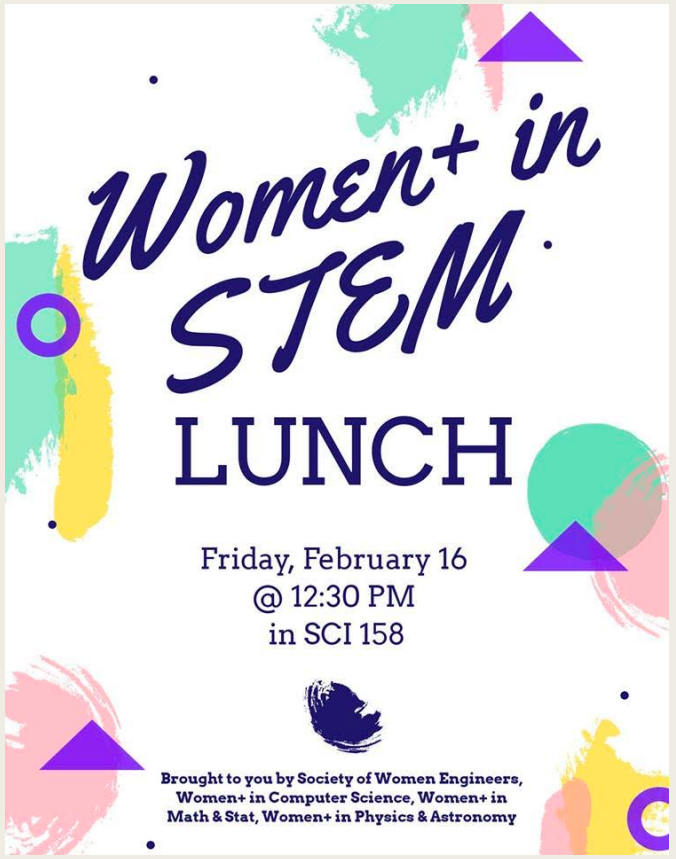
# CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

# Outline: Feb 16

- Finish: FM-index and practical read mapping
- Today: link between “getting data” and “using data”
- Introduce phylogenetic trees and evolutionary process of mutations




Notes:

- Remote edit new instructions on Piazza (and website)
- Welcome to use pandas (make sure it works on lab machines)

# Lab 4 Notes

# Notes about Lab 4

- For the reverse BWT part (i.e. getting back the original string from the BWT), you should not be creating any additional data structures beyond what you needed for mapping a read

- Optional: implement sorting the rotations (cyclic permutations) from the original string only

# Notes about Lab 4

- For the reverse BWT part (i.e. getting back the original string from the BWT), you should not be creating any additional data structures beyond what you needed for mapping a read

- Optional: implement sorting the rotations (cyclic permutations) from the original string only

# Notes about Lab 4

- For the reverse BWT part (i.e. getting back the original string from the BWT), you should not be creating any additional data structures beyond what you needed for mapping a read

- Optional: implement sorting the rotations (cyclic permutations) from the original string only

# Notes about Lab 4

- For the reverse BWT part (i.e. getting back the original string from the BWT), you should not be creating any additional data structures beyond what you needed for mapping a read

- Optional: implement sorting the rotations (cyclic permutations) from the original string only

`g > d therefore Rotation(i) > Rotation(j)`

*i* *j*

`.....abcg.....abcd.....$`

*A*

`abcd.....$.....abcd.....` *j*

⋮

`abcg.....abcg.....$.....` *i*

# Quick jump back to local alignment

From Lior Pachter's (professor at Caltech) blog

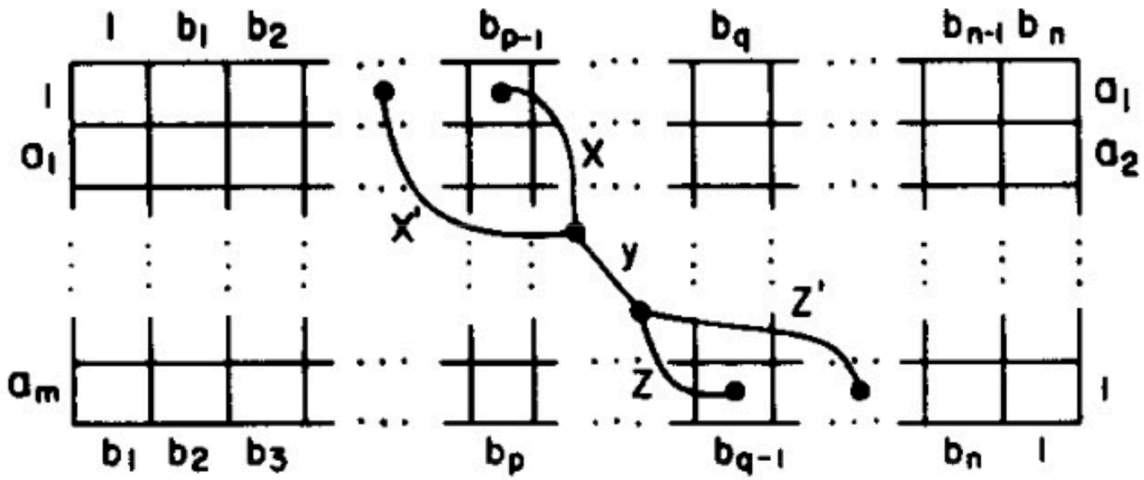
Feb. 15, 2018

Bioinformatics is not very different from mathematics; the literature is populated with many Amerindian eggs. My favorite example is the Smith-Waterman algorithm, an algorithm for local alignment published by Temple Smith and Michael Waterman in 1981. The Smith-Waterman algorithm is a simple modification of the Needleman-Wunsch algorithm:

1970

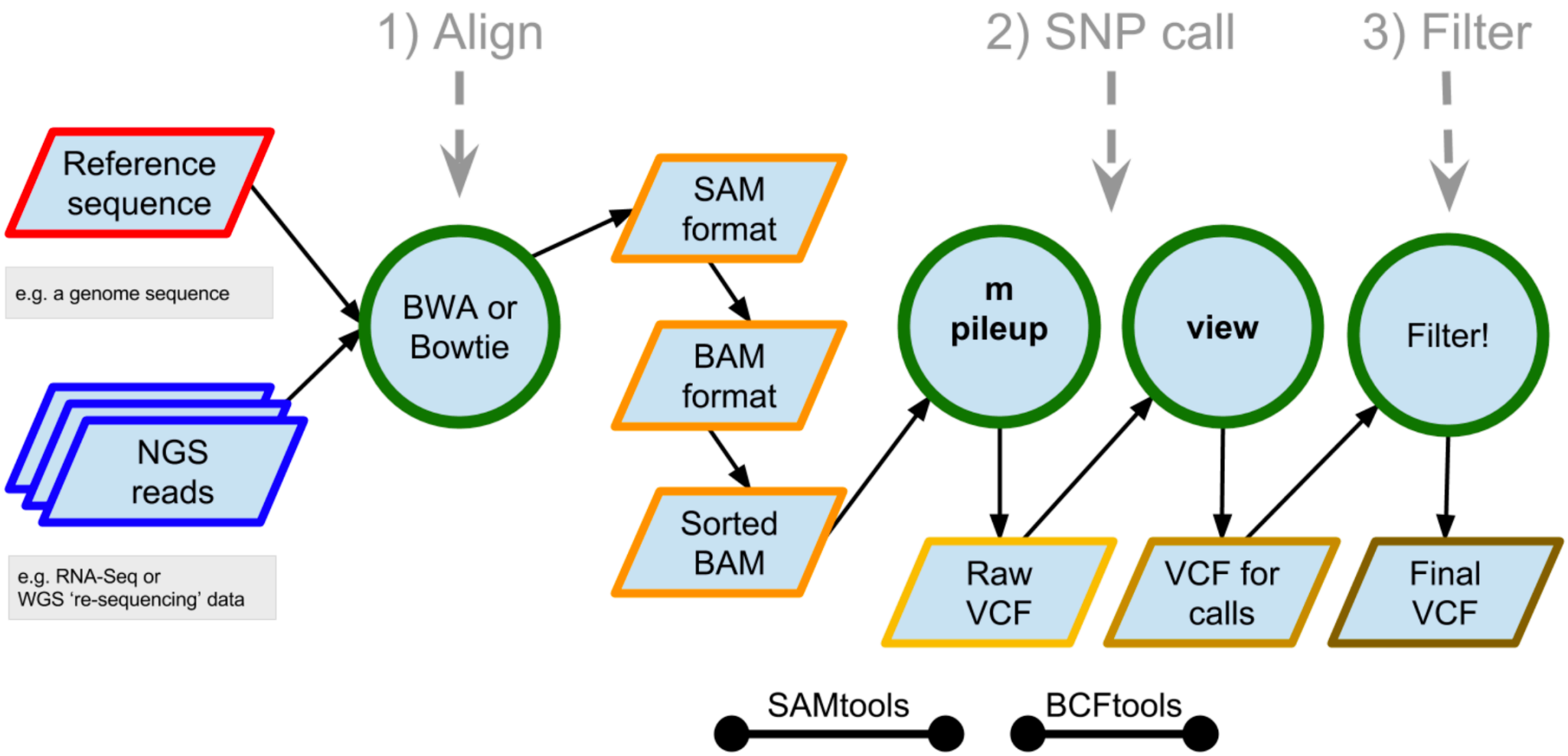
| | Smith–Waterman algorithm | Needleman–Wunsch algorithm |
|---|---|---|
| Initialization | First row and first column are set to 0 | First row and first column are subject to gap penalty |
| Scoring | Negative score is set to 0 | Score can be negative |
| Traceback | Begin with the highest score, end when 0 is encountered | Begin with the cell at the lower right of the matrix, end at top left cell |

The table above shows the differences. **That's it!** This table made for a (highly cited) paper. Just initialize the Needleman-Wunsch algorithm with zeroes instead of a gap penalty, set negative scores to 0, trace back from the highest score. In fact, it's such a minor modification that when I first learned the details of the algorithm I thought "This is obvious! After all, it's *just* the Needleman-Wunsch algorithm. Why does it even have a name?! Smith and Waterman got a highly cited paper?! For *this?!*" My skepticism lasted only as long as it took me to discover and read Peter Sellers' 1980 paper attempting to solve the same problem. It's a lot more complicated, relying on the idea of "inductive steps", and requires untangling mysterious diagrams such as:



# Practical Read Mapping

# Pipeline overview




Credit: Dan Bolster (EBI)

# Bowtie and BWA (posted reading)

- First practical read aligners to use the Burrows-Wheeler Transform

**Fast and accurate short read alignment with Burrows–Wheeler transform**

Heng Li, Richard Durbin ✉ Author Notes

*Bioinformatics*, Volume 25, Issue 14, 15 July 2009 Pages 1754–1760,
https://doi.org/10.1093/bioinformatics/btp324
**Published:** 18 May 2009 **Article history** ▾

BWA

Bowtie

**Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**

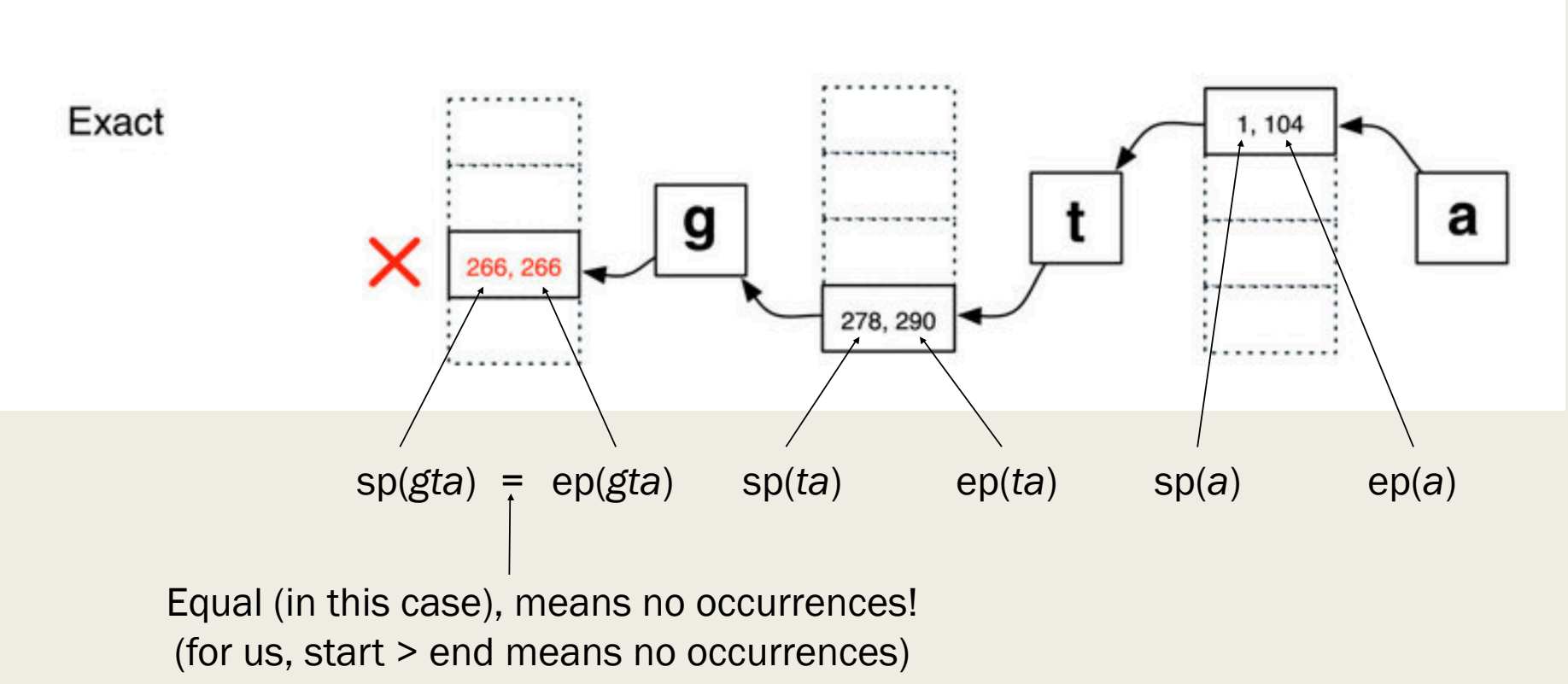
Ben Langmead ✉, Cole Trapnell, Mihai Pop and Steven L Salzberg

*Genome Biology* 2009 **10**:R25
https://doi.org/10.1186/gb-2009-10-3-r25 | © Langmead et al.; licensee BioMed Central Ltd. 2009
Received: 21 October 2008 | Accepted: 4 March 2009 | Published: 4 March 2009

# Bowtie: exact matching

- Figure 2 from the Bowtie paper: exactly what we have done in class, except exclusive of end-point

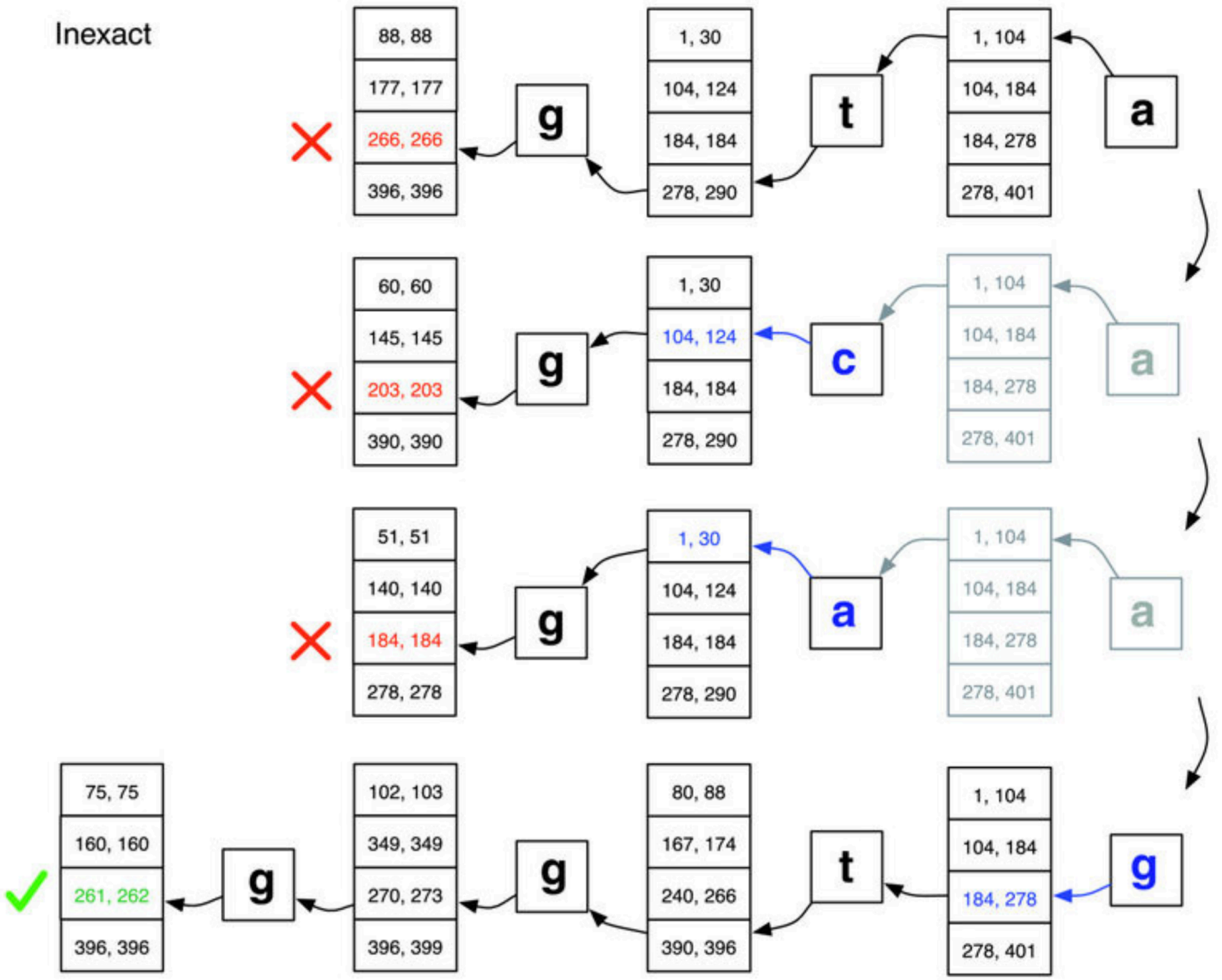
Read: 'ggta'



# Bowtie: inexact matching

Read: 'ggta'

Did not find the read, but we did find: 'ggtg'



# Comparison of Bowtie and BWA

**Table 2.**

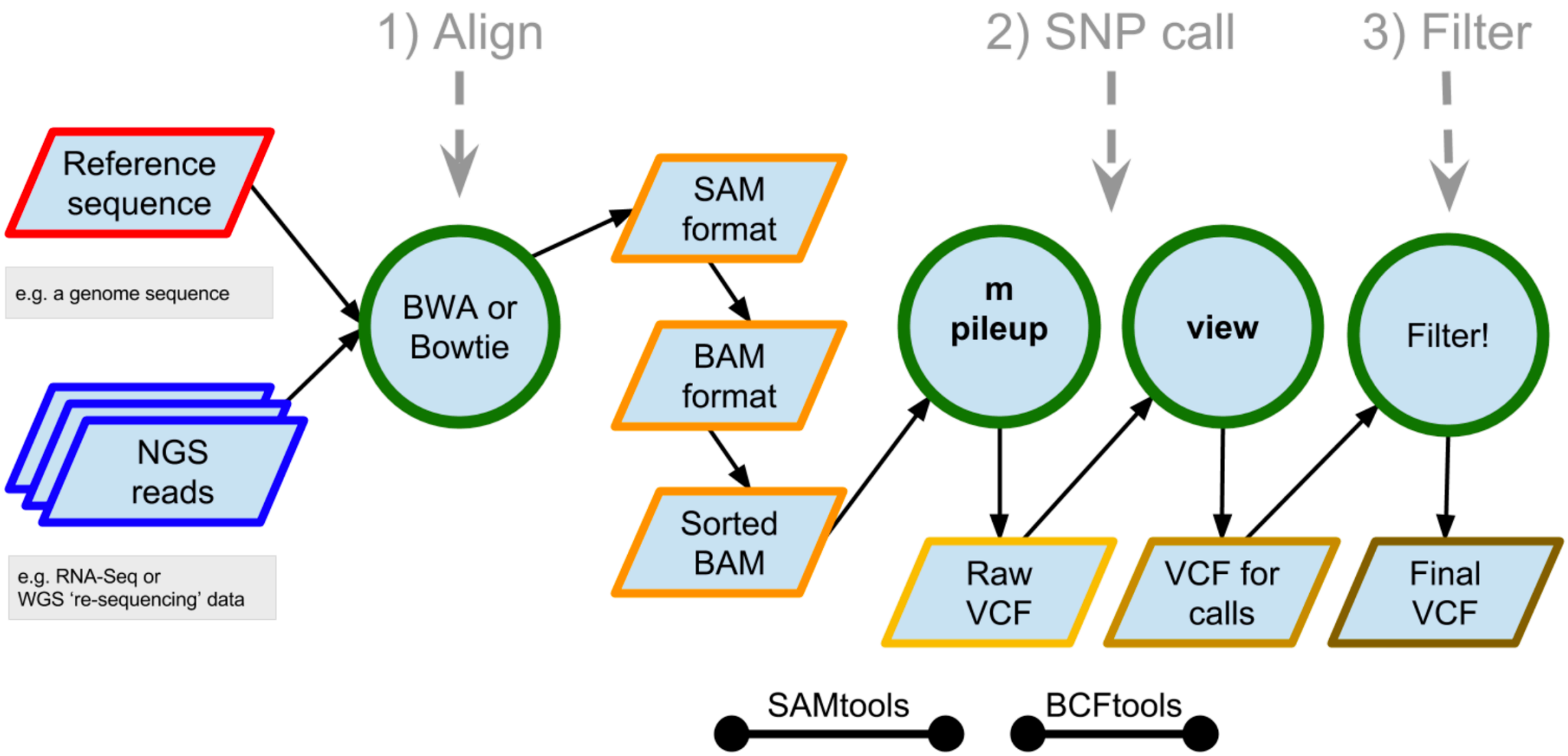
Evaluation on real data

| Program | Time (h) | Conf (%) | Paired (%) |
|---|---|---|---|
| Bowtie | 5.2 | 84.4 | 96.3 |
| BWA | 4.0 | 88.9 | 98.8 |
| MAQ | 94.9 | 86.1 | 98.7 |
| SOAP2 | 3.4 | 88.3 | 97.5 |

The 12.2 million read pairs were mapped to the human genome. CPU time in hours on a single core of a 2.5 GHz Xeon E5420 processor (Time), percent confidently mapped reads (Conf) and percent confident mappings with the mates mapped in the correct orientation and within 300 bp (Paired), are shown in the table.

# What comes after read-mapping?

# Pipeline overview

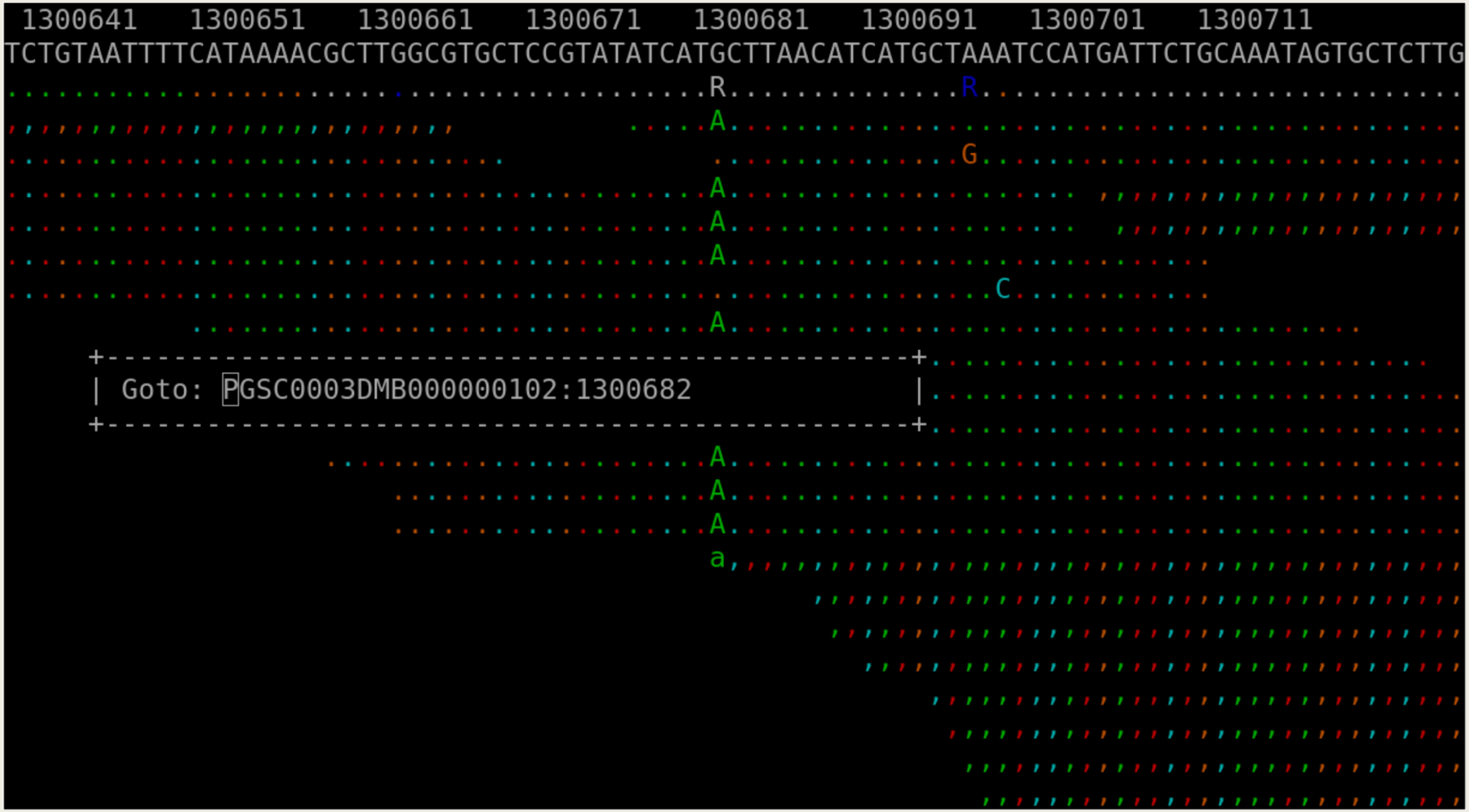



Credit: Dan Bolster (EBI)

# Variant calling: example of SAM “pileup”


Credit: Dan Bolster (EBI)

# Variant calling: example of SAM “pileup”

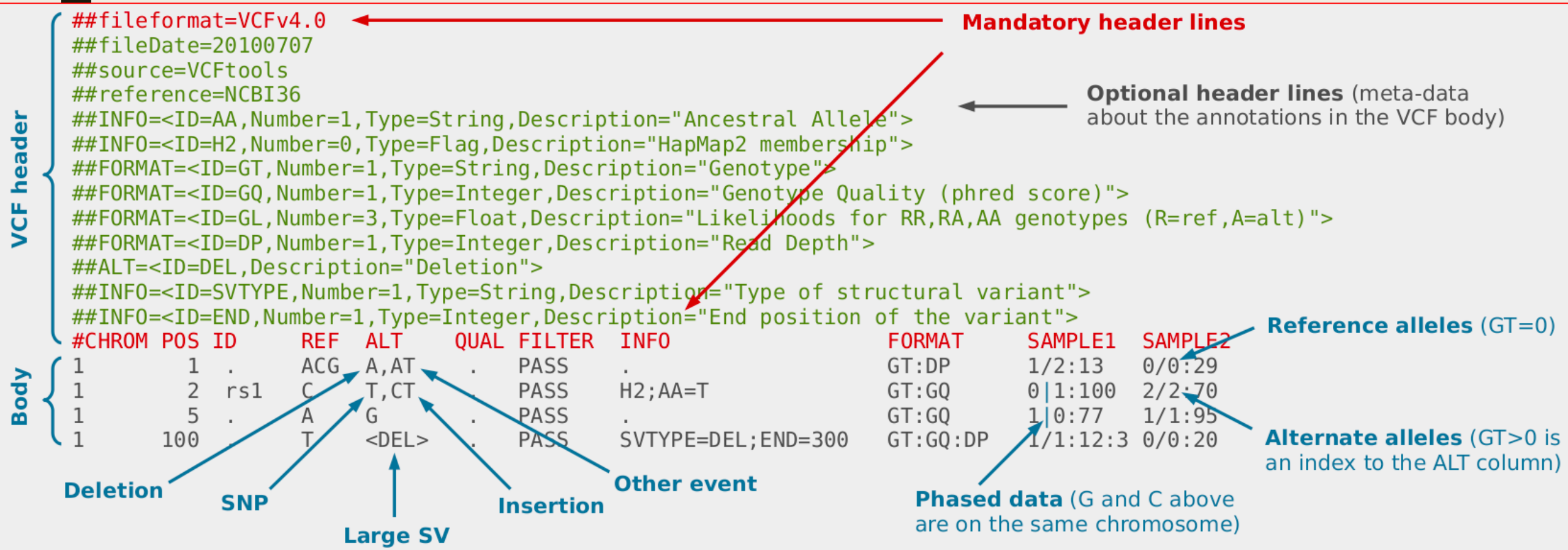



Credit: Dan Bolster (EBI)

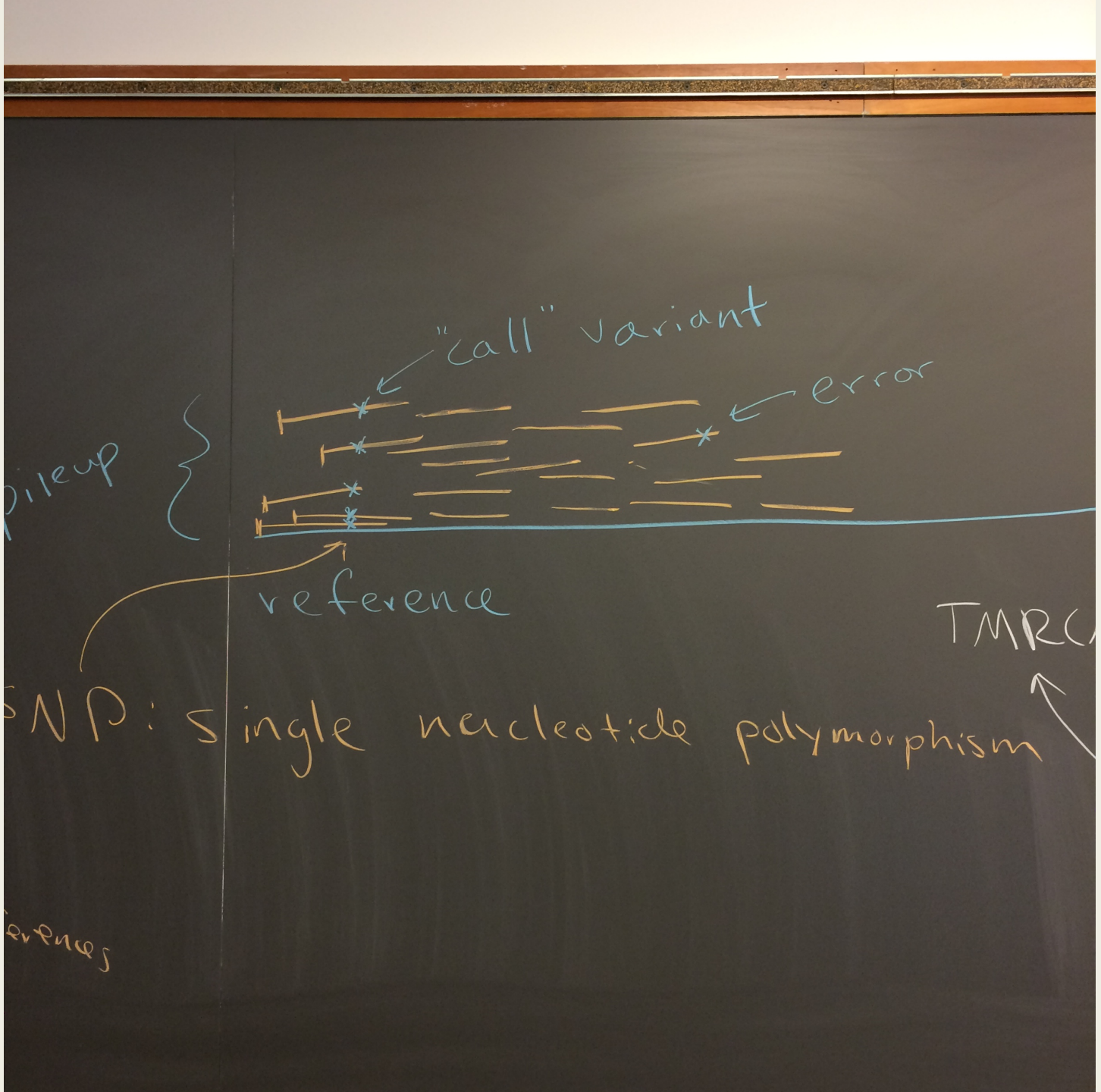
# SAM -> BAM -> sorted BAM -> VCF

- SAM: “Sequence Alignment/Map” file format
- Shows the locations of every read alignment, as well as the quality scores of each base of the alignment
- BAM: “Binary Alignment/Map” file format is a compressed version of SAM
- Next step: we need to sort the SAM or BAM file by the start position of each alignment
- Then we need to “call” the variation, finally giving us a VCF file
- VCF file: “Variant Call Format” shows the position of each SNP in order (usually we have one file for each chromosome)

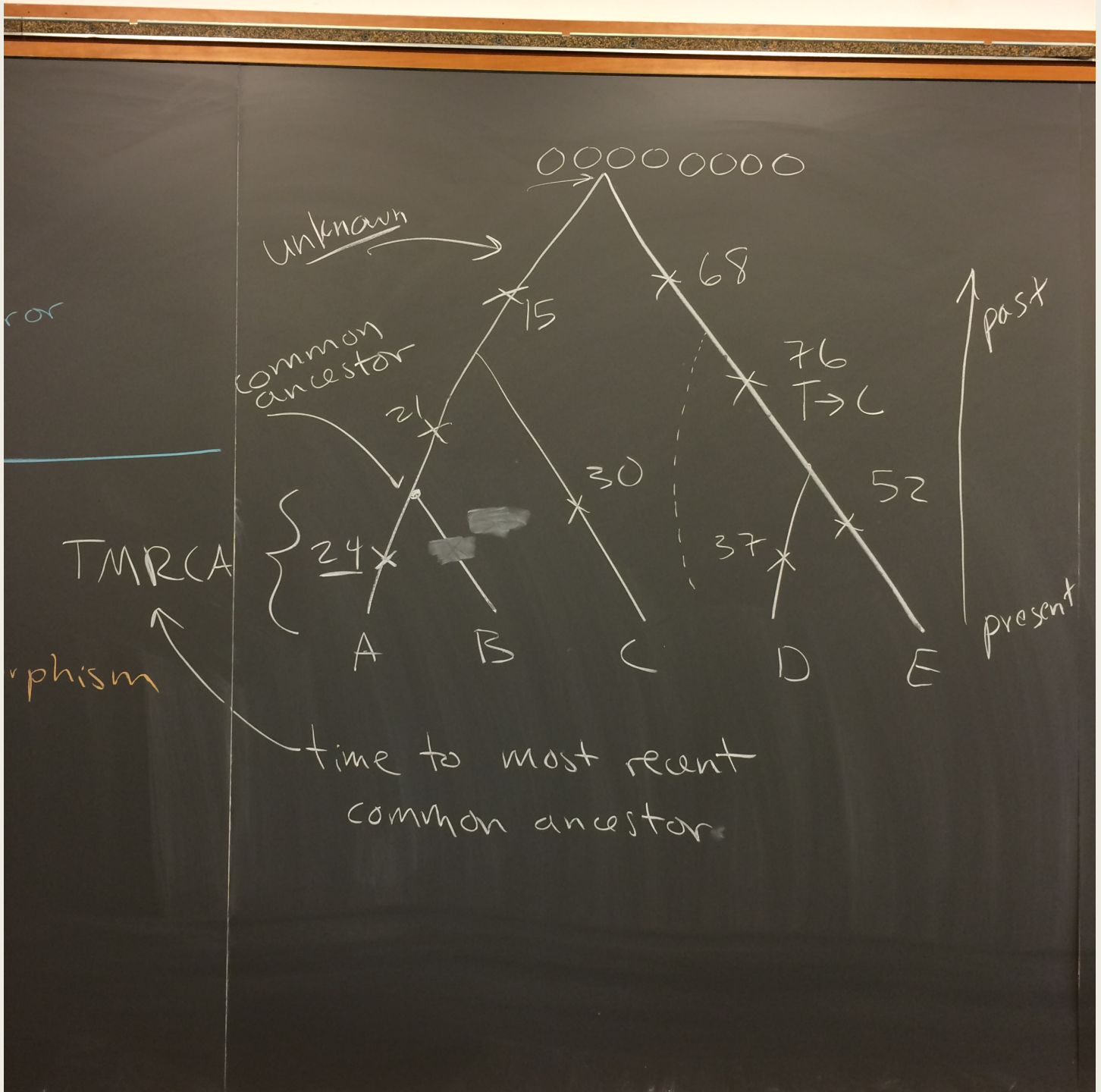
# VCF file format

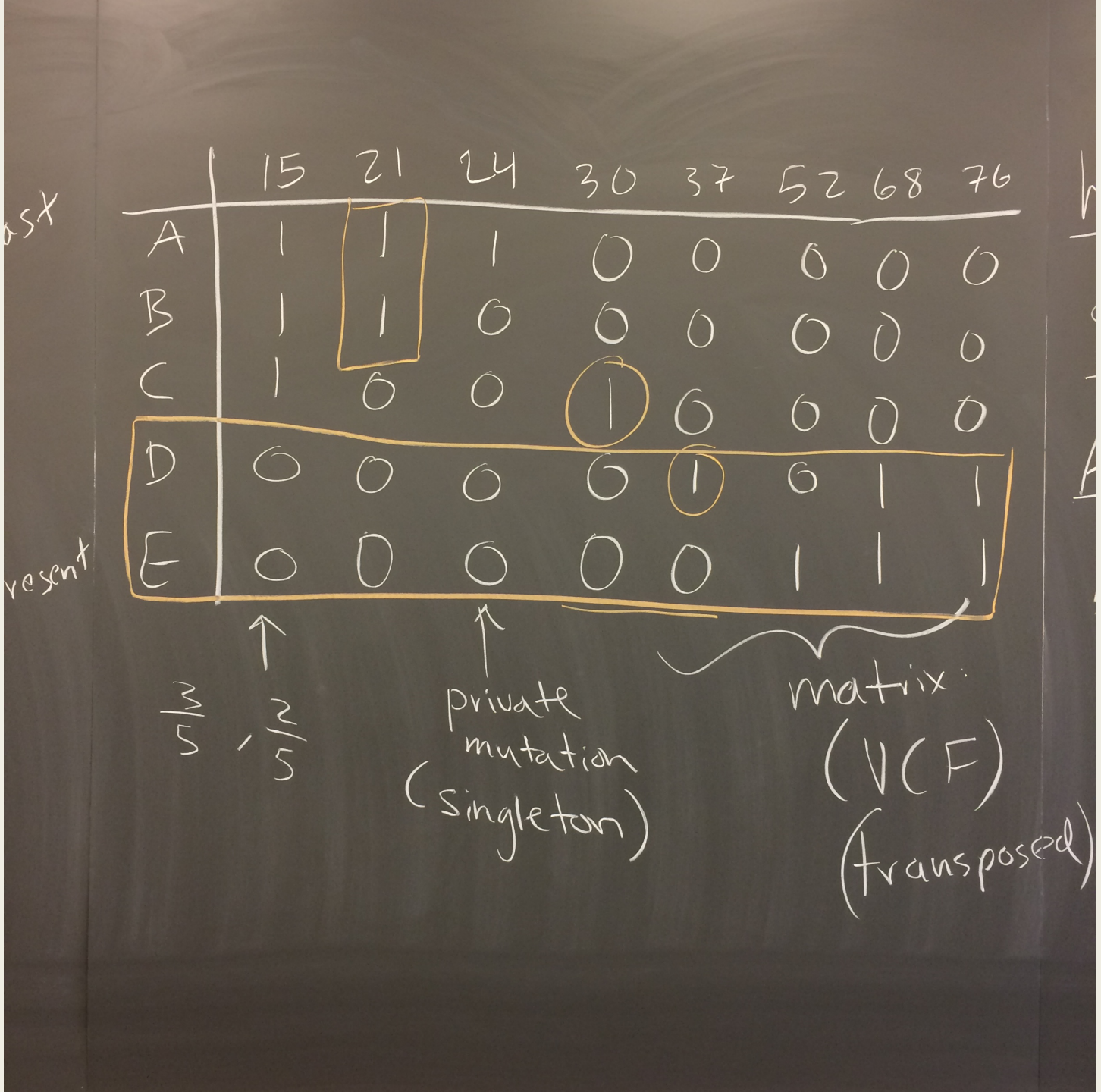



Credit: Dan Bolster (EBI)


"call" variant
error
pileup
reference
SNP: single nucleotide polymorphism
TMRC

# Begin: using variation to reconstruct evolutionary events


00000000
Unknown
15
68
common ancestor
21
76
T→C
30
52
24
37
TMRCA
A
B
C
D
E
past
present
time to most recent common ancestor


15 21 24 30 37 52 68 76
A 1 1 1 0 0 0 0 0
B 1 1 0 0 0 0 0 0
C 1 0 0 1 0 0 0 0
D 0 0 0 0 1 0 1 1
E 0 0 0 0 0 1 1 1
3/5
2/5
private mutation (singleton)
matrix (VCF) (transposed)

haplotype : a sequence from a single chromosome (humans are diploid)

SNP : single nucleotide polymorphism (caused by a single mutation)

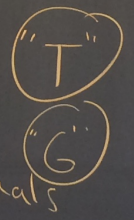
Allele : a type of observed variation

Ancestral : allele of the ancestor of all seqs in sample ("G")

Derived : allele of the variant observed in sample ("T")

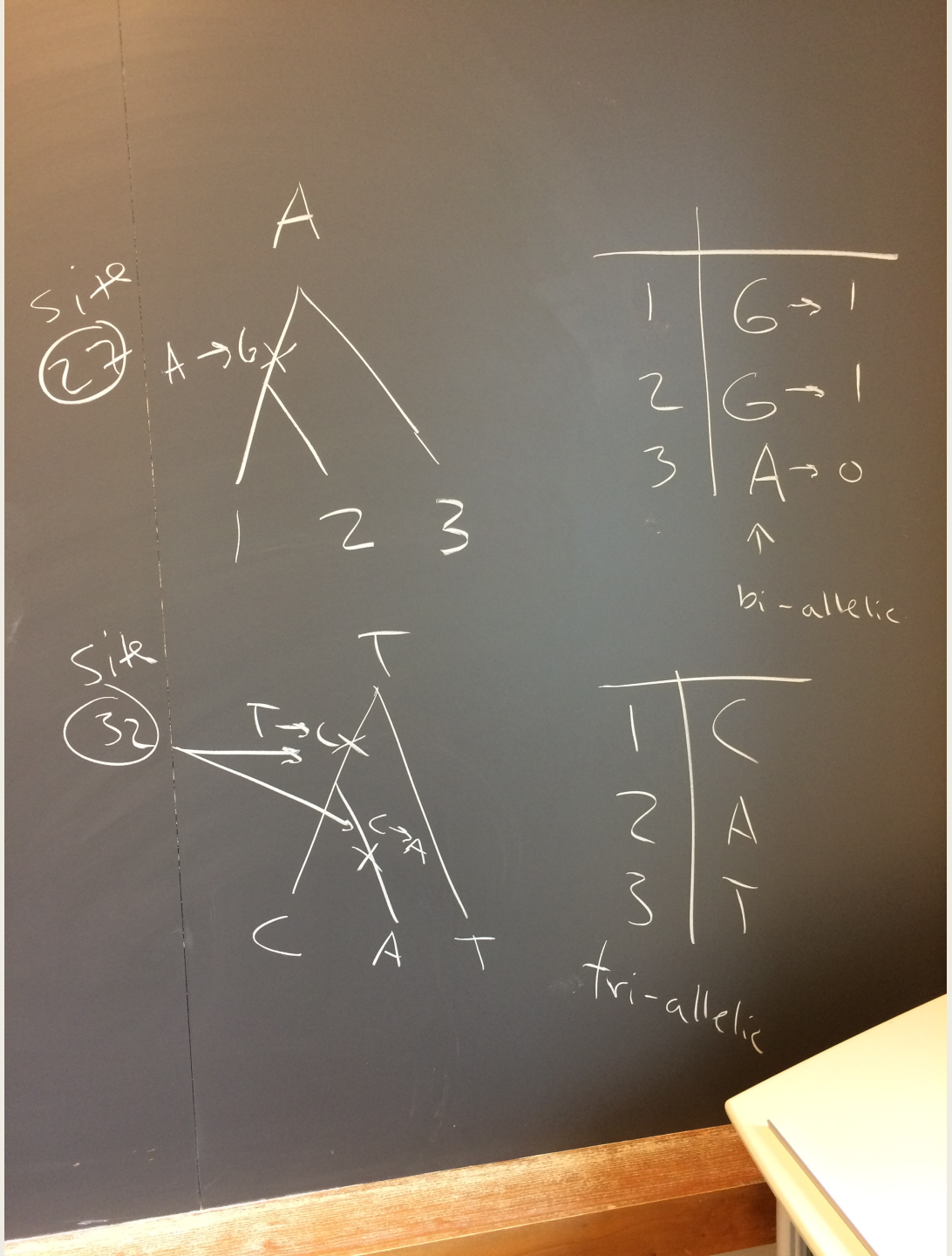
Ref : whoever we sequence first

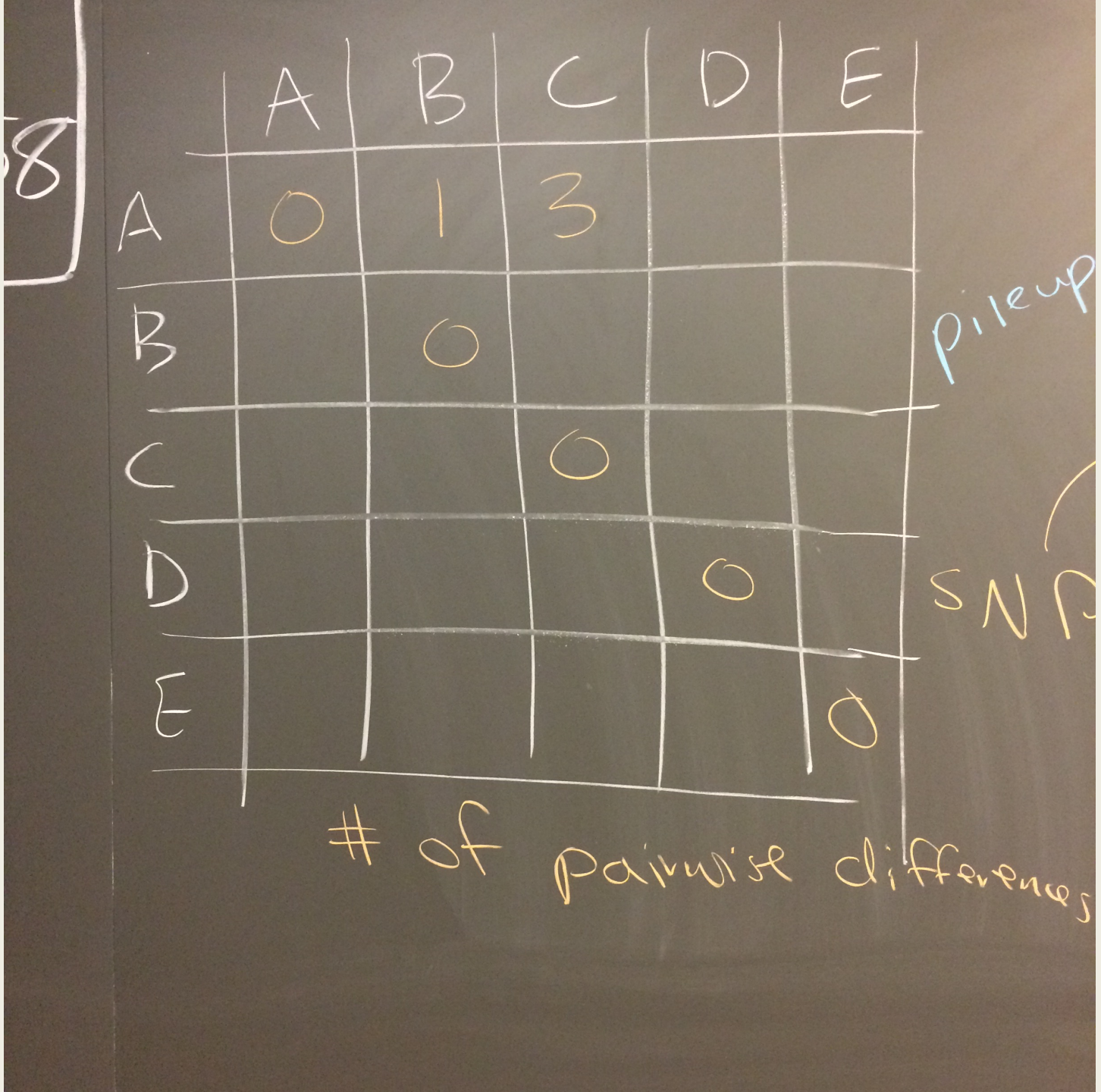
Alt : allele of observed non-ref individuals




most variation is biallelic

rare : triallelic


A
Site
27
A→G
1 2 3
1 G→1
2 G→1
3 A→0
bi-allelic
T
Site
32
T→C
C→A
C A T
1 C
2 A
3 T
tri-allelic


A B C D E
A 0 1 3
B 0
C 0
D 0
E 0
pileup
SNP
# of pairwise differences