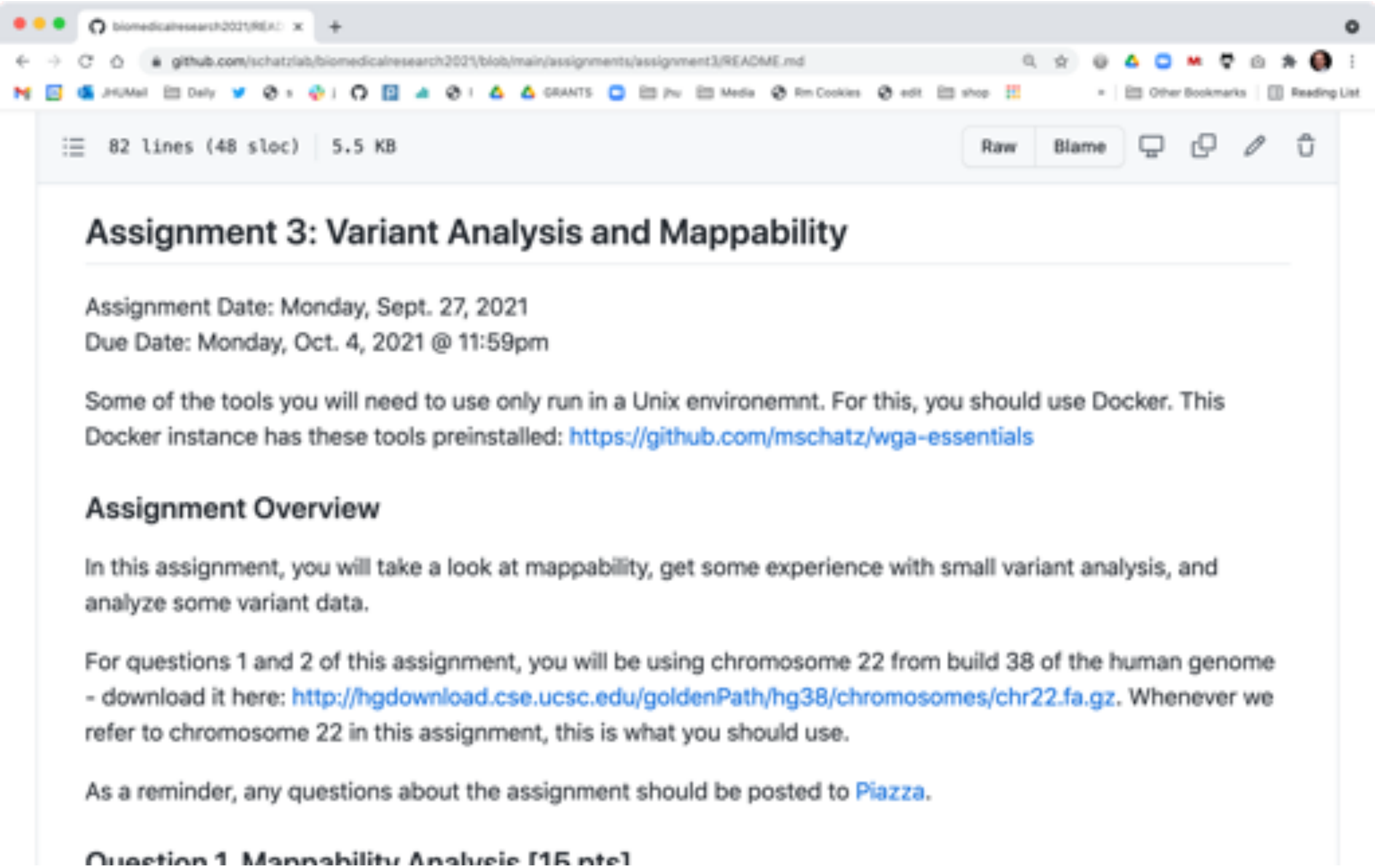
Variant Calling

Michael Schatz

Sept 29, 2021

Lecture 9: Computational Biomedical Research

Assignment 3: Variant Analysis & Mappability
Due Oct 4 @ 11:59pm


biomedicalresearch2021/READ
github.com/schatzlab/biomedicalresearch2021/blob/main/assignments/assignment3/README.md
JHUMail
Daily
GRANTS
jhu
Media
Rm Cookies
edit
shop
Other Bookmarks
Reading List
82 lines (48 sloc) 5.5 KB
Raw
Blame

Assignment 3: Variant Analysis and Mappability

Assignment Date: Monday, Sept. 27, 2021
Due Date: Monday, Oct. 4, 2021 @ 11:59pm

Some of the tools you will need to use only run in a Unix environemnt. For this, you should use Docker. This Docker instance 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.

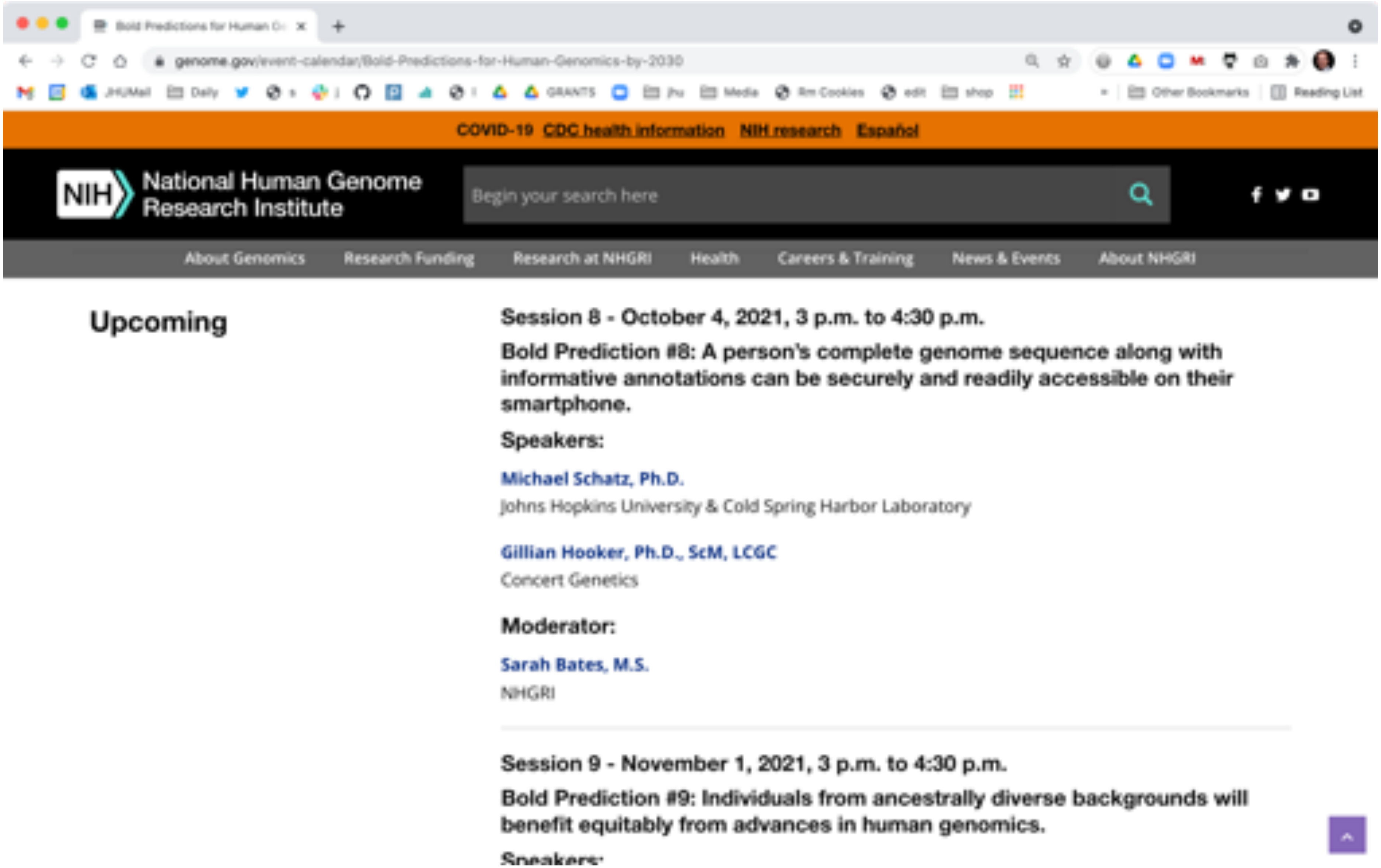
For questions 1 and 2 of this assignment, you will be using chromosome 22 from build 38 of the human genome - download it here: http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/chr22.fa.gz. Whenever we refer to chromosome 22 in this assignment, this is what you should use.

As a reminder, any questions about the assignment should be posted to Piazza.

Question 1. Mappability Analysis [15 pts]

https://github.com/schatzlab/biomedicalresearch2021

Monday's class



Upcoming

Session 8 - October 4, 2021, 3 p.m. to 4:30 p.m.

Bold Prediction #8: A person's complete genome sequence along with informative annotations can be securely and readily accessible on their smartphone.

Speakers:

Michael Schatz, Ph.D.
Johns Hopkins University & Cold Spring Harbor Laboratory

Gillian Hooker, Ph.D., ScM, LCGC
Concert Genetics

Moderator:

Sarah Bates, M.S.
NHGRI

Session 9 - November 1, 2021, 3 p.m. to 4:30 p.m.

Bold Prediction #9: Individuals from ancestrally diverse backgrounds will benefit equitably from advances in human genomics.

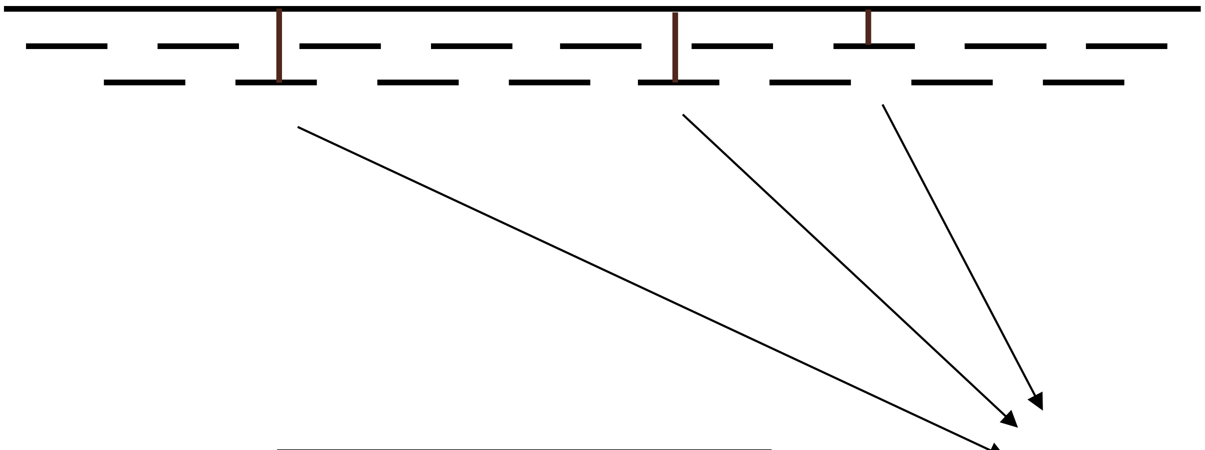
Speakers:

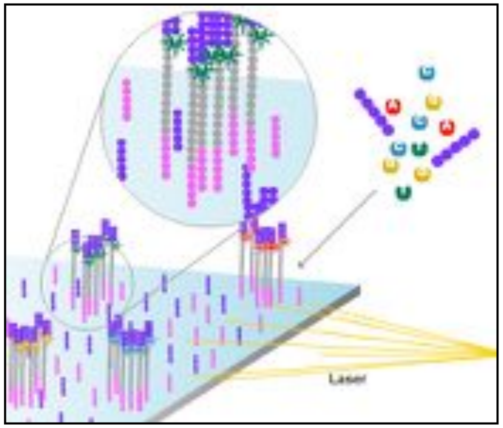
Registration: bit.ly/2XXhLYJ

Read Mapping

Personal Genomics

How does your genome compare to the reference?





Heart Disease

Cancer

Presidential smile

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
    3
```

Reverse Engineering Edit Distance

```
D(AGCACACA, 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…

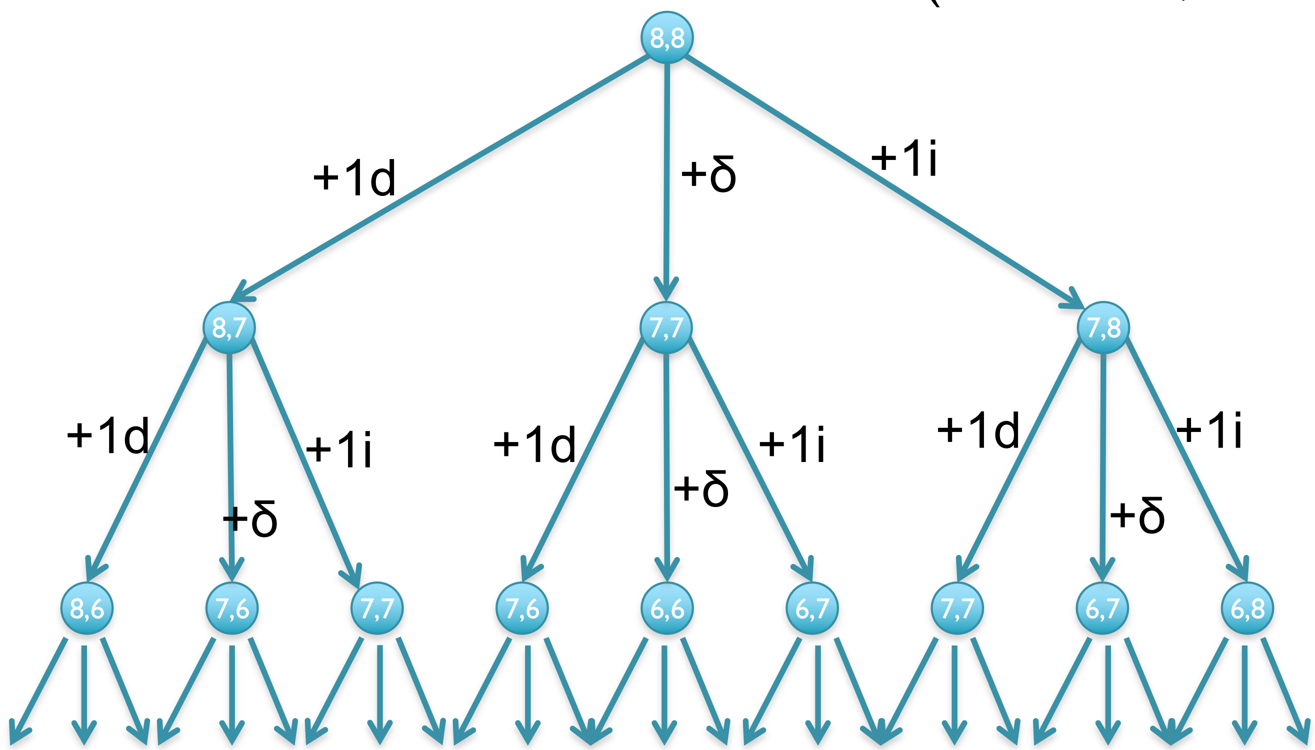
```
…M…     …I…     …D…
…X…     …-…     …X…
…Y…     …Y…     …-…
```

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’ ≤ i and j’ ≤ j.

$$\begin{aligned} D(\texttt{AGCACACA}, \texttt{ACACACTA}) = \min\{ & D(\texttt{AGCACACA}, \texttt{ACACACT}) + 1, \\ & D(\texttt{AGCACAC}, \texttt{ACACACTA}) + 1, \\ & D(\texttt{AGCACAC}, \texttt{ACACACT}) + \delta(\texttt{A}, \texttt{A})\} \end{aligned}$$



[What is the running time?]

Dynamic Programming

- We could code this as a recursive function call...

  ...with an exponential number of function evaluations

- There are only (n+1) x (m+1) pairs i and j
  - We are evaluating D(i,j) multiple times

- Compute D(i,j) bottom up.
  - Start with smallest (i,j) = (1,1).
  - Store the intermediate results in a table.
    - Compute D(i,j) *after* D(i-1,j), D(i,j-1), and D(i-1,j-1)

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

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

```
D(i,j) = min {
        D(i-1,j) + 1,            // align 0 chars from S, 1 from T
        D(i,j-1) + 1,            // align 1 chars from S, 0 from T
        D(i-1,j-1) + δ(S(i),T(j)) // align 1+1 chars
    }
```

[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								
G	2								
C	3								
A	4								
C	5								
A	6								
C	7								
A	8								

[What does the initialization mean?]

Dynamic Programming Matrix

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

$$D[A,A] = \min\{D[A,]+1,\ D[,A]+1,\ D[,]+\delta(A,A)\}$$

Dynamic Programming Matrix

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

D[A,AC] = min{D[A,A]+1, D[,AC]+1, D[,A]+δ(A,C)}

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					
G	2								
C	3								
A	4								
C	5								
A	6								
C	7								
A	8								

$D[A,ACA] = \min\{D[A,AC]+1,\ D[,ACA]+1,\ D[,AC]+\delta(A,A)\}$

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								
C	3								
A	4								
C	5								
A	6								
C	7								
A	8								

D[A,ACACACTA] = 7

```
-------A
*******|
ACACACTA
```

[What about the other A?]

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								
A	4								
C	5								
A	6								
C	7								
A	8								

D[AG,ACACACTA] = 7

```
----AG--
****|***
ACACACTA
```

Dynamic Programming Matrix

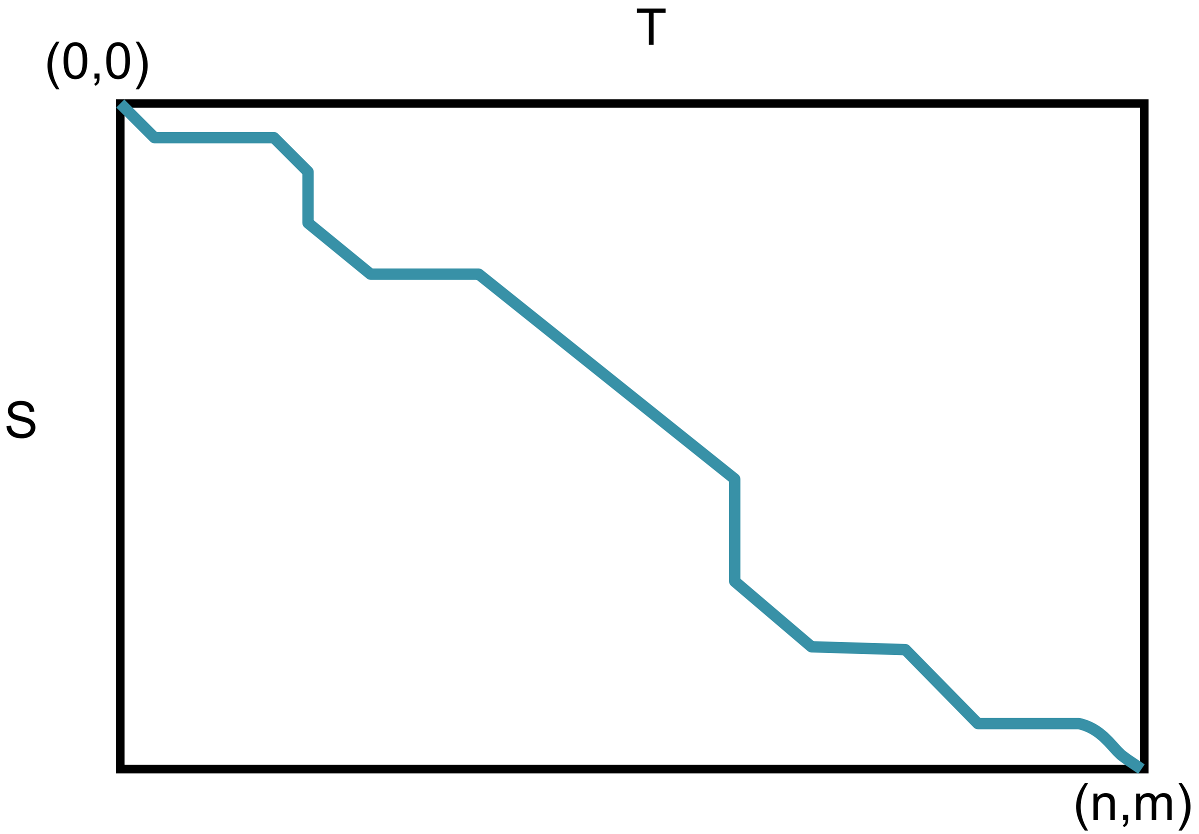
		A	C	A	C	A	C	T	A
	0	1	2	3	4	5	6	7	8
A	1	0	1	2	3	4	5	6	7
G	2	1	1	2	3	4	5	6	7
C	3	2	1	2	2	3	4	5	6
A	4	3	2	1	2	2	3	4	5
C	5	4	3	2	1	2	2	3	4
A	6	5	4	3	2	1	2	3	3
C	7	6	5	4	3	2	1	2	3
A	8	7	6	5	4	3	2	2	2

D[AGCACACA,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)

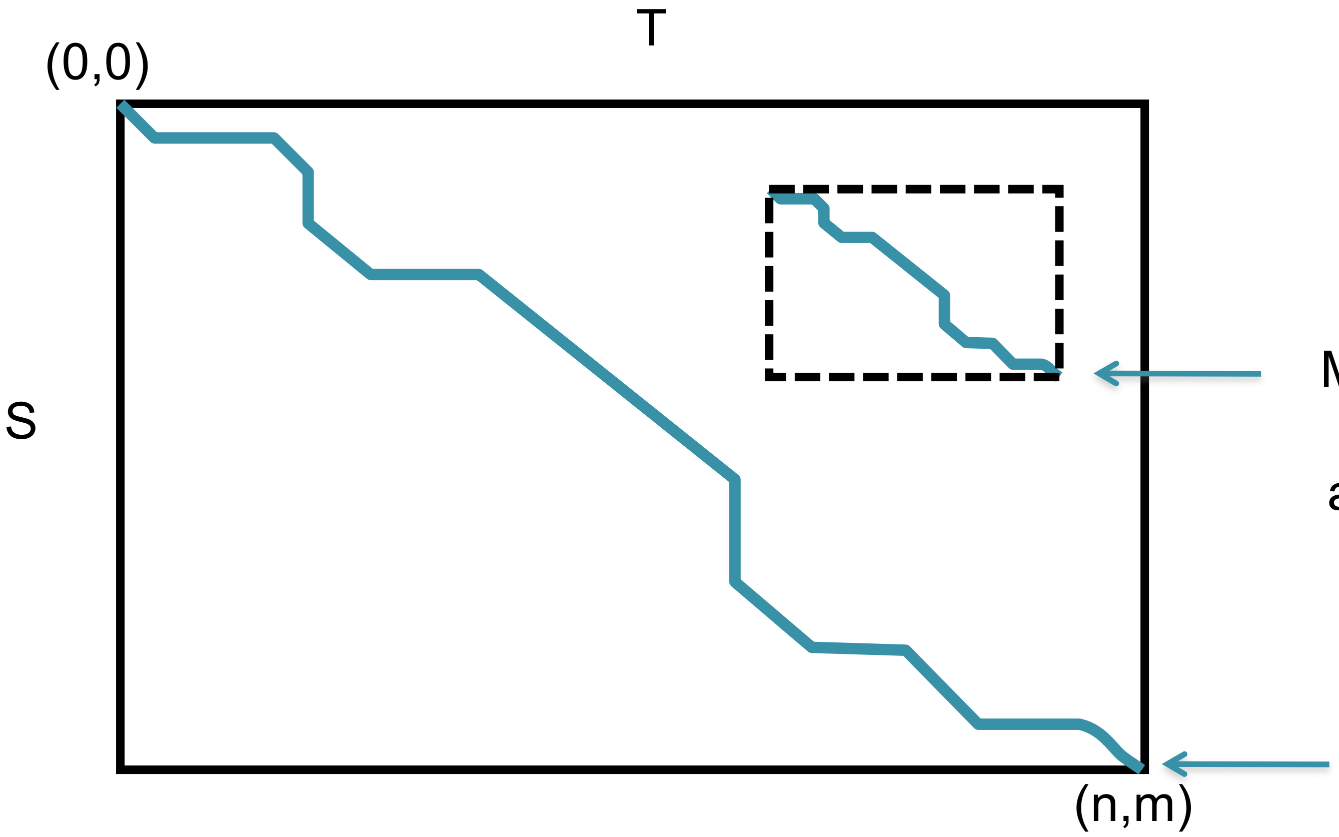
Nathan Edwards

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



Max score for local alignment

Global alignment always ends in the corner

Nathan Edwards

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
```

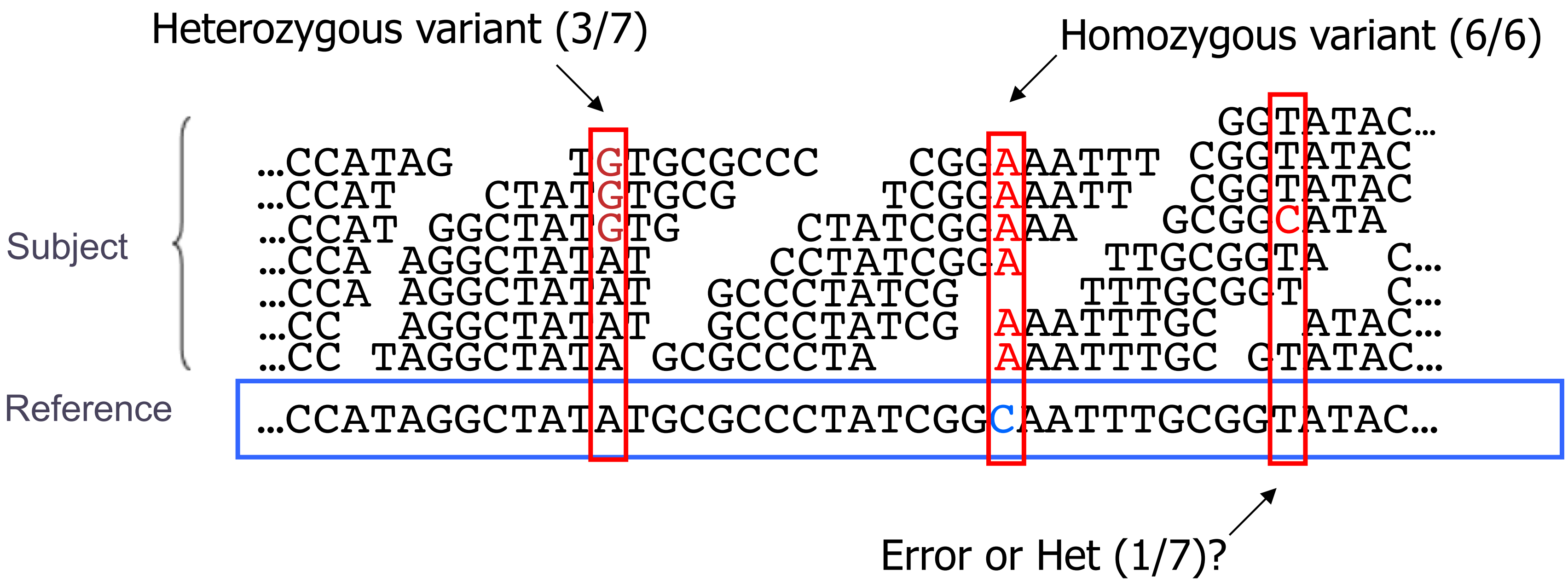
- Local Alignment—better alignment to find conserved segment

```
                  tccCAGTTATGTCAGgggacacgagcatgcagagac
                     ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

bioalgorithms.info

Part 2:Variant Calling

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

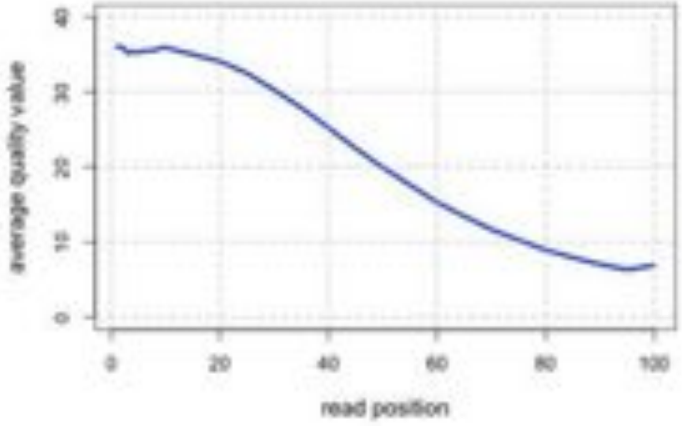
```
…CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC…
```

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



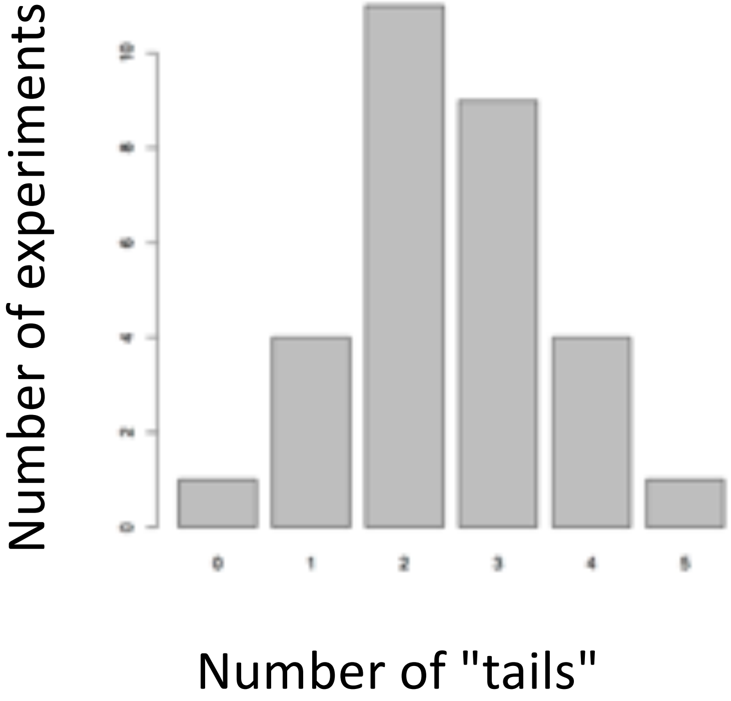
The Binomial Distribution: Adventures in Coin Flipping

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:

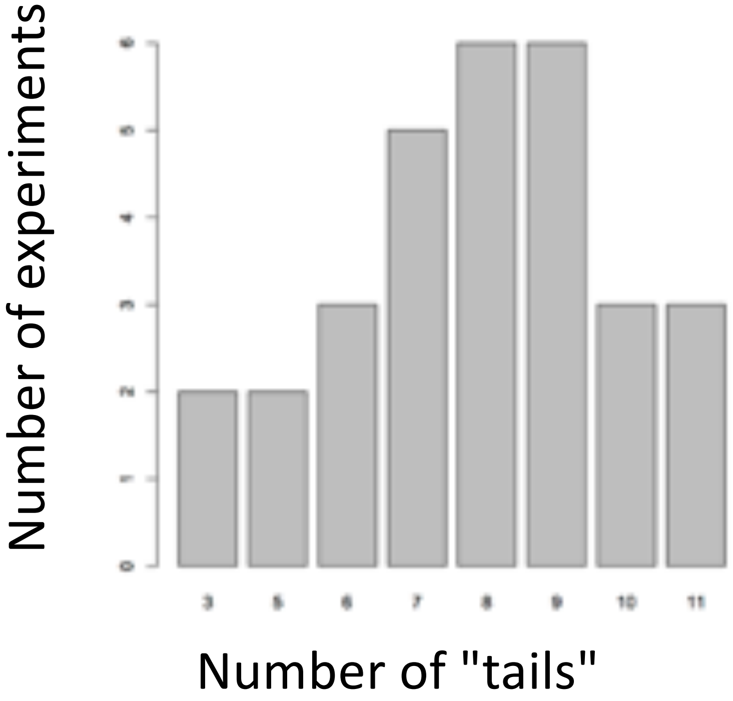
```
barplot(table(rbinom(30, 5, 0.5)))
```

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:

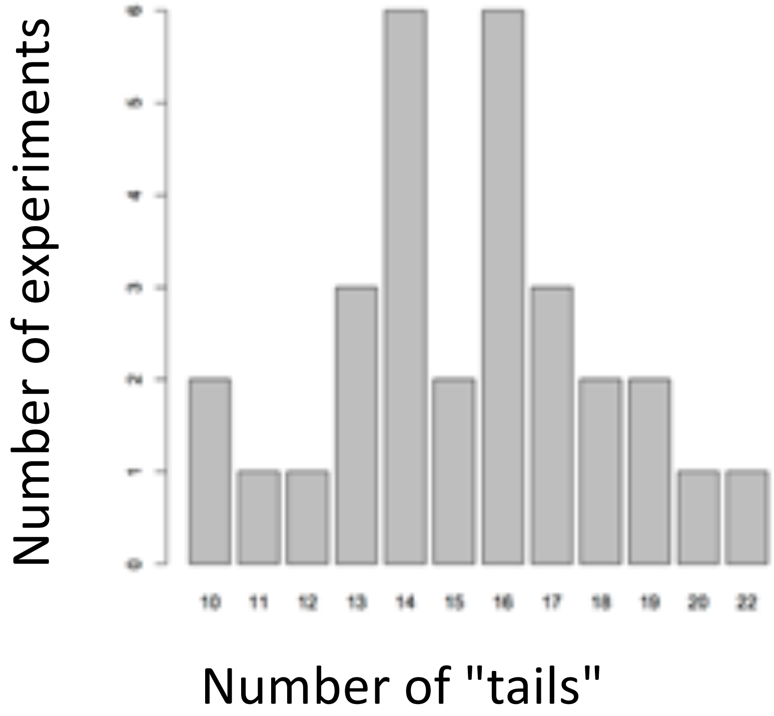
```
barplot(table(rbinom(30, 15, 0.5)))
```

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:

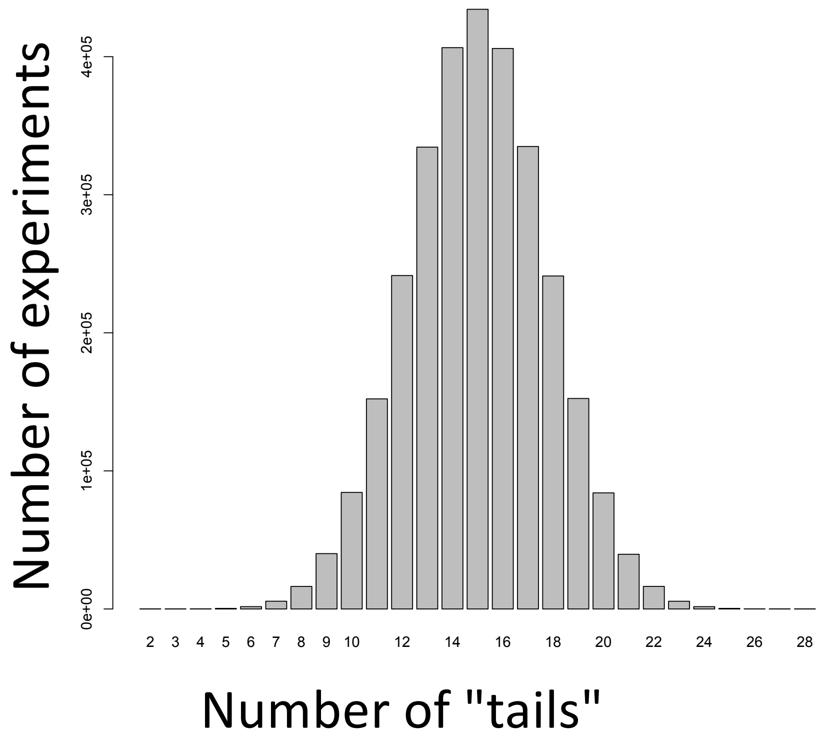
```
barplot(table(rbinom(30, 30, 0.5)))
```

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:

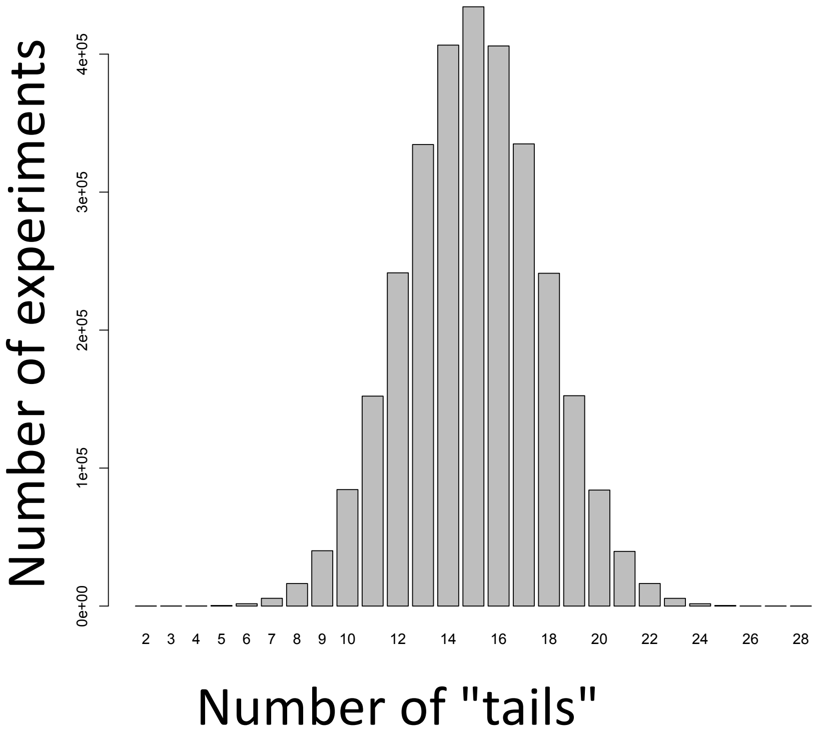
```
barplot(table(rbinom(3e6, 30, 0.5)))
```

3M experiments (students tossing coins)

30 tosses each

Probability of Tails

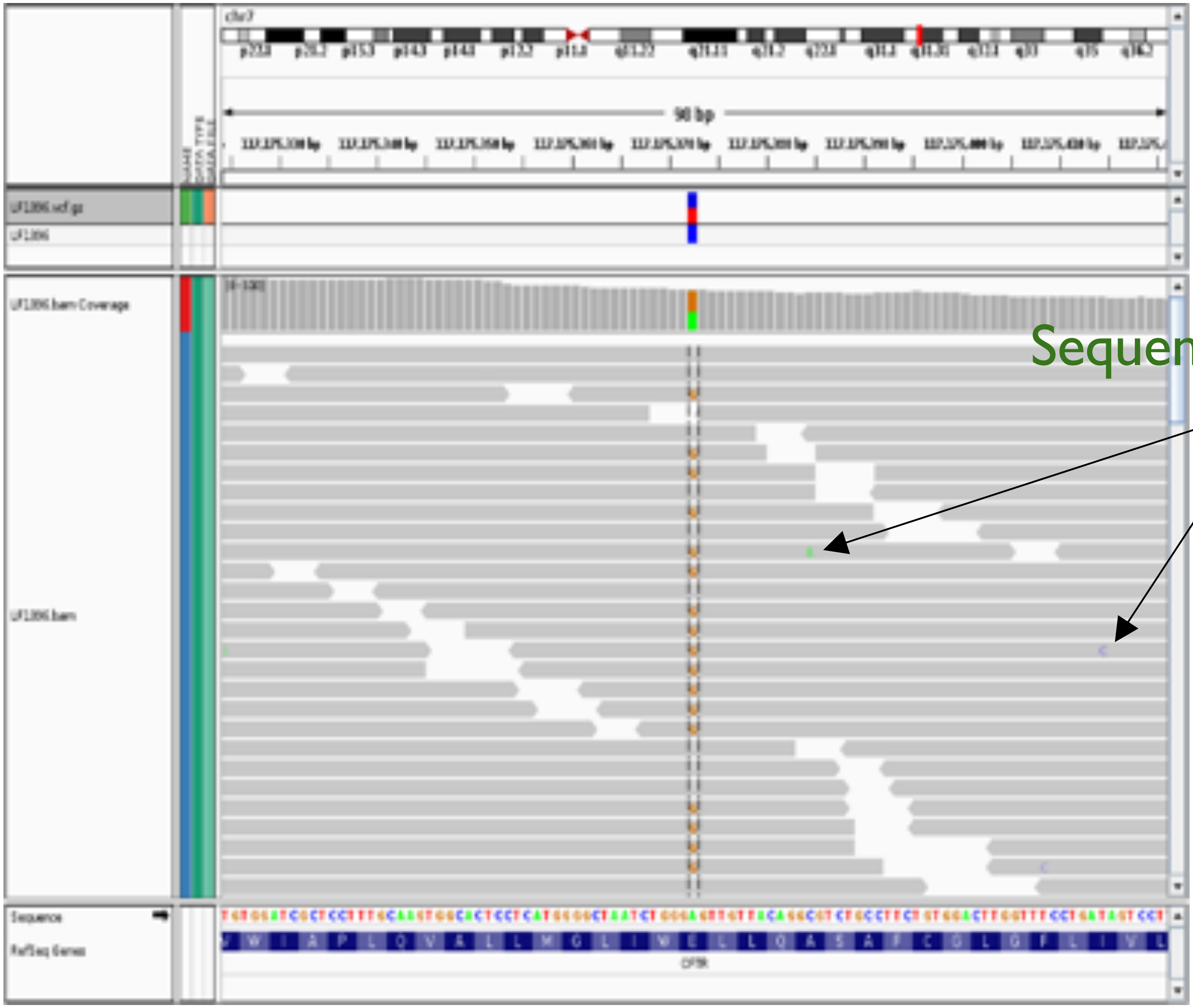
So, with 30 tosses (reads), we are much more likely to see an even mix of alternate and reference alleles at a heterozygous locus in a genome



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)

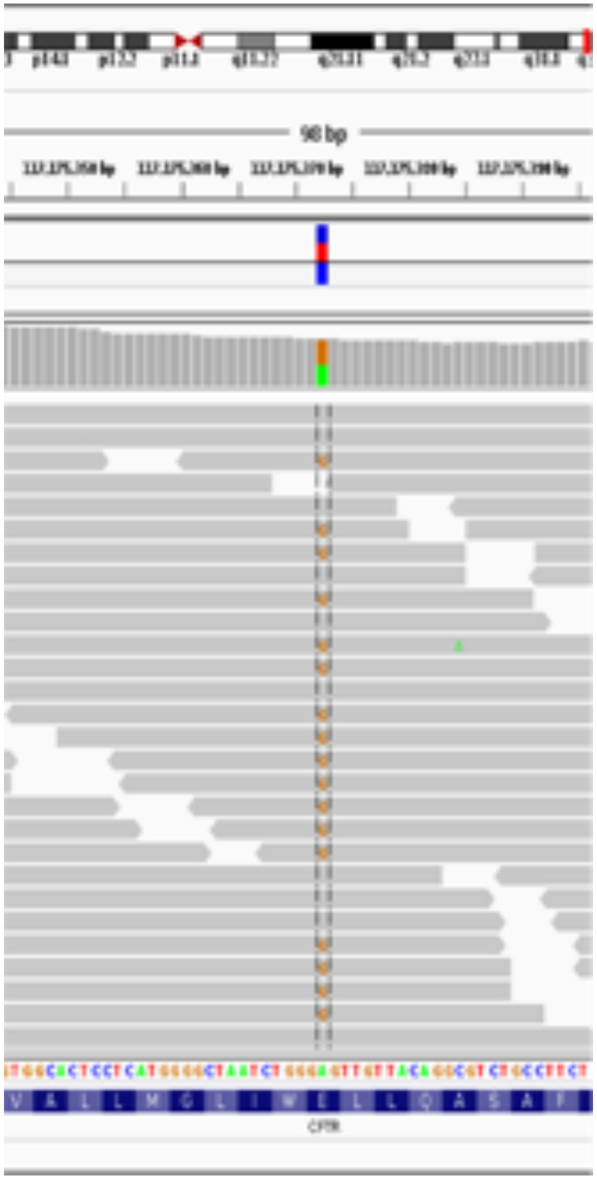
Sequencing errors fall out as noise (most of the time)



Sequencing errors

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

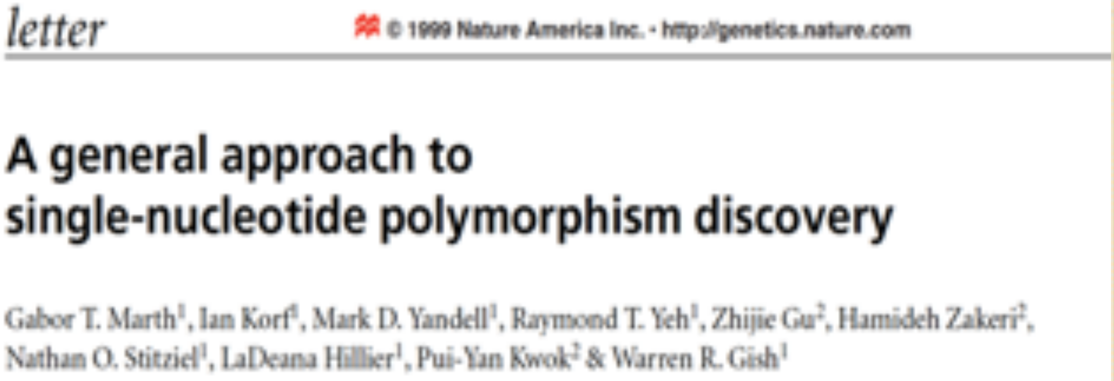
What information is needed to decide if a variant exists?



- 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 © 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]

Netscape: PolyBayes Web site
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: What's Related
PolyBayes
Site map
Evaluate
Reset default values

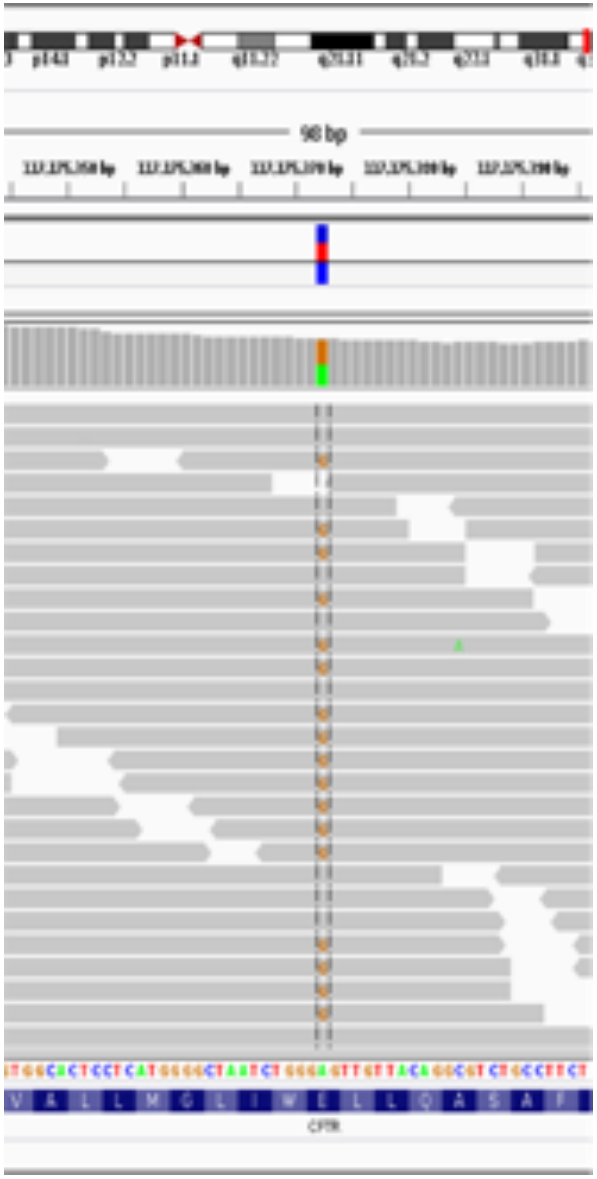
Results

Description	Symbol	Value
Probability of SNP	P(SNP)	[illegible]
Most likely variation	V(A)	A/G
Probability of variation	P(V(A))	[illegible]
Alignment depth	D	2

Its main innovation was the use of Bayes's theorem

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

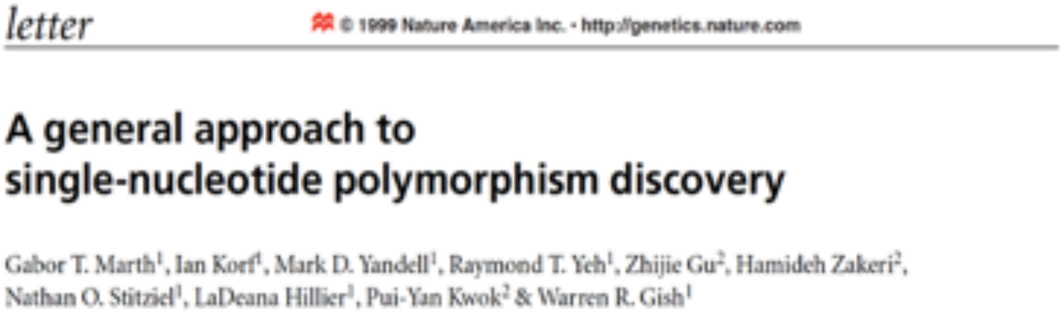
Bayesian SNP calling



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

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

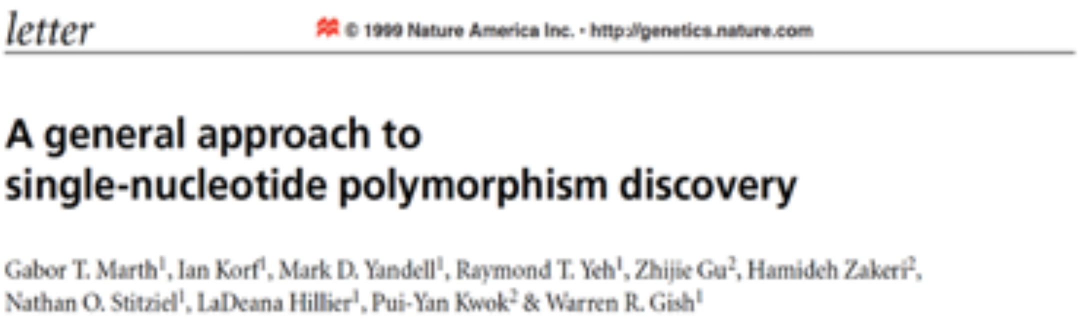
Expected (prior) polymorphism rate

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

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

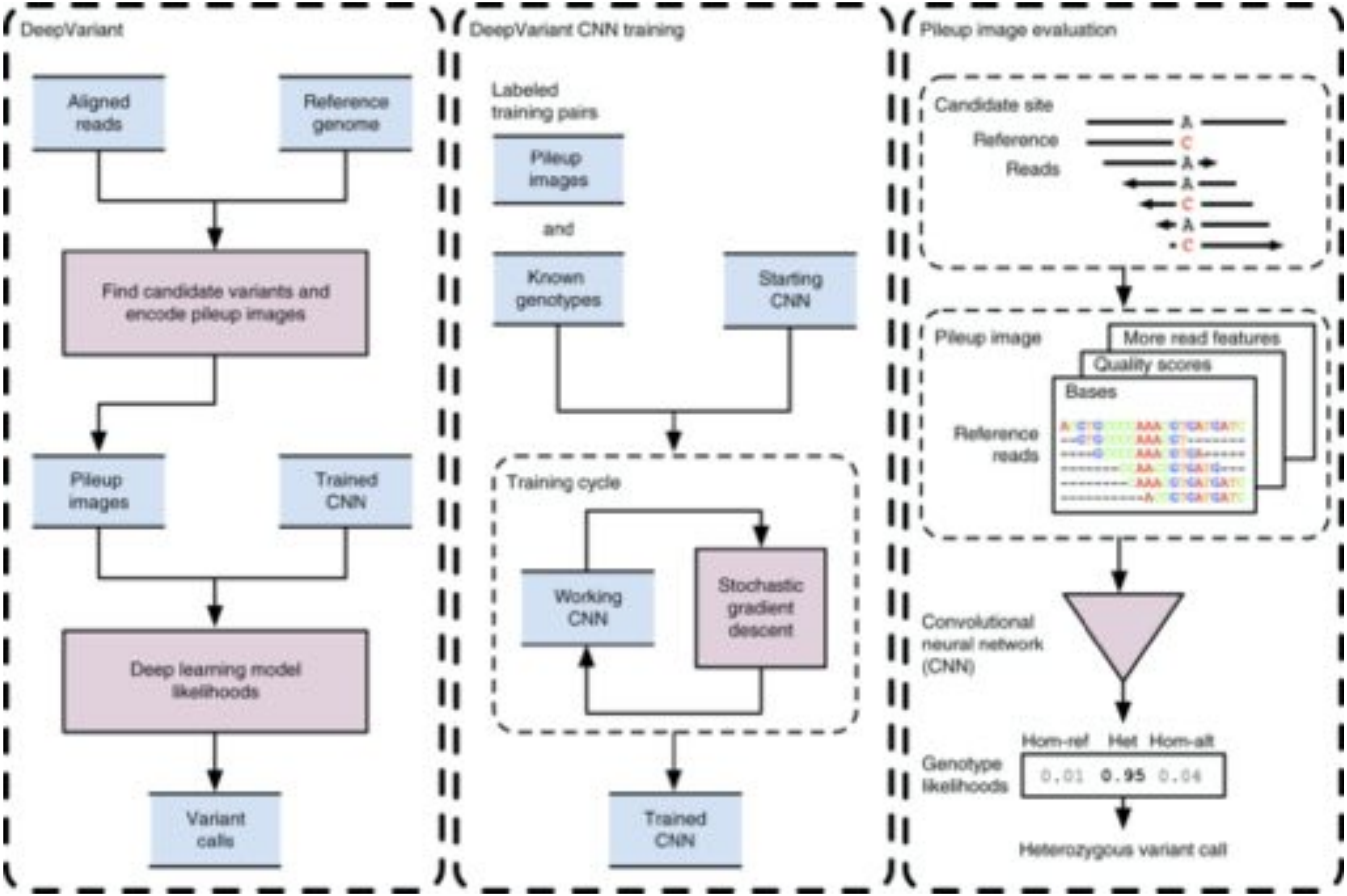
© 1999 Nature America Inc. • http://genetics.nature.com

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

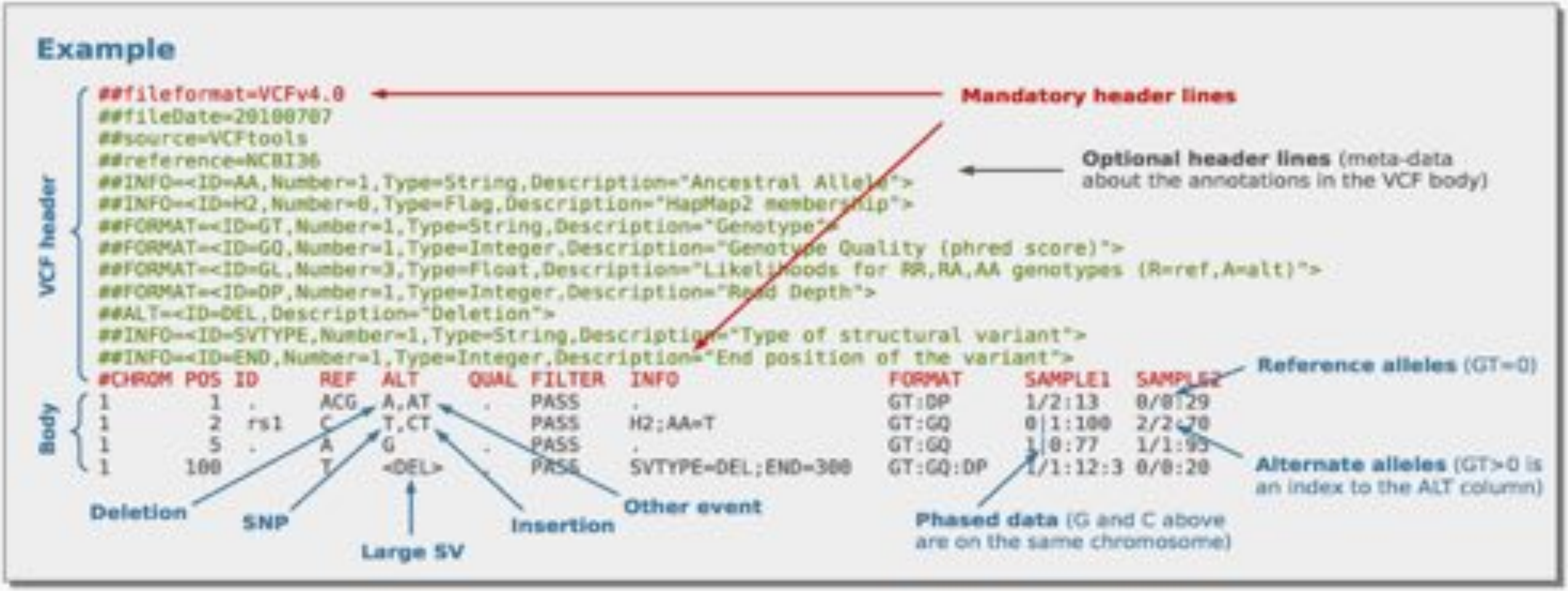
http://www.nature.com/ng/journal/v23/n4/full/ng1299_452.html



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

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

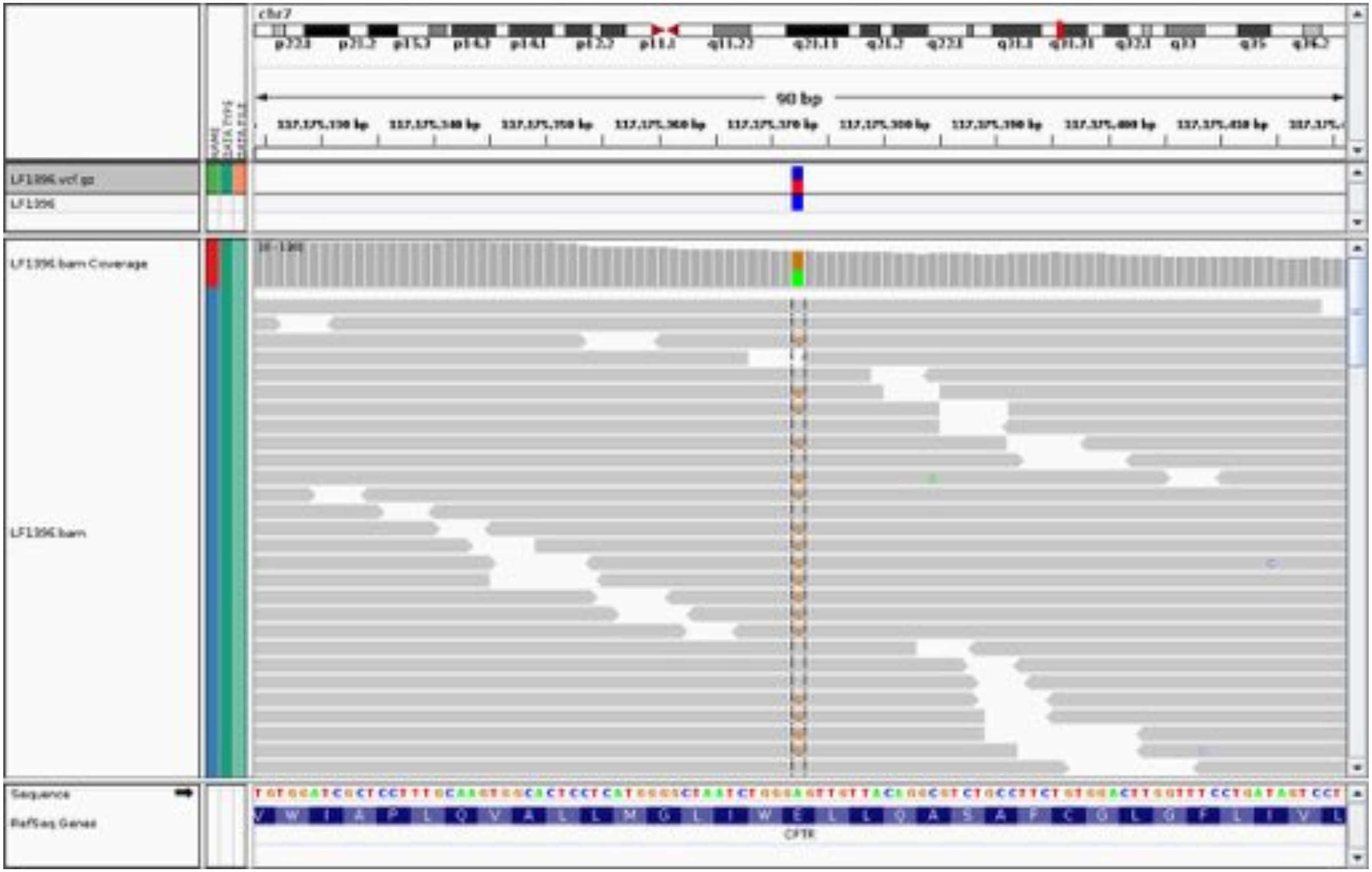
VCF Format



Example
VCF header
Body
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

```
##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
```

VCF Format



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