

Structural Variant Calling

Michael Schatz

Feb 22, 2018

Lecture 9: Applied Comparative Genomics



Assignment 3: Due Thursday Feb 22

Assignment 3: Genome Assembly, Phylogenetics, and the SWT

Assignment Date: Thursday, Feb. 16, 2017
Due Date: Thursday, Feb. 22, 2017 @ 11:59pm

Question 1: de Bruijn Graph construction [10 pts]

- Q1a. 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).

AFF
ATT
CAT
CTA
GAT
TAT
TGT
TCG
TCA
TAA
TTG
TCT
TTA

- Q1b. Assume that the maximum number of occurrences of any 3-mer in the actual genome is 3 using the k-mers from Q1a. Write one possible genome sequence.
- Q1c. What is the longest repeat?

Question 2: Phylogenetics Analysis [10 pts]

Your colleague is developing an experimental and computational protocol to determine the species present in food samples based on DNA sequencing. (See here for a technology working towards making this a reality!) She extracted DNA from a mixed-meal sausage of 100% lamb sequencing. When the data returns, she uses a short-read aligner such as Bowtie2 or BWA to align the sequencing reads. As the references, she chose several genomes of animals whose meat is commonly consumed, including chicken and pig and cow common genomes. Next, she extracts the unmapped reads and runs a short-read assembler such as Spades on those reads. She only gets a few contigs that are longer than a few hundred base pairs.

1. Suggest two reasons there are only a few, short contigs assembled from non-mapping reads. (2)

She asks for your help in finding the origin of these "mystery meat" contigs. Fortunately you are familiar with genome databases and offer to help her out. You use query the NCBI's database of reference genome assemblies with the longest contig using the BLAST+ k-mer alignments between your sequence and a database. One contig you examine has several high E-value alignments to scaffolds in the *Mus musculus* genome assembly. Two of the alignments are in annotated gene regions. However, the *Mus musculus* genome assembly (see

2. Based on the link above, give two indicators that this genome assembly is poor quality. (2)

Because the assembly is rough, you are suspicious that the contig has more than one alignment, it overlaps more than one annotated gene. Could there be a duplicated region or misassembly in the reference genome? Or does the tammar wallaby actually have genes that align to both?

Homologous genes are genes with a shared evolutionary history. Homologous genes in the same genome arise from a gene duplication event long ago in evolution. Homologous genes in the same genome are called paralogs. Paralogs usually have detectable sequence differences (ENSGALU000000000886 and ENSMUSLU000000000891). The annotated genes within (or of) this contig's alignments are paralogs. You decide to build a phylogenetic tree of these genes, as well as some sequences from other species to see whether these genes are paralogs.

Here are some protein sequences of some hits from a blastx search including the two sequences from *M. musculus*: [Mus musculus](#)NP_001162.2. Some proteins are annotated "hemoglobin epsilon" and others are annotated "hemoglobin beta". *ab* and *bb* in the sequence names in the file

3. Use the web version of MUSCLE to create a multiple sequence alignment. The tool outputs a neighbor-joining, binary phylogenetic tree. Because MUSCLE's built-in tree graphics is very poor, download the data in Newick format, and open the file in visualization software such as Treeo. Include an image of the tree in your report. Feel free to explore a variety of visualization options, but just make sure the leaf labels are readable and the branches have appropriate length.

4. What do the leaves of the tree represent? Is the tree rooted or unrooted? (2)

5. Propose a location for the root of the tree, and justify your answer. (Mark it on the image of the tree) (2)

6. Do you think the "ab" and "bb" genes are paralogs? Justify your answer by referring to the tree. (2)

Here is the output from MrBayes, a Bayesian MCMC tree algorithm, run on the same protein sequences.

Assignment 4: Due Thursday March 1

Assignment 4: Read mapping and variant calling

Assignment Date: Thursday, Feb. 22, 2018

Due Date: Thursday, Mar. 1, 2018 @ 11:59pm

Assignment Overview

In this assignment, you will align reads to a reference genome to call SNPs and short indels. Then, you will perform an experiment to empirically determine the “mappability” of a genomic region. Finally, you will investigate some empirical behavior of the binomial test for heterozygous variant calling.

As a reminder, any questions about the assignment should be posted to [Piazza](#). Don't forget to read the Resources section at the bottom of the page!

Question 1. Small Variant Analysis [XX pts]

Download chromosome 22 from build 38 of the human genome from here:

<http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/chr22.fa.gz>

Download the read set from here:

<http://schatzlab.cshl.edu/data/teaching/sample.tgz>

For this question, you may find this tutorial helpful:

<http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html>

- * 1a. How many reads align to the reference? How many reads did not align? How many aligned reads had a mate that did not align (AKA singletons)? Count each read in a pair separately.
[Hint: Build the index using `bowtie2-build`, align reads using `bowtie2`, analyze with `samtools flagstat`.]
- * 1b. How many reads are mapped to the reverse strand? Count each read in a pair separately.
[Hint: Find out what SAM flags mean [here](#) and use `samtools view`.]
- * 1c. How many high-quality (QUAL > 20) single nucleotide and indel variants does the sample have? Of the high-quality SNPs, what is the transition / transversion ratio? Of the indels, how many are insertions and how many are deletions?
[Hint: Identify variants using `freebayes -` sort the SAM file first. Filter using `bcftools filter`, and summarize using `bcftools stats`.]
- * 1d. Does the sample have any nonsense or missense mutations?
[Hint: try the [Variant Effect Predictor](#) using the [Genode basic transcripts](#)]

Question 2. Read Mapping Uncertainty [XX pts]

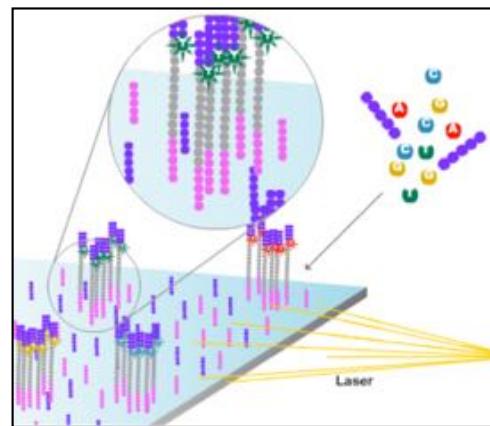
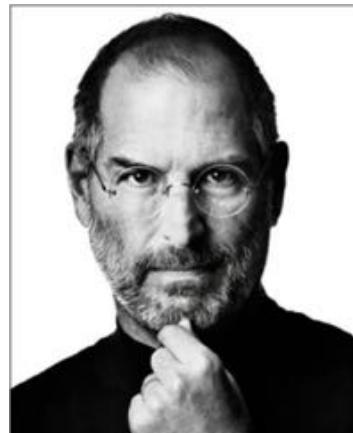
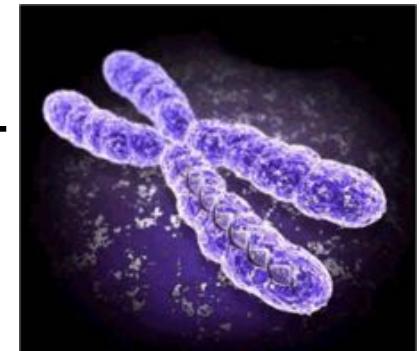
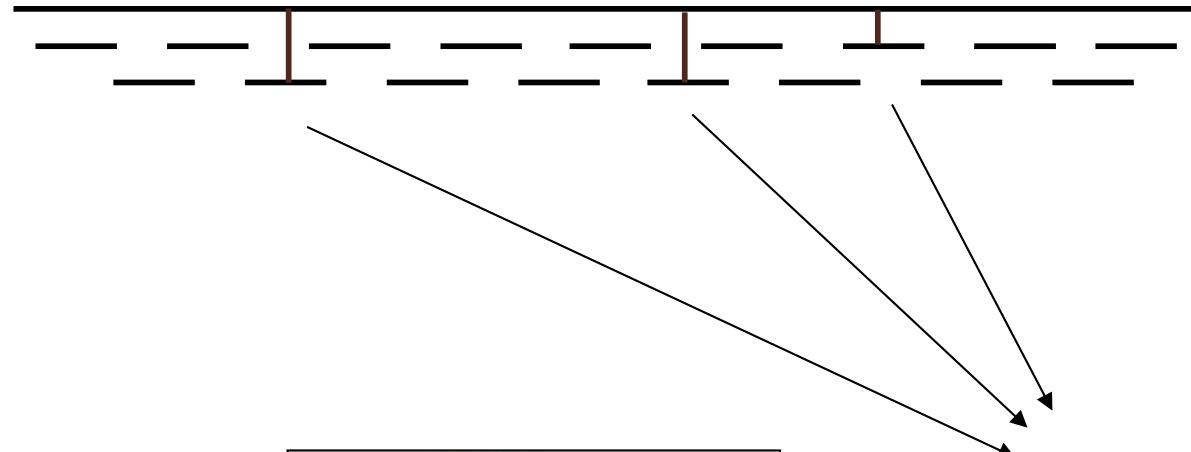
For the region chr22:21000000-22000000 of the reference sequence for chromosome 22, extract every substring of length 35. Format the substrings as a FASTA file and use read names that indicate the origin. (No need to construct quality values or read pairs: use `bowtie2` with `-f` and `-d` respectively). Make a new index and align these “reads” to chr22:21000000-22000000.

[Hint: On the command line or in a script, load the sequence once and extract substrings in a loop.]

- * 2a. How many reads align more than one time to the reference? How many reads did not align?

Personal Genomics

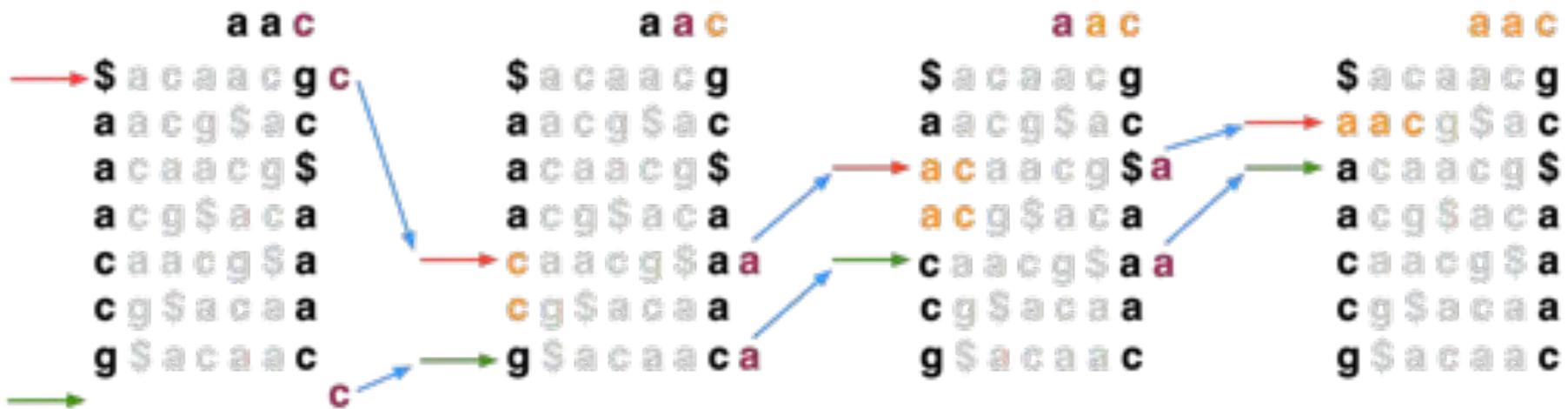
How does your genome compare to the reference?



Heart Disease
Cancer
Creates magical
technology

BWT Exact Matching

- Start with a range, (**top**, **bot**) encompassing all rows and repeatedly apply **LFc**:
top = **LFc**(**top**, **qc**); **bot** = **LFc**(**bot**, **qc**)
qc = the next character to the left in the query



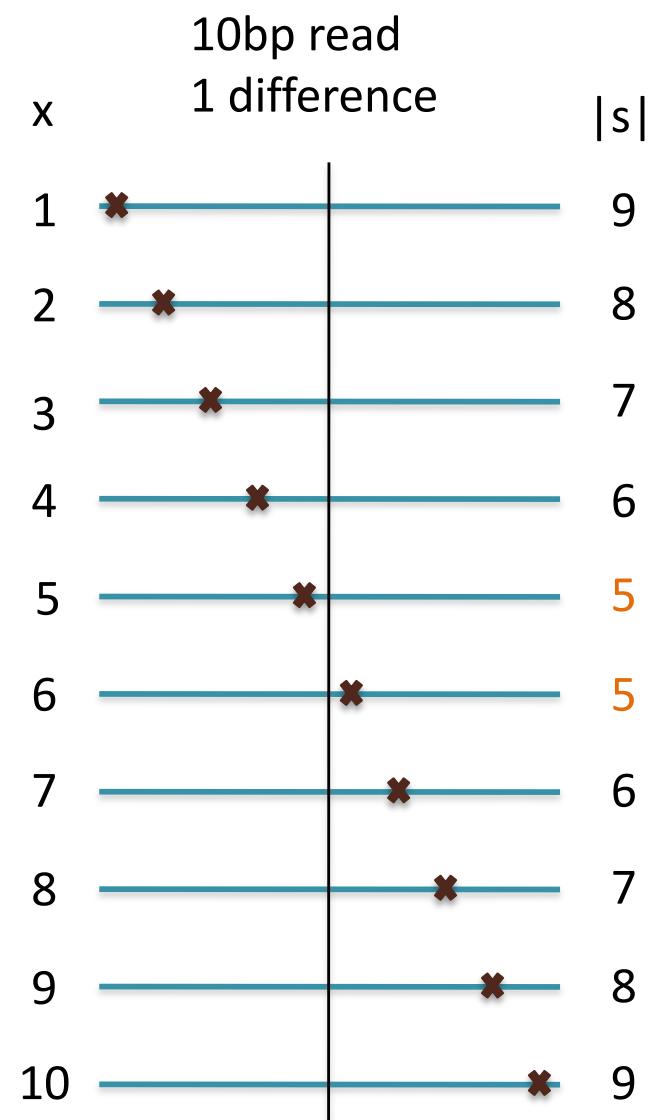
Ferragina P, Manzini G: Opportunistic data structures with applications. FOCS. IEEE Computer Society; 2000.

[Search for TTA this BWT string: ACTGA\$TTA]

Seed-and-Extend Alignment

Theorem: An alignment of a sequence of length m with at most k differences **must** contain an exact match at least $s=m/(k+1)$ bp long
(Baeza-Yates and Perleberg, 1996)

- Proof: Pigeonhole principle
 - 1 pigeon can't fill 2 holes
- Seed-and-extend search
 - Use an index to rapidly find short exact alignments to seed longer in-exact alignments
 - BLAST, MUMmer, Bowtie, BWA, SOAP, ...
 - Specificity of the depends on seed length
 - Guaranteed sensitivity for k differences
 - Also finds some (but not all) lower quality alignments <- heuristic



Algorithm Overview

1. Split read into segments

Read
CCAGTAGCTCTCAGCCTTATTTACCCAGGCCTGTA Read (reverse complement)
TACAGGCCTGGGTAAAATAAGGCTGAGAGCTACTGG

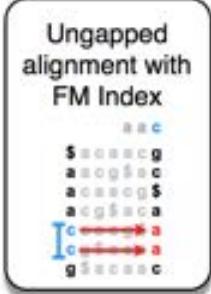
Policy: extract 16 nt seed every 10 nt

Seeds

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

2. Lookup each segment and prioritize

Seeds

+ , 0: CCAGTAGCTCTCAGCC	→	Ungapped alignment with FM Index	→	Seed alignments (as B ranges)
+ , 10: TCAGCCTTATTTACC				{ [211, 212], [212, 214] }
+ , 20: TTTACCCAGGCCTGTA				{ [653, 654], [651, 653] }
- , 0: TACAGGCCTGGGTAAA				{ [684, 685] }
- , 10: GGTAAAATAAGGCTGA				{ }
- , 20: GGCTGAGAGCTACTGG				{ }
				{ [624, 625] }

3. Evaluate end-to-end match

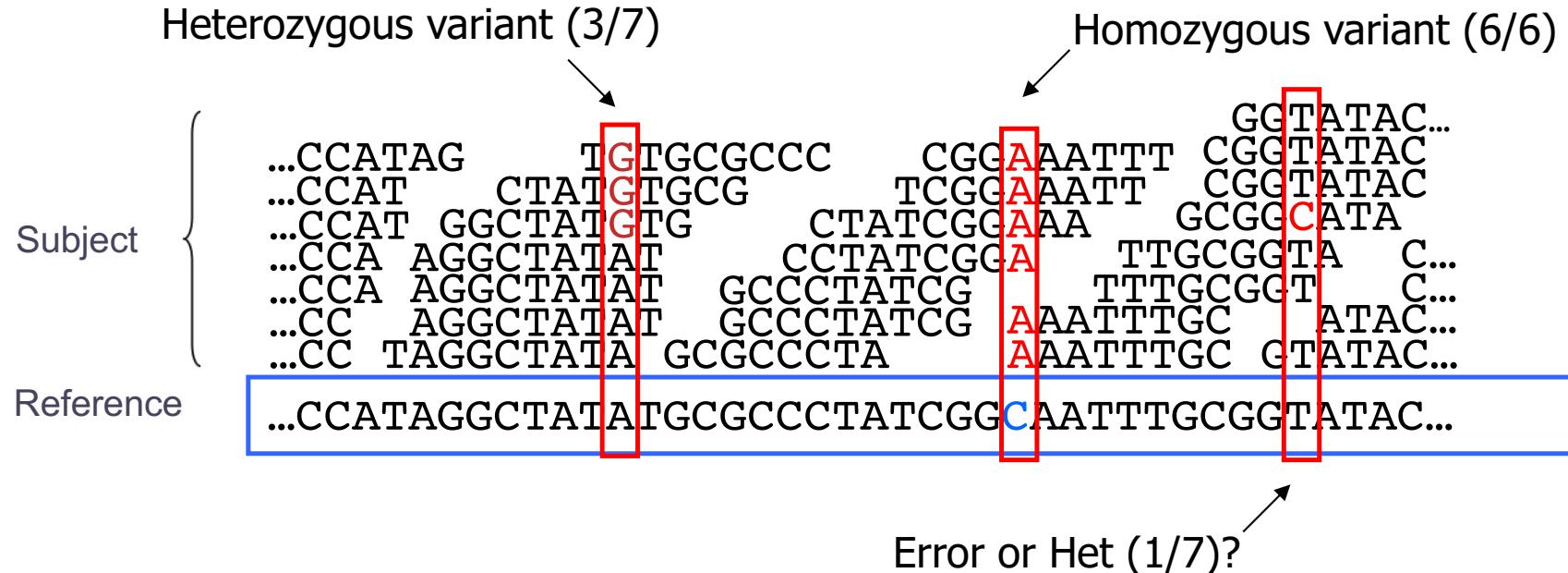
Extension candidates

SA:684, chr12:1955	→	SIMD dynamic programming aligner	→	SAM alignments
SA:624, chr2:462				r1 0 chr12 1936 0
SA:211: chr4:762				36M * 0 0
SA:213: chr12:1935				CCAGTAGCTCTCAGCCTTATTTACCCAGGCCTGTA
SA:652: chr12:1945				II

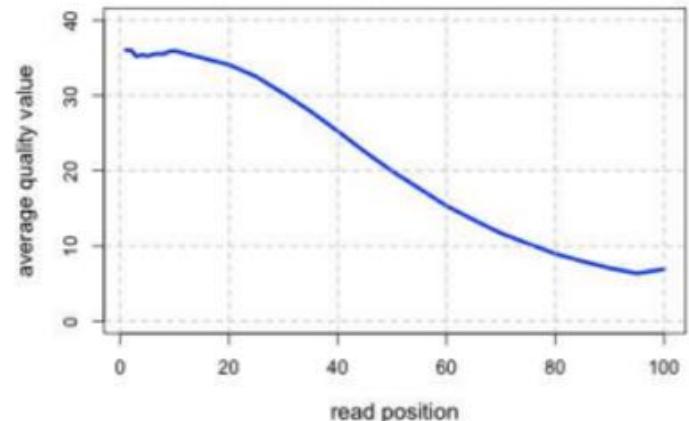
...

(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

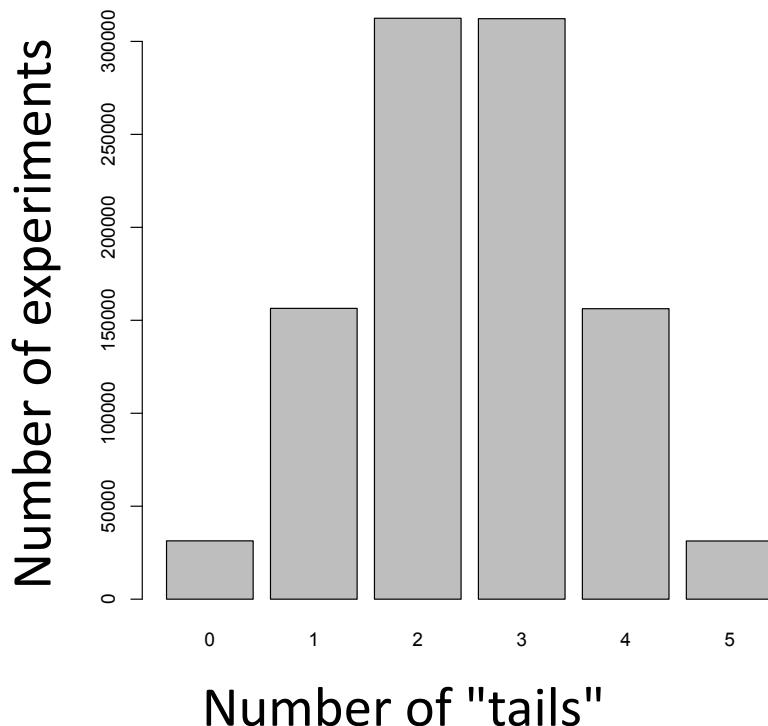


$P(\text{heads}) = 0.5$



$P(\text{tails}) = 0.5$

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



R code:

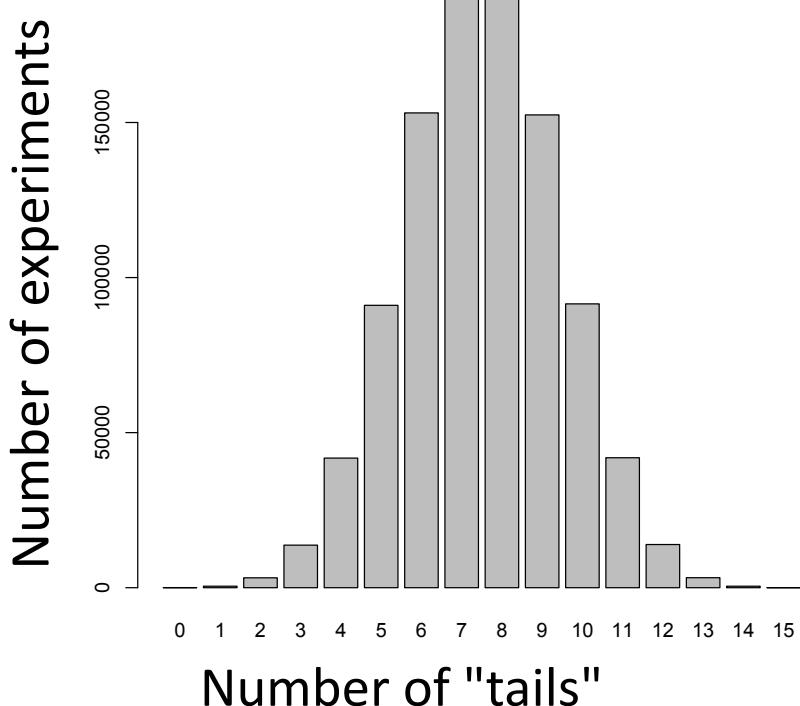
```
barplot(table(rbinom(1e6, 5, 0.5)))
```

1M 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:

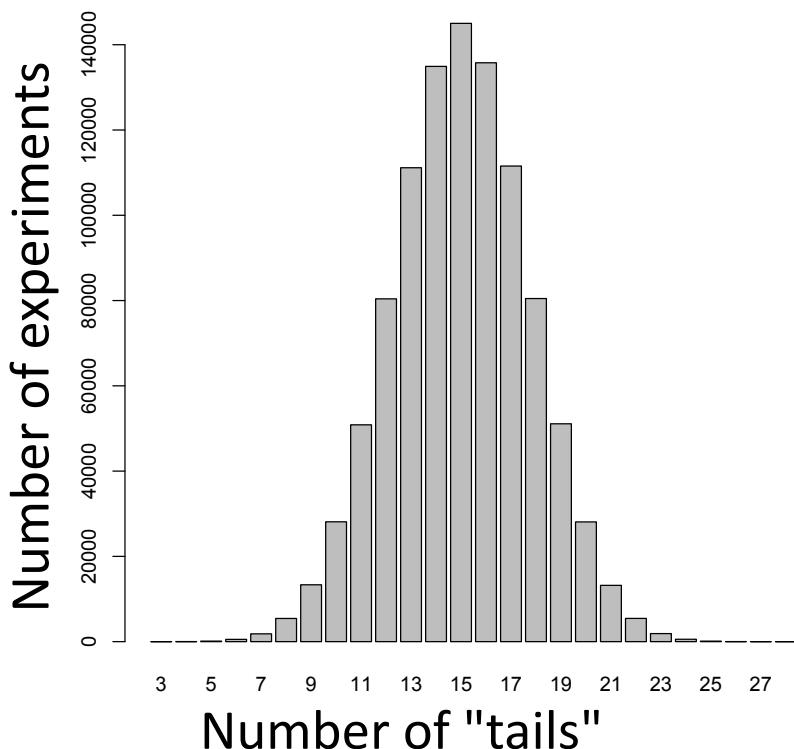
```
barplot(table(rbinom(1e6, 15, 0.5)))
```

1M 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:

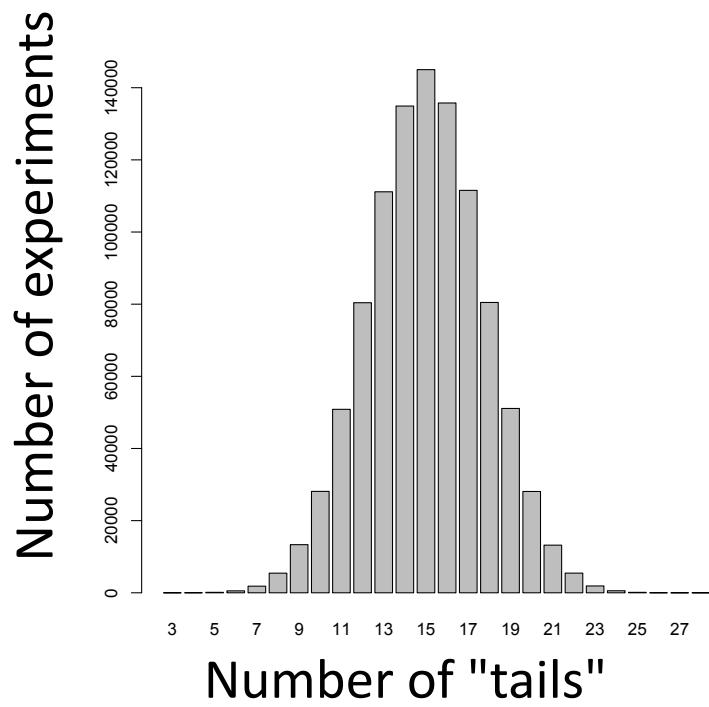
```
barplot(table(rbinom(1e6, 30, 0.5)))
```

1M 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



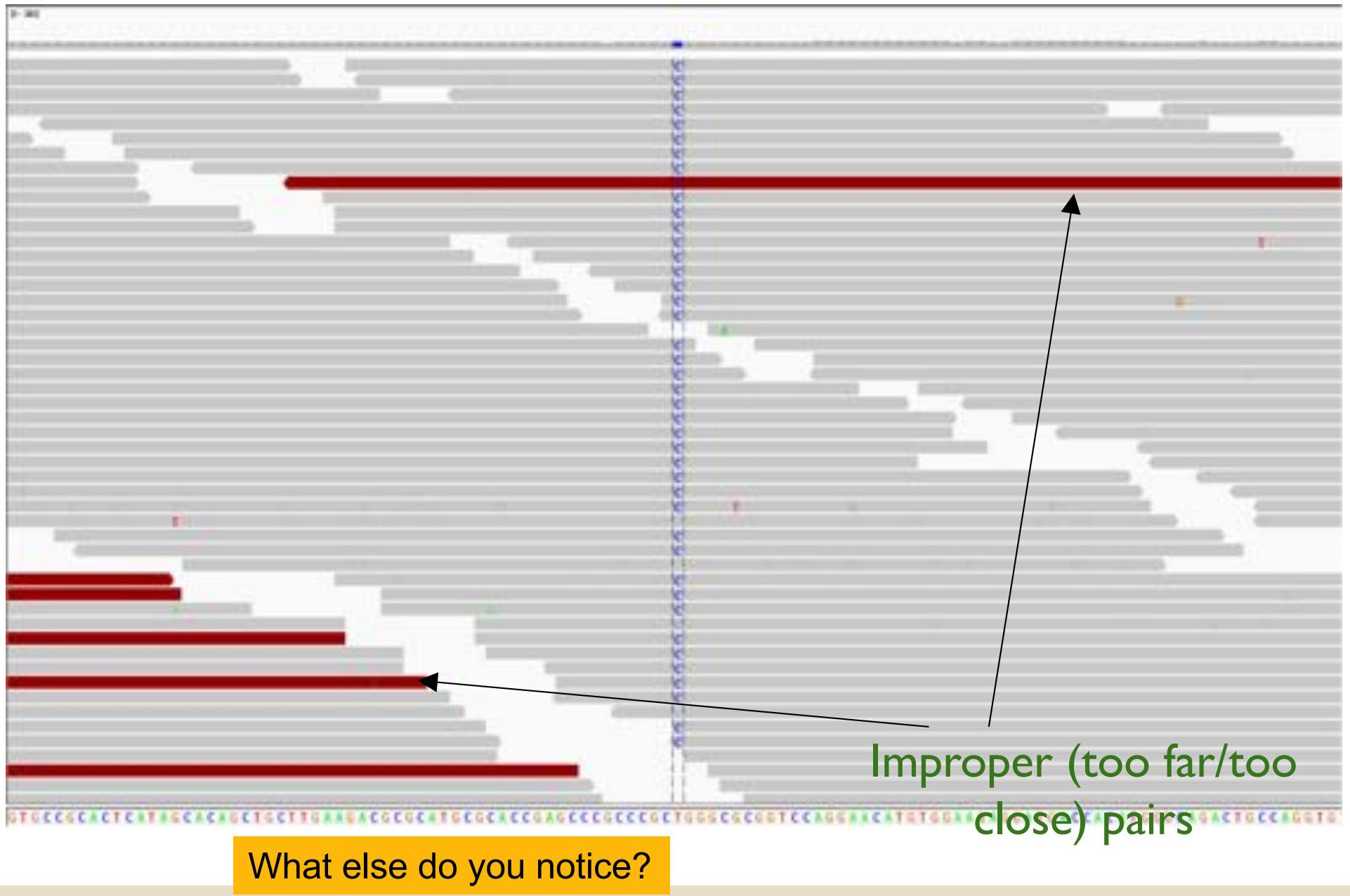
This is why at least a "30X" (30 fold sequence coverage) genome is recommended: it confers sufficient power to distinguish heterozygous alleles and from mere sequencing errors

$P(3/30 \text{ het}) <?> P(3/30 \text{ err})$

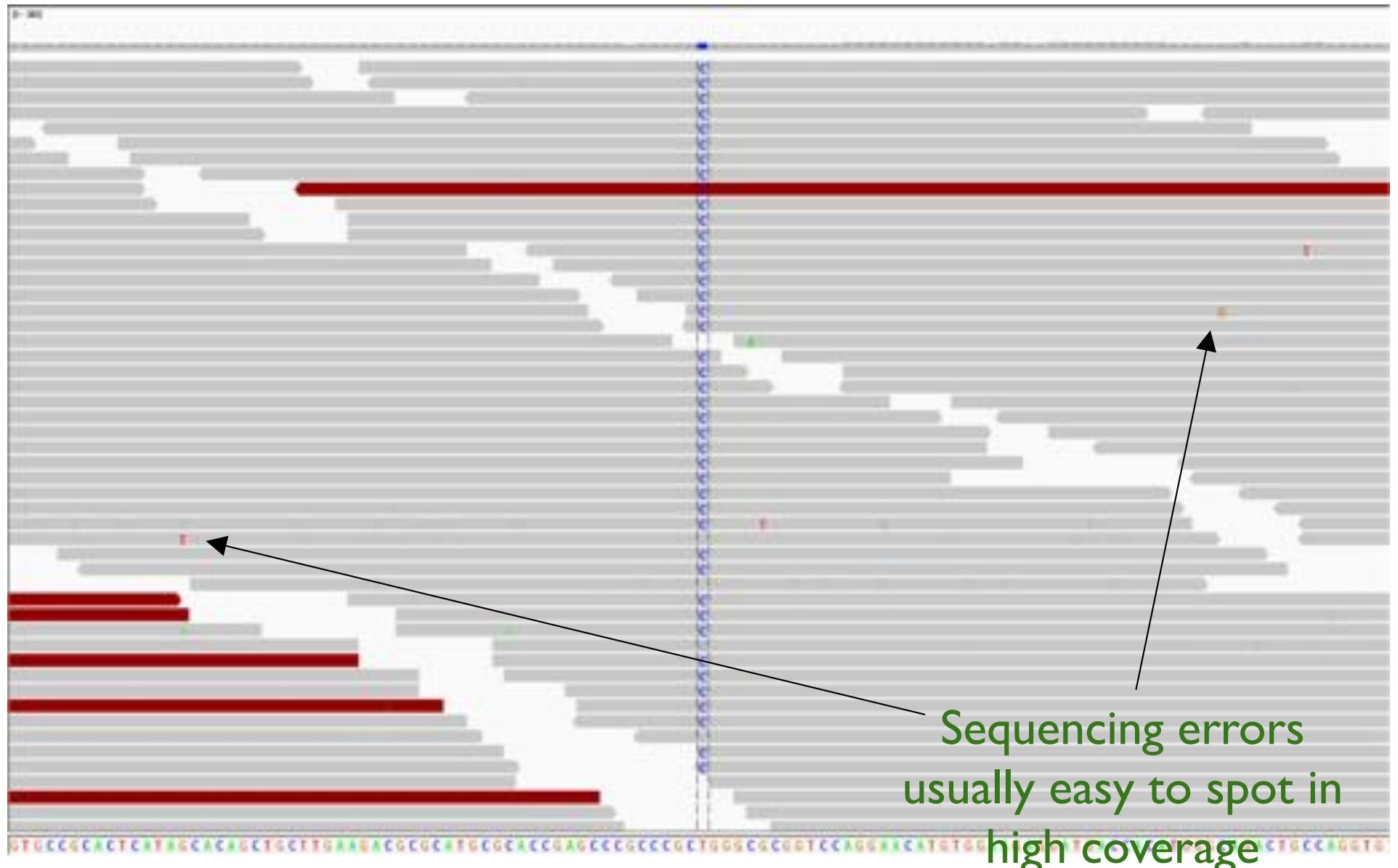
Some real examples of SNPs in IGV



Homozygous for the "C" allele

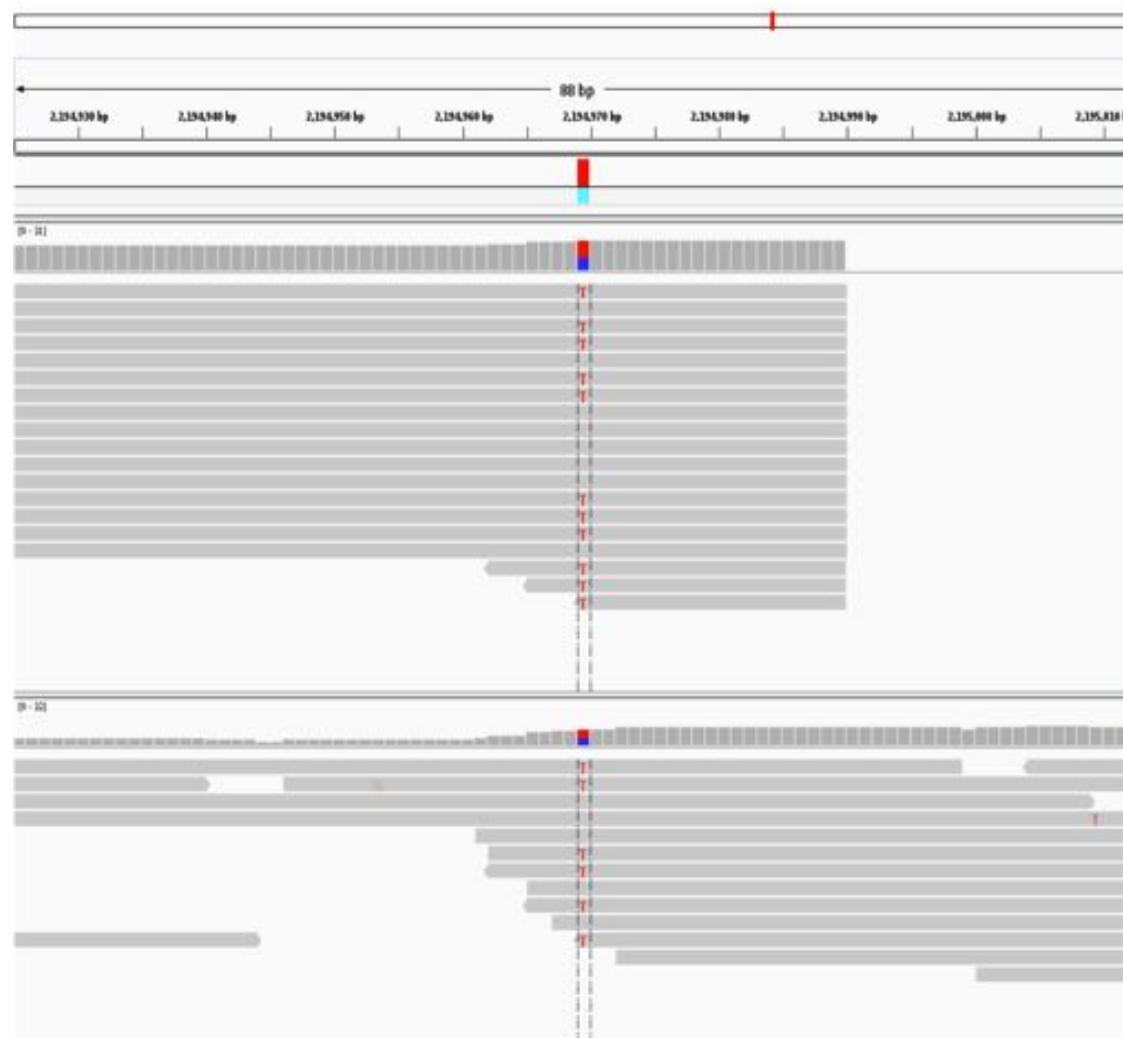


Homozygous for the "C" allele

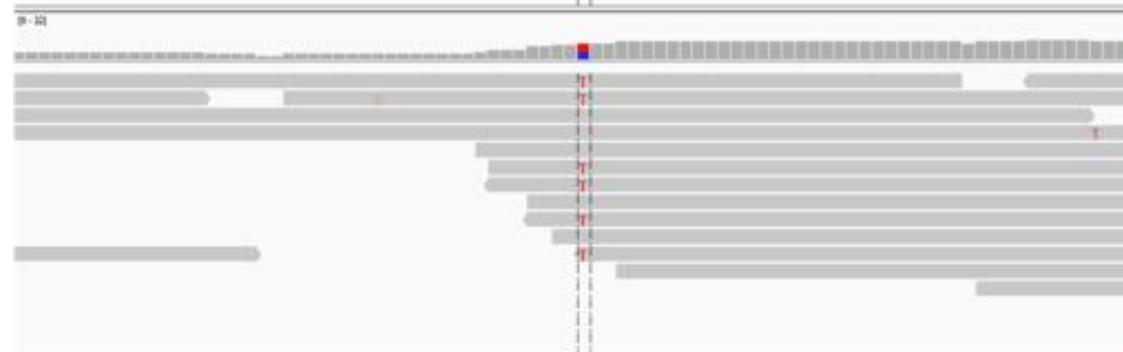


Heterozygous for the alternate allele

Individual
1



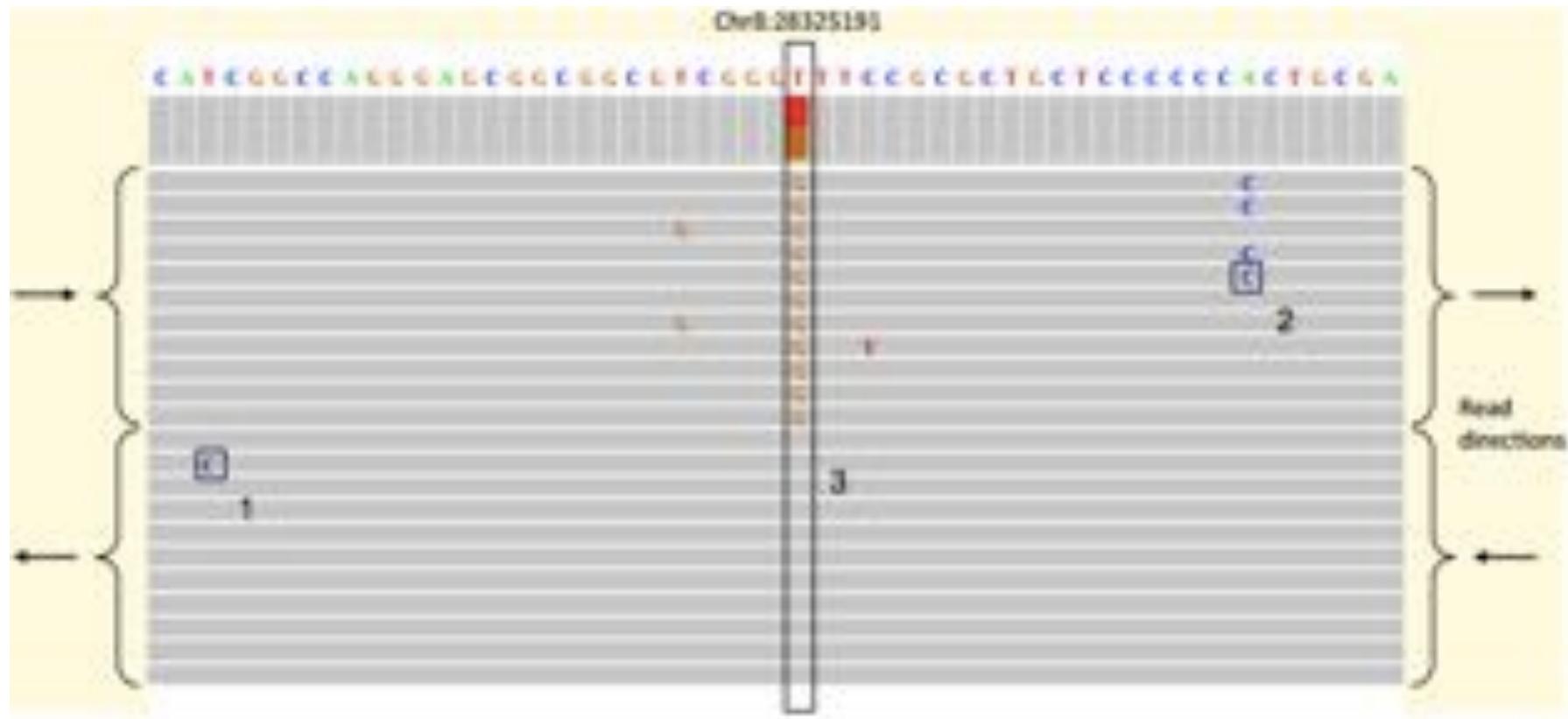
Individual
2



Which genotype prediction do you have more confidence in?

It is not always so easy ☹

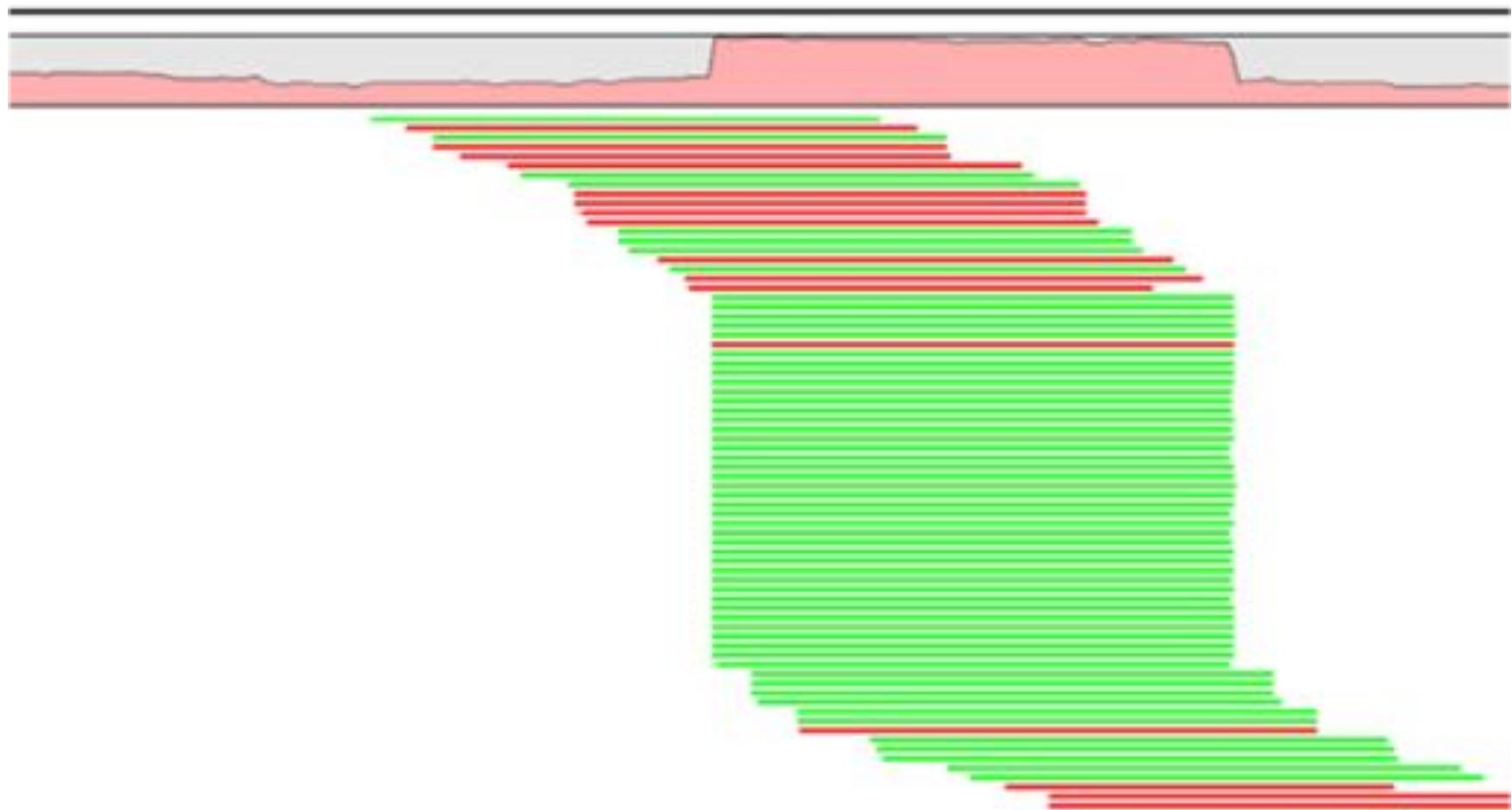
Beware of Systematic Errors



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

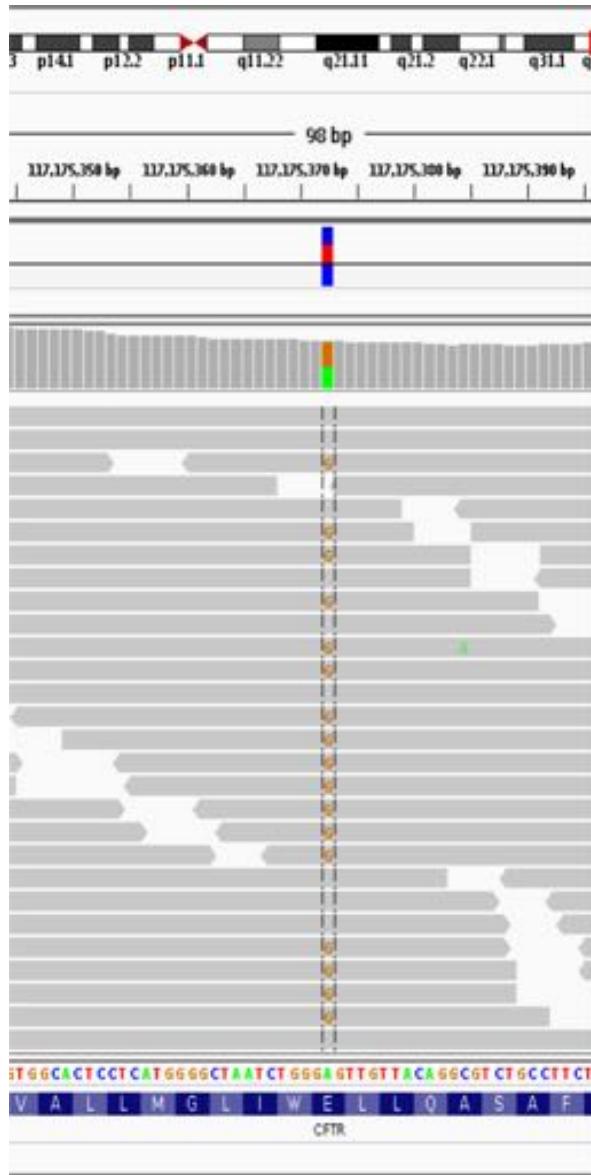
Beware of Duplicate Reads



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

Picard: <http://picard.sourceforge.net>

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

PolyBayes: The first statistically rigorous variant detection tool.

letter

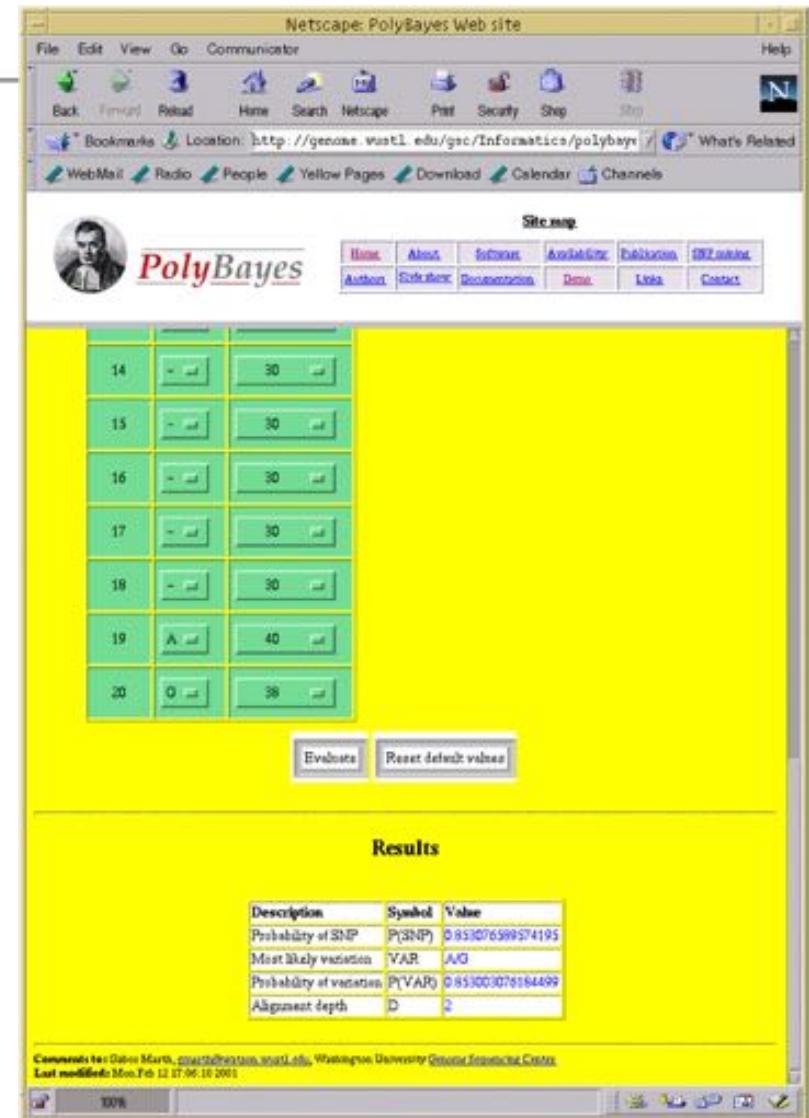


© 1999 Nature America Inc. • <http://genetics.nature.com>

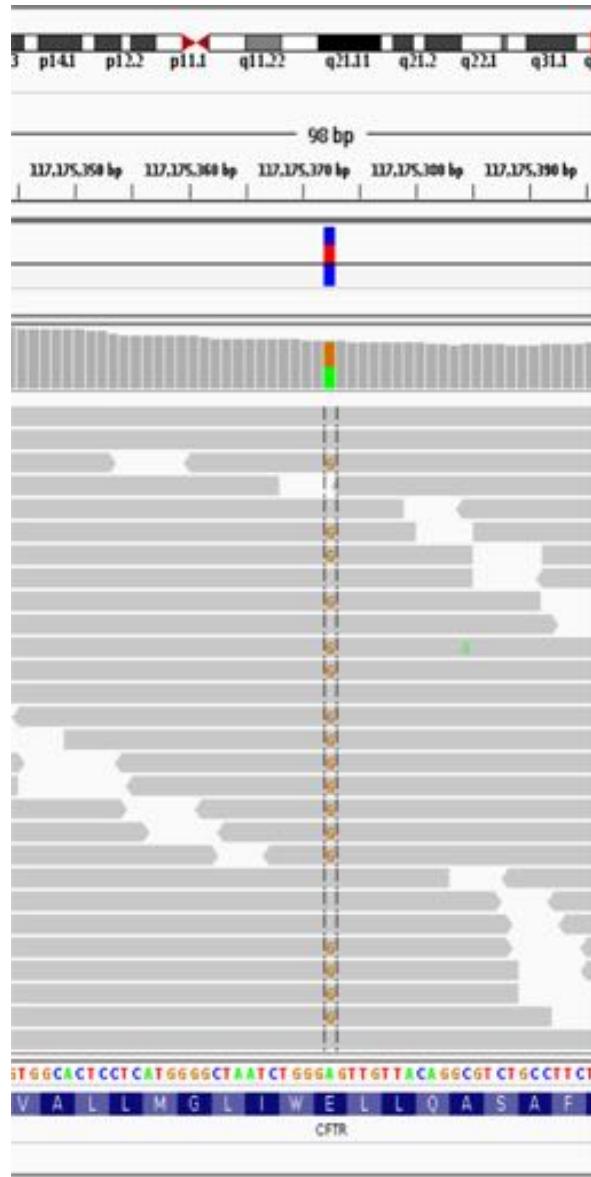
A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth¹, Ian Korf¹, Mark D. Yandell¹, Raymond T. Yeh¹, Zhijie Gu², Hamideh Zakeri², Nathan O. Stitzel¹, LaDeana Hillier¹, Pui-Yan Kwok² & Warren R. Gish¹

Its main innovation was the use of Bayes's theorem

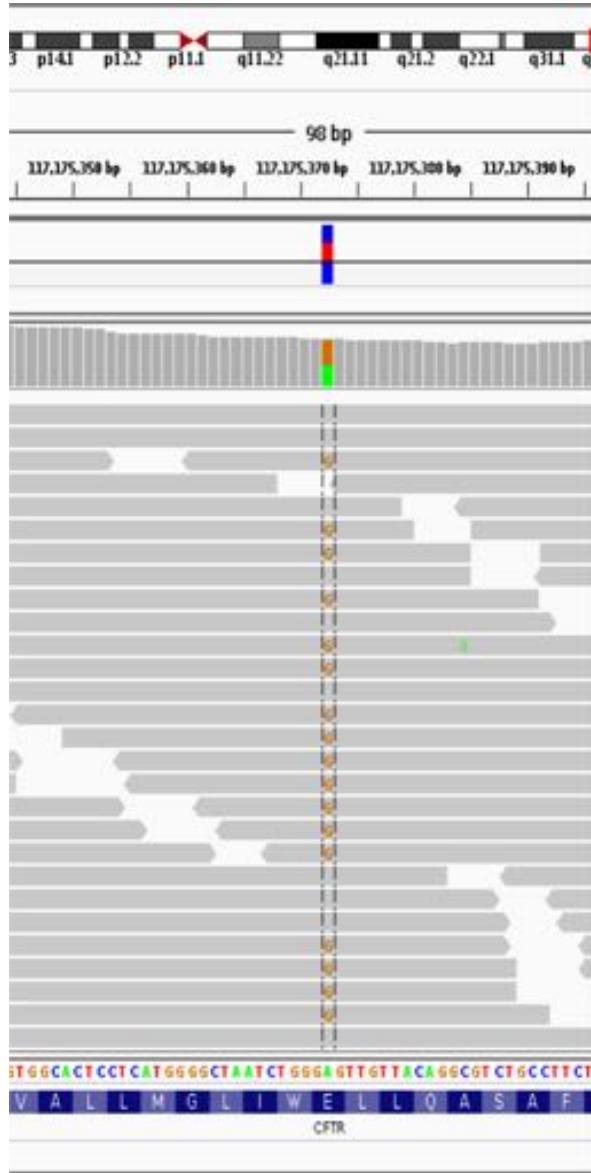


Bayesian SNP calling



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

Bayesian SNP calling



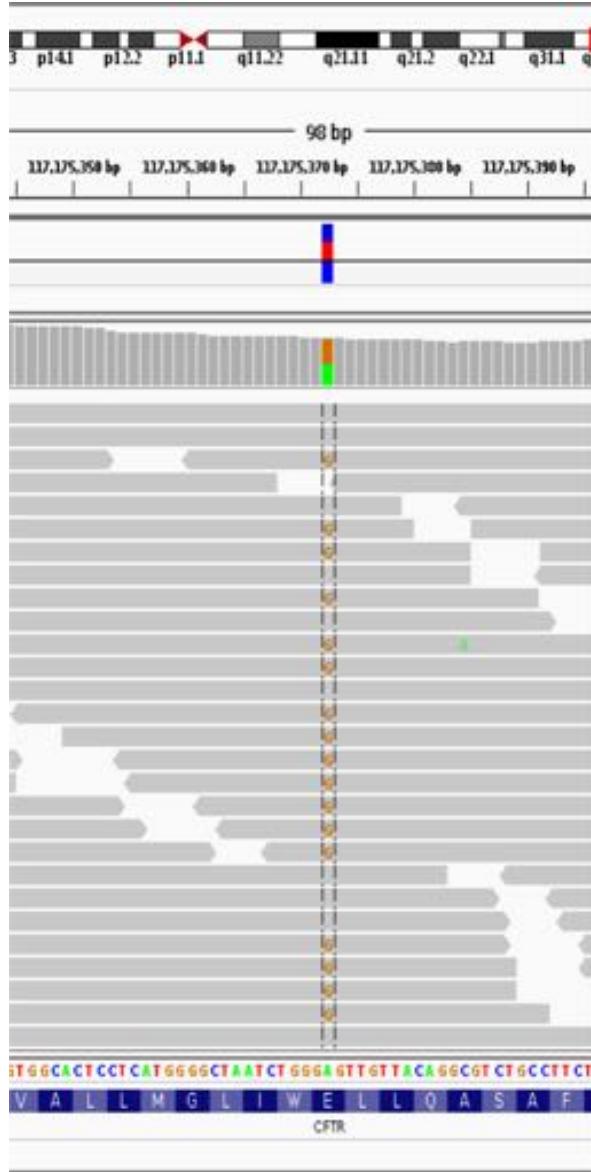
Hard to compute

Much easier

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

See bonus slides for more info

Bayesian SNP calling



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

- 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

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. Marth¹, Ian Korf¹, Mark D. Yandell¹, Raymond T. Yeh¹, Zhijie Gu², Hamideh Zakeri², Nathan O. Stitzel¹, LaDeana Hillier¹, Pui-Yan Kwok² & Warren R. Gish¹

Bayesian posterior probability

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

Base call + Base quality
Expected (prior) polymorphism rate

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

PolyBayes: The first statistically rigorous variant detection tool.

letter

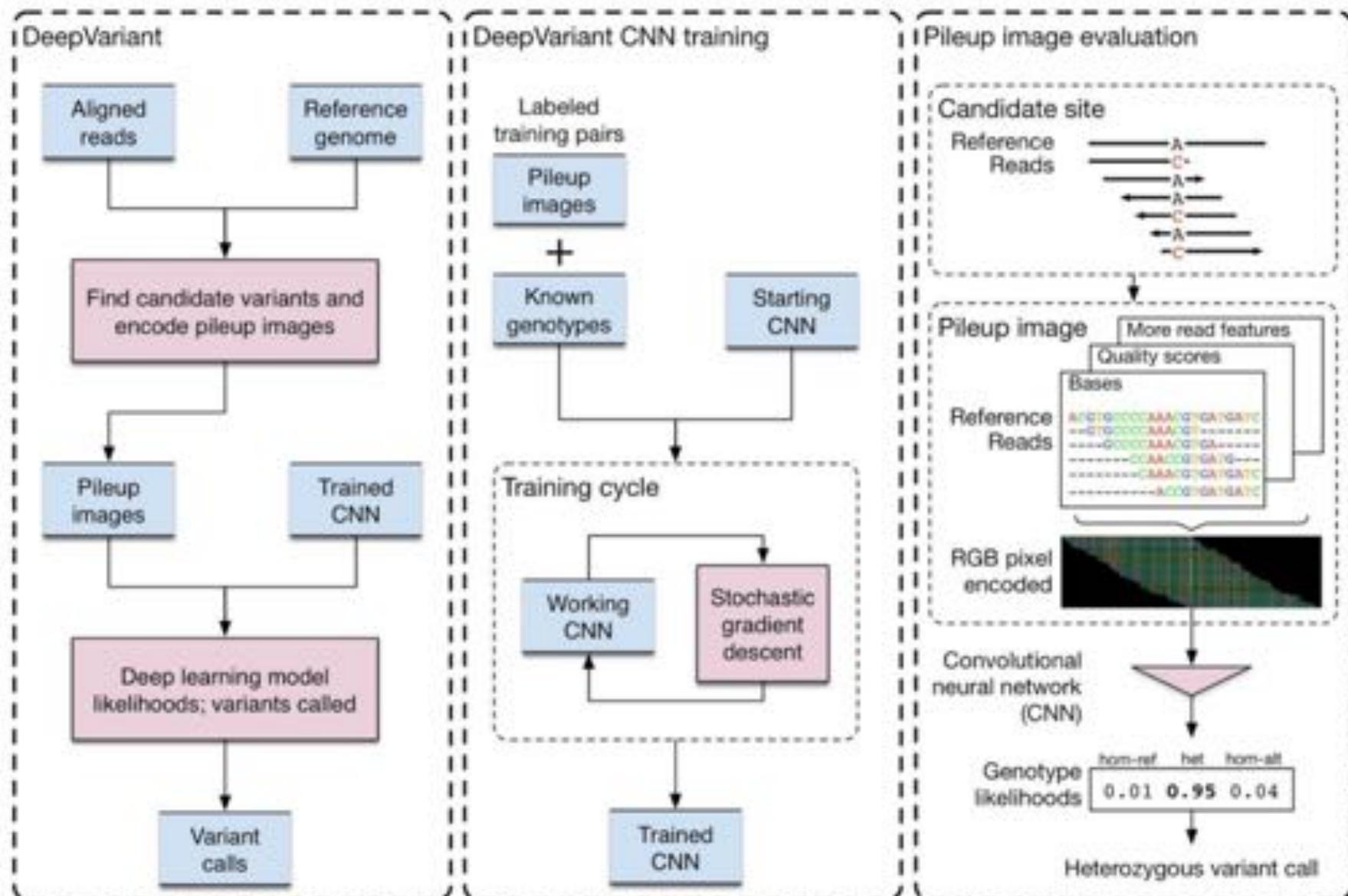
 © 1999 Nature America Inc. • <http://genetics.nature.com>

A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth¹, Ian Korf¹, Mark D. Yandell¹, Raymond T. Yeh¹, Zhijie Gu², Hamideh Zakeri², Nathan O. Stitzel¹, LaDeana Hillier¹, Pui-Yan Kwok² & Warren R. Gish¹

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

Deep Variant



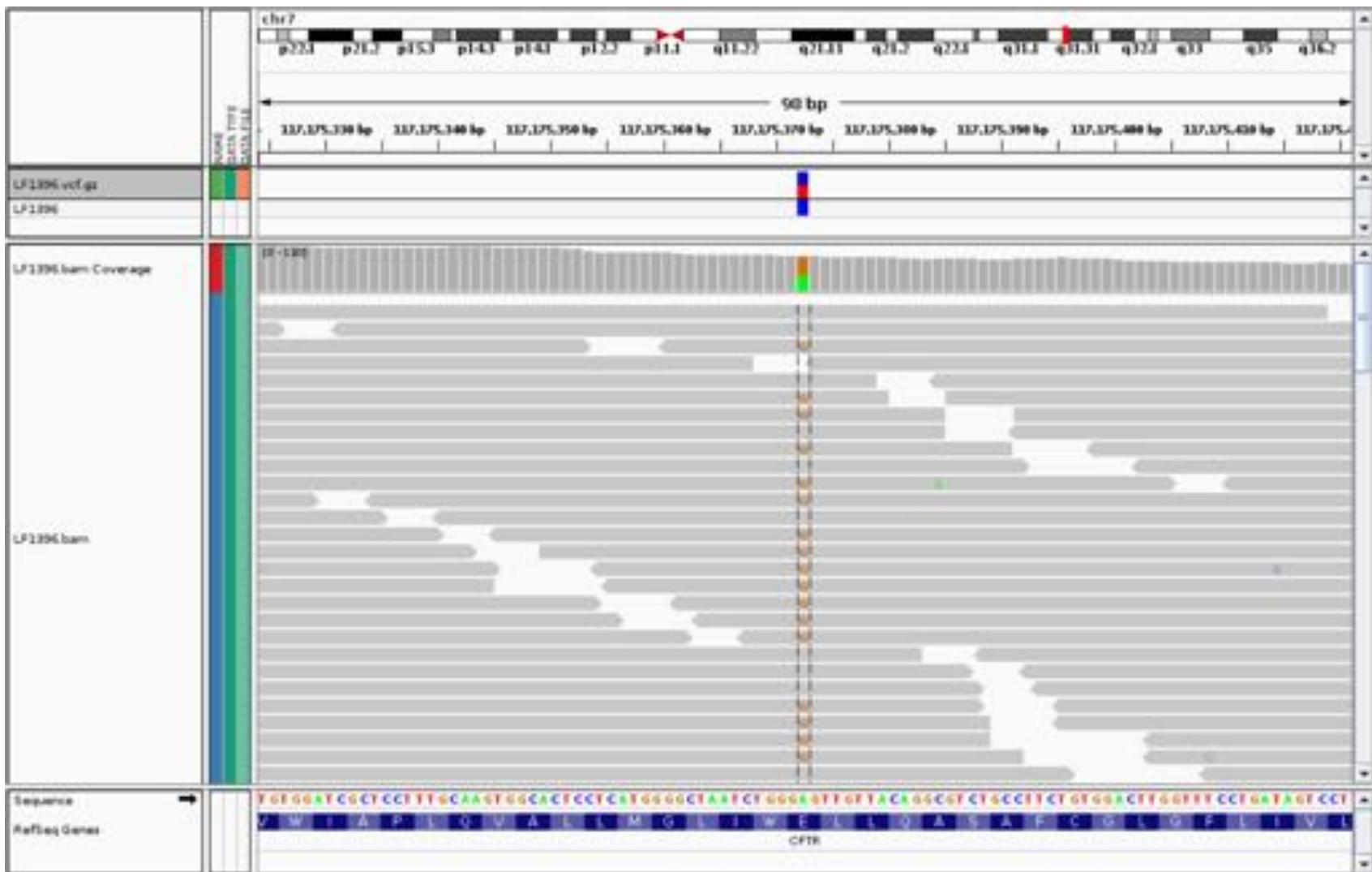
Creating a universal SNP and small indel variant caller with deep neural networks

Poplin et al. (2016) bioRxiv. doi: <https://doi.org/10.1101/092890>

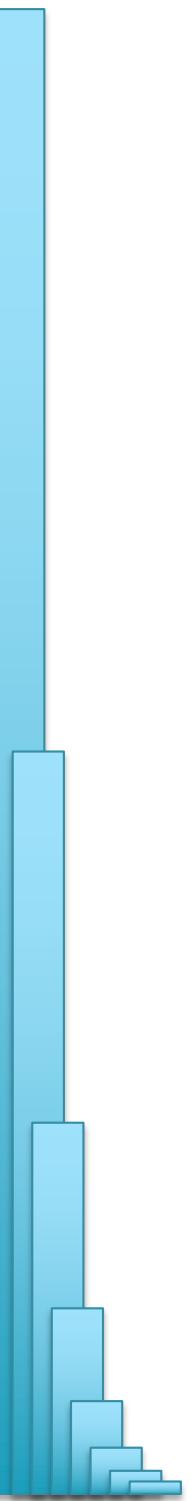
VCF Format

Example

VCF Format

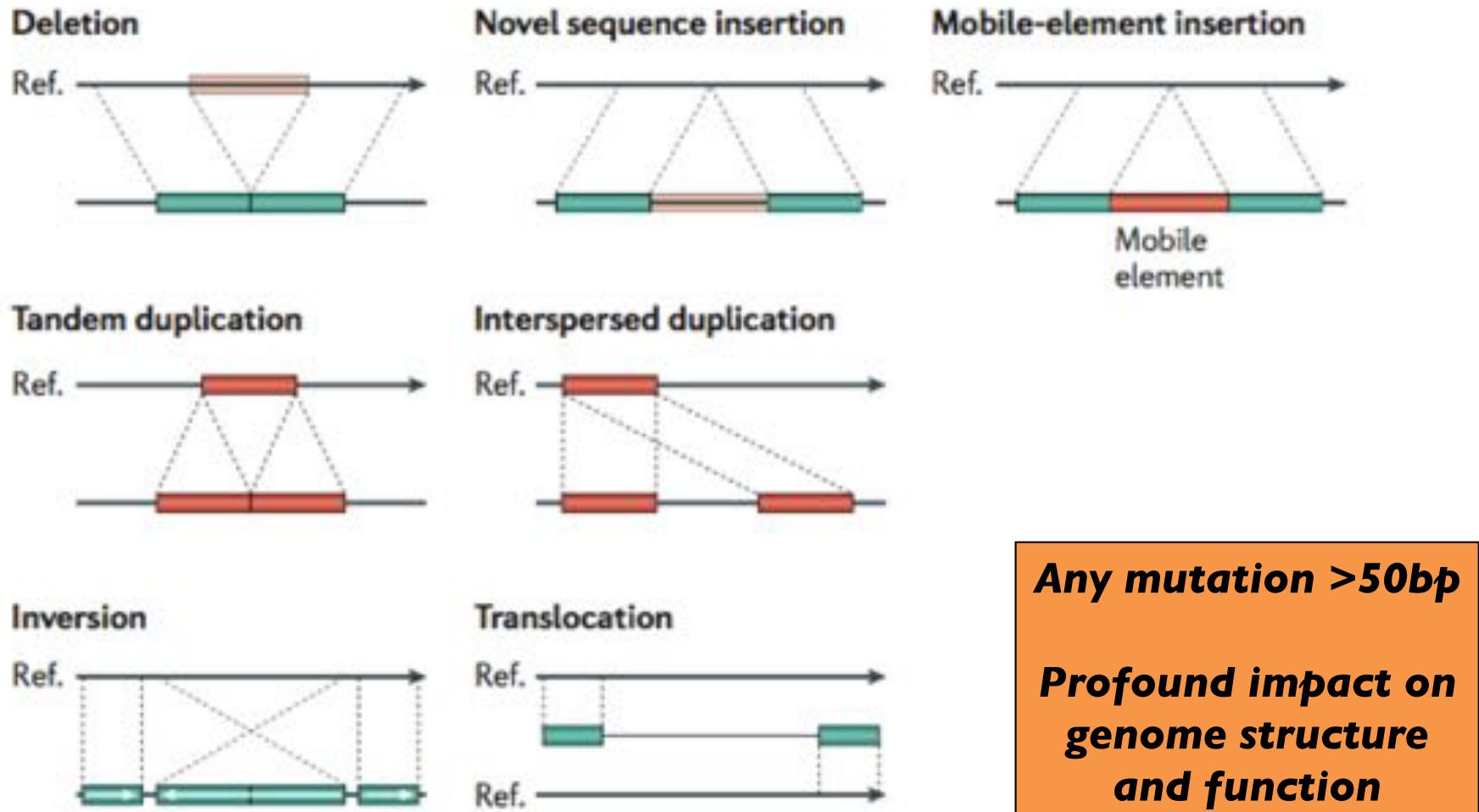


#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	LF1396
chr7	117175373	.	A	G	90	PASS	AF=0.5	GT	0/1



Part 2: What about indels & structural variants

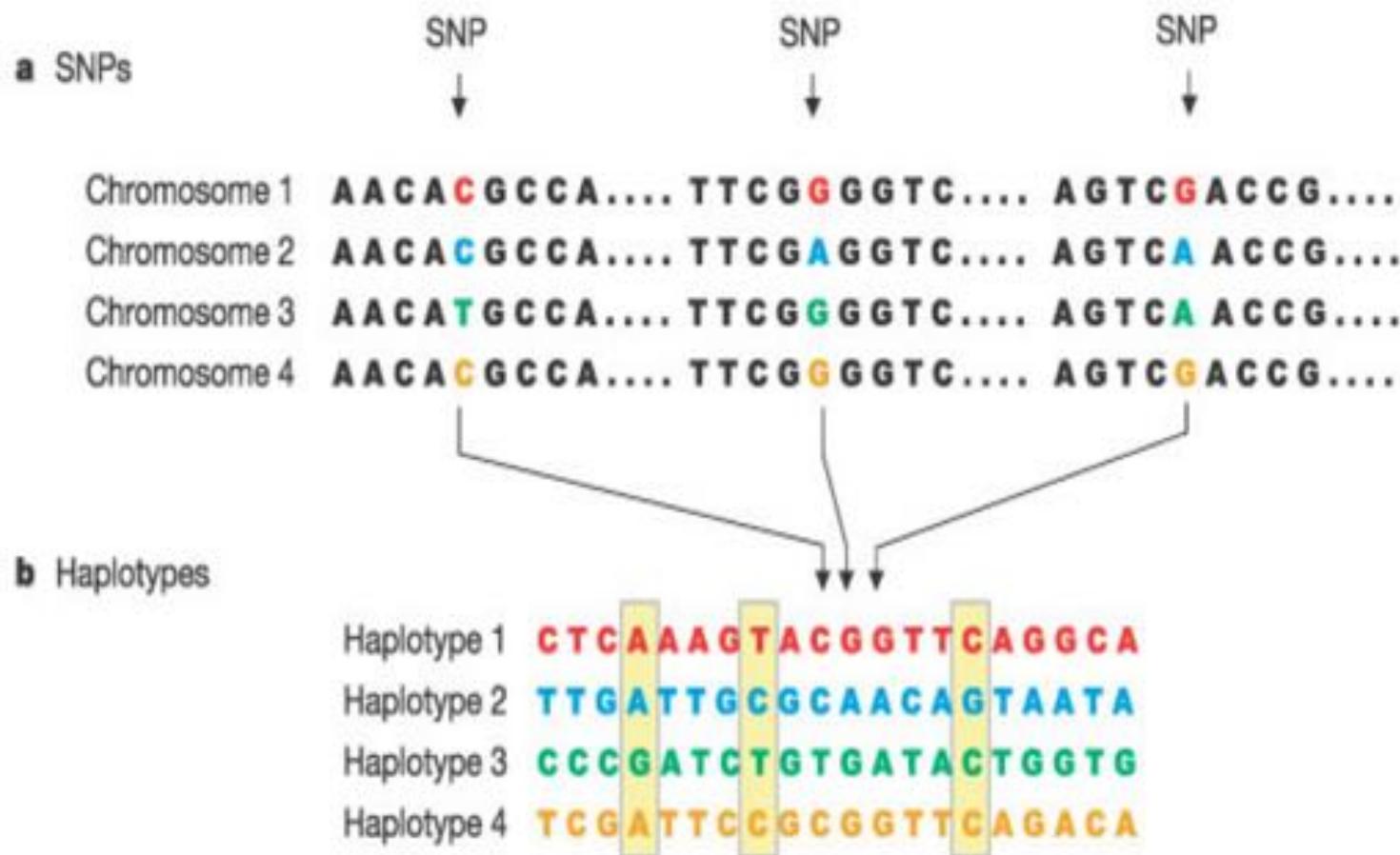
Structural Variations



Genome structural variation discovery and genotyping

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

Early 2000s dogma: SNPs account for most human genetic variation



Discovery of abundant copy-number variation

Science, July 2004

Large-Scale Copy Number Polymorphism in the Human Genome

Jonathan Sebat,¹ B. Lakshmi,¹ Jennifer Troge,¹ Joan Alexander,¹ Janet Young,² Pär Lundin,³ Susanne Mänér,³ Hillary Massa,² Megan Walker,² Maoyen Chi,¹ Nicholas Navin,¹ Robert Lucito,¹ John Healy,¹ James Hicks,¹ Kenny Ye,⁴ Andrew Reiner,¹ T. Conrad Gilliam,⁵ Barbara Trask,² Nick Patterson,⁶ Anders Zetterberg,³ Michael Wigler^{1*}

76 CNVs in 20 individuals

70 genes

Nature Genetics, Aug. 2004

Detection of large-scale variation in the human genome

A John Iafrate^{1,2}, Lars Feuk³, Miguel N Rivera^{1,2}, Marc L Listewnik¹, Patricia K Donahoe^{2,4}, Ying Qi³, Stephen W Scherer^{3,5} & Charles Lee^{1,2,5}

255 CNVs in 55 individuals

127 genes

- 331 CNVs, only 11 in common
- Half observed in only 1 individual
- Impact "plenty" of genes
- Correlated with segmental duplications in the reference genome

Why is structural variation relevant / important?

- ▶ They are common and affect a large fraction of the genome
 - ▶ In total, SVs impact more base pairs than all single-nucleotide differences.
- ▶ They are a major driver of genome evolution
 - ▶ Speciation can be driven by rapid changes in genome architecture
 - ▶ Genome instability and aneuploidy: hallmarks of solid tumor genomes

SV and human disease phenotypes

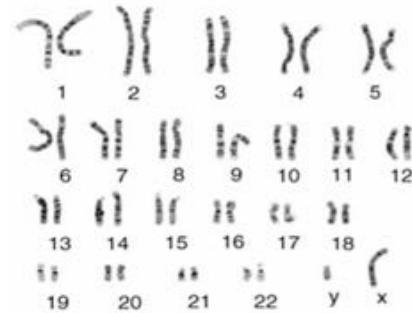
Mendelian (X-linked)

Hemophilia A	306700	<i>F8</i>	inv/del
Hunter syndrome	309900	<i>IDS</i>	del/inv
Ichthyosis	308100	<i>STS</i>	del
Mental retardation	300706	<i>HUWE1</i>	dup
Pelizaeus-Merzbacher disease	312080	<i>PLP1</i>	del/dup/tri
Progressive neurological symptoms (MR+SZ)	300260	<i>MECP2</i>	dup
Red-green color blindness	303800	opsin genes	del

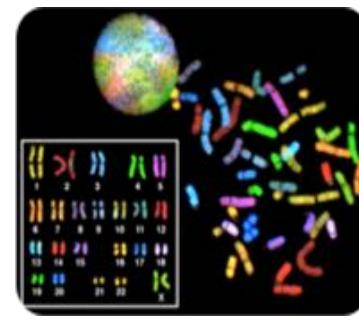
Complex traits

Alzheimer disease	104300	<i>APP</i>	dup
Autism	612200	3q24	inherited homozygous del
	611913	16p11.2	del/dup
Crohn disease	266600	<i>HBD-2</i>	copy number loss
	612278	<i>IRGM</i>	del
HIV susceptibility	609423	<i>CCL3LI</i>	copy number loss
Mental retardation	612001	15q13.3	del
	610443	17q21.31	del
	300534	Xp11.22	dup
Pancreatitis	167800	<i>PRSSI</i>	tri
Parkinson disease	168600	<i>SNCA</i>	dup/tri
Psoriasis	177900	<i>DEFB</i>	copy number gain
Schizophrenia	612474	1q21.1	del
	181500	15q11.2	del
	612001	15q13.3	del
Systemic lupus erythematosus	152700	<i>FCGR3B</i>	copy number loss
	120810	<i>C4</i>	copy number loss

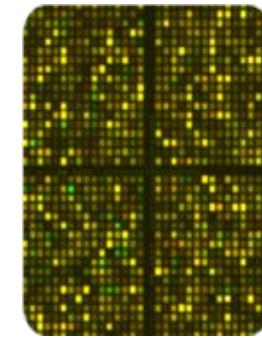
Our understanding of structural variation is driven by technology



1940s - 1980s
Cytogenetics / Karyotyping



1990s
CGH / FISH /
SKY / COBRA



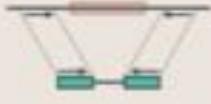
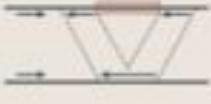
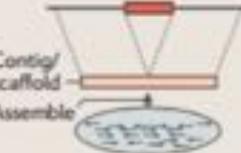
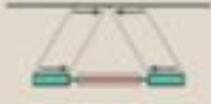
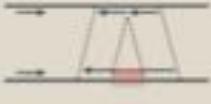
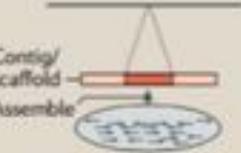
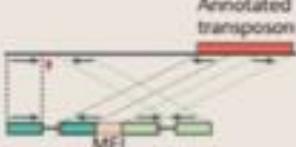
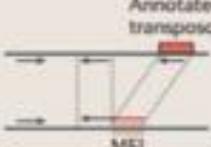
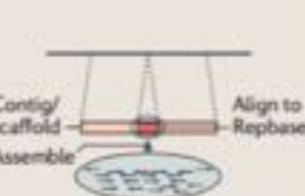
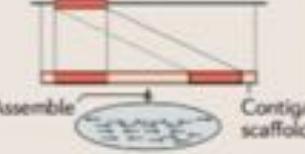
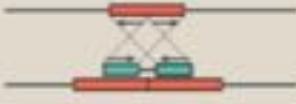
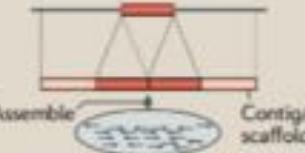
2000s
Genomic microarrays
BAC-aCGH / oligo-aCGH

Today
High throughput
DNA sequencing



Tomorrow
Long Read
DNA sequencing

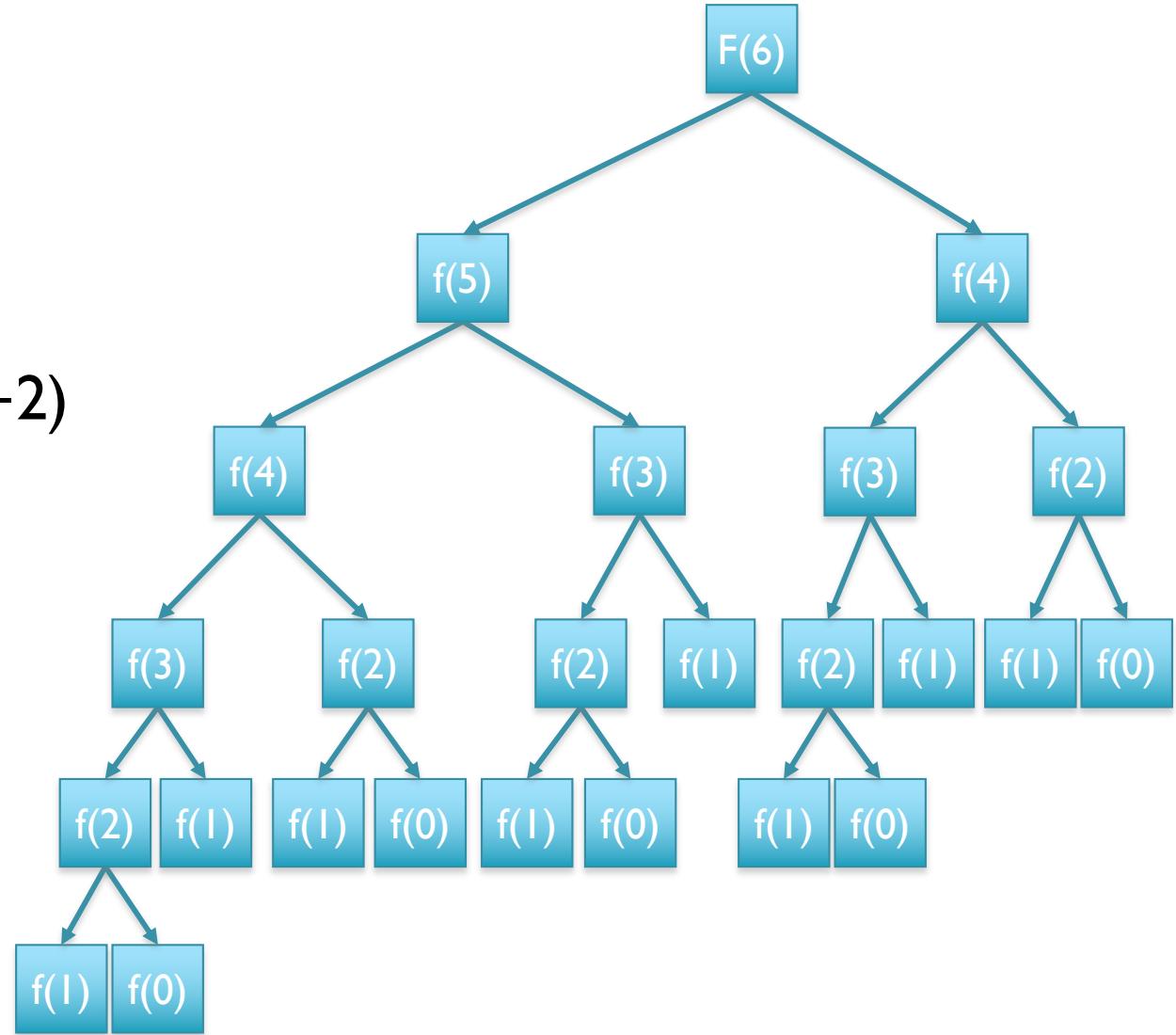
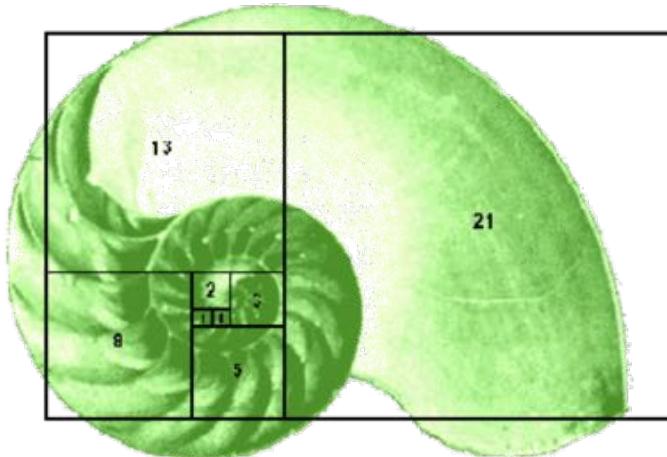
Structural Variation Sequence Signatures

SV classes	Read pair	Read depth	Split read	Assembly
Deletion				
Novel sequence insertion		Not applicable		
Mobile-element insertion		Not applicable		
Inversion		Not applicable		
Interspersed duplication				
Tandem duplication				

Understanding Dynamic Programming

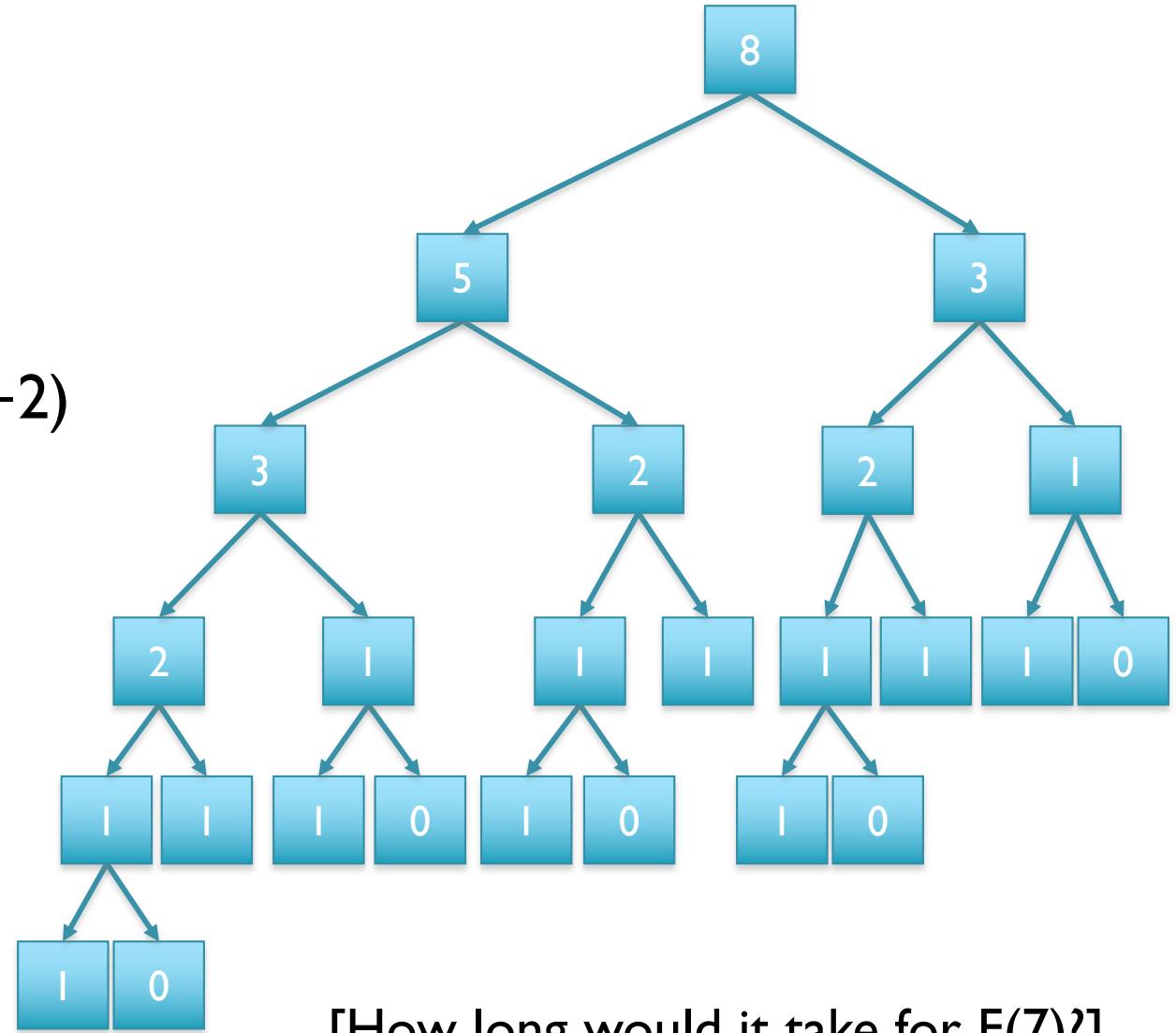
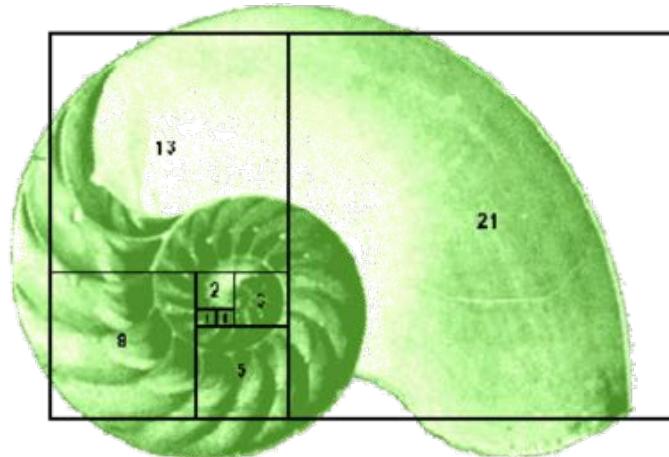
Fibonacci Sequence

```
def fib(n):  
    if n == 0 or n == 1:  
        return n  
  
    else:  
        return fib(n-1) + fib(n-2)
```



Fibonacci Sequence

```
def fib(n):  
    if n == 0 or n == 1:  
        return n  
  
    else:  
        return fib(n-1) + fib(n-2)
```



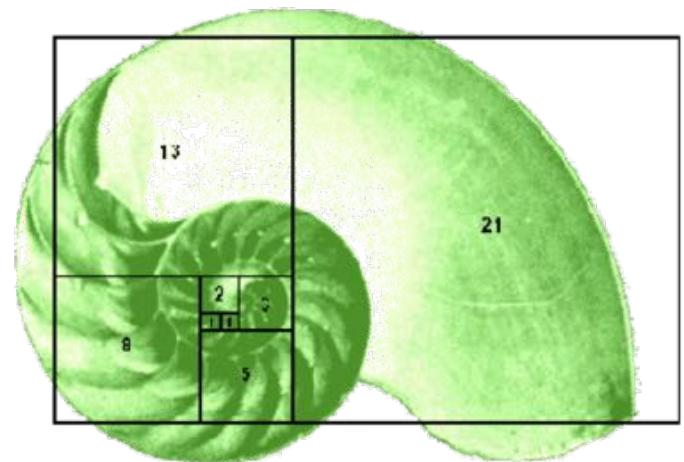
[How long would it take for $F(7)$?]
[What is the running time?]

Bottom-up Fibonacci Sequence

```
def fib(n):  
    table = [0] * (n+1)  
    table[0] = 0  
    table[1] = 1  
    for i in range(2,n+1):  
        table[i] = table[i-2] + table[i-1]  
    return table[n]
```

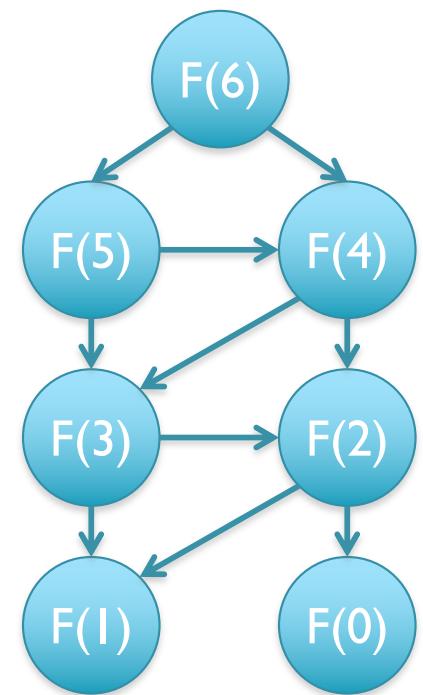
0	1	2	3	4	5	6
0	1	1	2	3	5	8

[How long will it take for F(7)?]
[What is the running time?]



Dynamic Programming

- General approach for solving (some) complex problems
 - When applicable, the method takes far less time than naive methods.
 - Polynomial time ($O(n)$ or $O(n^2)$) instead of exponential time ($O(2^n)$ or $O(3^n)$)
- Requirements:
 - Overlapping subproblems
 - Optimal substructure
- Applications:
 - Fibonacci
 - Longest Increasing Subsequence (Bonus Slides!)
 - Sequence alignment, Dynamic Time Warp, Viterbi
- Not applicable:
 - Traveling salesman problem, Clique finding, Subgraph isomorphism, ...
 - The cheapest flight from airport A to airport B involves a single connection through airport C, but the cheapest flight from airport A to airport C involves a connection through some other airport D.



And now for the main event!

In-exact alignment

- Where is GATTACA *approximately* in the human genome?
 - And how do we efficiently find them?
- It depends...
 - Define 'approximately'
 - Hamming Distance, Edit distance, or Sequence Similarity
 - Ungapped vs Gapped vs Affine Gaps
 - Global vs Local
 - All positions or the single 'best'?
 - Efficiency depends on the data characteristics & goals
 - Bowtie: BWT alignment for short read mapping
 - Smith-Waterman: Exhaustive search for optimal alignments
 - BLAST: Hash based homology searches
 - MUMmer: Suffix Tree based whole genome alignment

Similarity metrics

- Hamming distance
 - Count the number of substitutions to transform one string into another

MIKESCHATZ
| | x | | xxxx |
MICESHATZZ
5

- Edit distance
 - The minimum number of substitutions, insertions, or deletions to transform one string into another

MIKESCHAT-Z
| | x | | x | | | x |
MICES-HATZZ

Edit Distance Example

AGCACACCA → ACACACTA in 4 steps

AGCACACCA → (1. change G to C)

ACCACACCA → (2. delete C)

ACACACCA → (3. change A to T)

ACACACTT → (4. insert A after T)

ACACACTA → done

[Is this the best we can do?]

Edit Distance Example

AGCACACCA → ACACACTA in 3 steps

AGCACACCA → (1. change G to C)

ACCACACCA → (2. delete C)

ACACACCA → (3. insert T after 3rd C)

ACACACTA → done

[Is this the best we can do?]



Welcome to Applied Comparative Genomics
<https://github.com/schatzlab/appliedgenomics2018>

Questions?

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

T. Bayes.

Statement of theorem

Bayes' theorem is stated mathematically as the following equation^[3]

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

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.
- + $P(A | B)$, a conditional probability, is the probability of observing event A given that B is true.
- + $P(B | A)$ is the probability of observing event B given that A is true.

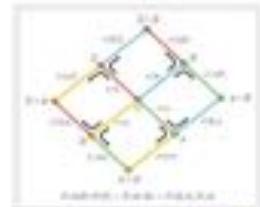
History

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^[4] Bayes' major work "An Essay towards solving a Problem in the Doctrine of Chances"^[5] (1763), which appeared in "Philosophical Transactions",^[6] and contains Bayes' Theorem. Price wrote an introduction to the paper which provides some of the philosophical basis of Bayesian statistics. In 1766 he was elected a Fellow of the Royal Society in recognition of his work on the legacy of Bayes.^{[7][8]}

The French mathematician Pierre-Simon Laplace reproduced and extended Bayes' results in 1774, apparently quite unaware of Bayes' work.^{[7][9]} The Bayesian interpretation of probability was developed mainly by Laplace.^[10]

Stephen Stigler suggested in 1969 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]



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

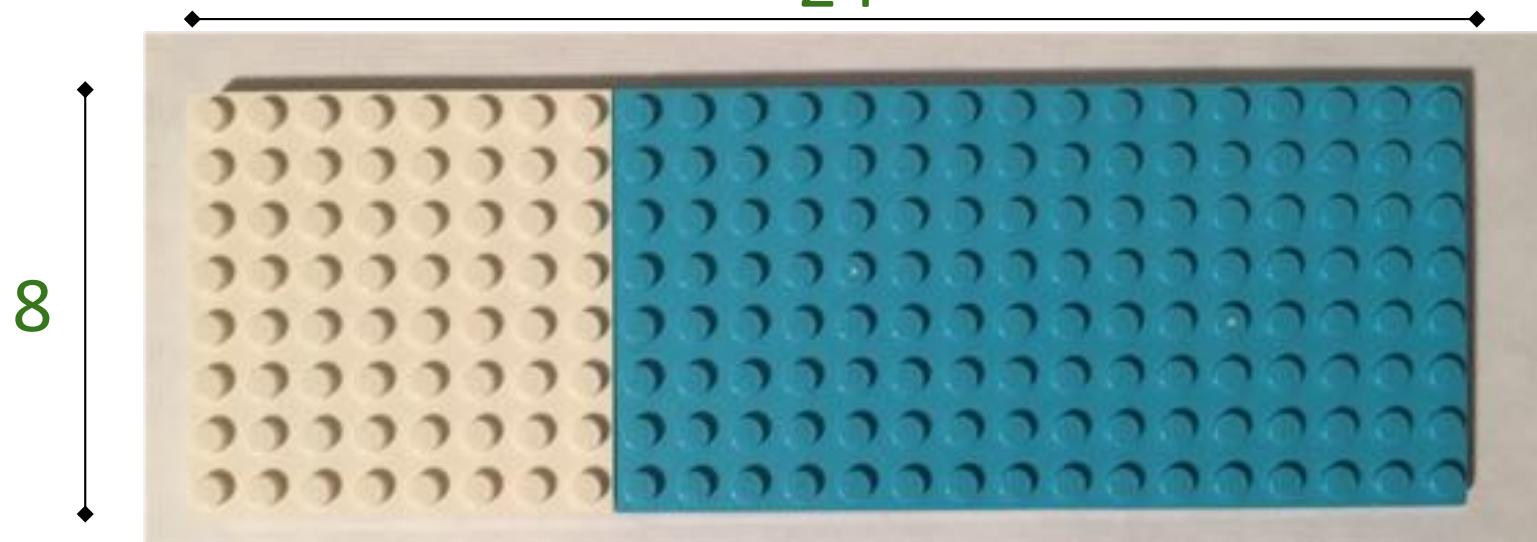
Bayes theorem

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

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

Bayes' theorem with legos

24

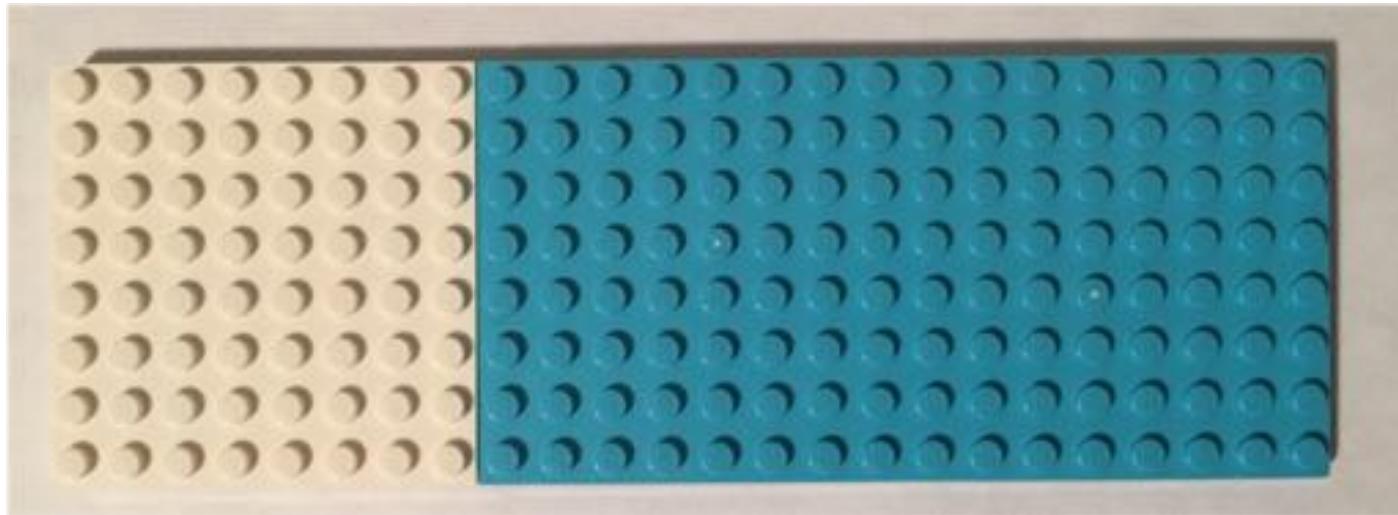


$8 \times 24 = 192$ pegs, 64 are white, 128 are blue.

$$P(\text{White}) = 64 / 192 = 0.33$$

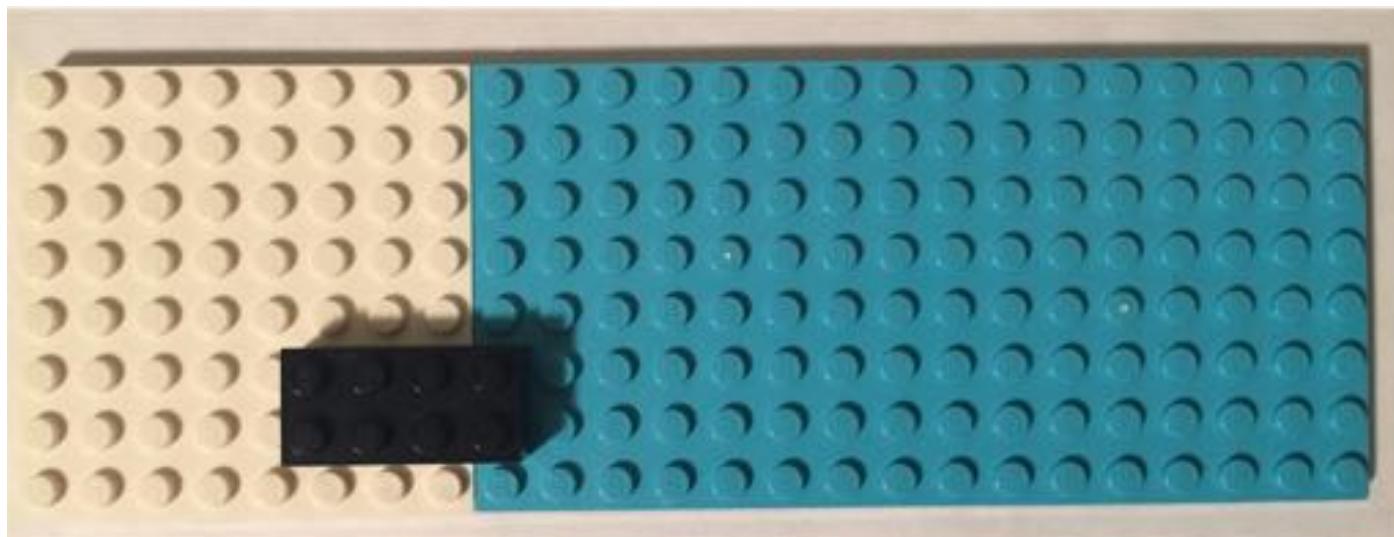
$$P(\text{Blue}) = 128 / 192 = 0.67$$

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



$$P(\text{White}) + P(\text{Blue}) = 1$$

What is the probability of black?

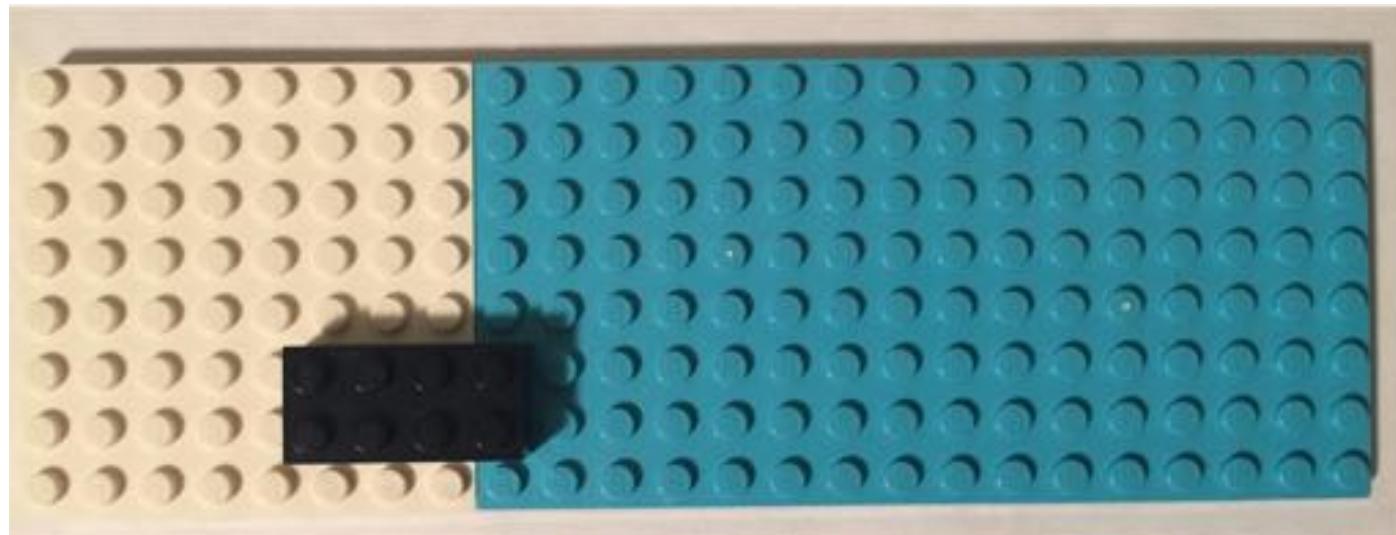


$$P(\text{Black}) = 8 / 192 = 0.042$$

1

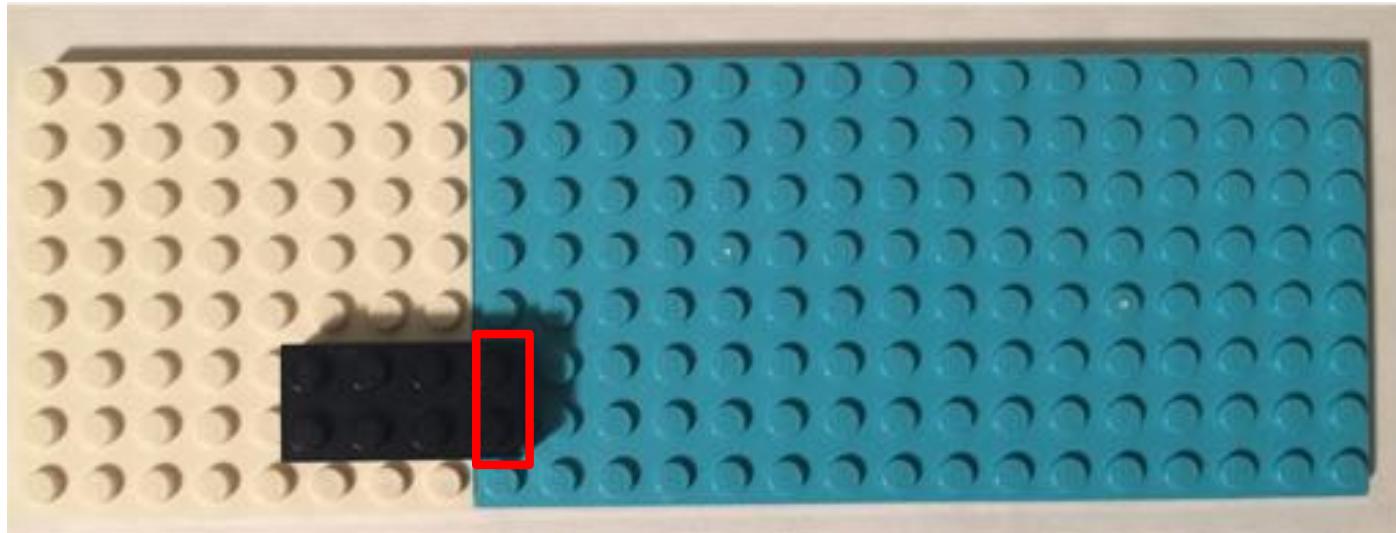
Inspired by
<https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego>

No, probability space is > 1 .
 $P(\text{Black})$ is conditional on $P(\text{White})$ and $P(\text{Blue})$.



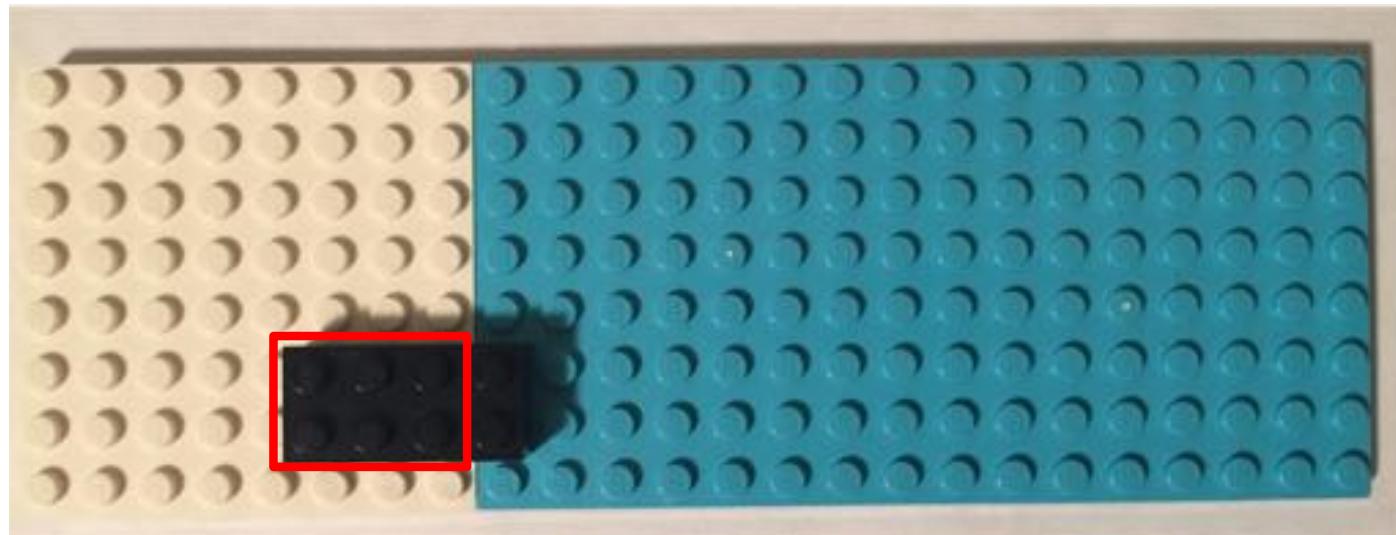
$$P(\text{White}) + P(\text{Blue}) + P(\text{Black}) = 1.042$$

$P(\text{black} \mid \text{blue})$: "probability of black given that we are on a blue peg"



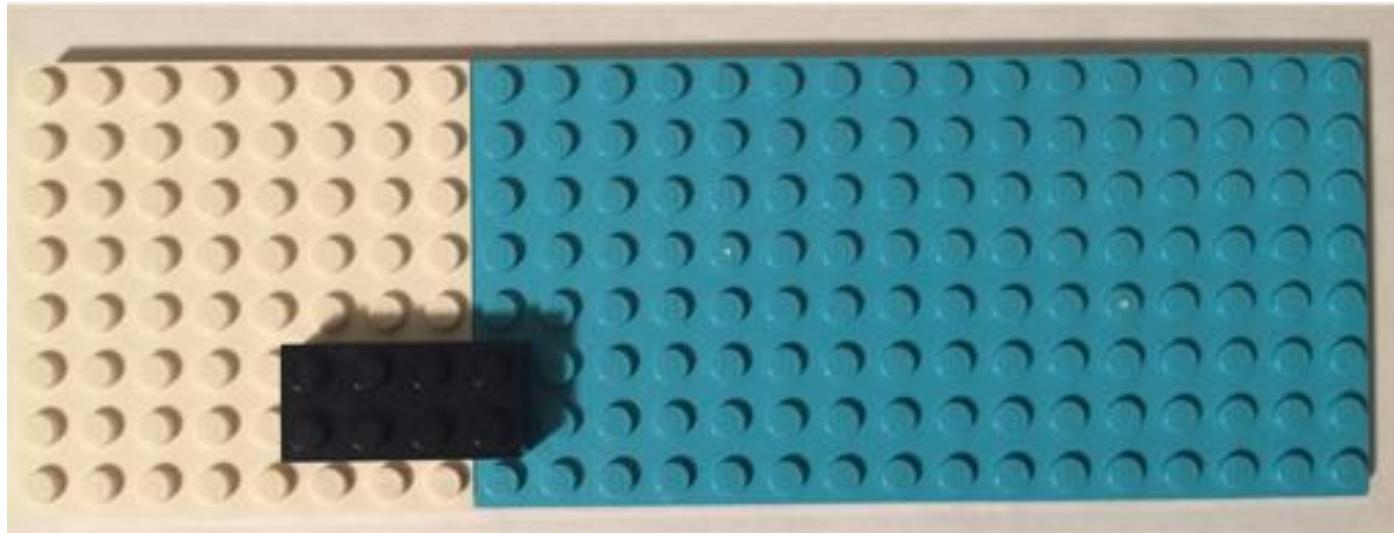
$$P(\text{black} \mid \text{blue}) = 2 / 128 = 0.015625$$

$P(\text{black} \mid \text{white})$: "probability of black given that we are on a white peg"



$$P(\text{black} \mid \text{white}) = 6 / 64 = 0.09375$$

But what about the $P(\text{blue} \mid \text{black})$?



$$P(\text{blue} \mid \text{black}) = 2 / 8 = 0.25$$

This intuition is formalized with Bayes' theorem.

Bayes theorem

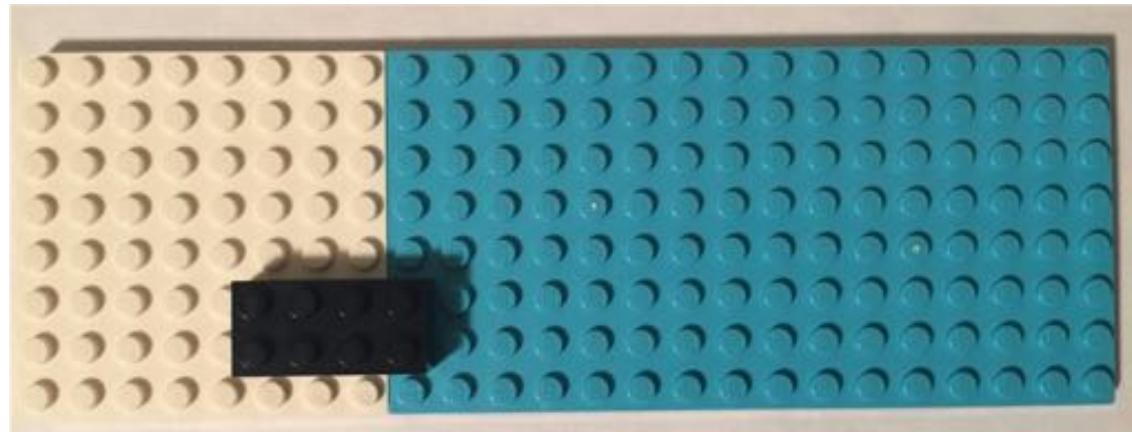
Prior
Probability
Of A

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

Posterior
probability

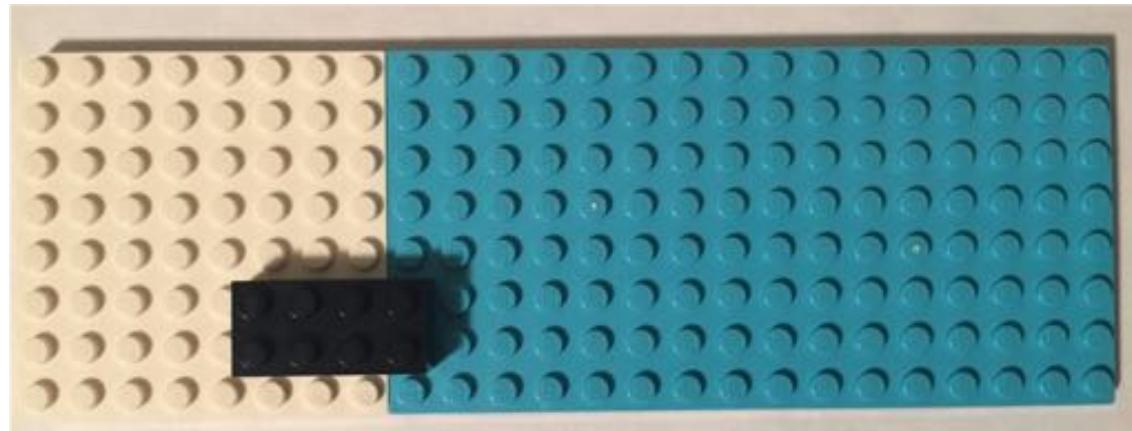
Bayes theorem

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



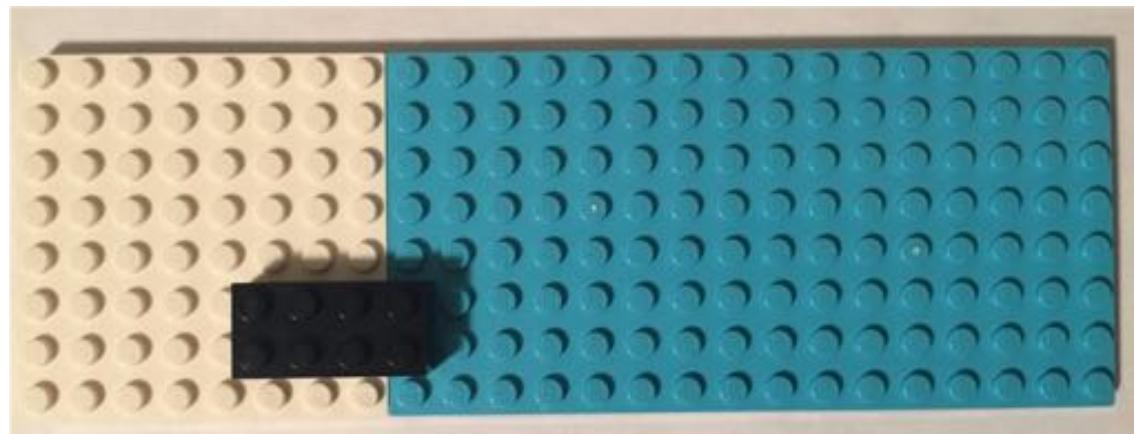
Bayes theorem

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



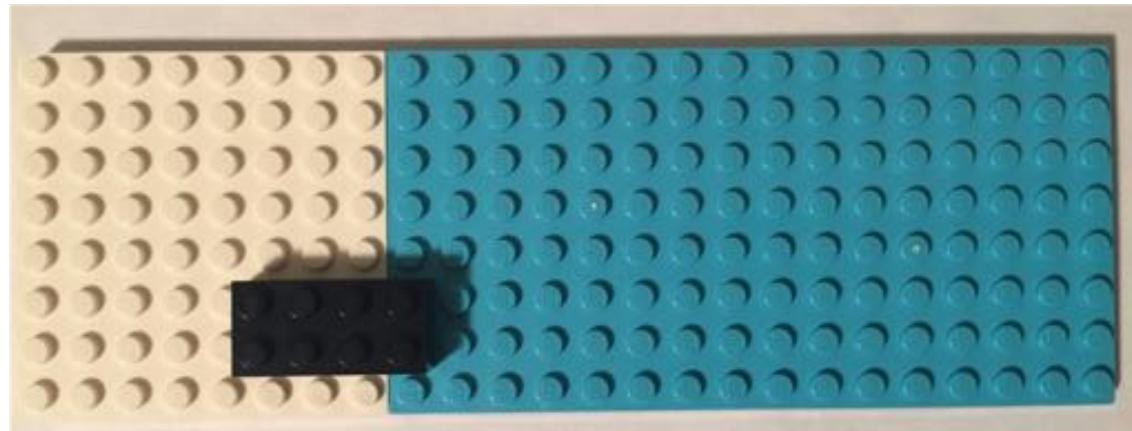
Bayes theorem

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



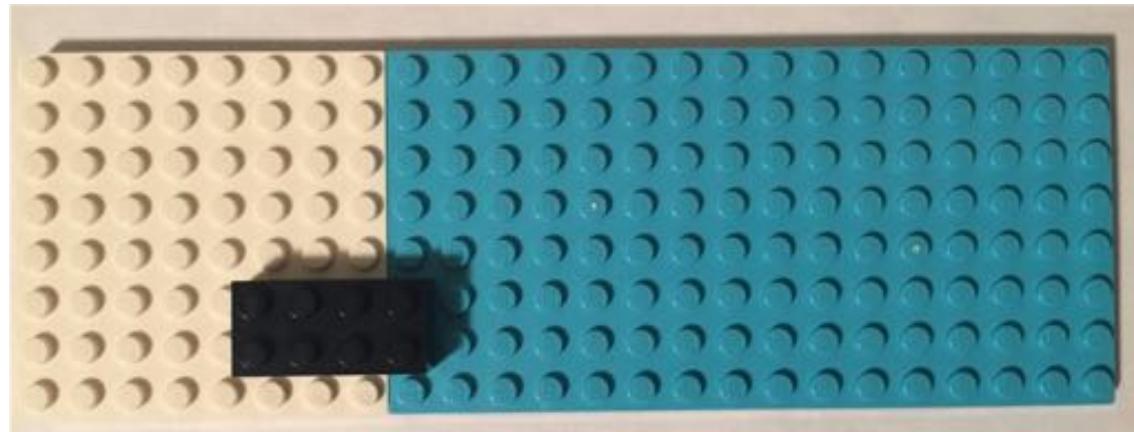
Bayes theorem

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



Bayes theorem

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



Bayes theorem

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

