GoSeqIt

Read mapping to identify contaminating reads and SNPs

Learning objective:

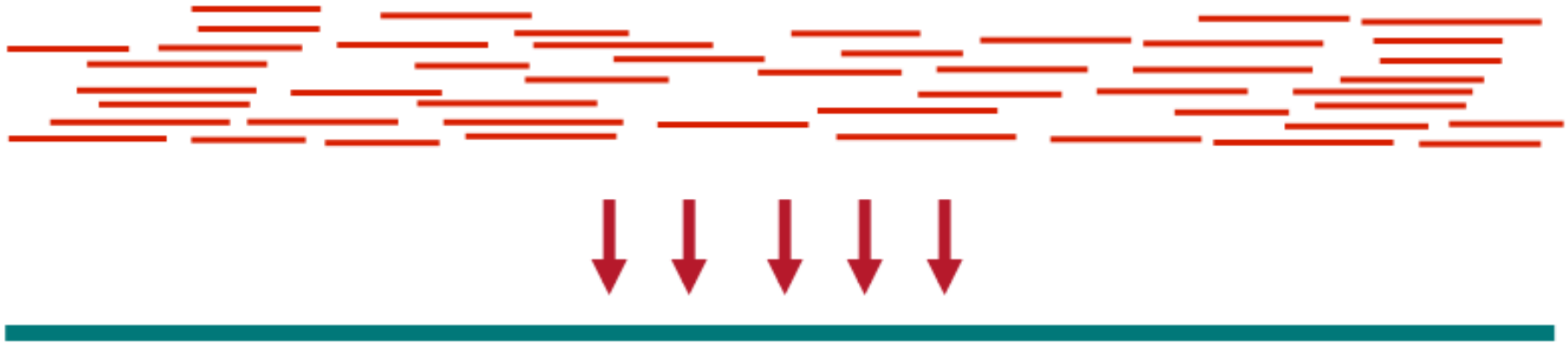
After this lecture, you should be able to…

…in general terms describe how read mapping is performed

…describe the SAM format for storing information about reads mapped to a reference sequence

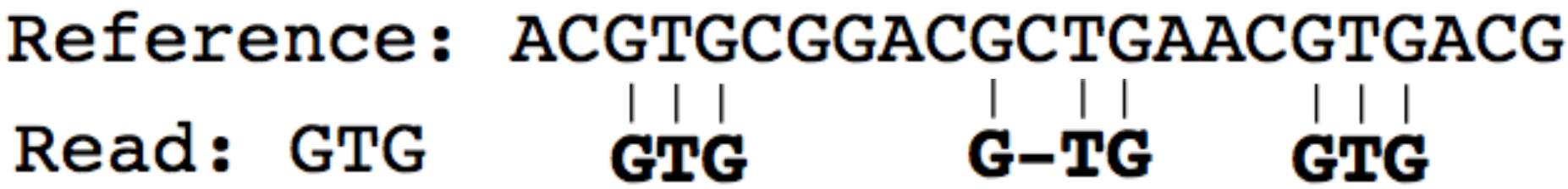
…describe the VCF format for storing information about variants

Read Mapping



```
                CATCGACCGAGCGCGATGCTAGCTAGGTGATCGT......
              TGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATCGT...
            GCATGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATCGT
          GTGCATGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATC
.......AGGTGCATGCCGCATCGATCGAGCGCGATGCTAGCTAGCTGATCGT.......
```

Simplest approach - exact string match



- You have to allow mismatches/INDELS

- On one of the Worlds most powerfull computers, K computer (Riken, Japan), mapping 20 mill reads of 100 bp length vs. human genome would take ~ 1 month

Smart approach - reduce the search space

Burrows Wheeler Transformation (BWT)

Invented in 1994 for the purpose of text compression

BWT rearranges a character string into runs of similar characters

Original genome: ACGTTAGAT$

BWT(ACGTTAGAT$) = TT$GAACCATG

The transformation is reversible!

All possible transformations of the string

ACGTTACGAT$
$ACGTTACGAT
T$ACGTTACGA
AT$ACGTTACG
GAT$ACGTTAC
CGAT$ACGTTA
ACGAT$ACGTT
TACGAT$ACGT
TTACGAT$ACG
GTTACGAT$AC
CGTTACGAT$A

Sorted lexicographically

$ACGTTACGAT
ACGAT$ACGTT
ACGTTACGAT$
AT$ACGTTACG
CGAT$ACGTTA
CGTTACGAT$A
GAT$ACGTTAC
GTTACGAT$AC
T$ACGTTACGA
TACGAT$ACGT
TTACGAT$ACG

Sequence Alignments

```
Coor      12345678901234  5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1         TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                       ATAGCT..............TCAGC
-r003                              ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

Read r001/1 and r001/2 is a read pair

Bases in lower cases are “clipped” (unaligned parts of the reads)

r003 is a chimeric read (non-linear alignment)

r004 represents a split alignment

How can be store this information in a condensed and easily accessible way?

Sequence Alignment Map (SAM) format, I

Format for storing biological sequences aligned to a reference sequence

TAB-delimited text with optional headers section (lines start with "@") and alignment section

```
Coor     12345678901234  56789012345678901234567890123456
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                      ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                         CAGCGGCAT
```

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

tags (optional)

Phred quality score

read name

flag

reference name

posi-tion

mapping quality

cigar

name of mate pair

position of mate read

length between read pairs

sequence of alignment

Sequence Alignment Map (SAM) format - flags

Flags explained: http://broadinstitute.github.io/picard/explain-flags.html

Decoding SAM flags

This utility makes it easy to identify what are the properties of a read based on its SAM flag value, or conversely, to find what the SAM Flag value would be for a given combination of properties.

To decode a given SAM flag value, just enter the number in the field below. The encoded properties will be listed under Summary below, to the right.

SAM Flag: 99 Explain

Switch to mate Toggle first in pair / second in pair

Find SAM flag by property:

To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- [x] read paired
- [x] read mapped in proper pair
- [ ] read unmapped
- [ ] mate unmapped
- [ ] read reverse strand
- [x] mate reverse strand
- [x] first in pair
- [ ] second in pair
- [ ] not primary alignment
- [ ] read fails platform/vendor quality checks
- [ ] read is PCR or optical duplicate
- [ ] supplementary alignment

Summary:

read paired (0x1)
read mapped in proper pair (0x2)
mate reverse strand (0x20)
first in pair (0x40)

Sequence Alignment Map (SAM) format - cigar

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

For example:

```
RefPos:     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C
Read: ACTAGAATGGCT
```

Aligning these two:

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                  A  C  T  A  G  A  A     T  G  G  C  T
```

With the alignment above, you get:

```
POS: 5
CIGAR: 3M1I3M1D5M
```

Finally, to save space, SAM -> BAM (binary format)

Manipulating SAM/BAM files - Samtools

Samtools flagstat: mapping statistics

```
$ samtools flagstat NZ_CP009361_SRR4114395.bam
951327 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
3081 + 0 supplementary
0 + 0 duplicates
926390 + 0 mapped (97.38% : N/A)
948246 + 0 paired in sequencing
474123 + 0 read1
474123 + 0 read2
920548 + 0 properly paired (97.08% : N/A)
923164 + 0 with itself and mate mapped
145 + 0 singletons (0.02% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Variant calling

Variants are called with samtools mpileup followed by bcftools

The output is a file in Variant Calling Format (VCF)

The header lines of VCF files start with "#"

The lines containing information on the variants have 8 mandatory columns and additional optional columns

(a) VCF example

Header

```
##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">
```

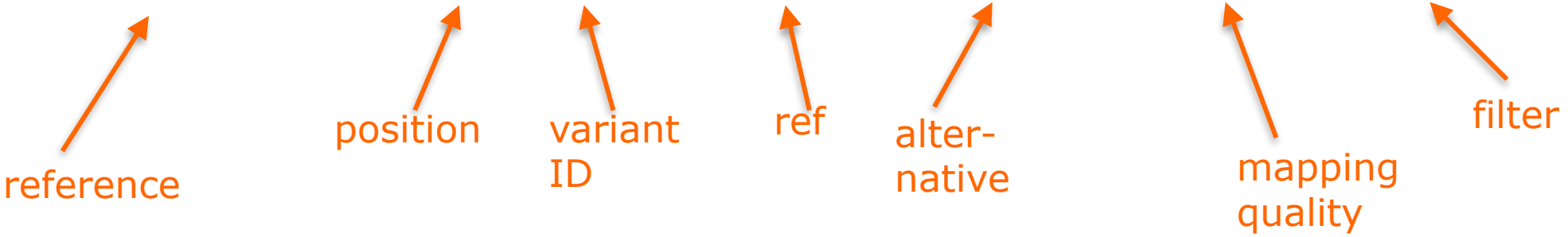
Body

#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

Real data from exercise 5

The first 7 columns:

NZ_CP009361.1	631465	.	G	A	5.28287	.
NZ_CP009361.1	732437	.	A	T	5.25638	.
NZ_CP009361.1	775157	.	T	C	11.8646	.
NZ_CP009361.1	847365	.	T	A	17.9624	.
NZ_CP009361.1	1012394	.	ATAGTAGT	ATAGT	225	
NZ_CP009361.1	1012397	.	GT	GTCAGCAT	225	
NZ_CP009361.1	1012398	.	T	TCAGCAA	228	.
NZ_CP009361.1	1041461	.	A	G	18.2112	.
NZ_CP009361.1	1053737	.	G	C	51	.



Real data from exercise 5

Column 8 - the INFO column

```
DP=14;SGB=-0.379885;RPB=1;MQB=1;BQB=1;MQOF=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=0,7,0,1;MQ=60
DP=43;VDB=0.52;SGB=-0.453602;RPB=0.8;MQB=1;MQSB=1;BQB=0.56;MQOF=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=9
DP=12;SGB=-0.379885;RPB=1;MQB=1;BQB=1;MQOF=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=0,6,0,1;MQ=60
DP=12;SGB=-0.379885;RPB=1;MQB=1;MQSB=1;BQB=1;MQOF=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=1,5,0,1;MQ=60
DP=6;SGB=-0.379885;RPB=1;MQB=1;MQSB=1;BQB=1;MQOF=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=2,2,0,1;MQ=60
DP=6;SGB=-0.379885;RPB=1;MQB=1;MQSB=1;BQB=1;MQOF=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=0,2,1,0;MQ=60
DP=5;SGB=-0.379885;RPB=1;MQB=1;MQSB=1;BQB=1;MQOF=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=2,0,0,1;MQ=60
```

- In the file header, you can see what each of the abbreviations mean
- Some important ones:
  - ✳ DP=raw read depth
  - ✳ MQ=Average mapping quality

Column 9,10 - the FORMAT column and its values

```
GT:PL    0/1:89,0,255
GT:PL    0/1:255,0,255
GT:PL    1/1:255,27,0
GT:PL    1/1:255,126,0
GT:PL    1/1:255,61,0
GT:PL    1/1:96,42,0
```

GT:Genotype

0/1=heterozygous

1/1=homozygous

PL: List of Phred-scaled genotype likelihoods