

Variant Calling (part 3)

Michael Schatz

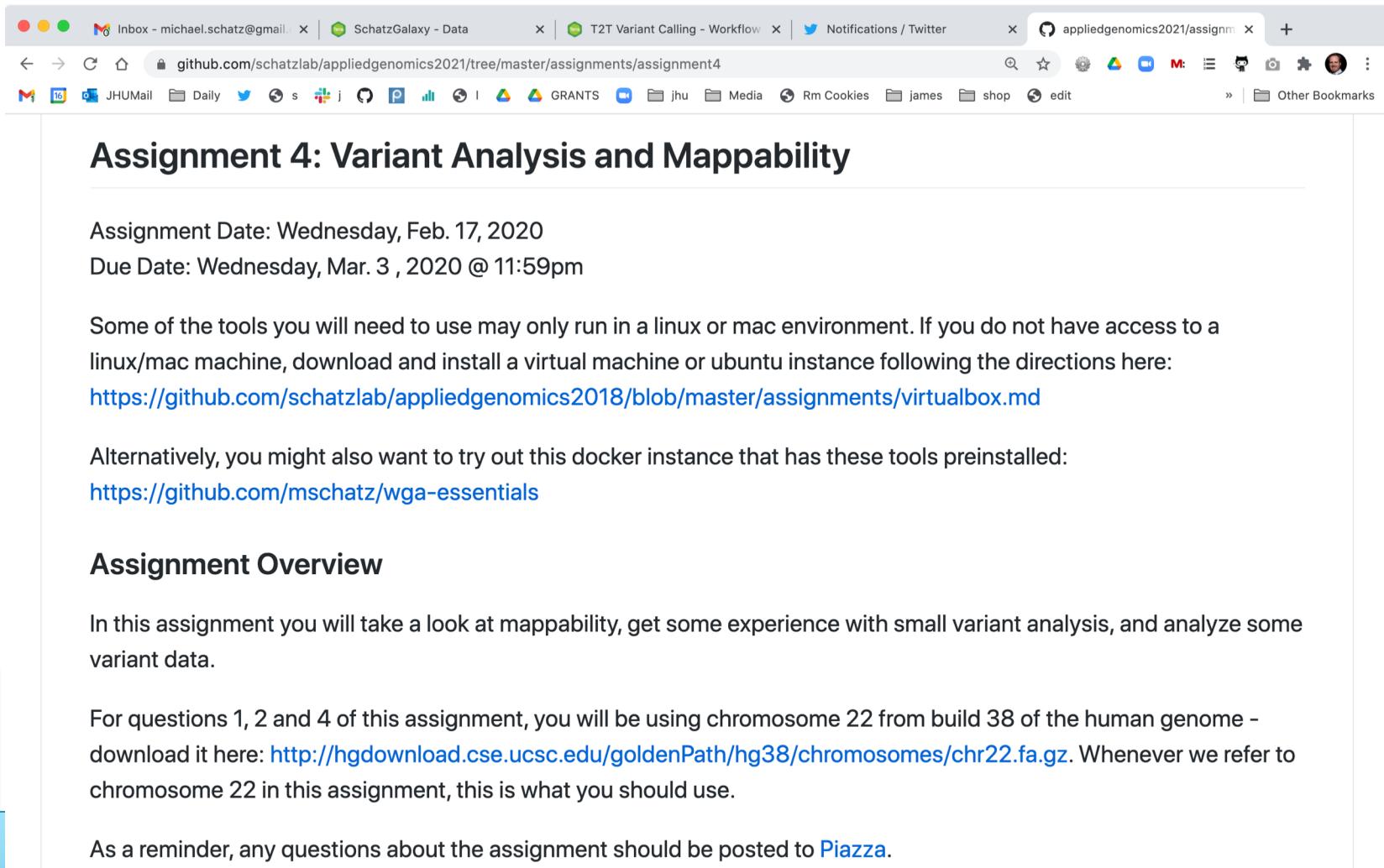
March 1, 2021

Lecture 11: Computational Biomedical Research



Assignment 4: Variant Analysis & Mappability

Due March 3 @ 11:59pm



A screenshot of a web browser window. The address bar shows the URL github.com/schatzlab/appliedgenomics2021/tree/master/assignments/assignment4. The browser has multiple tabs open, including "Inbox - michael.schatz@gmail.com", "SchatzGalaxy - Data", "T2T Variant Calling - Workflow", "Notifications / Twitter", and "appliedgenomics2021/assignm". The main content area displays the assignment details.

Assignment 4: Variant Analysis and Mappability

Assignment Date: Wednesday, Feb. 17, 2020
Due Date: Wednesday, Mar. 3 , 2020 @ 11:59pm

Some of the tools you will need to use may only run in a linux or mac environment. If you do not have access to a linux/mac machine, download and install a virtual machine or ubuntu instance following the directions here:
<https://github.com/schatzlab/appliedgenomics2018/blob/master/assignments/virtualbox.md>

Alternatively, you might also want to try out this docker instance that has these tools preinstalled:
<https://github.com/mschatz/wga-essentials>

Assignment Overview

In this assignment you will take a look at mappability, get some experience with small variant analysis, and analyze some variant data.

For questions 1, 2 and 4 of this assignment, you will be using chromosome 22 from build 38 of the human genome - download it here: <http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/chr22.fa.gz>. Whenever we refer to chromosome 22 in this assignment, this is what you should use.

As a reminder, any questions about the assignment should be posted to [Piazza](#).

<https://github.com/schatzlab/appliedgenomics2021>

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

Reverse Engineering Edit Distance

$$D(\text{AGCACACA}, \text{ACACACTA}) = ?$$

Imagine we already have the optimal alignment of the strings, the last column can only be 1 of 3 options:

...M	...I	...D
...A	...-	...A
...A	...A	...-

The optimal alignment of last two columns is then 1 of 9 possibilities

...MM ...IM ...DM	...MI ...II ...DI	...MD ...ID ...DD
...CA ...-A ...CA	...A- ...-- ...A-	...CA ...-A ...CA
...TA ...TA ...-A	...TA ...TA ...-A	...A- ...A-

The optimal alignment of the last three columns is then 1 of 27 possibilities...

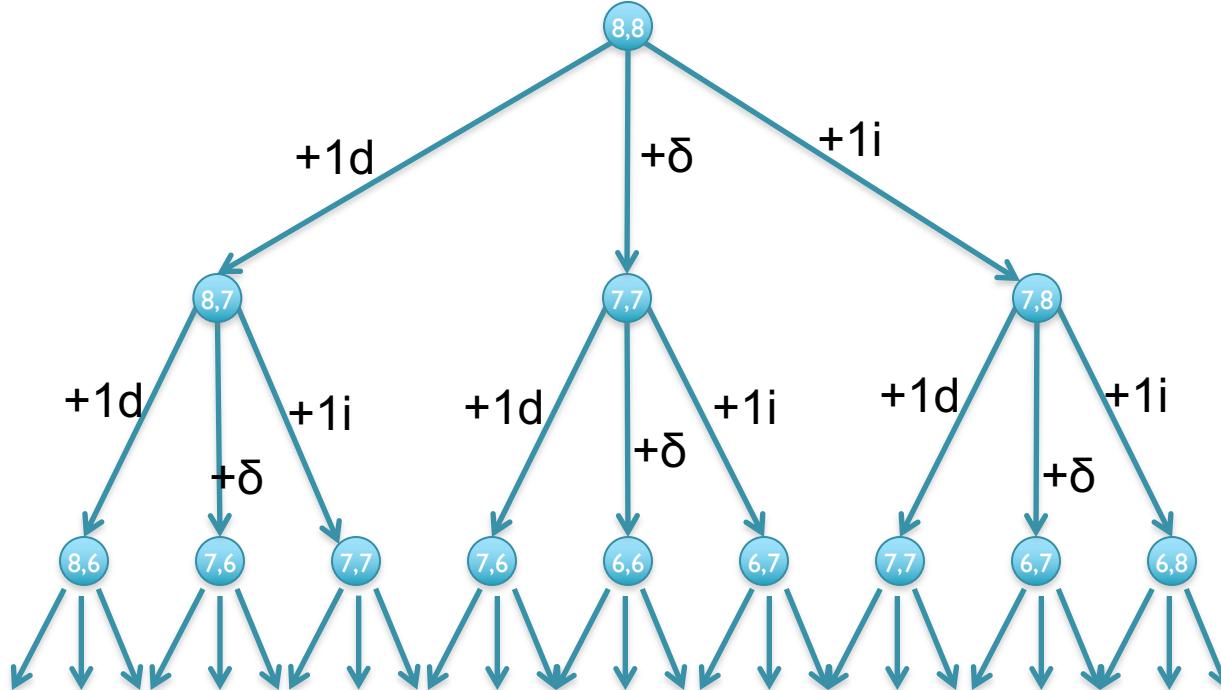
...M...	...I...	...D...
...X...	...-...	...X...
...Y...	...Y...	...-...

Eventually spell out every possible sequence of {I,M,D}

Recursive solution

- Computation of D is a recursive process.
 - At each step, we only allow matches, substitutions, and indels
 - $D(i,j)$ in terms of $D(i',j')$ for $i' \leq i$ and $j' \leq j$.

$$D(\text{AGCACACA}, \text{ACACACTA}) = \min\{D(\text{AGCACACA}, \text{ACACACT}) + 1, \\ D(\text{AGCACAC}, \text{ACACACTA}) + 1, \\ D(\text{AGCACAC}, \text{ACACACT}) + \delta(A, A)\}$$



[What is the running time?]

Recurrence Relation for D

Find the edit distance (minimum number of operations to convert one string into another) in $O(mn)$ time

- Base conditions:

- $D(i,0) = i$, for all $i = 0, \dots, n$
- $D(0,j) = j$, for all $j = 0, \dots, m$

- For $i > 0, j > 0$:

$$D(i,j) = \min \{ \begin{aligned} & D(i-1,j) + 1, && // \text{align 0 chars from } S, 1 \text{ from } T \\ & D(i,j-1) + 1, && // \text{align 1 chars from } S, 0 \text{ from } T \\ & D(i-1,j-1) + \delta(S(i),T(j)) && // \text{align 1+1 chars} \end{aligned} \}$$

[Why do we want the min?]

Dynamic Programming Matrix

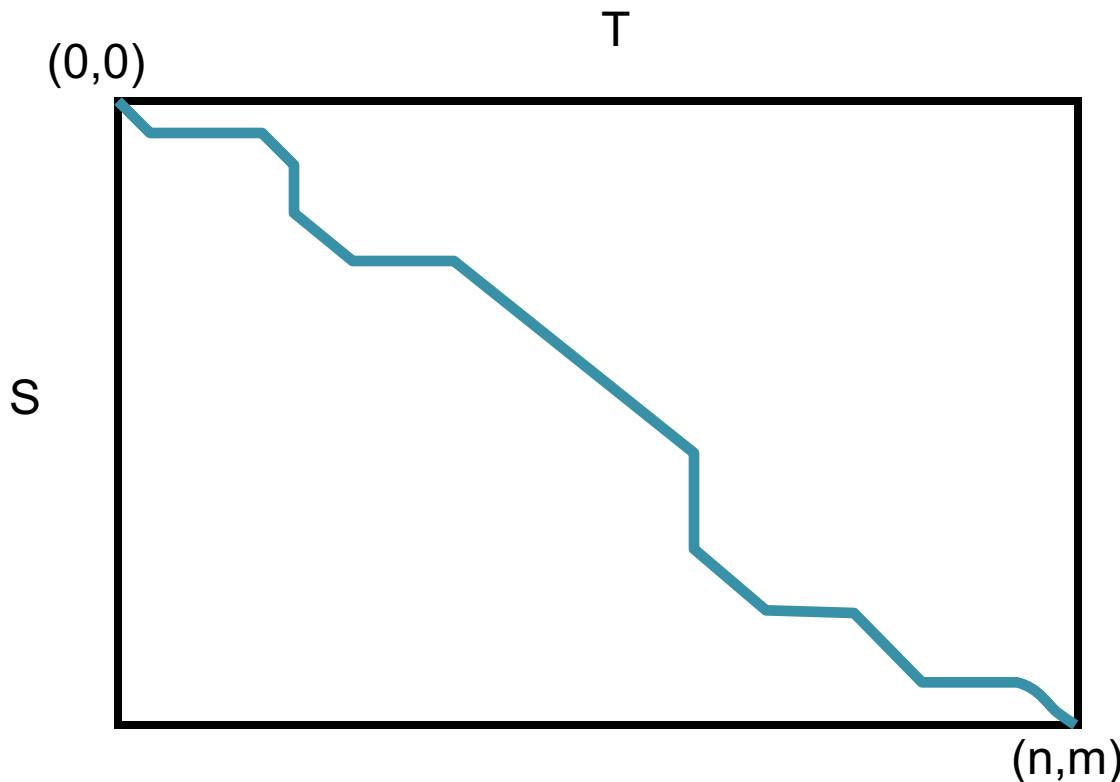
		A	C	A	C	A	C	T	A
	0	1	2	3	4	5	6	7	8
A	1	0	1	2	3	4	5	6	7
G	2	1	1	2	3	4	5	6	7
C	3	2	1	2	2	3	4	5	6
A	4	3	2	1	2	2	3	4	5
C	5	4	3	2	1	2	2	3	4
A	6	5	4	3	2	1	2	3	3
C	7	6	5	4	3	2	1	2	3
A	8	7	6	5	4	3	2	2	2

$$D[AGCACACACA, ACACACTA] = 2$$

AGCACAC-A
 | * | | | | * |
 A-CACACTA

[Can we do it any better?]

Global Alignment Schematic

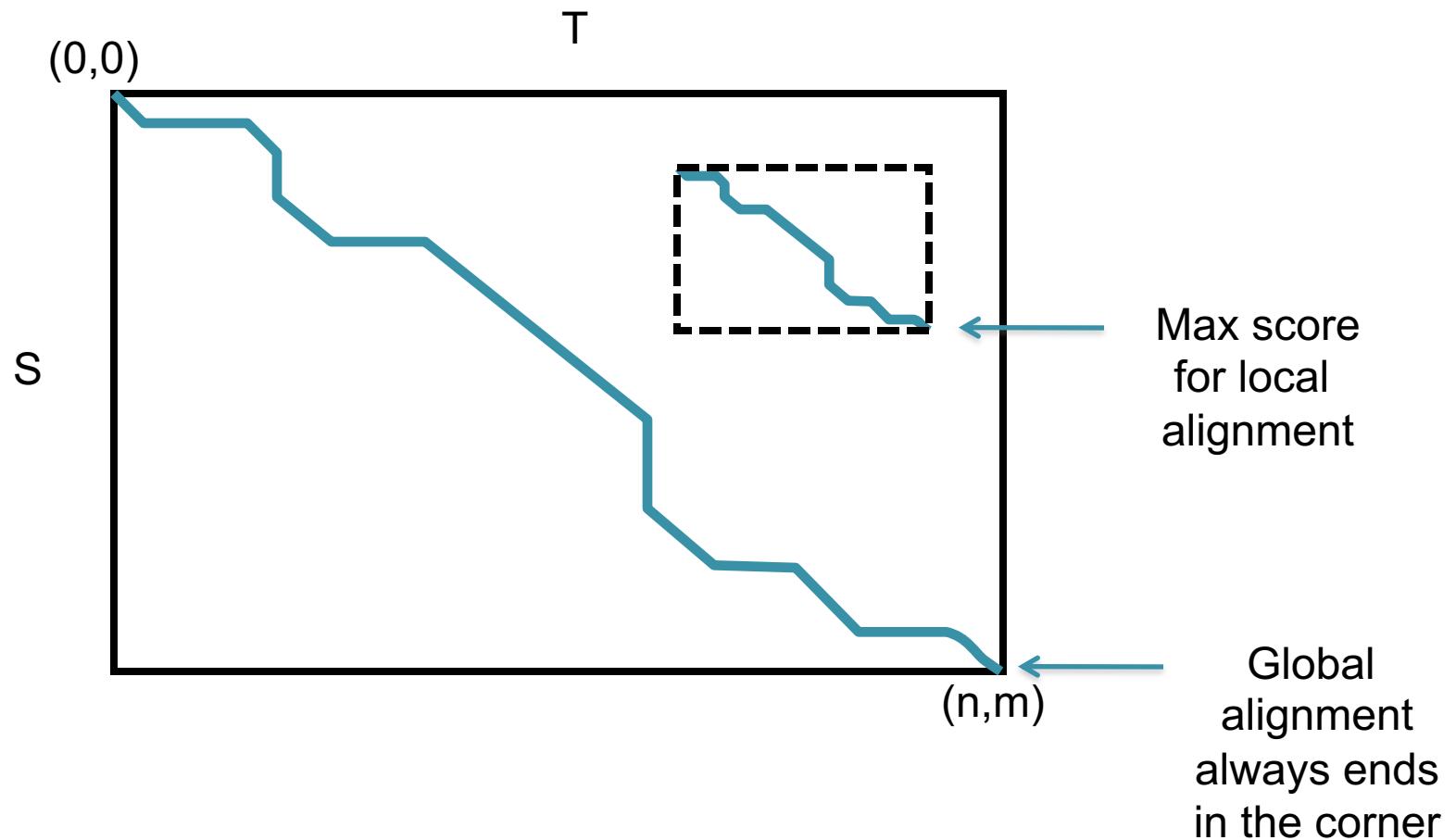


- A high quality alignment will stay close to the diagonal
 - If we are only interested in high quality alignments, we can skip filling in cells that can't possibly lead to a high quality alignment
 - Find the global alignment with at most edit distance d : $O(2dn)$

Local vs. Global Alignment

- The Global Alignment Problem tries to find the best end-to-end alignment between the two strings
 - Only applicable for very closely related sequences
- The Local Alignment Problem tries to find pairs of **substrings** with highest similarity.
 - Especially important if one string is substantially longer than the other
 - Especially important if there is only a distant evolutionary relationship

Global vs Local Alignment Schematic



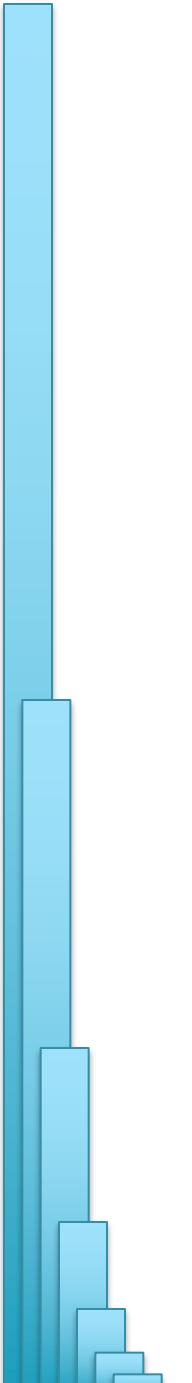
Local vs. Global Alignment (cont' d)

- **Global Alignment**

```
--T--CC-C-AGT--TATGT-CAGGGGACACG-A-GCATGCAGA-GAC  
| | | | | | | | | | | | | | | | | | | | | | | | | | | |  
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG-T-CAGAT--C
```

- **Local Alignment**—better alignment to find conserved segment

```
tccCAGTTATGTCAGggacacgagcatgcagagac  
| | | | | | | | |  
aattgccgcgtcgtttcagCAGTTATGTCAGatc
```

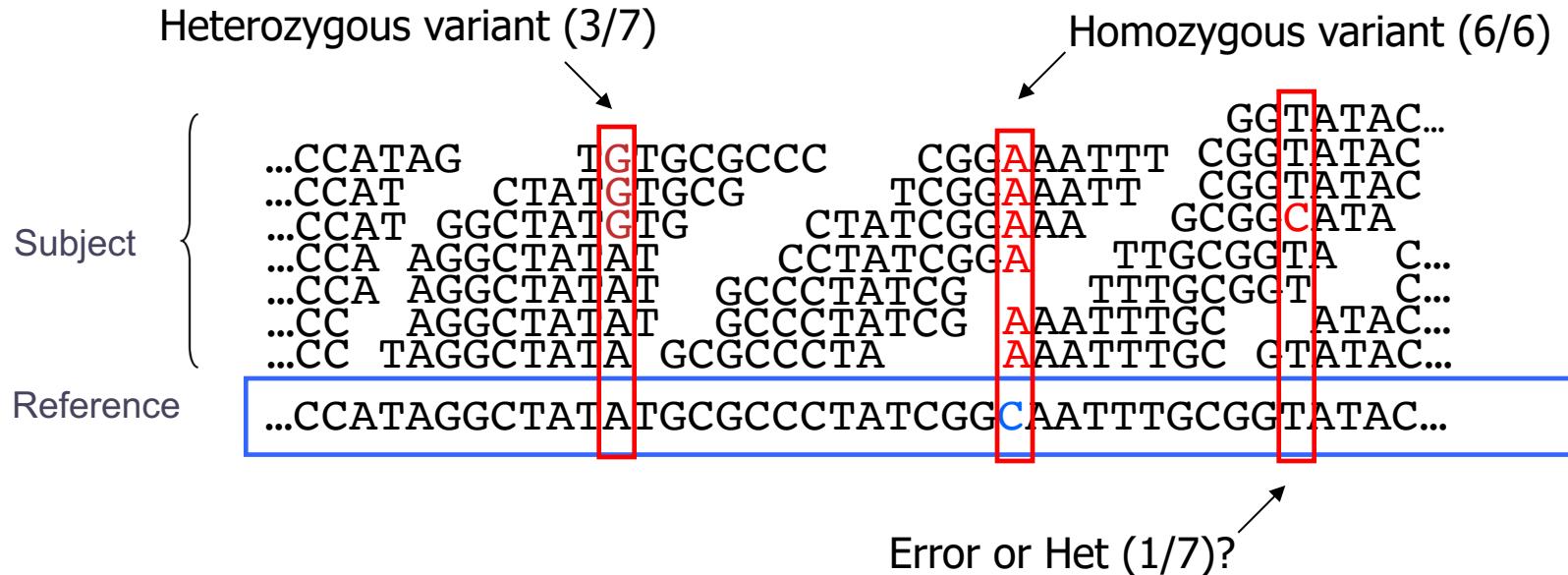


Part 3: Variant Calling

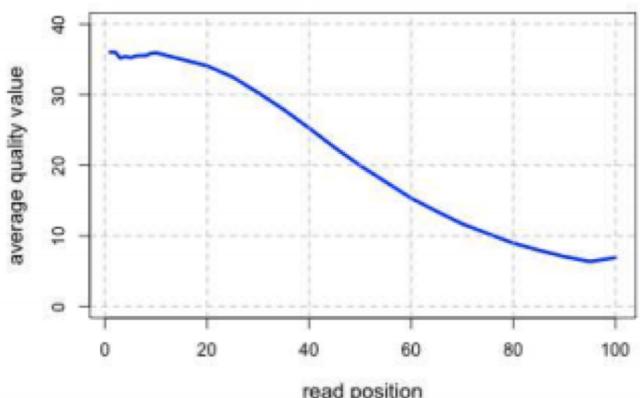
Variant Calling Overview



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

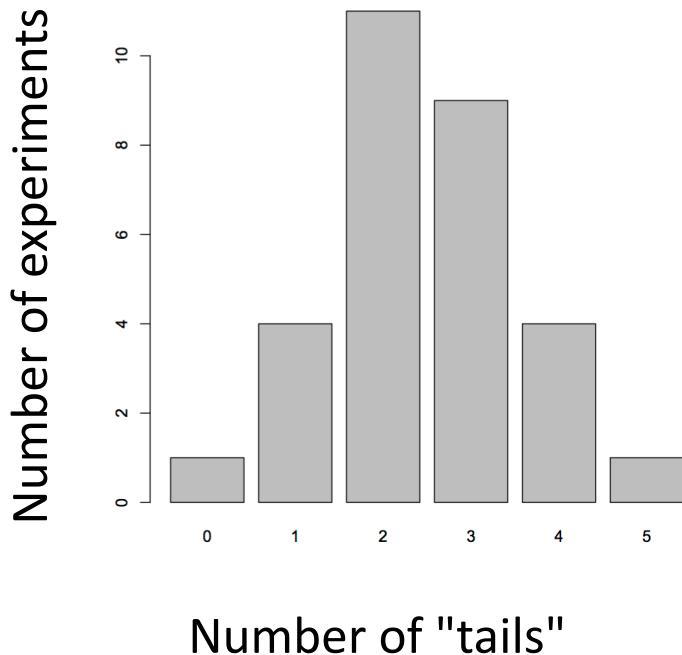


$P(\text{heads}) = 0.5$



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

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



R code:

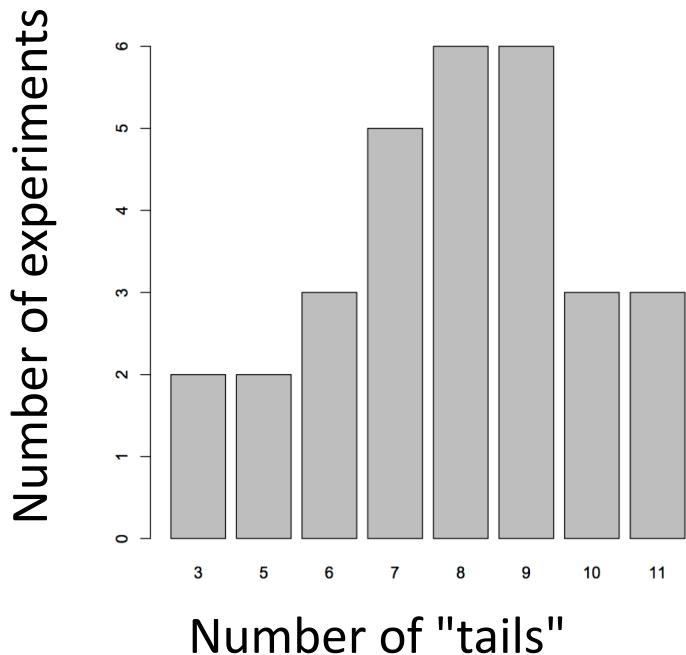
```
barplot(table(rbinom(30, 5, 0.5)))
```

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:

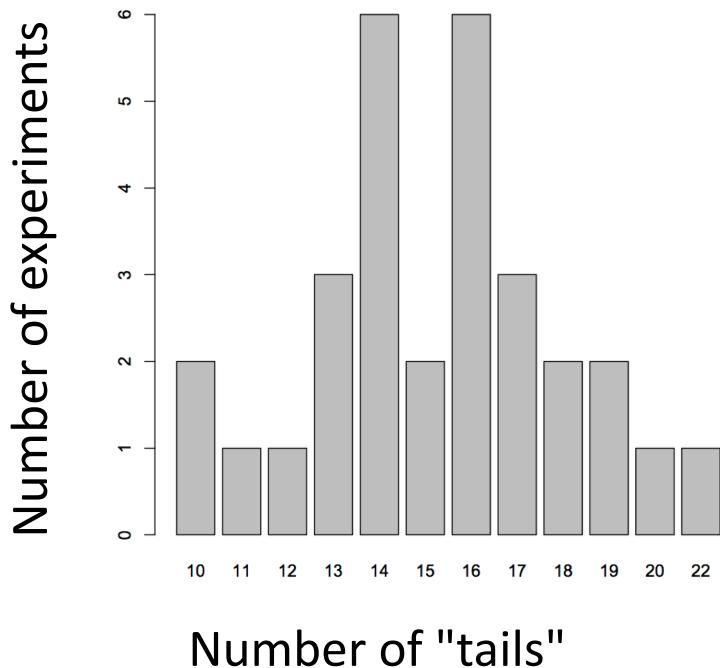
```
barplot(table(rbinom(30, 15, 0.5)))
```

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:

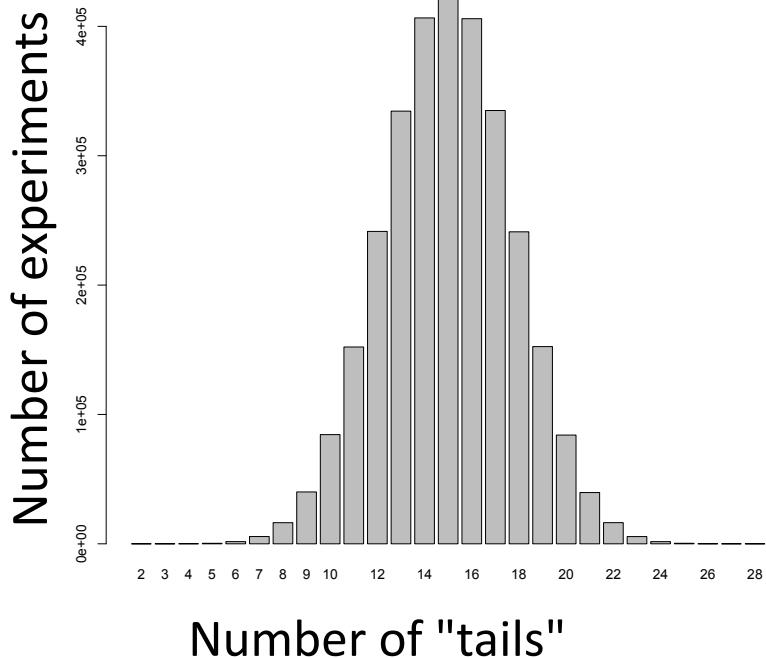
```
barplot(table(rbinom(30, 30, 0.5)))
```

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:

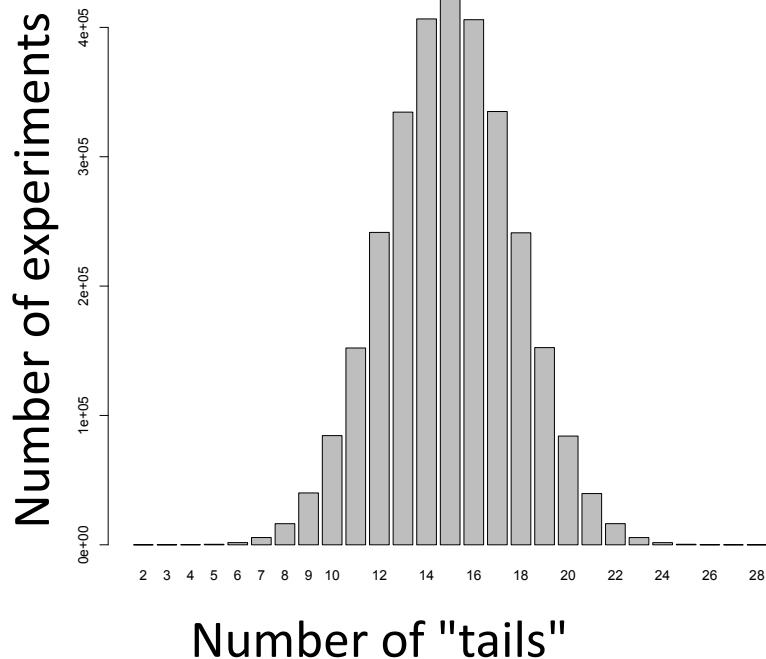
```
barplot(table(rbinom(3e6, 30, 0.5)))
```

3M experiments (students tossing coins)

30 tosses each

Probability of Tails

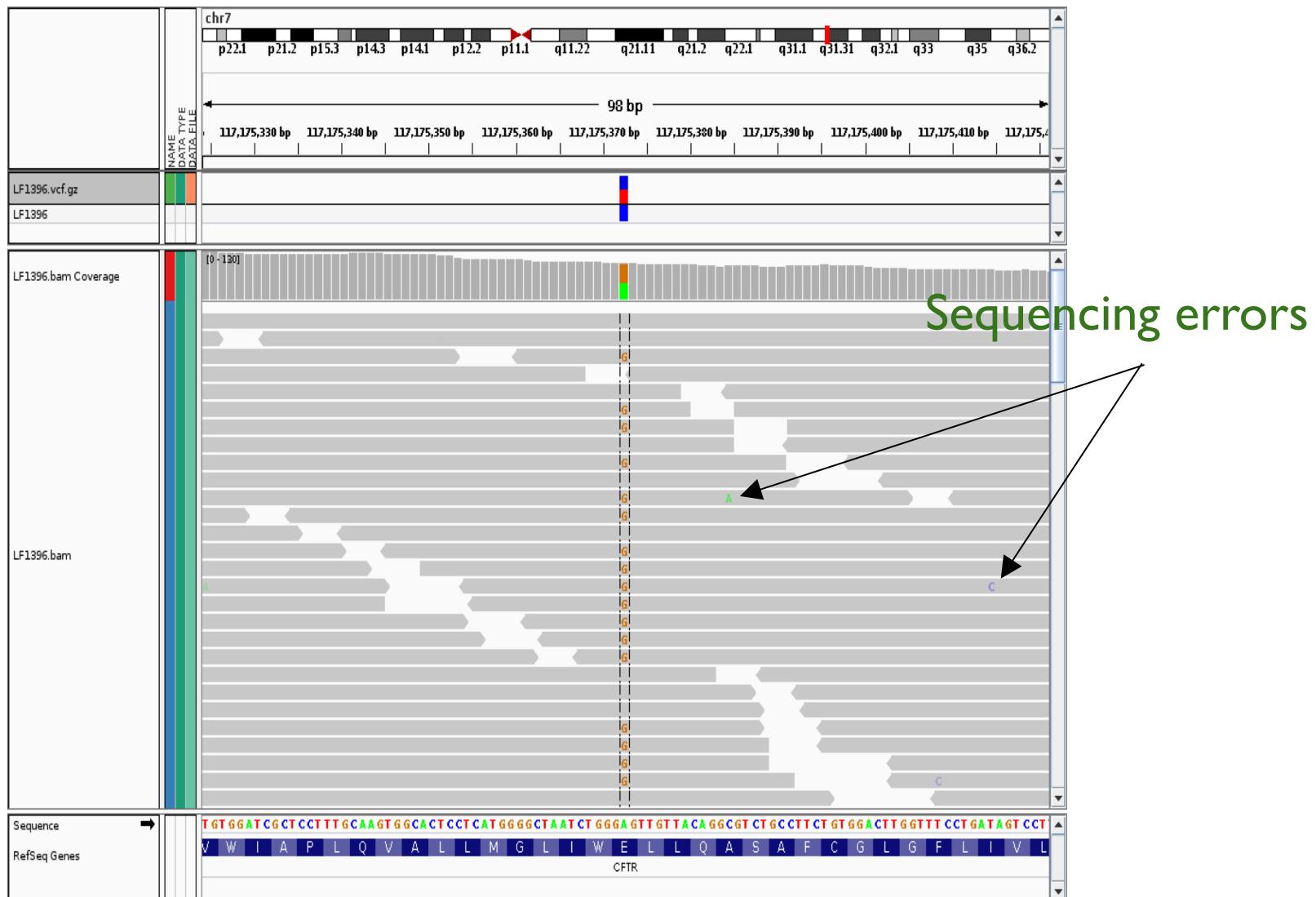
So, with 30 tosses (reads), we are much more likely to see an even mix of alternate and reference alleles at a heterozygous locus in a genome



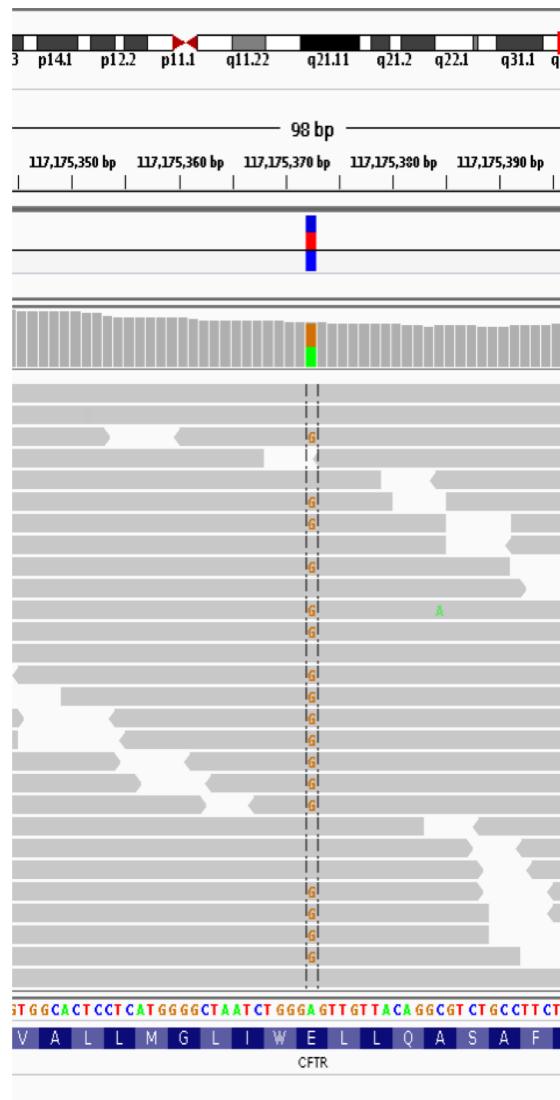
This is why at least a "30X" (30 fold sequence coverage) genome is recommended: it confers sufficient power to distinguish heterozygous alleles and from mere sequencing errors

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

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

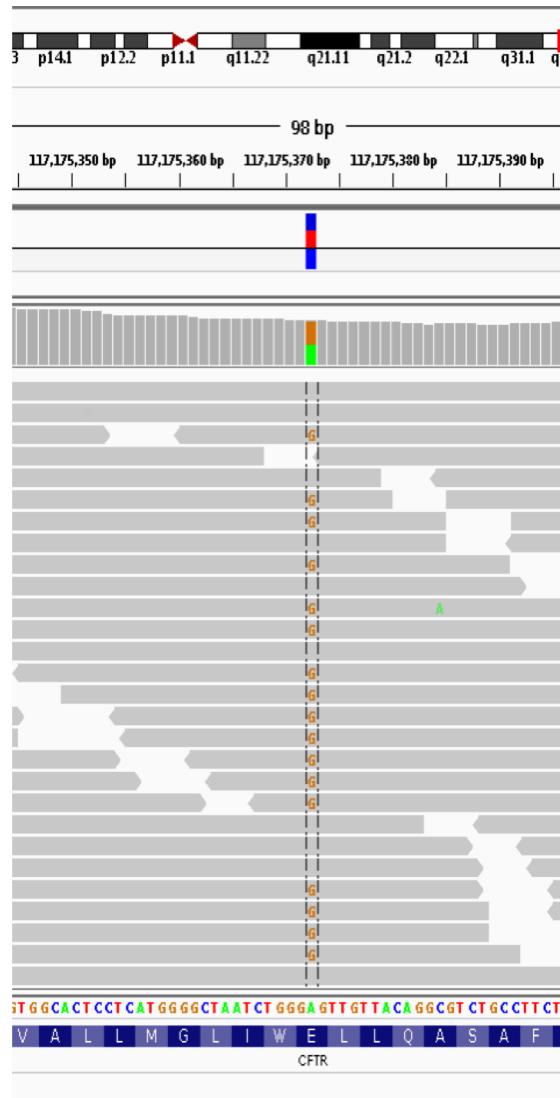


What information is needed to decide if a variant exists?



- 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

Bayesian SNP calling



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

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

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

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

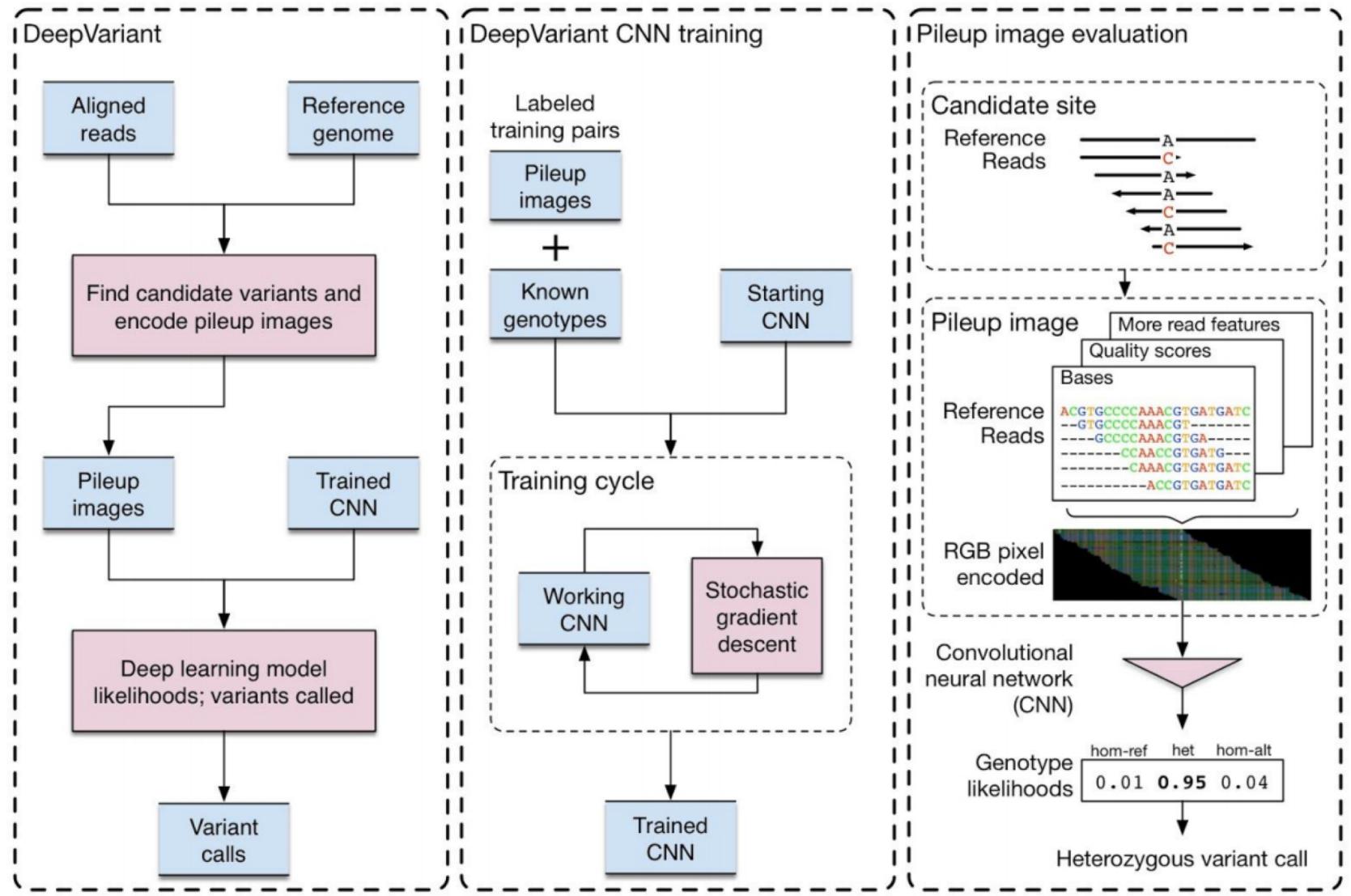
 © 1999 Nature America Inc. • <http://genetics.nature.com>

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



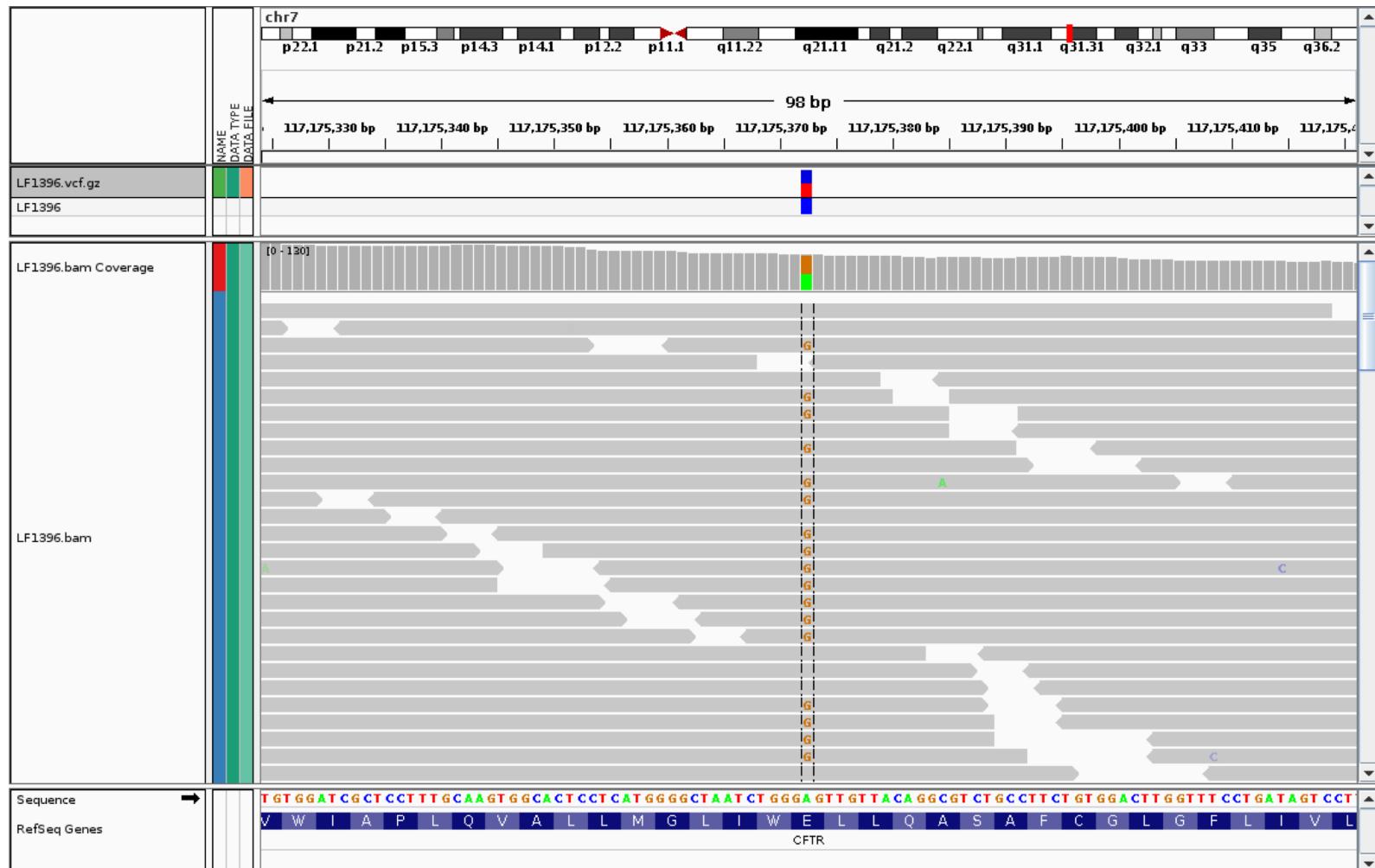
A universal SNP and small-indel variant caller using deep neural networks

Poplin et al. (2018) Nature Biotechnology. <https://doi.org/10.1038/nbt.4235>

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