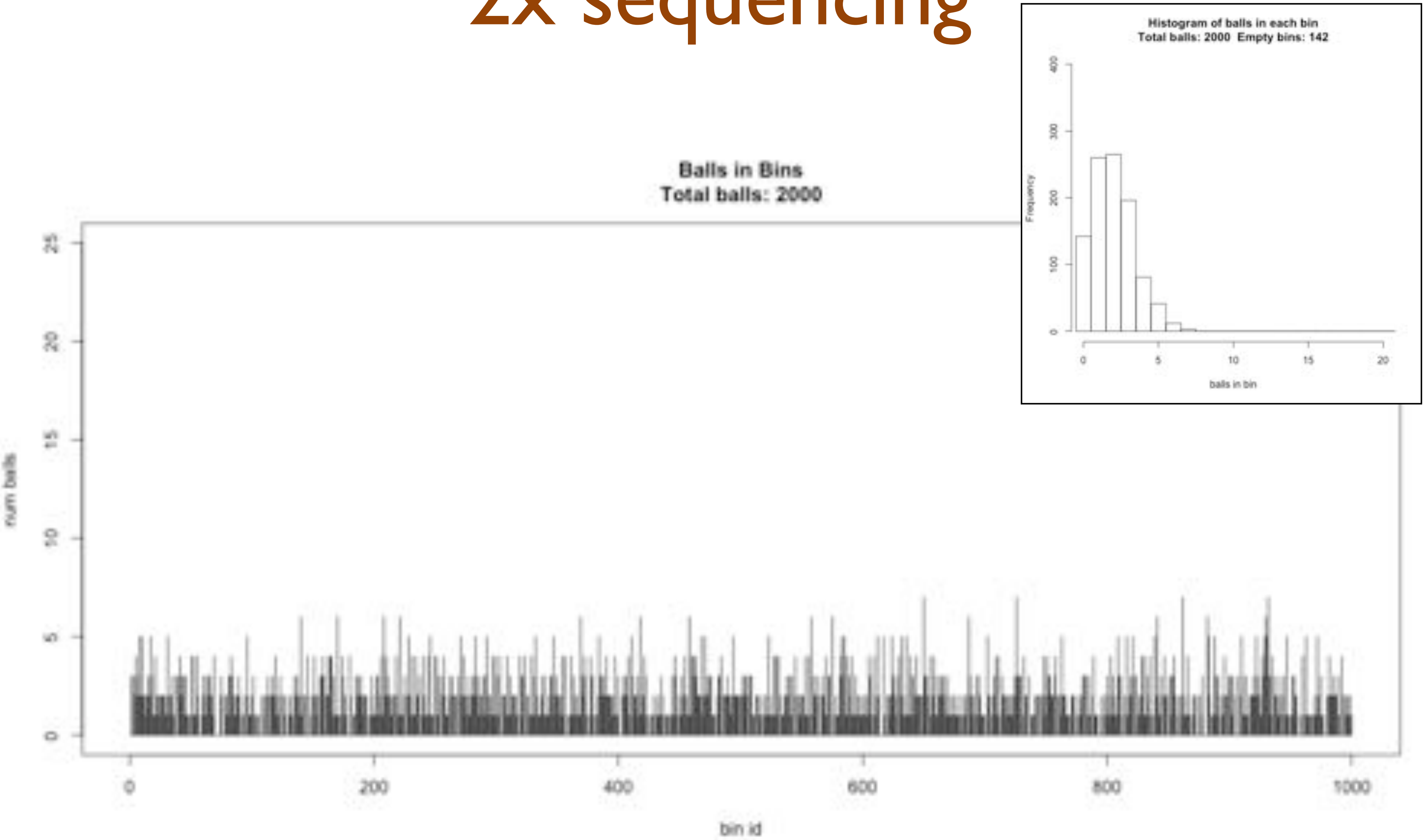
We want to cover the entire sidewalk but each drop costs $1

If the genome is 100 Mbp, should we sequence 1M 100bp reads?

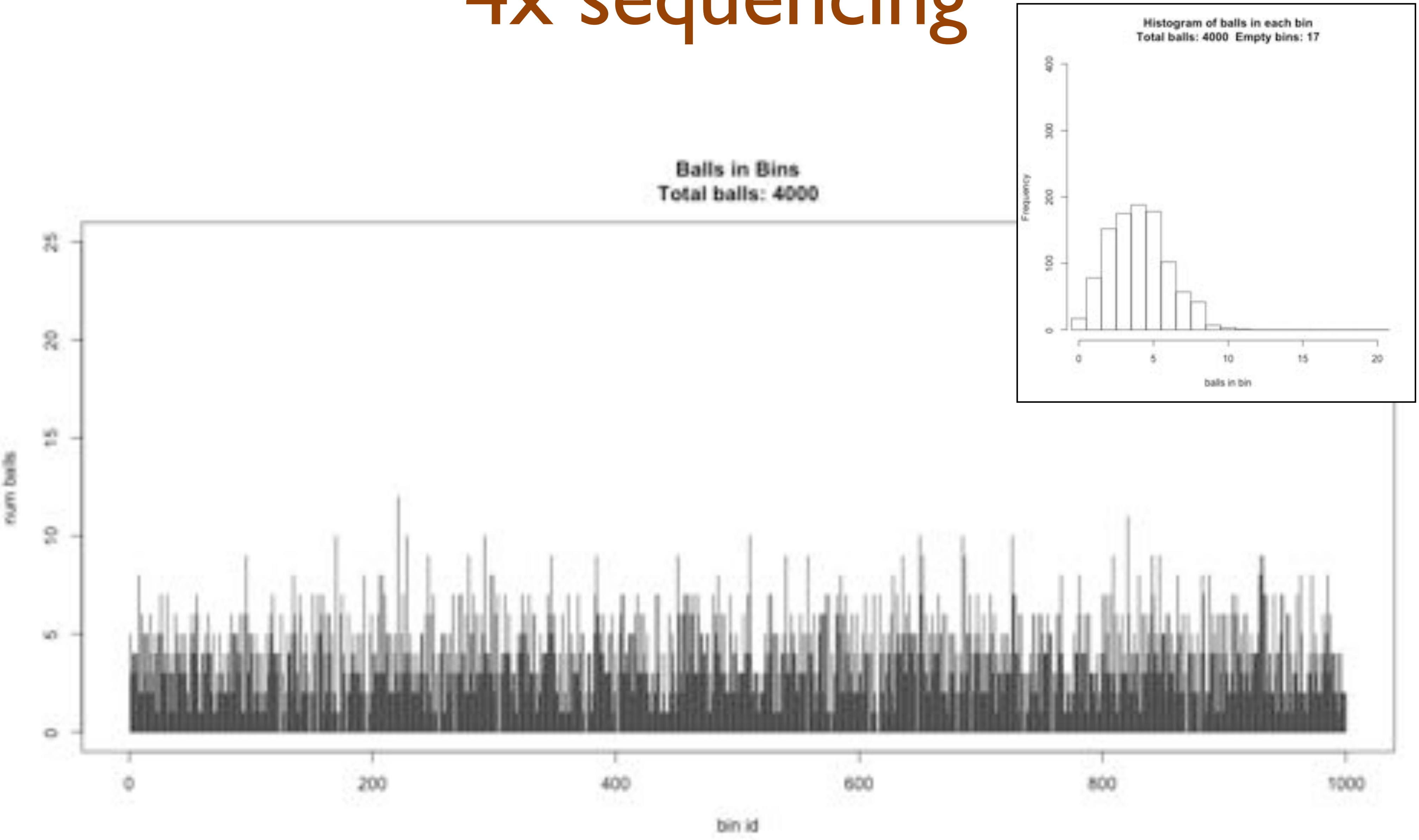
1x sequencing



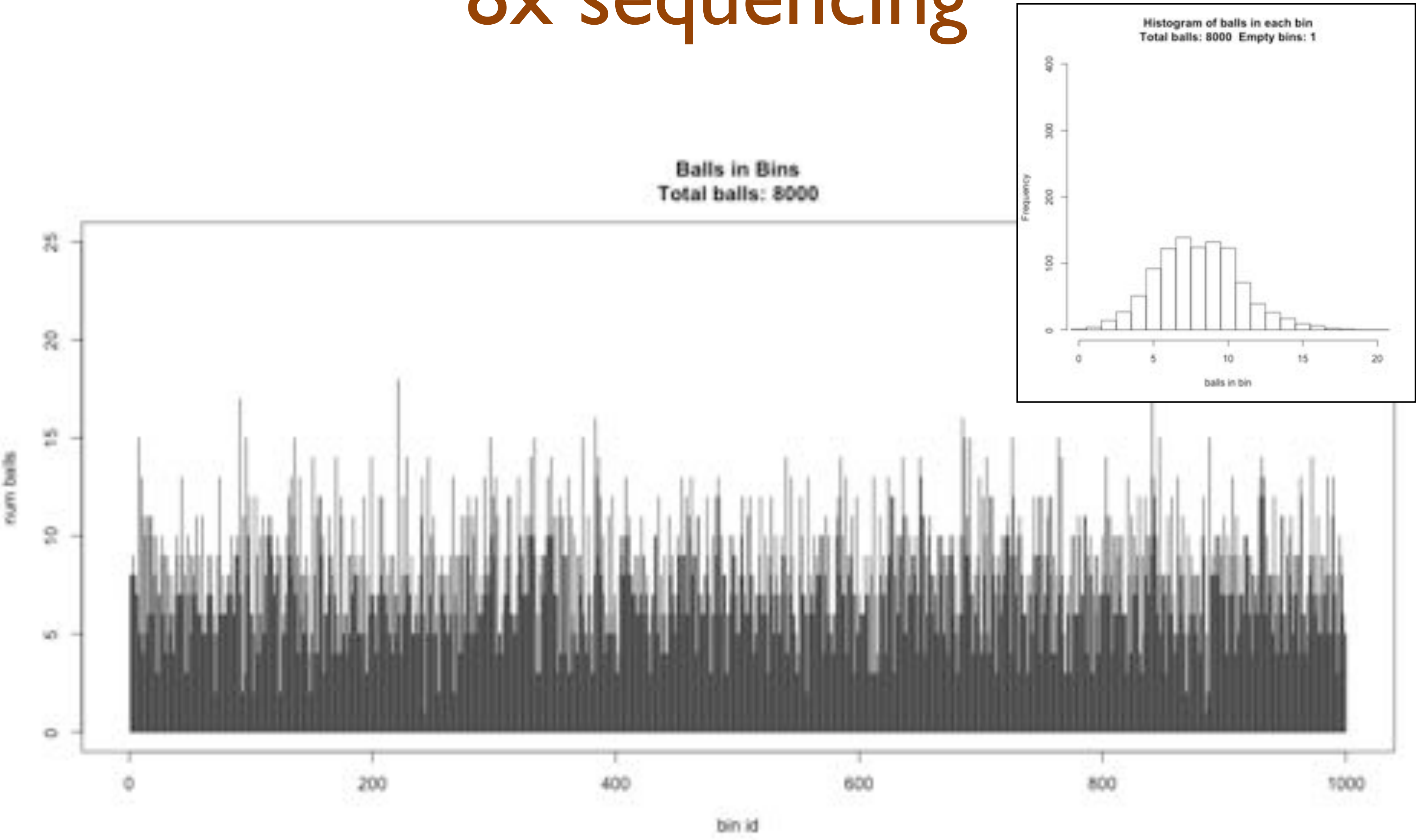
2x sequencing



4x sequencing



8x sequencing



Poisson Distribution

The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

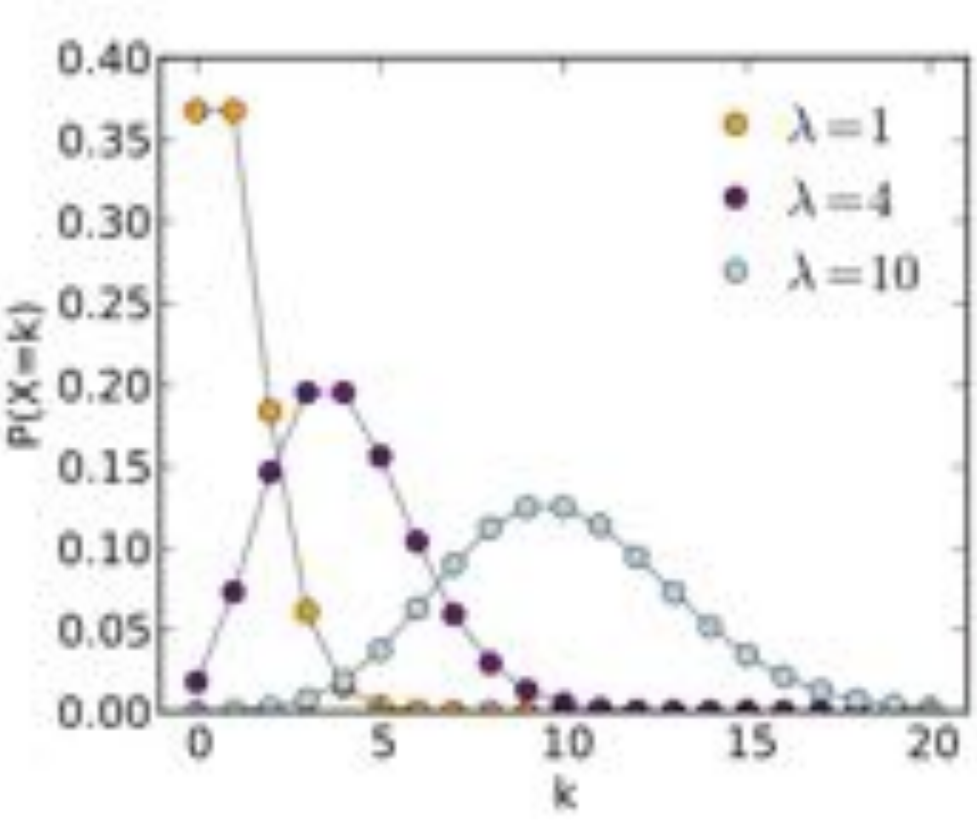
$$P(k) = \frac{\lambda^k}{k!}e^{-\lambda}$$

Formulation comes from the limit of the binomial equation

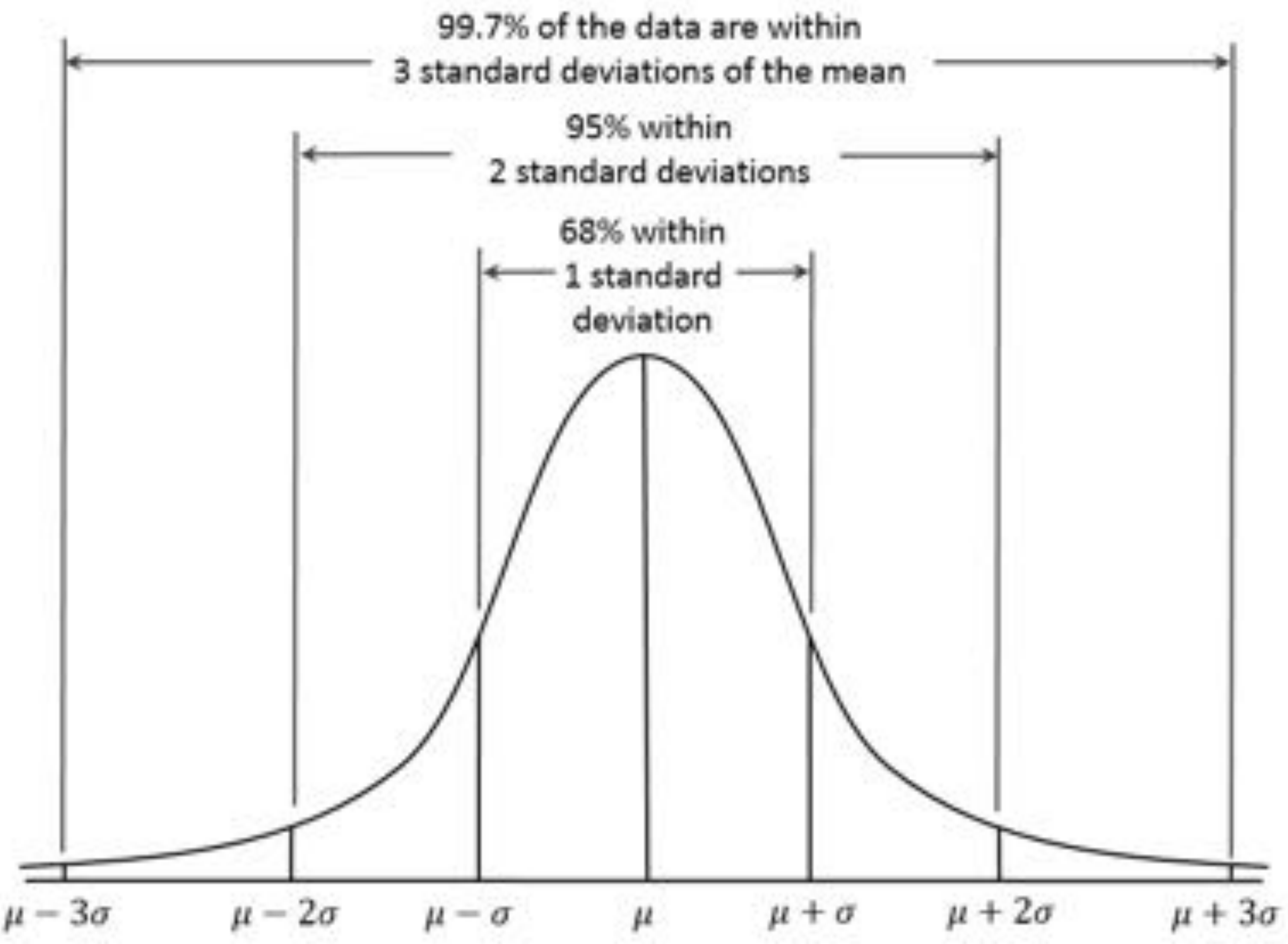
Resembles a normal distribution, but over the positive values, and with only a single parameter.

Key properties:

- ***The standard deviation is the square root of the mean.***
- ***For mean > 5, well approximated by a normal distribution***



Normal Approximation



Can estimate Poisson distribution as a normal distribution when λ > 10

Pop Quiz!

I want to sequence a 10Mbp genome to 24x coverage.
How many 150bp reads do I need?

I need 10Mbp x 24x = 240Mbp of data
240Mbp / 150bp / read = 1.6M reads

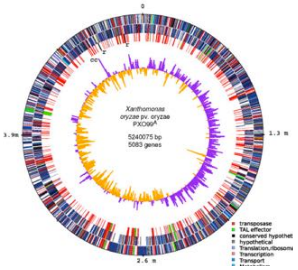
I want to sequence a 10Mbp genome so that
>97.5% of the genome has at least 24x coverage.
How many 150bp reads do I need?

Find X such that X-2*sqrt(X) = 24

36-2*sqrt(36) = 24

I need 10Mbp x 36x = 360Mbp of data
360Mbp / 150bp / read = 2.4M reads

Assembly Summary



Assembly quality depends on

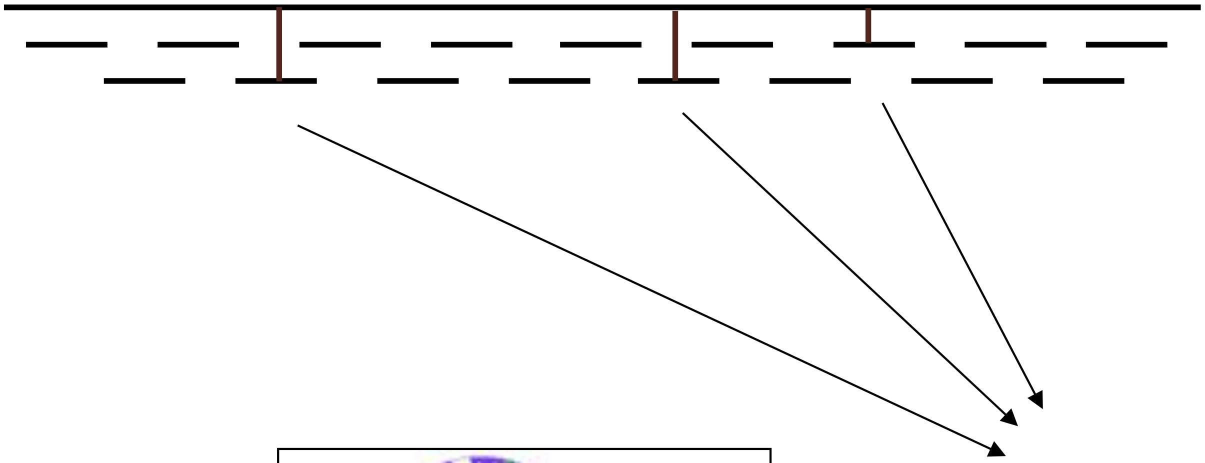
1. ***Coverage***: low coverage is mathematically hopeless
2. ***Repeat composition***: high repeat content is challenging
3. ***Read length***: longer reads help resolve repeats
4. ***Error rate***: errors reduce coverage, obscure true overlaps

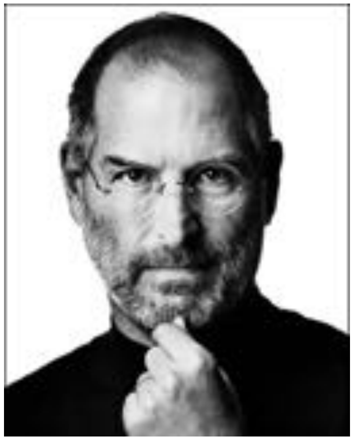
- Assembly is a hierarchical, starting from individual reads, build high confidence contigs/unitigs, incorporate the mates to build scaffolds
  - Extensive error correction is the key to getting the best assembly possible from a given data set

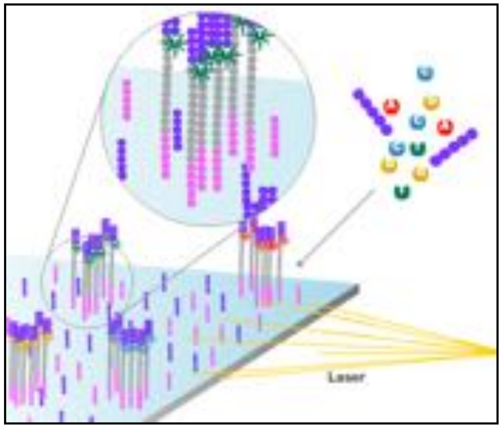
- Watch out for collapsed repeats & other misassemblies
  - Globally/Locally reassemble data from scratch with better parameters & stitch the 2 assemblies together

Personal Genomics

How does your genome compare to the reference?





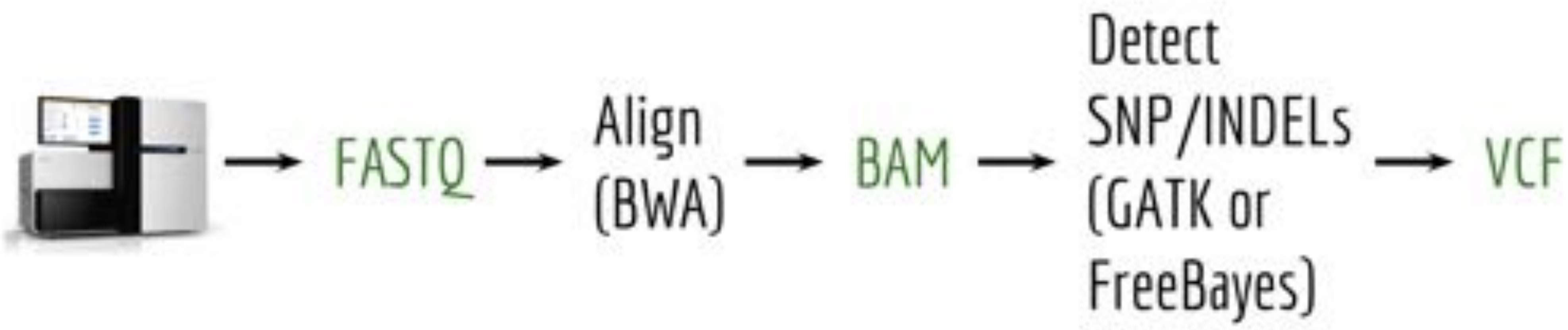


Heart Disease

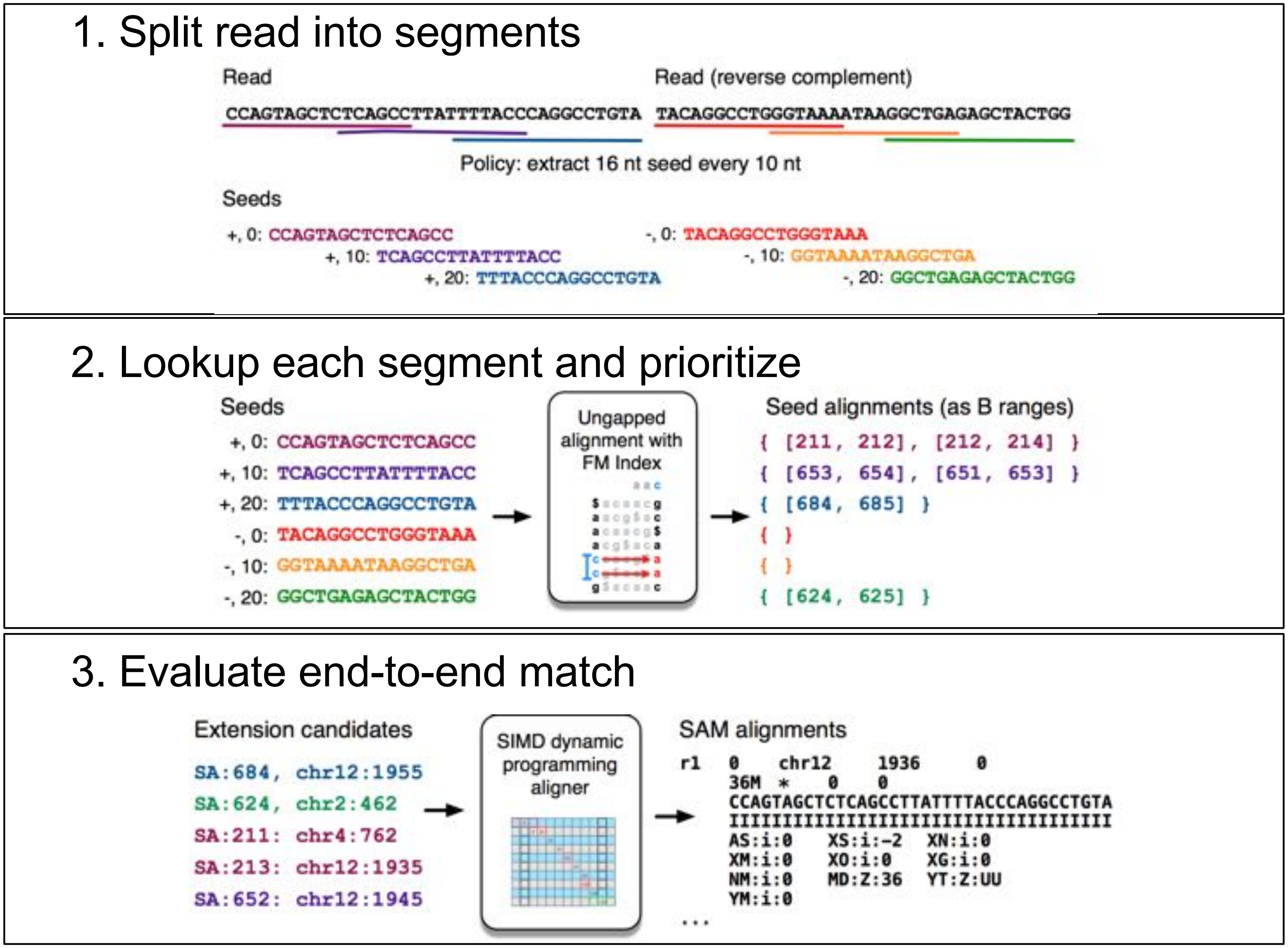
Cancer

Creates magical technology

Variant Calling Overview



Read Mapping 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
```

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
```

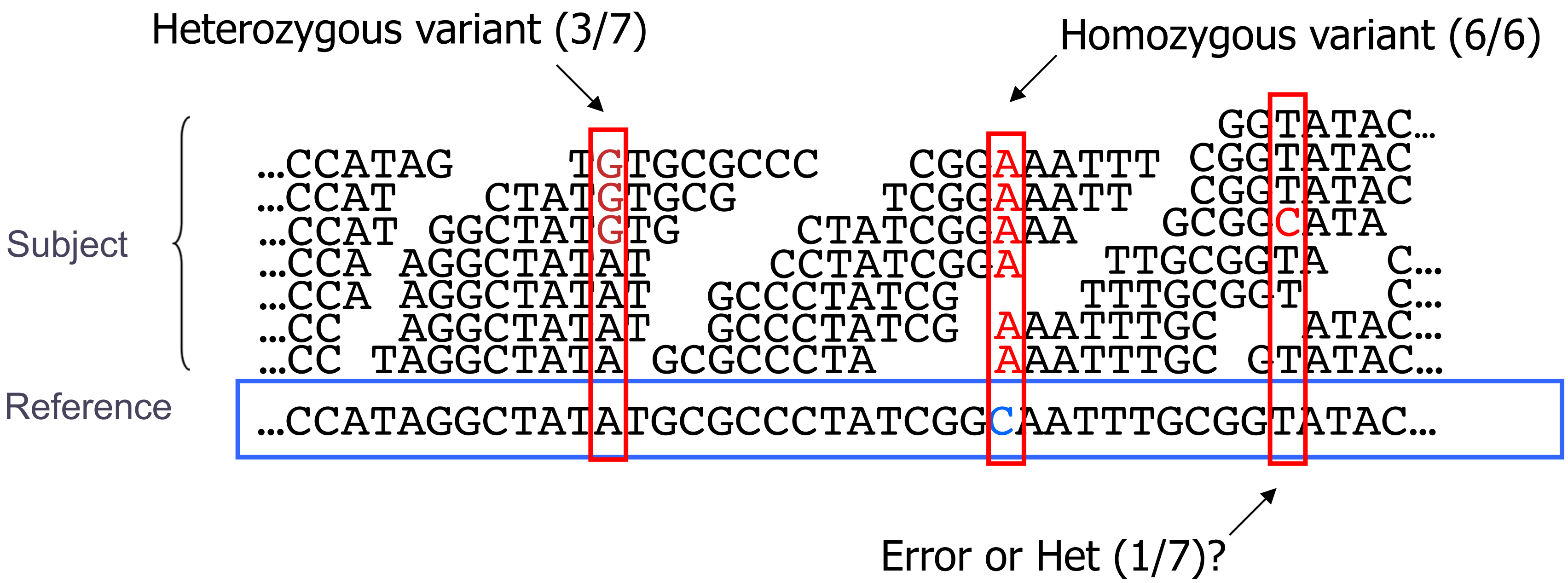
SAM alignments

```
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
...
```

Fast gapped-read alignment with Bowtie 2

Langmead & Salzberg (2012) Nature Methods. doi:10.1038/nmeth.1923

Genotyping Theory



Heterozygous variant (3/7)

Homozygous variant (6/6)

Subject

```
                                  GGTATAC…
…CCATAG    TGTGCGCCC   CGGAAATTT CGGTATAC
…CCAT   CTATGTGCG     TCGGAAATT  CGGTATAC
…CCAT GGCTATGTG    CTATCGGAAA   GCGGCATA
…CCA AGGCTATAT    CCTATCGGA   TTGCGGTA  C…
…CCA AGGCTATAT  GCCCTATCG    TTTGCGGT   C…
…CC  AGGCTATAT  GCCCTATCG AAATTTGC   ATAC…
…CC TAGGCTATA GCGCCCTA    AAATTTGC GTATAC…
```

Reference

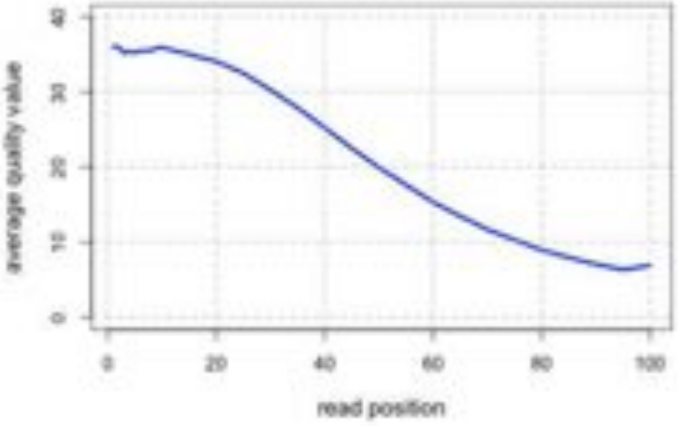
```
…CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC…
```

Error or Het (1/7)?

- 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



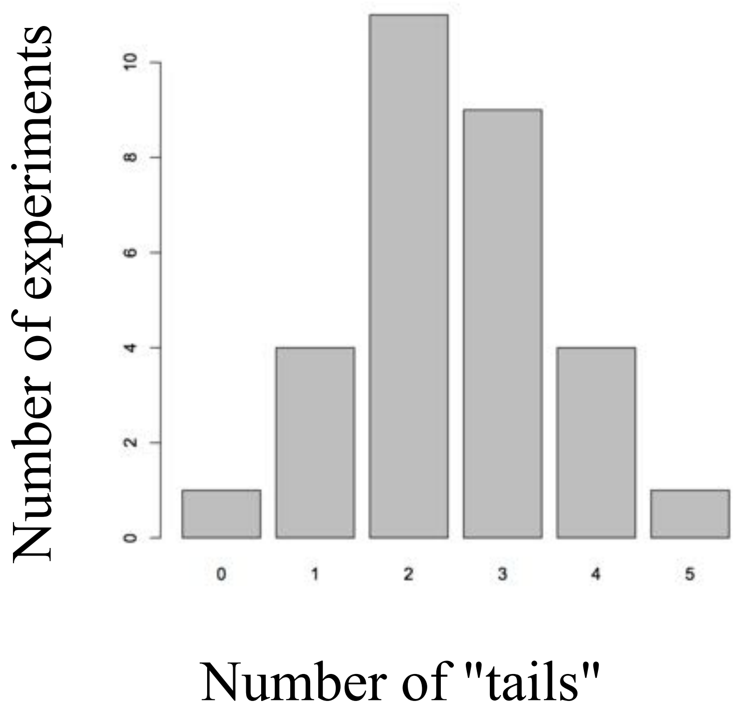
The Binomial Distribution: Adventures in Coin Flipping

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:

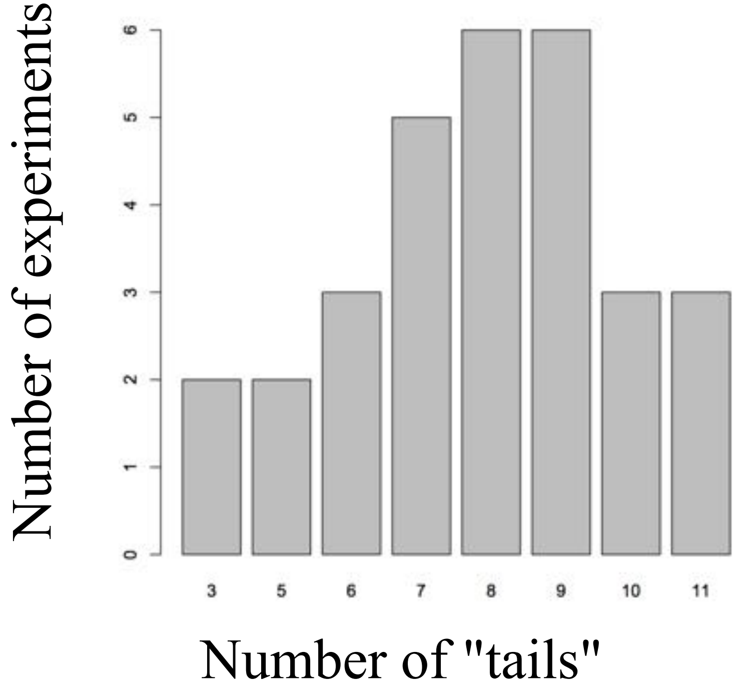
```
barplot(table(rbinom(30, 5, 0.5)))
```

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:

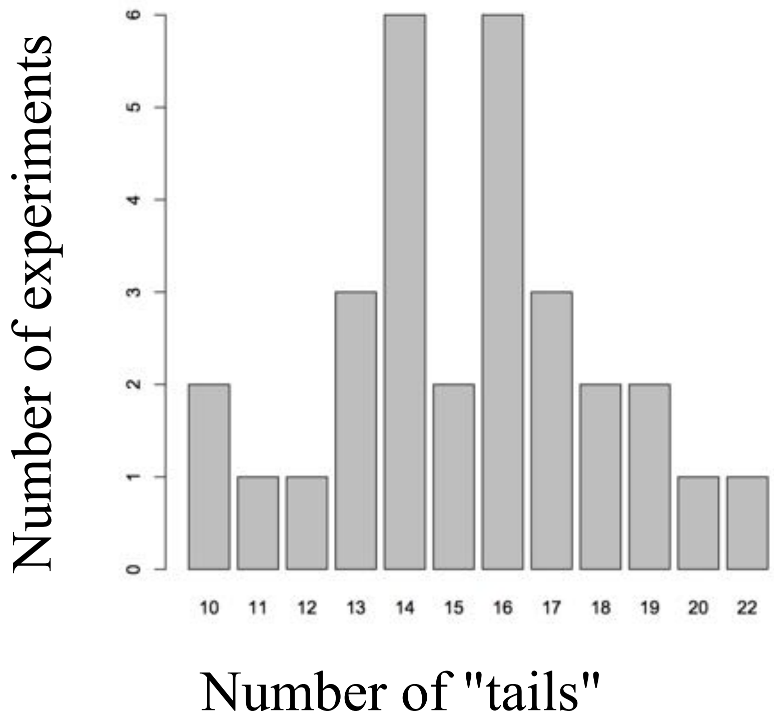
```
barplot(table(rbinom(30, 15, 0.5)))
```

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:

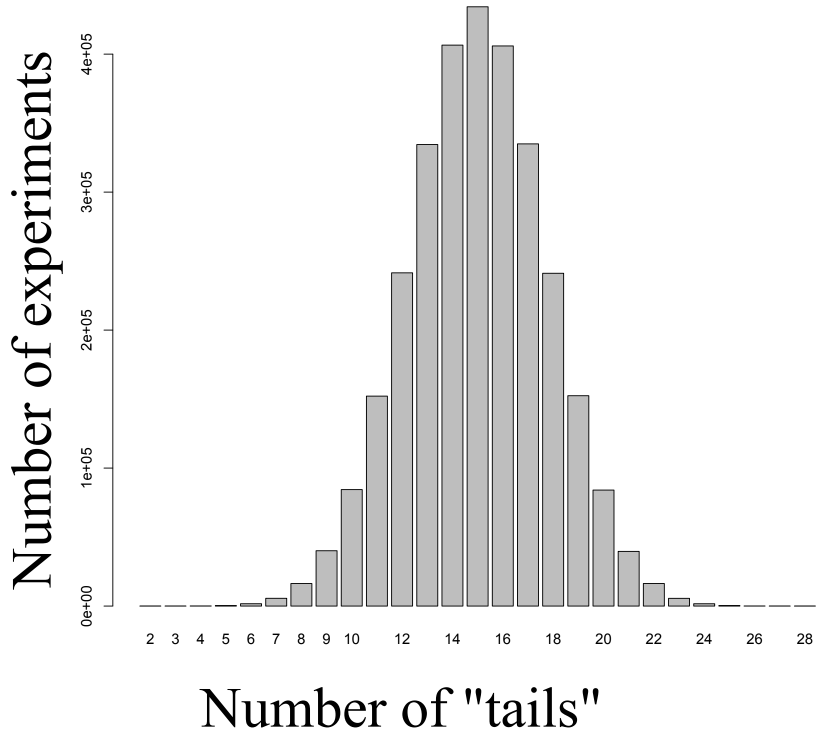
```
barplot(table(rbinom(30, 30, 0.5)))
```

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:

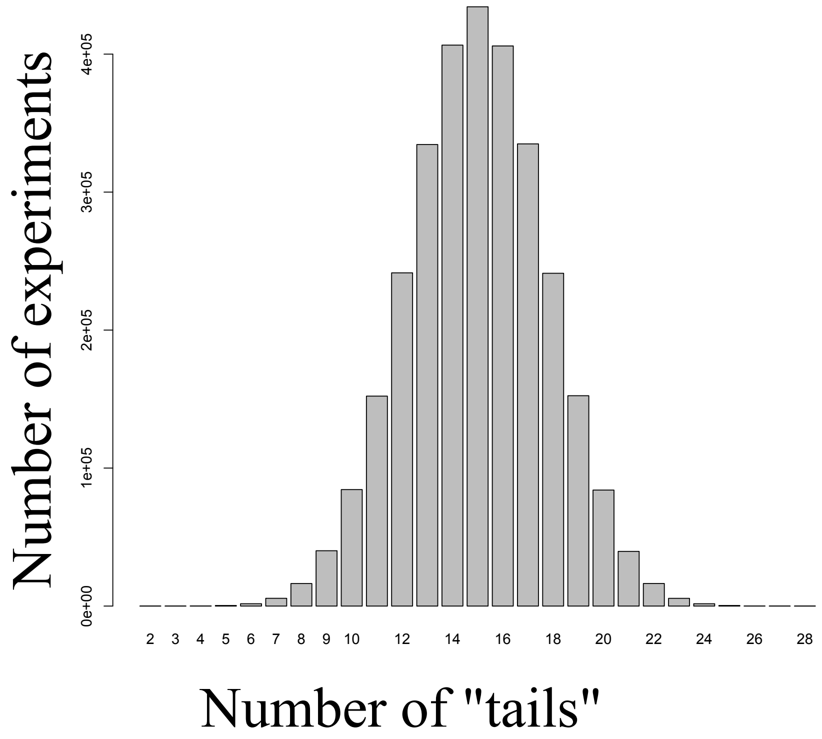
```
barplot(table(rbinom(3e6, 30, 0.5)))
```

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

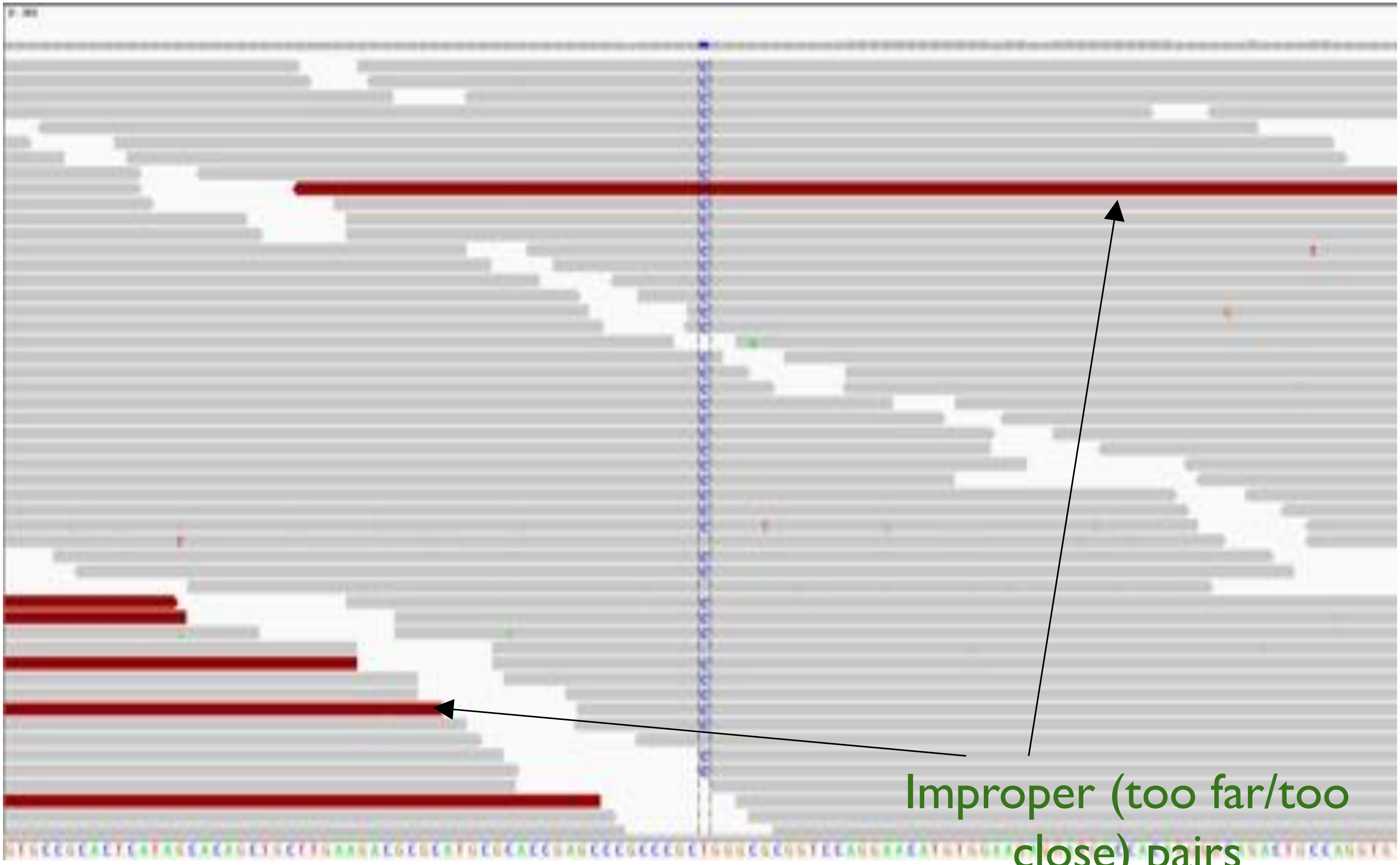


This is why at least a "30X" (30 fold sequence coverage) genome is recommended: it confers sufficient power to distinguish heterozygous alleles and from mere sequencing errors

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

Some real examples of SNPs in IGV

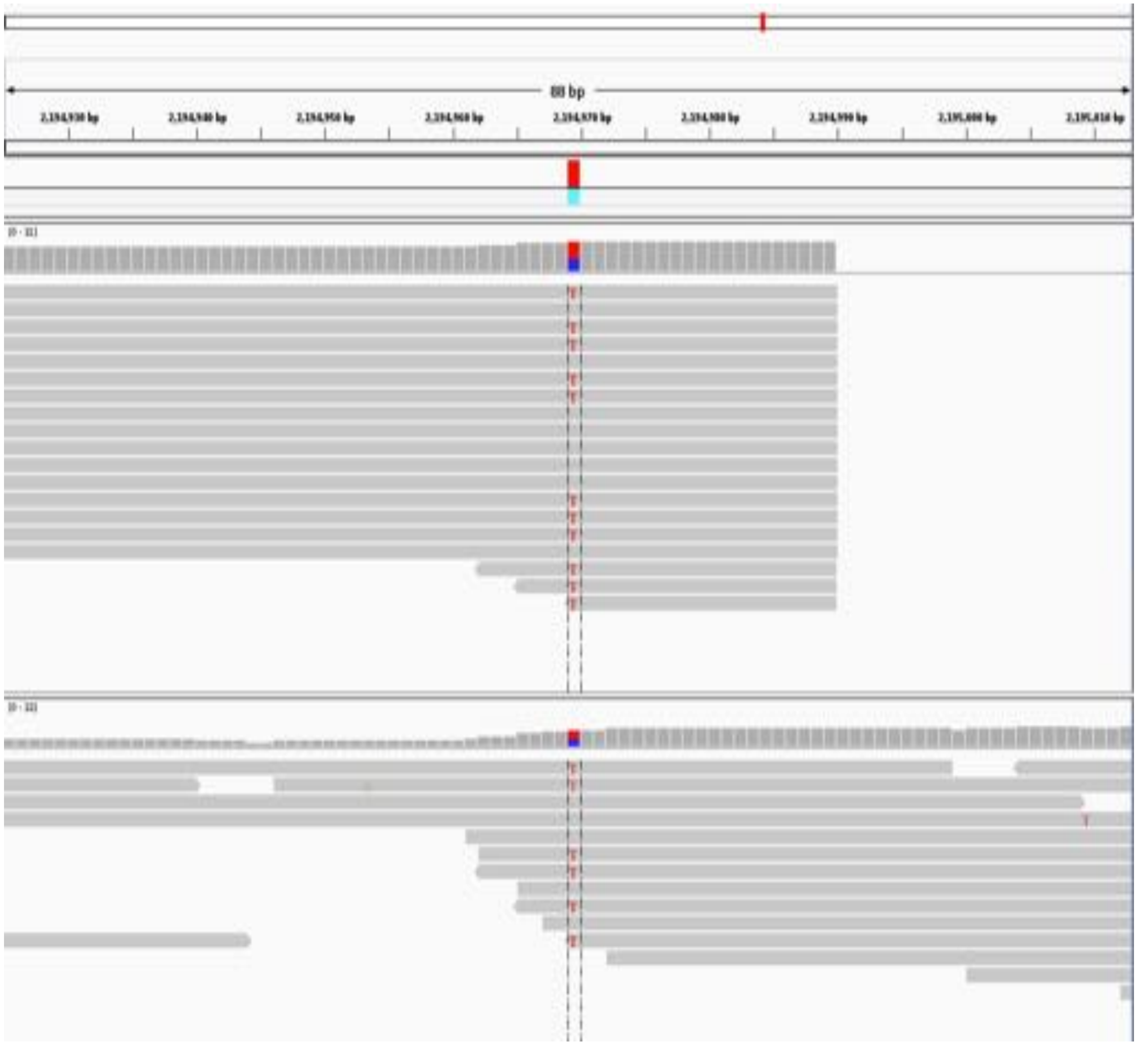
Homozygous for the "C" allele



Improper (too far/too close) pairs

What else do you notice?

Heterozygous for the alternate allele

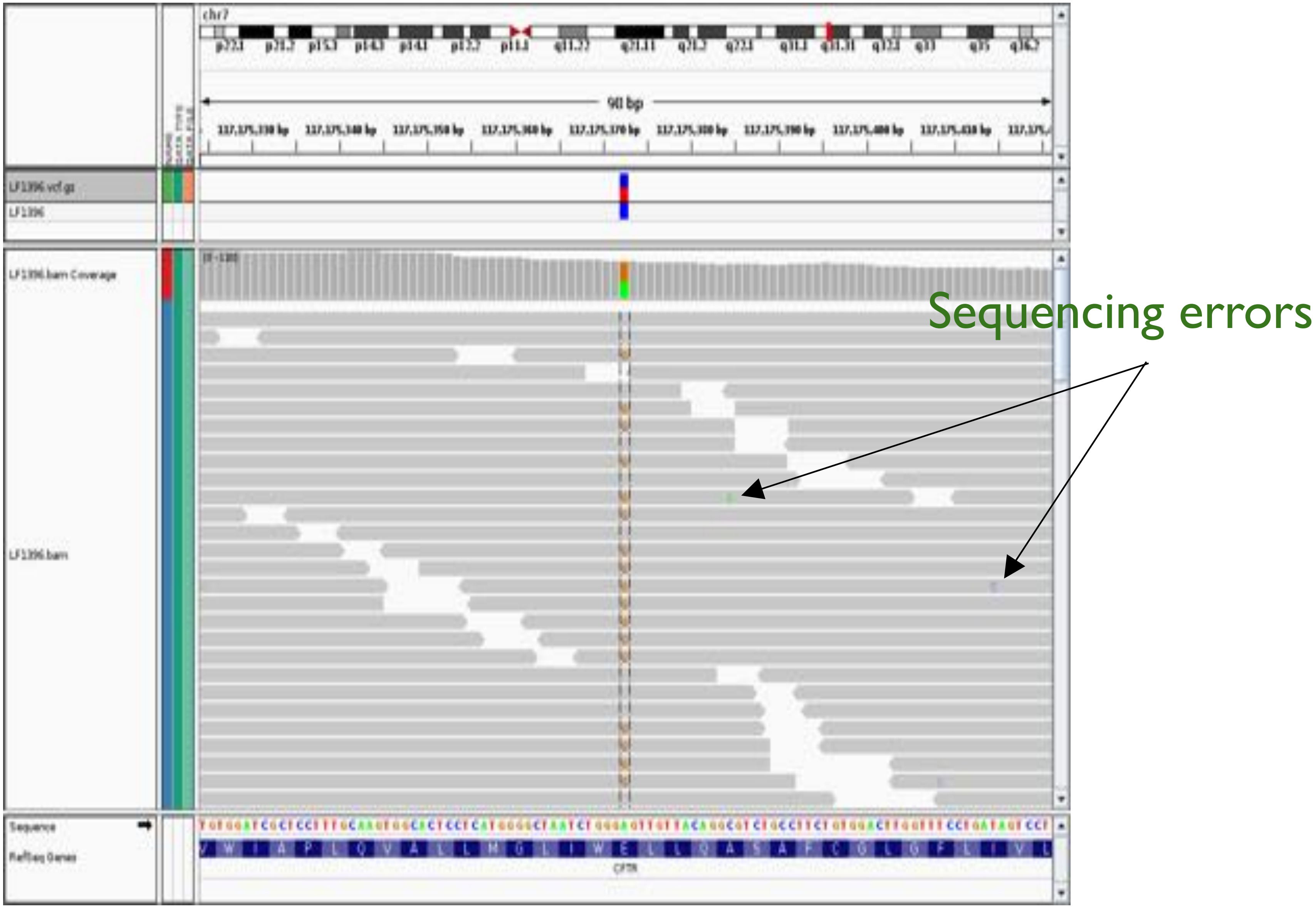


Individual 1

Individual 2

Which genotype prediction do you have more confidence in?

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

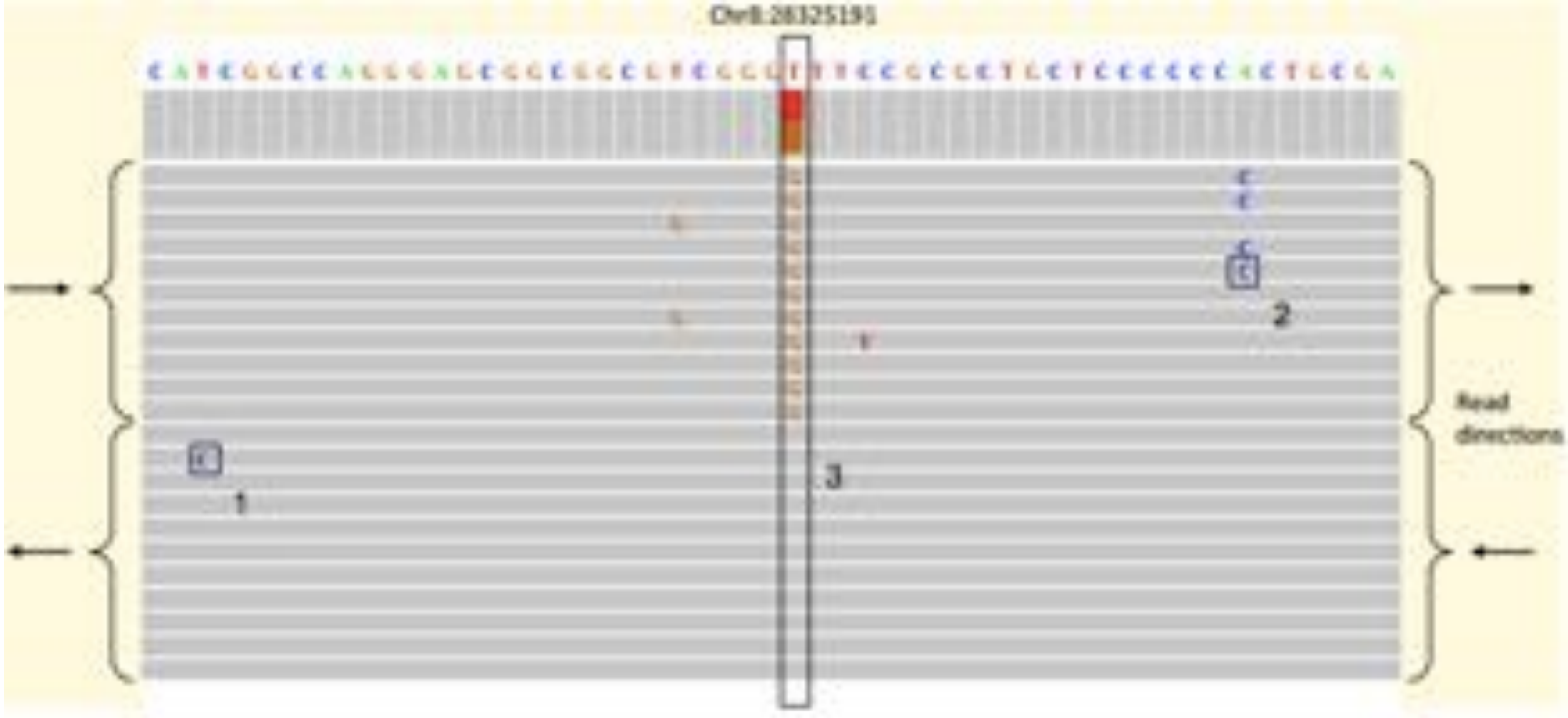


Sequencing errors

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

It is not always so easy ☹

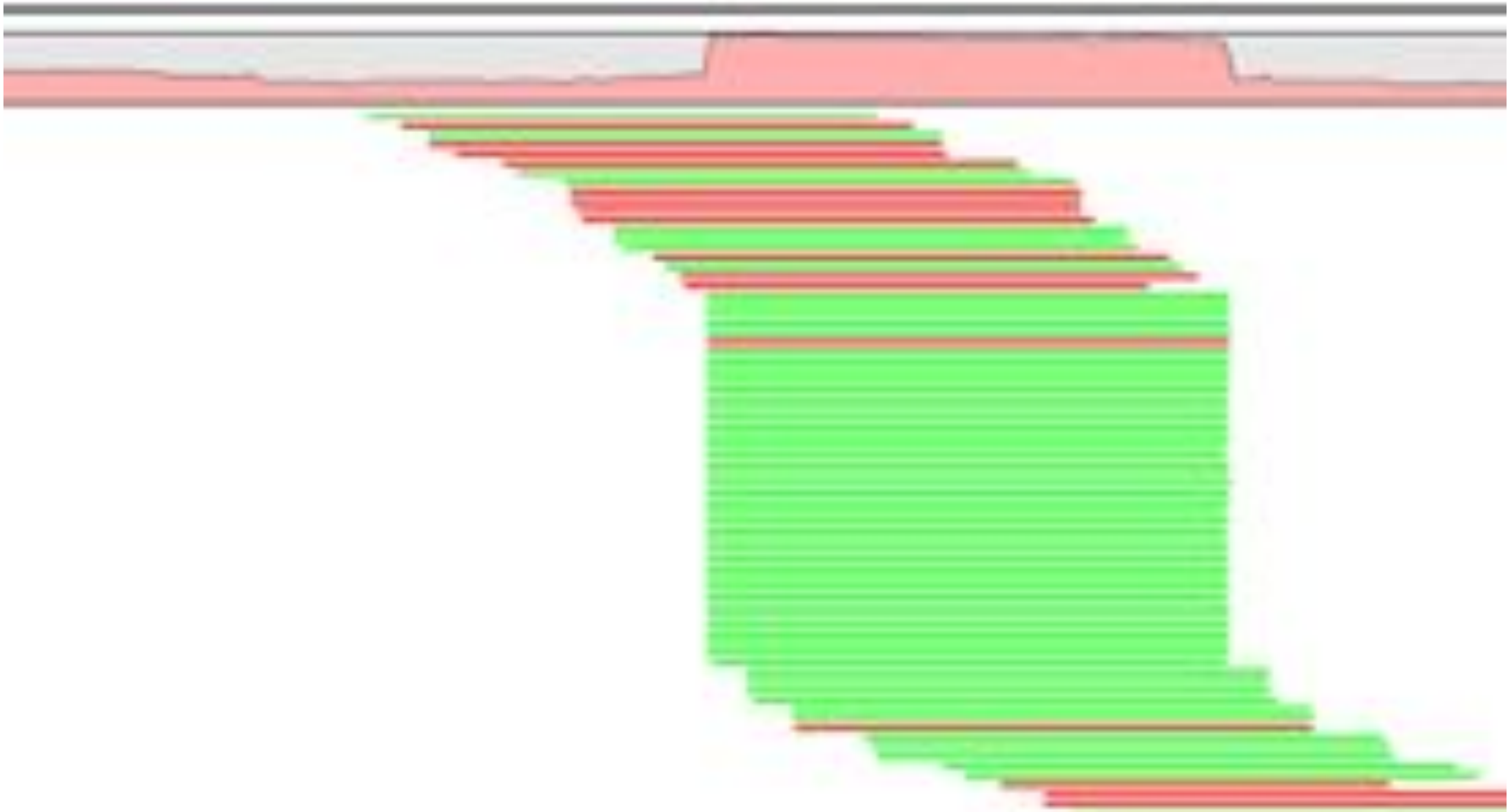
Beware of Systematic Errors



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

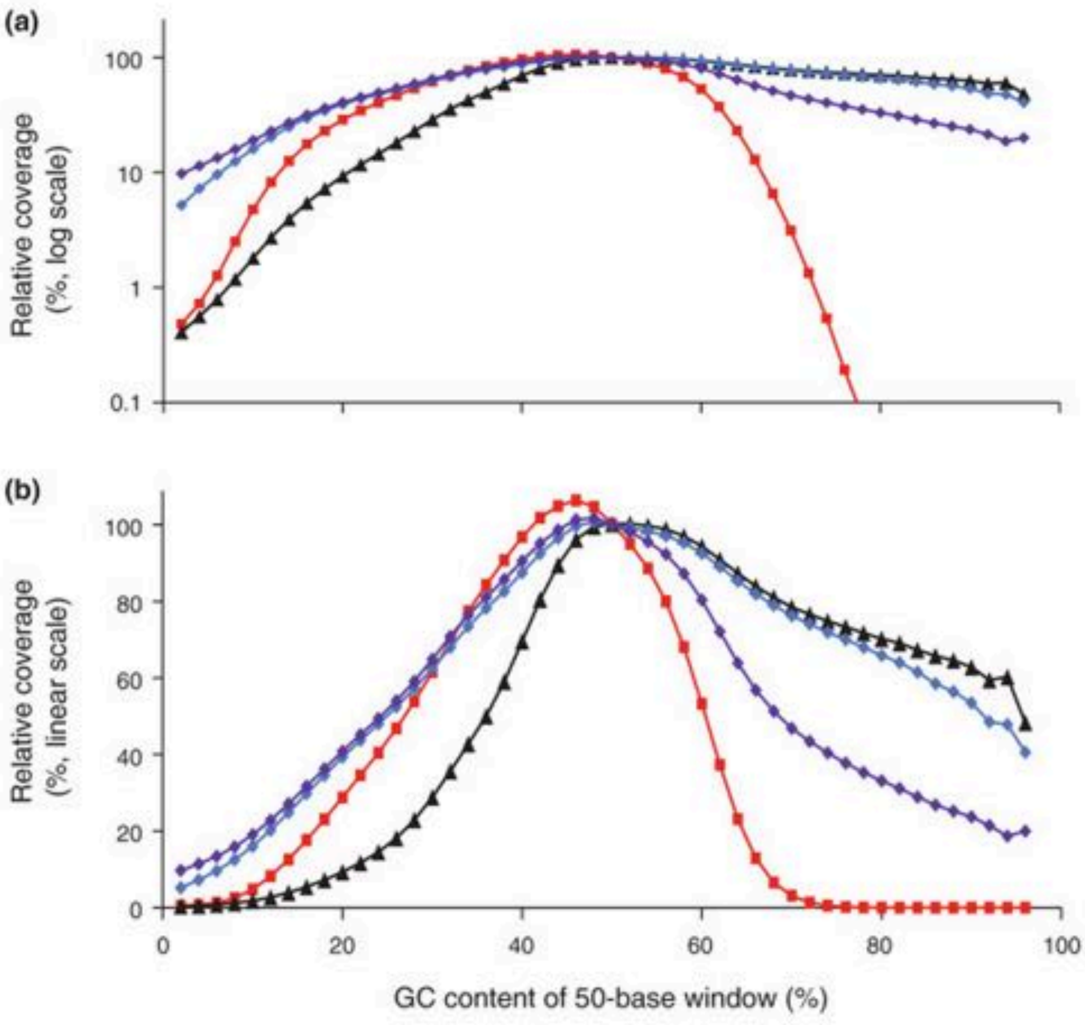
Beware of Duplicate Reads



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

Picard: http://picard.sourceforge.net

Beware of GC Biases



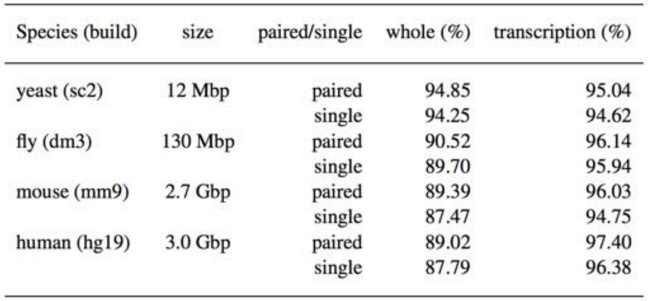
Illumina sequencing does not produce uniform coverage over the genome

- 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

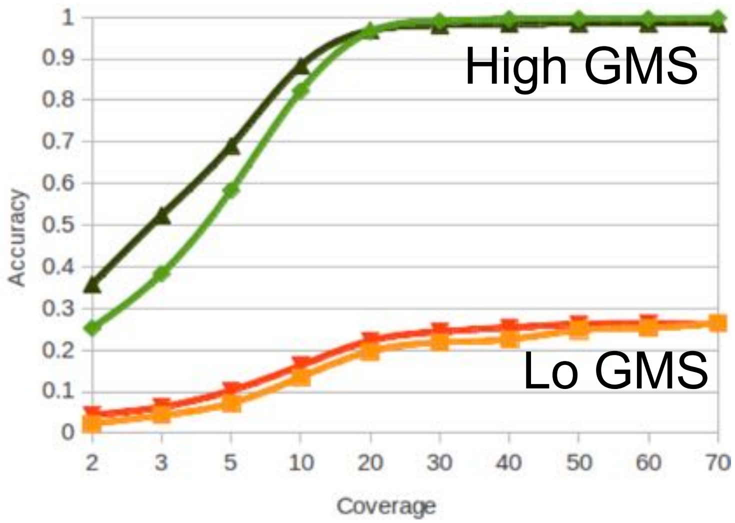
Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.
Aird et al. (2011) *Genome Biology.* 12:R18.

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

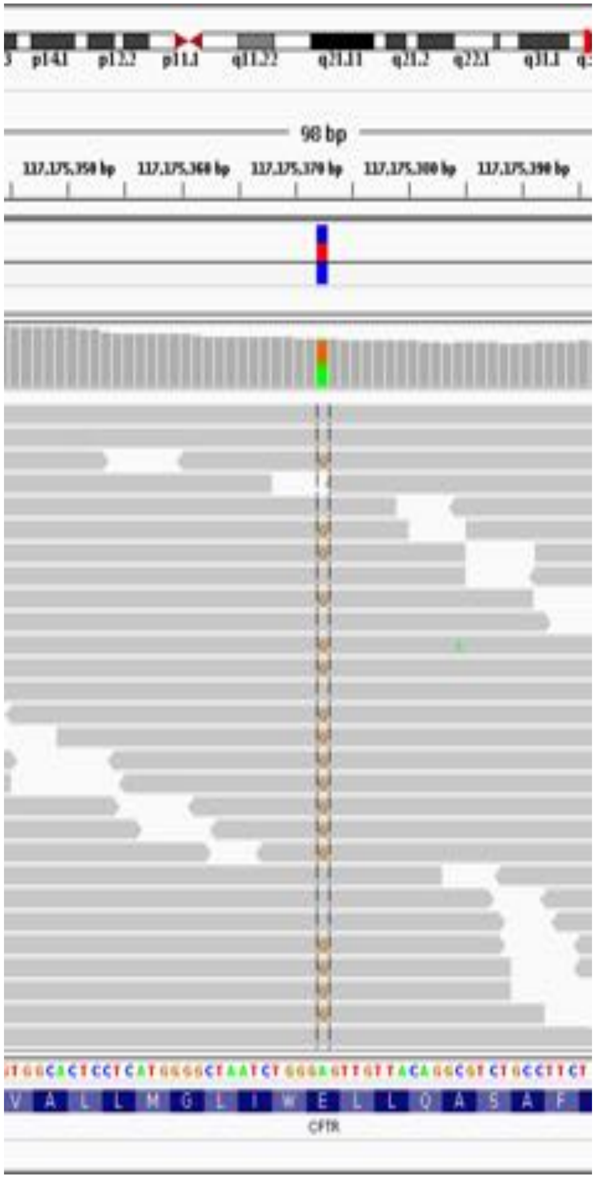


Species (build)	size	paired/single	whole (%)	transcription (%)
yeast (sc2)	12 Mbp	paired	94.85	95.04
		single	94.25	94.62
fly (dm3)	130 Mbp	paired	90.52	96.14
		single	89.70	95.94
mouse (mm9)	2.7 Gbp	paired	89.39	96.03
		single	87.47	94.75
human (hg19)	3.0 Gbp	paired	89.02	97.40
		single	87.79	96.38



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

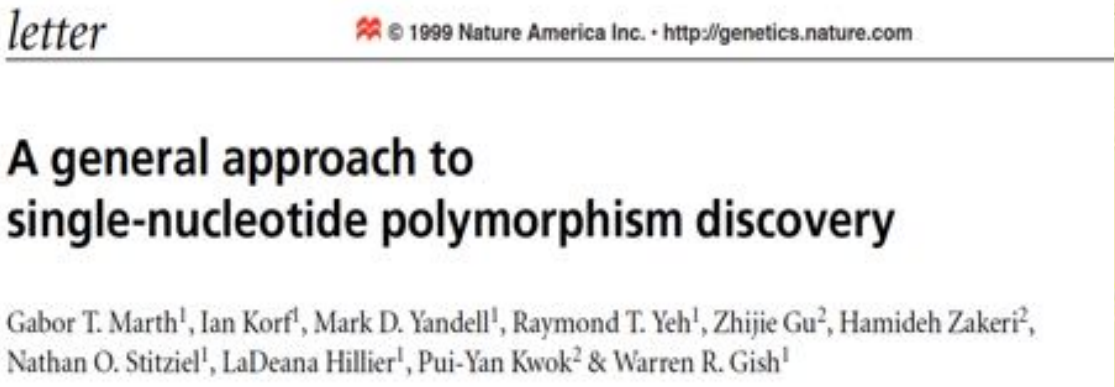
What information is needed to decide if a variant exists?



- 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

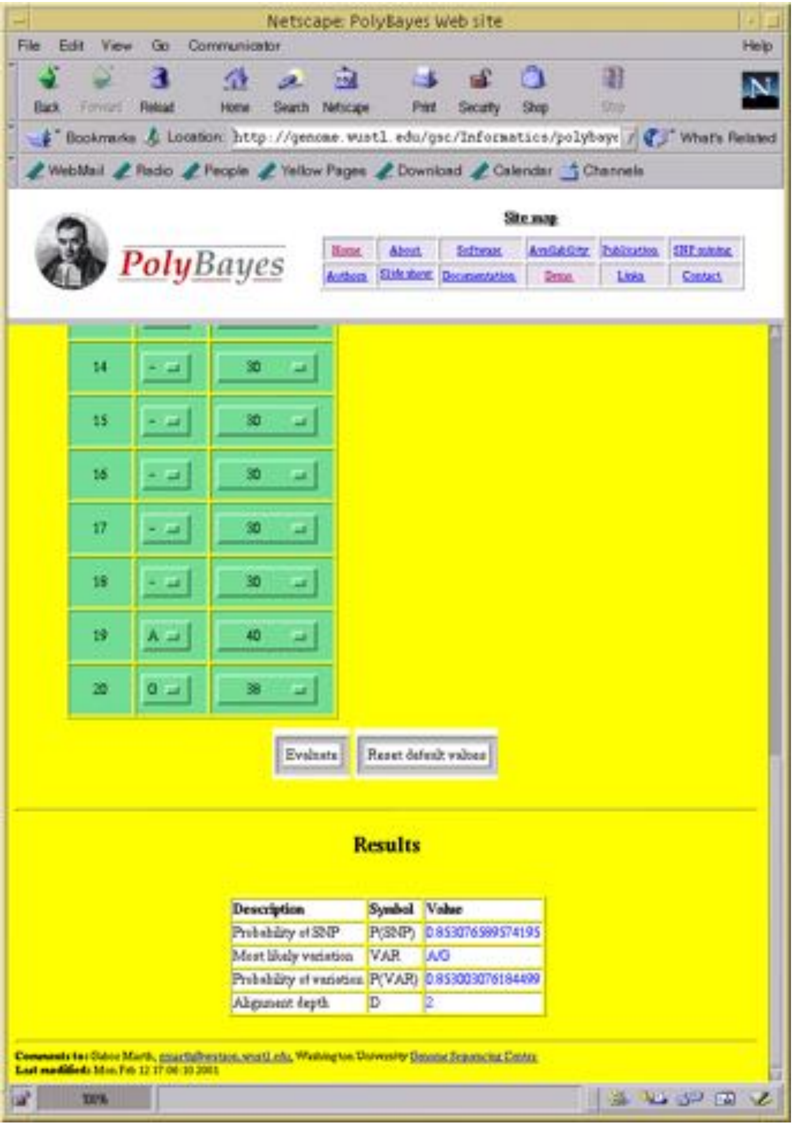
PolyBayes: The first statistically rigorous variant detection tool.



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]



Netscape: PolyBayes Web site
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://genome.wustl.edu/gsc/Informatics/polybaye
What's Related
WebMail Radio People Yellow Pages Download Calendar Channels
PolyBayes
Site map
Home About Software Availability Publication SNP mining
Authors Slide show Documentation Demo Links Contact
Evaluate
Reset default values
100%

14	-	30
15	-	30
16	-	30
17	-	30
18	-	30
19	A	40
20	G	38

Results

Description	Symbol	Value
Probability of SNP	P(SNP)	0.853076589574195
Most likely variation	VAR	A/G
Probability of variation	P(VAR)	0.853003076184499
Alignment depth	D	2

Comments to: Gabor Marth, gmarth@watson.wustl.edu, Washington University Genome Sequencing Center
Last modified: Mon Feb 12 17:06:10 2001

Its main innovation was the use of Bayes's theorem

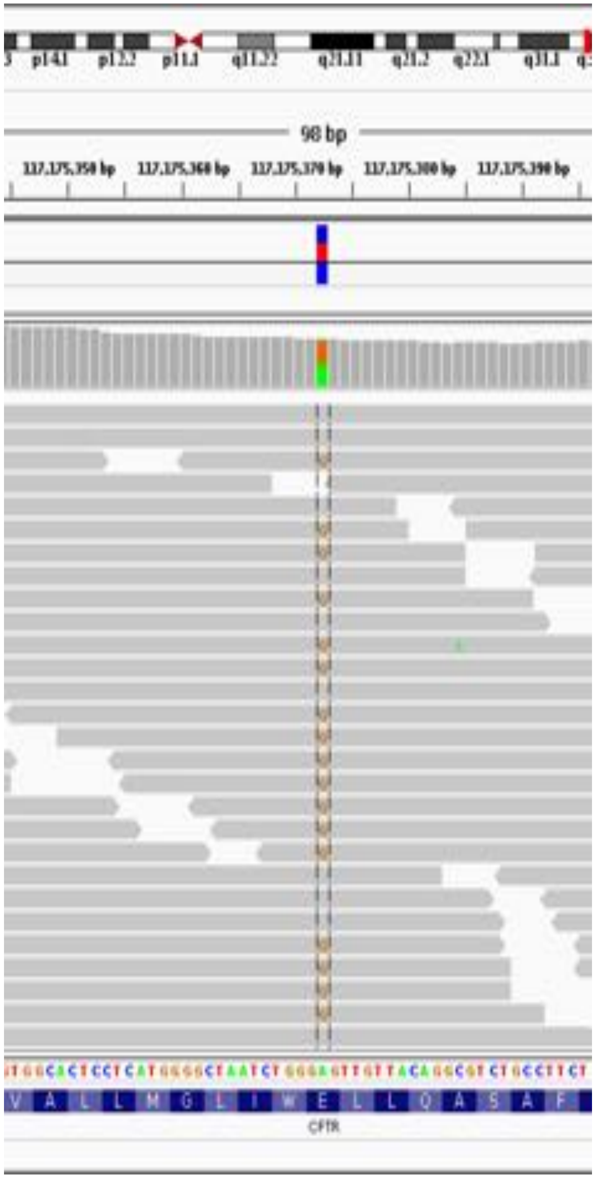
http://www.nature.com/ng/journal/v23/n4/full/ng1299_452.html

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

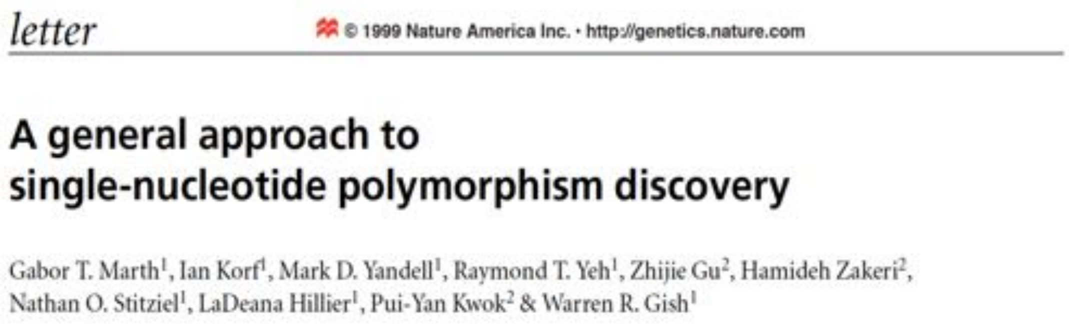


$$P(SNP|Data) = \frac{P(Data|SNP) * P(SNP)}{P(Data)}$$

- 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 © 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]

Bayesian posterior probability

Base call + Base quality

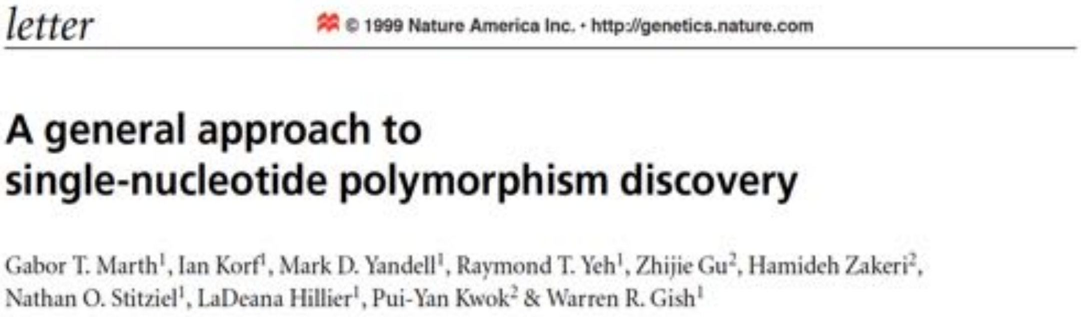
Expected (prior) polymorphism rate

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

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

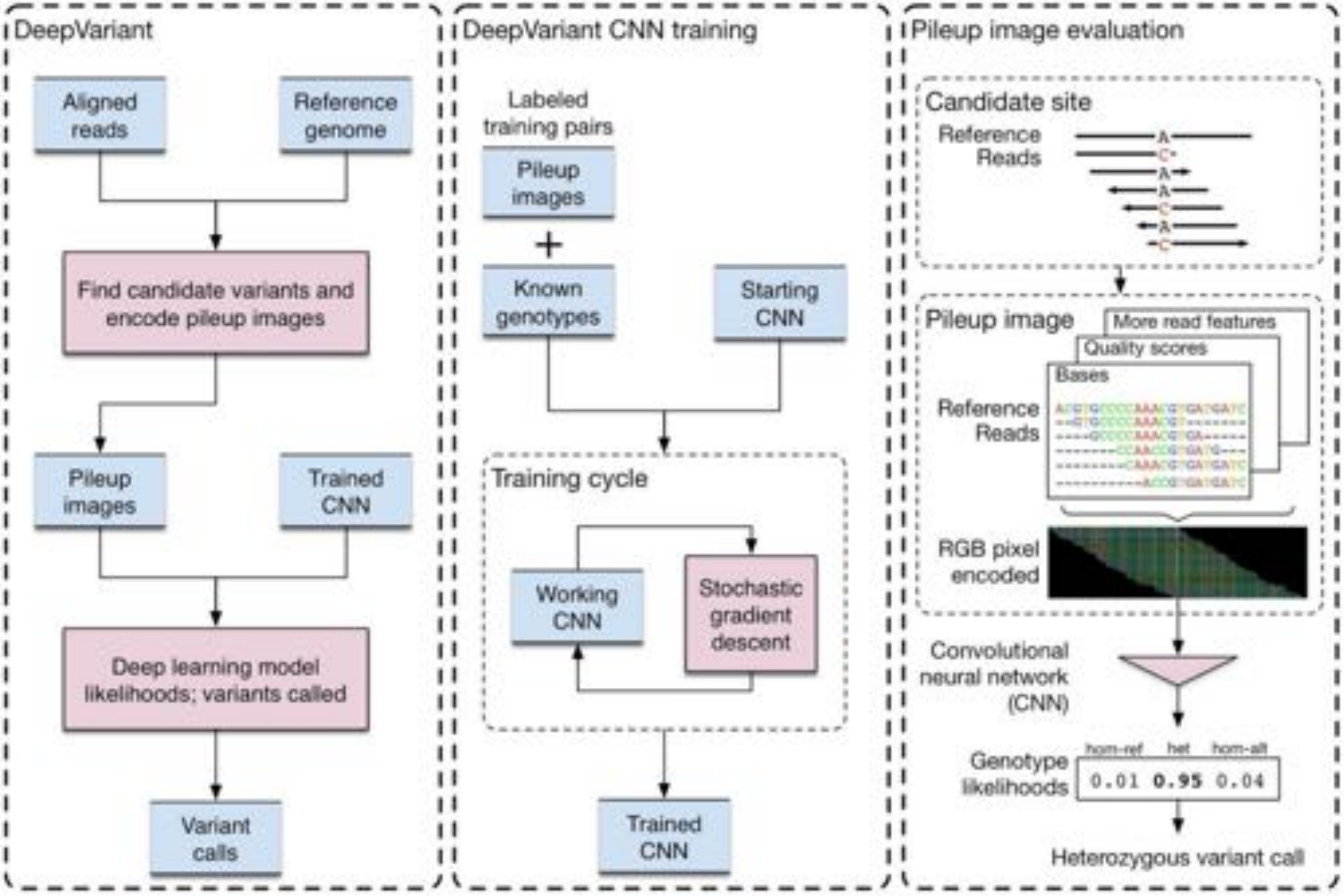
A general approach to single-nucleotide polymorphism discovery

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

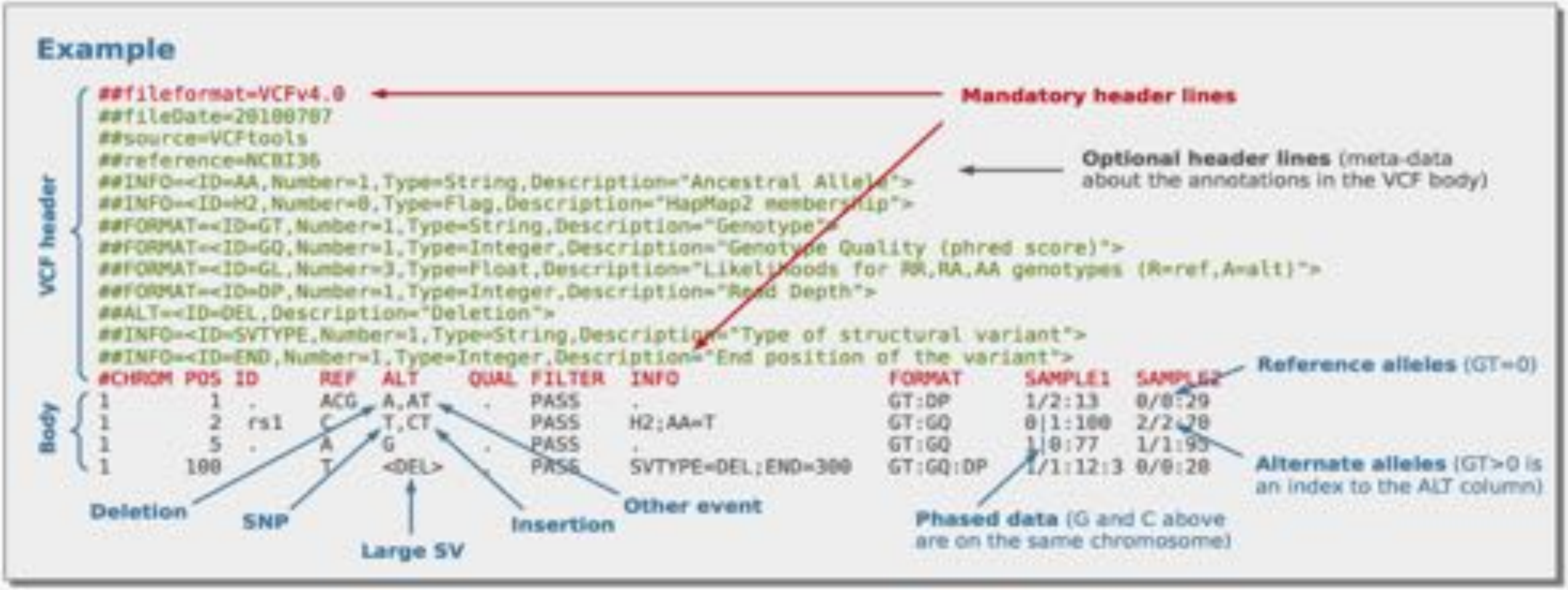
Deep Variant



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

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)

Phased data (G and C above are on the same chromosome)

Deletion

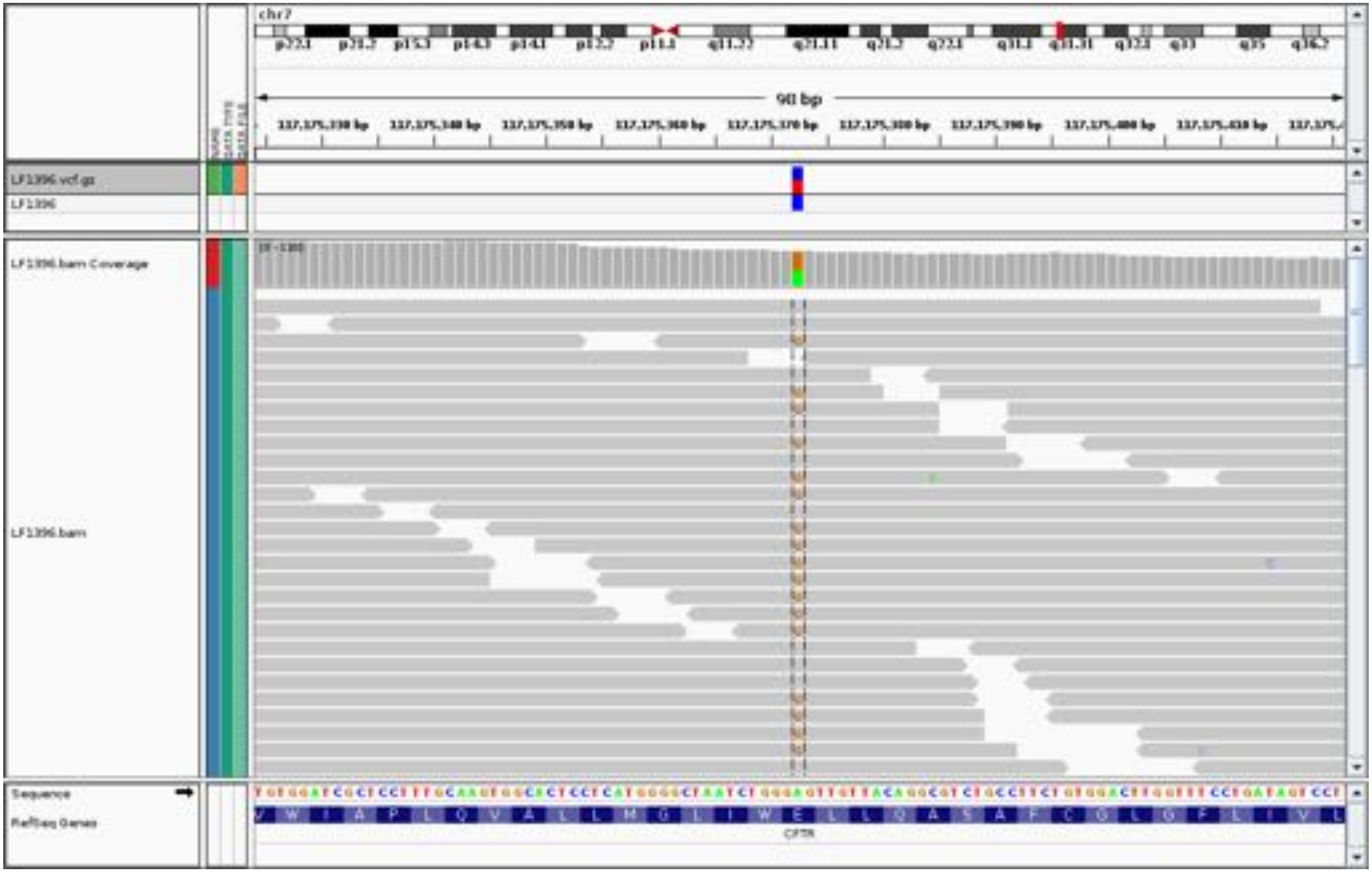
SNP

Large SV

Insertion

Other event

VCF Format



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