Week 4: GATK

Assignment #1 Lessons

1. Seemingly simple tasks can get complex quickly
2. The best way to understand something is to dive in
3. ...and it's best to dive in early before the day something is due
4. The same will go for everything this quarter.
    a. Try early
    b. Fail early
    c. Debug early
    d. Stress less

Lessons to come:

1. Painful software installation (w/ command line bullshittery) is a thing
2. Online documentation and tutorials are awesome
3. Trying things out (on a small dataset) is a great way to get things moving forward

Discussion Questions

- What is GATK? What is it used for generally?
- What are the different file types GATK works with? For what is each file format used?
- What are the main capabilities of the GATK software that you anticipate using this quarter (meaning: which will be helpful for what we're trying to do)?

Student Questions: File / Pipeline Confusion

I am confused on the data pipeline we are creating. I am confused how we are going to clean and analyze the data and which format would be the best to achieve this. I am not sure what exactly we are trying to do with the data we are collecting.

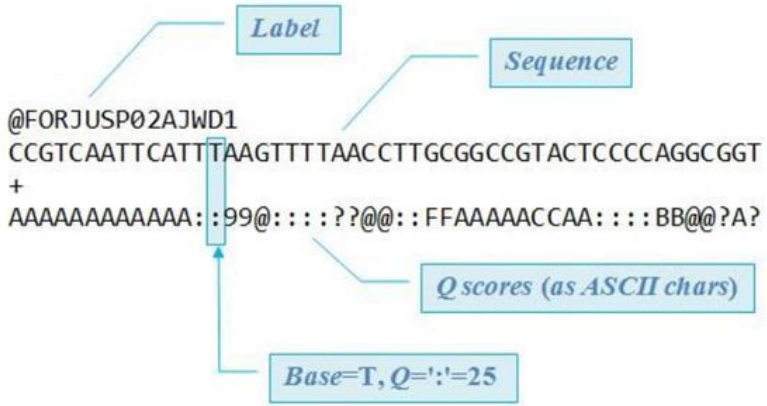
I'm a bit confused about how to correctly process the data. I would like to go more in depth about the file formats and how to process them correctly

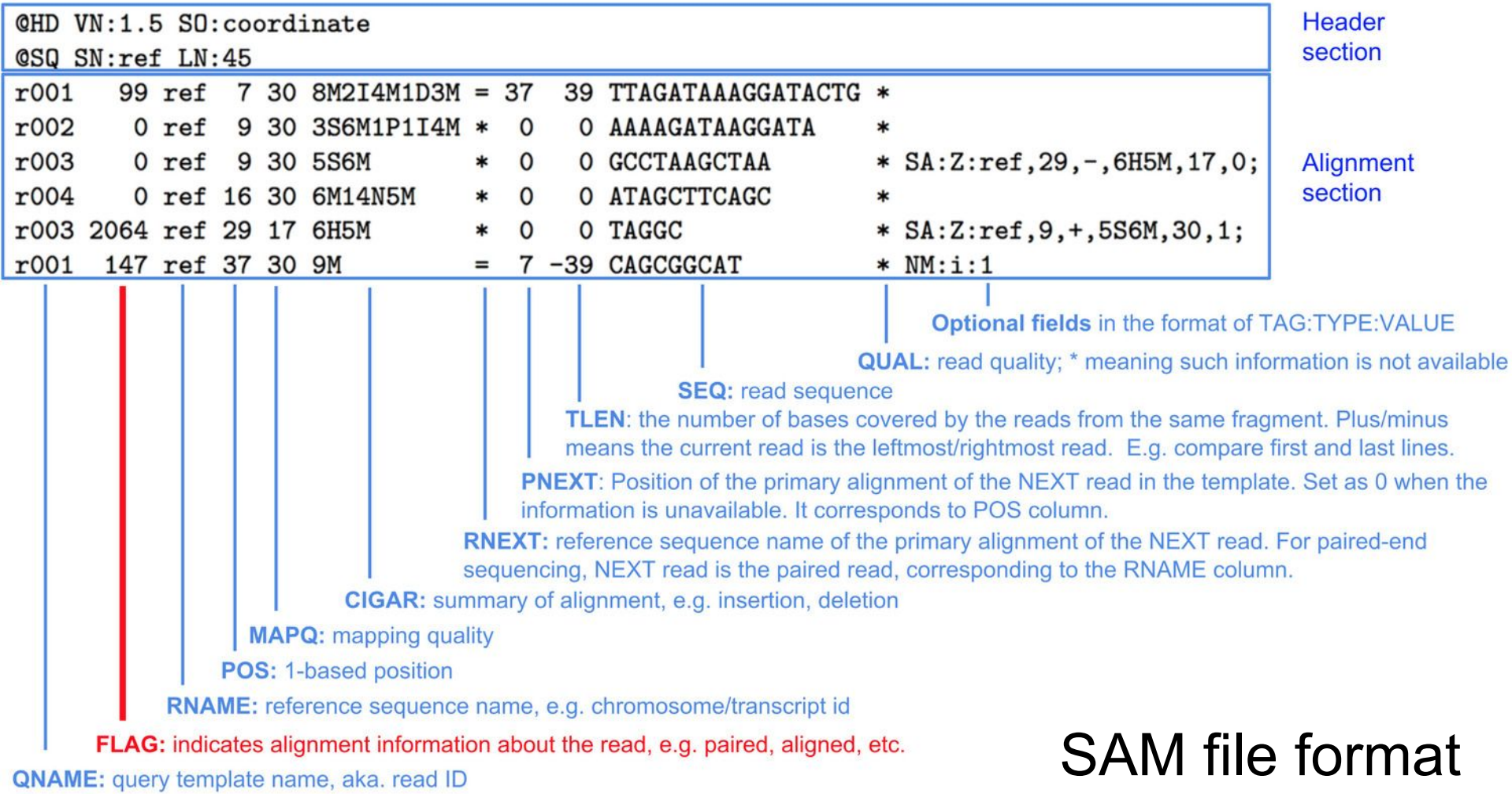
Also: a (totally reasonable) request for a demo that I'm going to hold off on for now.

FASTQ format

The FASTQ format was invented at the turn of the century at the Wellcome Trust Sanger Institute by Jim Mullikin, gradually disseminated, but never formally documented (Antony V. Cox, Sanger Institute, personal communication 2009).

```
@HWI-M01141:63:A4NDL:1:1101:14849:1418 1:N:0:TATAGCGAGACACCGT
NACGAAGGGTGCAAGCGTTACTCGGAATTACTGGGCGTAAAGCGTGCGTAGGTGGTGGTTTA
+
#>>>>A??AFAA1BGGEGGAAFGGCA0BFF1D2BCF/EEG/DBEE/E?GAEEFGAEFAEFG1
@HWI-M01141:63:A4NDL:1:1101:13802:1421 1:N:0:TATAGCGAGACACCGT
NACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTA
+
#>>AAABBBABBGGGGGGG?FGHGGGGGHHHHHHHHHGGGGHI
@HWI-M01141:63:A4NDL:1:1101:15928:1426 1:I
NACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAA
+
#>>AABFB@FBBGGGGGGGGGGHGGGGFHHHHHHHHHGGGGHI
@HWI-M01141:63:A4NDL:1:1101:14861:1431 1:I
NACGAAGGGTGCAAGCGTTACTCGGAATTACTGGGCGTAAA
+
#>>AAAABBFABGGGGGGECEGHGGEFFHHHHHHHGHGGGGHI
@HWI-M01141:63:A4NDL:1:1101:15264:1465 1:I
NACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAA
+
```





SAM file format

VCF files

(a) VCF example

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##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">
##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     40  PASS    .                                 GT:DP       1/1:13    2/2:29
1      2   .       C   T,CT      .  PASS    H2;AA=T                           GT          0|1       2/2
1      5   rs12    A   G        67  PASS    .                                 GT:DP       1|0:16    2/2:20
X      100 .       T   <DEL>    .  PASS    SVTYPE=DEL;END=299                GT:GQ:DP    1:12:.    0/0:20:36
```

(b) SNP

Alignment

```
1234
ACGT
ATGT
 ^
```

VCF representation

```
POS REF ALT
2   C   T
```

(c) Insertion

```
12345
AC-GT
ACTGT
  ^
```

```
POS REF ALT
2   C   CT
```

(d) Deletion

```
1234
ACGT
A--T
 ^^
```

```
POS REF ALT
1   ACG A
```

(e) Replacement

```
1234
ACGT
A-TT
 ^^
```

```
POS REF ALT
1   ACG AT
```

(f) Large structural variant

Alignment

```
   100       110       120       290       300
    .         .         .         .         .
ACGTACGTACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC
ACGT------------------------[...]----------GTAC
```

VCF representation

```
POS REF  ALT    INFO
100 T    <DEL>  SVTYPE=DEL;END=299
```

(g) Resolving ambiguity

Alignment

```
1234567890
TTTCCCTCTA
CTTACCT--A
^  ^   ^^
```

Possible representation

```
POS  REF        ALT
1    TTTCCCTCT  CTTACCTA
```

Possible representation

```
POS REF  ALT
1   T    C
4   C    A
7   TCT  T
```

Recommended VCF representation

```
POS  REF  ALT
1    T    C
4    C    A
5    CCT  C
```

https://samtools.github.io/hts-specs/VCFv4.2.pdf

(Great) Student GATK Question:

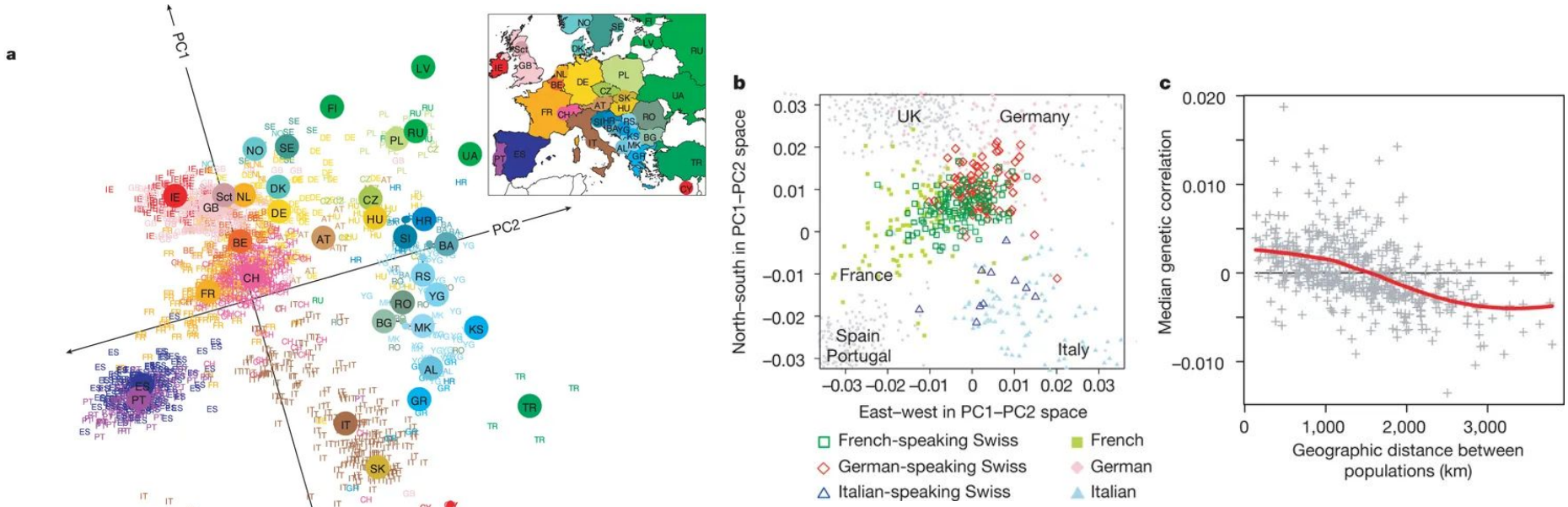
In Basic Protocol 1, the commands in #1 and #2 result in the creation of the same file so I'm unsure what the need for #2 is.

BWA index:

- use the indexing algorithm that is capable of handling the whole human genome
- Result: collection of files used by BWA to perform the alignment.
    - BWT (Burrows-Wheeler Transformation) encoded sequences that BWA (aligner) can understand
    - BWA: encode long sequences, like a genome, into a form that is easy to search

FASTA index:

- one record per line for each of the contigs in the FASTA reference file



https://www.nature.com/articles/nature07331/figures/1

Additional software: PLINK2

PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

The focus of **PLINK** is purely on *analysis* of genotype/phenotype data, so there is no support for steps prior to this (e.g. study design and planning, generating genotype or CNV calls from raw data). Through integration with gPLINK and Haploview, there is some support for the subsequent visualization, annotation and storage of results.

PLINK (one syllable) is being developed by Shaun Purcell whilst at the Center for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH), and the Broad Institute of Harvard & MIT, with the support of others.

https://www.cog-genomics.org/plink2/

PLINK: QC Considerations

These files are BIG.

Our goal is to look at human genetic variation across the globe.

We don’t need every single variant to accomplish this.

So, let’s make our lives easier by only including informative variants:

1. Filter out low frequency variants (while doing VCF -> BED/BIM/FAM conversion) ; SLOW
2. Filter out SNPs with high missingness

Running things in parallel will almost certainly be critical; chromosomes help with this.