Mapping and Calling Variants

Eve198

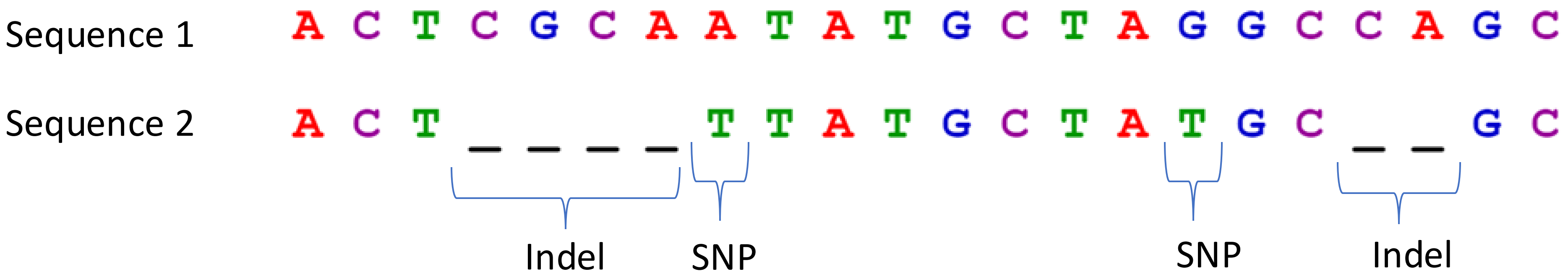
Week 4: April 23rd

Maddie Armstrong & Rachael Bay

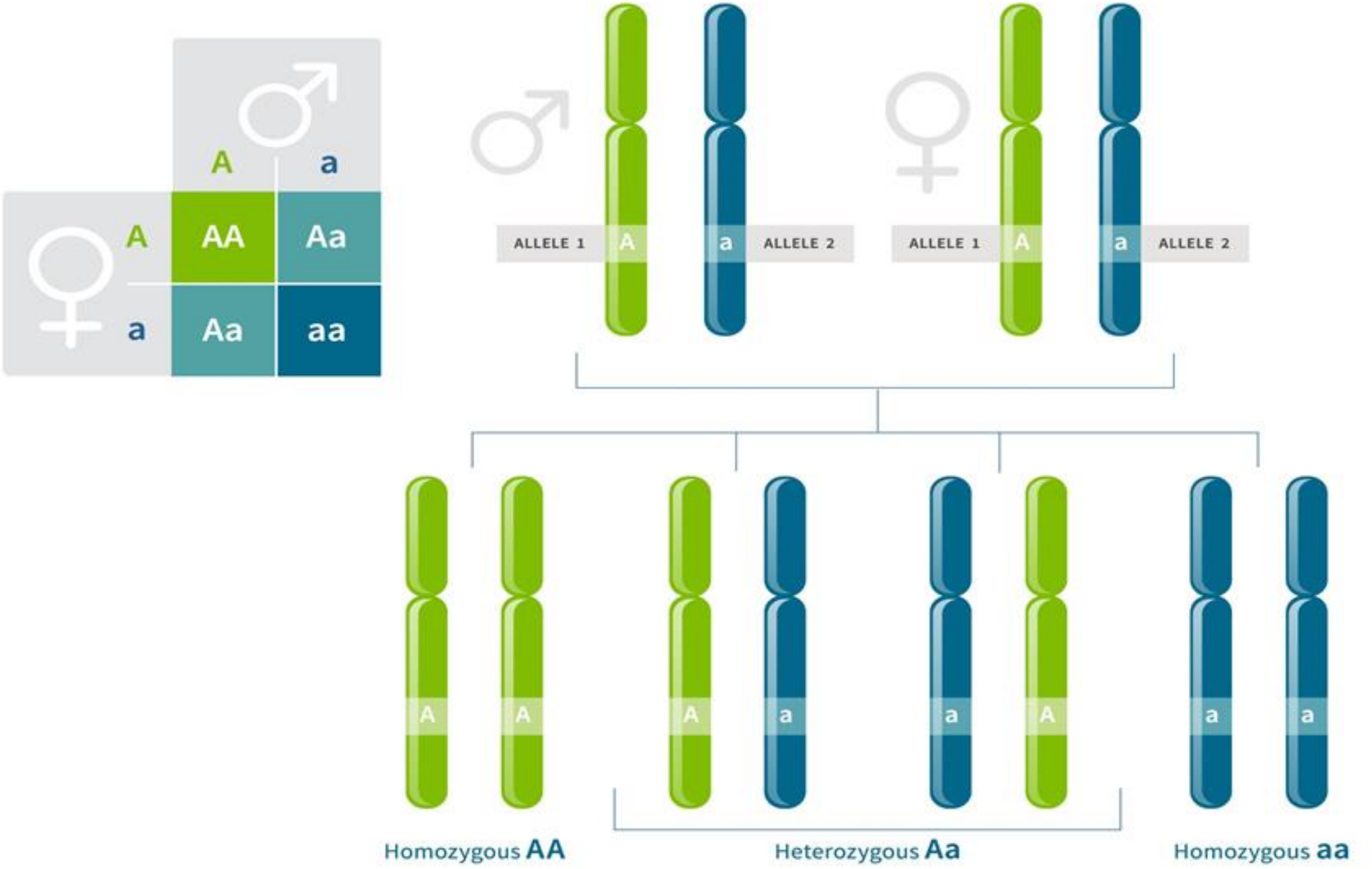
What is a genetic variant?

A region of the genome that differs from the reference (or another genome)

Signifies a mutation and can be a single base-pair, or larger insertion and/or deletion of several base-pairs.


Sequence 1
A C T C G C A A T A T G C T A G G C C A G C
Sequence 2
A C T _ _ _ _ T T A T G C T A T G C _ _ G C
Indel
SNP
SNP
Indel

What is a genotype?



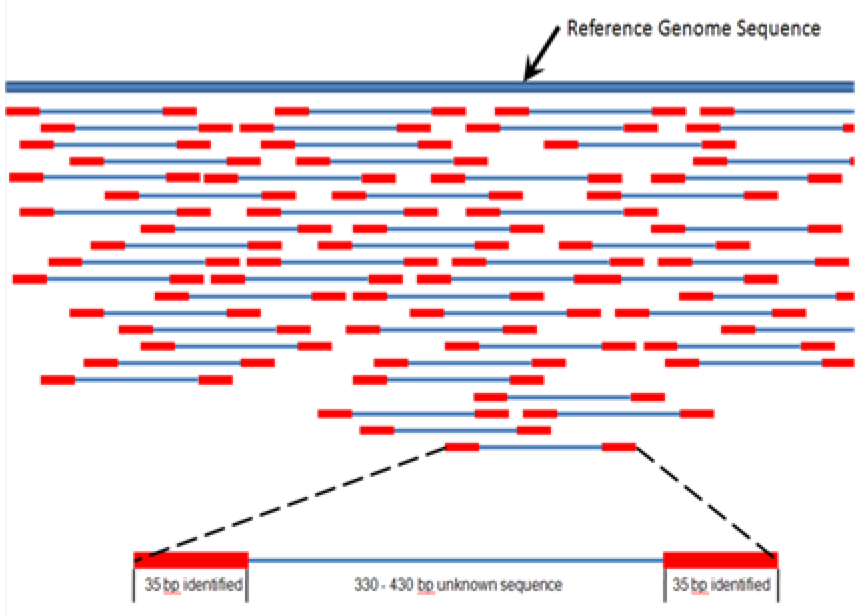
	♂ A	♂ a
♀ A	AA	Aa
♀ a	Aa	aa

Image credit: https://www.ancestry.com/lp/genotype

How do we find a variant?

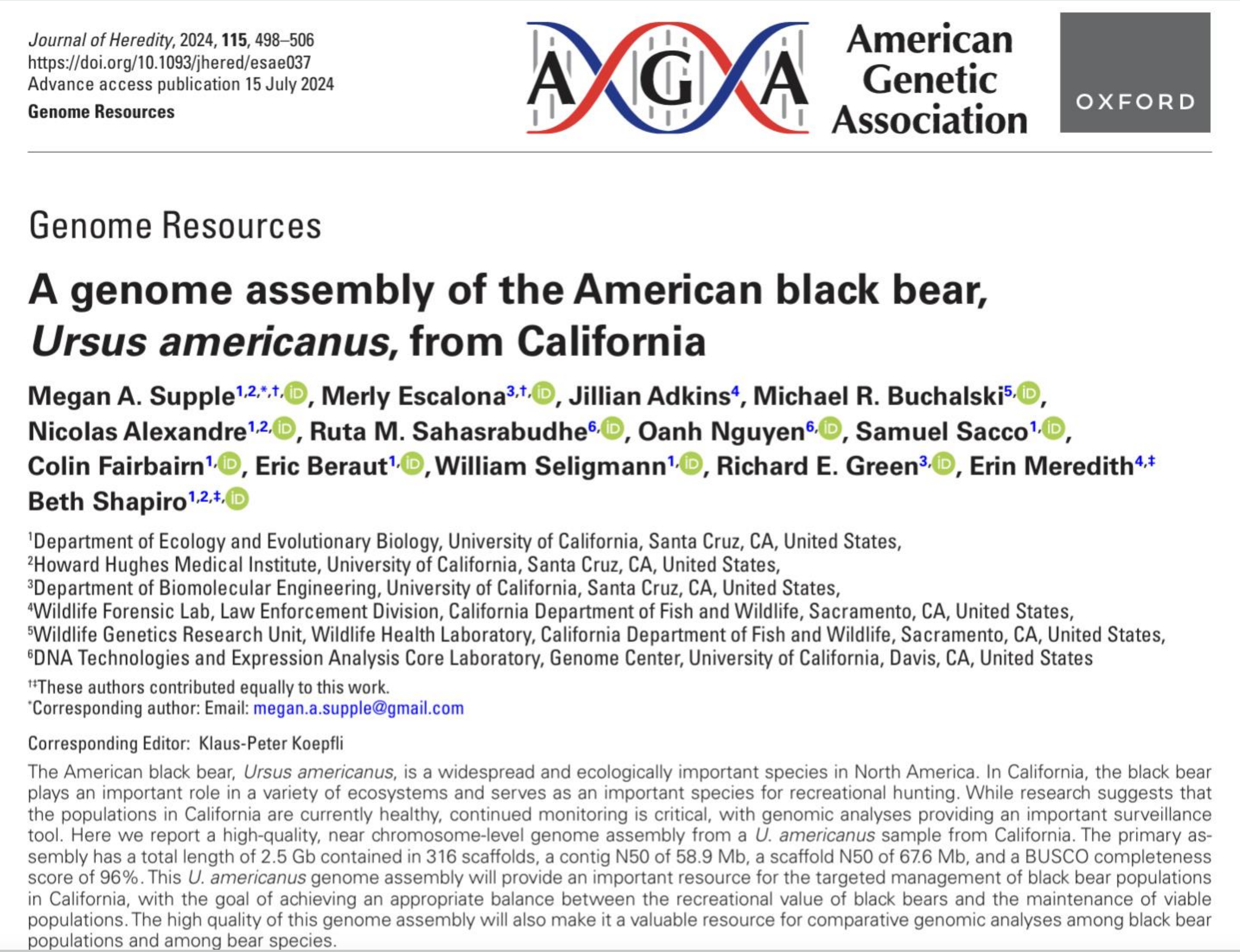
Map and align sequences from other individuals to a reference genome

- Does It matter what your reference genome is?
  - Is it the same or different species?
  - Is it from the same population?



- Short answer: Yes, it matters!

Genomes are continually being improved & sequenced all the time!



Journal of Heredity, 2024, **115**, 498–506
https://doi.org/10.1093/jhered/esae037
Advance access publication 15 July 2024
Genome Resources

Genome Resources

A genome assembly of the American black bear, *Ursus americanus*, from California

Megan A. Supple[1,2,*,†], Merly Escalona[3,†], Jillian Adkins[4], Michael R. Buchalski[5], Nicolas Alexandre[1,2], Ruta M. Sahasrabudhe[6], Oanh Nguyen[6], Samuel Sacco[1], Colin Fairbairn[1], Eric Beraut[1], William Seligmann[1], Richard E. Green[3], Erin Meredith[4,‡] Beth Shapiro[1,2,‡]

[1]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA, United States,
[2]Howard Hughes Medical Institute, University of California, Santa Cruz, CA, United States,
[3]Department of Biomolecular Engineering, University of California, Santa Cruz, CA, United States,
[4]Wildlife Forensic Lab, Law Enforcement Division, California Department of Fish and Wildlife, Sacramento, CA, United States,
[5]Wildlife Genetics Research Unit, Wildlife Health Laboratory, California Department of Fish and Wildlife, Sacramento, CA, United States,
[6]DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California, Davis, CA, United States

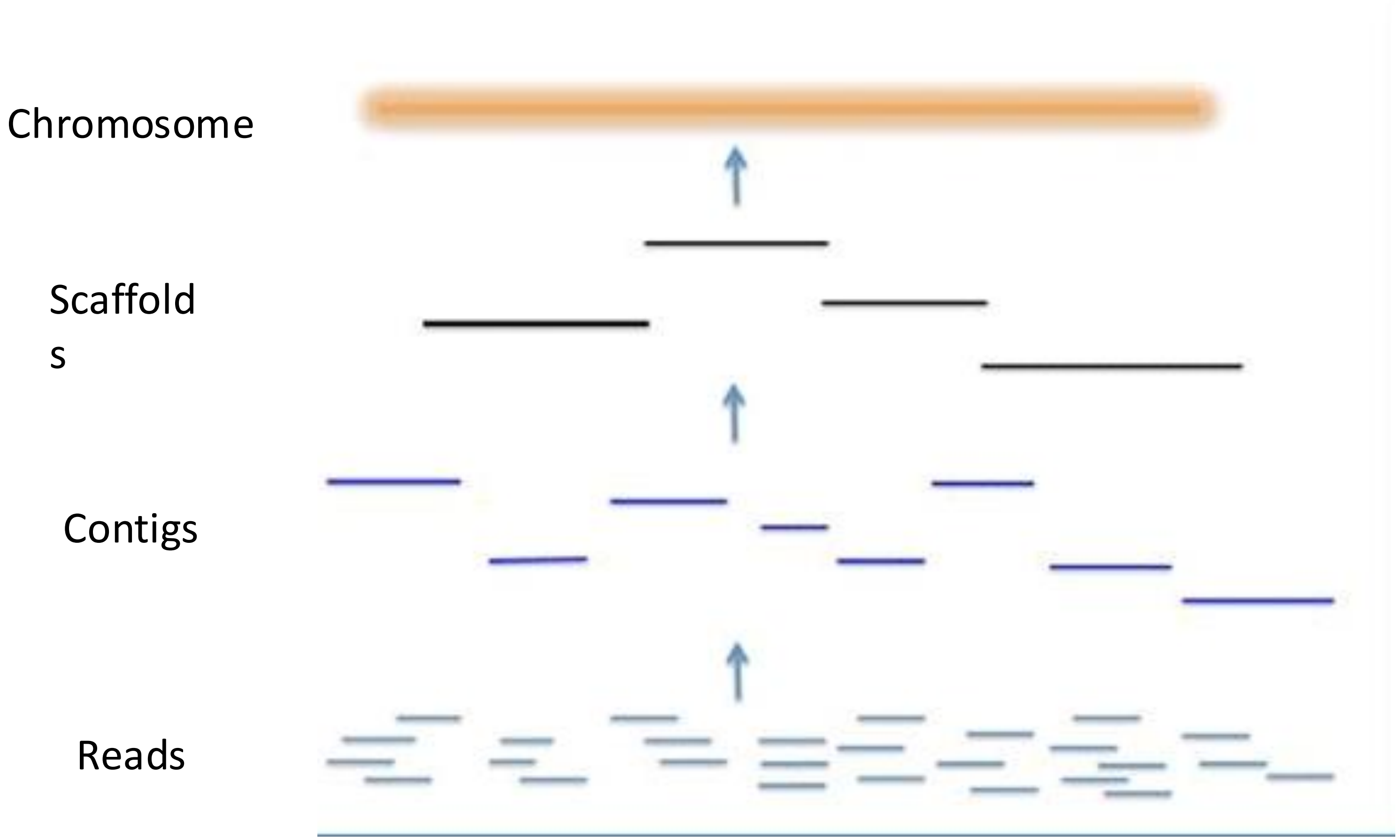
†‡These authors contributed equally to this work.
*Corresponding author: Email: megan.a.supple@gmail.com

Corresponding Editor: Klaus-Peter Koepfli

The American black bear, *Ursus americanus*, is a widespread and ecologically important species in North America. In California, the black bear plays an important role in a variety of ecosystems and serves as an important species for recreational hunting. While research suggests that the populations in California are currently healthy, continued monitoring is critical, with genomic analyses providing an important surveillance tool. Here we report a high-quality, near chromosome-level genome assembly from a *U. americanus* sample from California. The primary assembly has a total length of 2.5 Gb contained in 316 scaffolds, a contig N50 of 58.9 Mb, a scaffold N50 of 67.6 Mb, and a BUSCO completeness score of 96%. This *U. americanus* genome assembly will provide an important resource for the targeted management of black bear populations in California, with the goal of achieving an appropriate balance between the recreational value of black bears and the maintenance of viable populations. The high quality of this genome assembly will also make it a valuable resource for comparative genomic analyses among black bear populations and among bear species.

https://academic.oup.com/jhered/article/115/5/498/7713838

Finding variants – some terminology



Chromosome

Scaffold
s

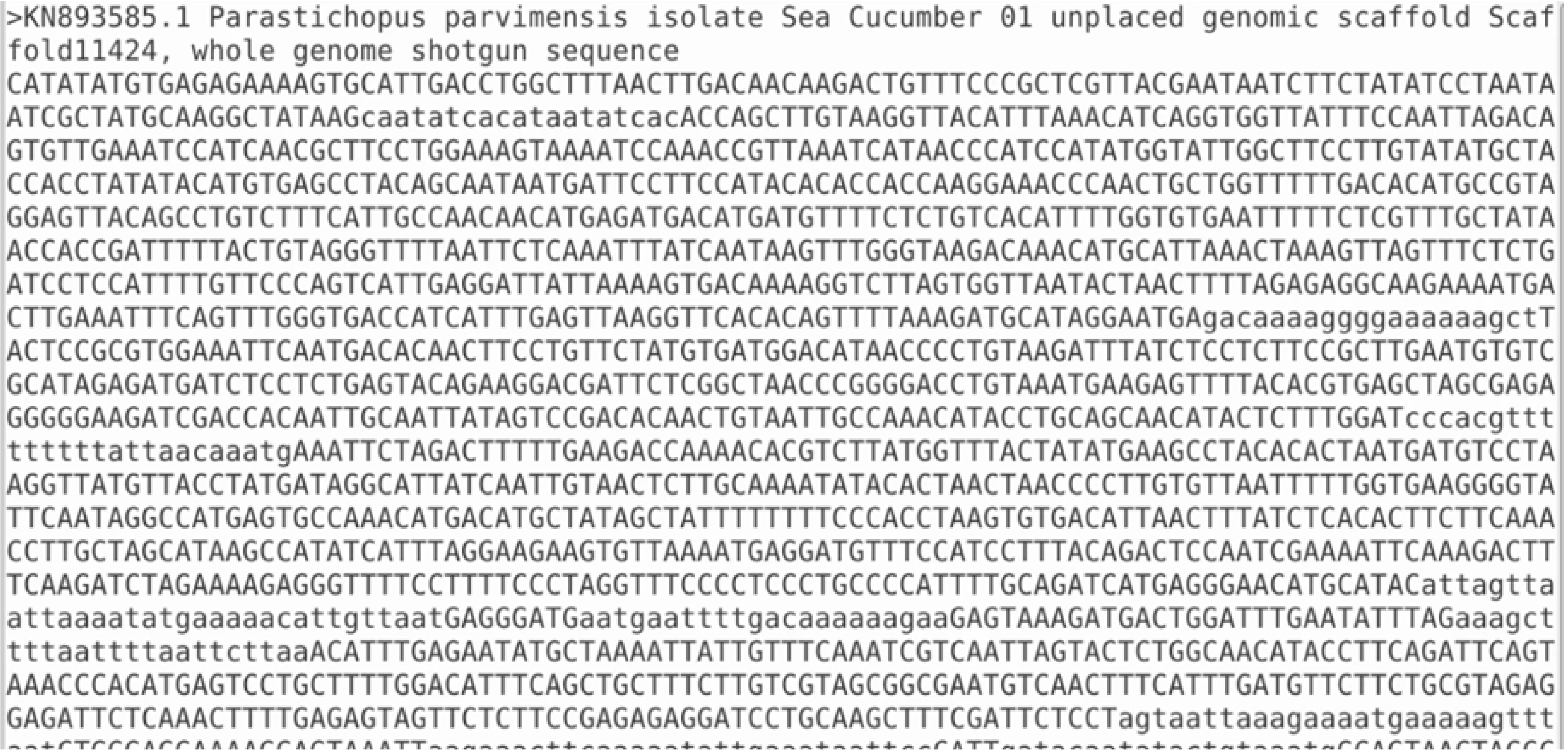
Contigs

Reads

A reference genome is a collection of contigs

Figure modified from here https://www.ddbj.nig.ac.jp/ddbj/assembly-e.html & S. Caplins slides

Finding variants – some terminology

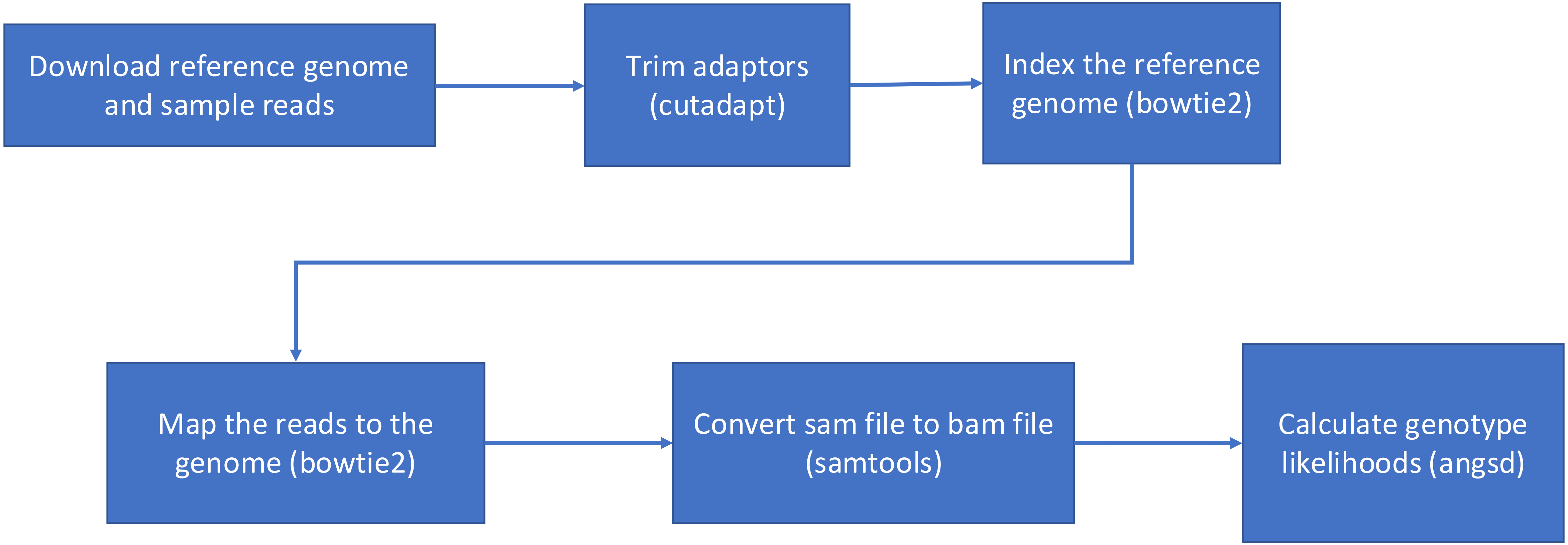


```
>KN893585.1 Parastichopus parvimensis isolate Sea Cucumber 01 unplaced genomic scaffold Scaf
fold11424, whole genome shotgun sequence
CATATATGTGAGAGAAAAGTGCATTGACCTGGCTTTAACTTGACAACAAGACTGTTTCCCGCTCGTTACGAATAATCTTCTATATCCTAATA
ATCGCTATGCAAGGCTATAAGcaatatcacataatatcacACCAGCTTGTAAGGTTACATTTAAACATCAGGTGGTTATTTCCAATTAGACA
GTGTTGAAATCCATCAACGCTTCCTGGAAAGTAAAATCCAAACCGTTAAATCATAACCCATCCATATGGTATTGGCTTCCTTGTATATGCTA
CCACCTATATACATGTGAGCCTACAGCAATAATGATTCCTTCCATACACACCACCAAGGAAACCCAACTGCTGGTTTTTGACACATGCCGTA
GGAGTTACAGCCTGTCTTTCATTGCCAACAACATGAGATGACATGATGTTTTCTCTGTCACATTTTGGTGTGAATTTTTCTCGTTTGCTATA
ACCACCGATTTTTACTGTAGGGTTTTAATTCTCAAATTTATCAATAAGTTTGGGTAAGACAAACATGCATTAAACTAAAGTTAGTTTCTCTG
ATCCTCCATTTTGTTCCCAGTCATTGAGGATTATTAAAAGTGACAAAAGGTCTTAGTGGTTAATACTAACTTTTAGAGAGGCAAGAAAATGA
CTTGAAATTTCAGTTTGGGTGACCATCATTTGAGTTAAGGTTCACACAGTTTTAAAGATGCATAGGAATGAgacaaaaggggaaaaaagctT
ACTCCGCGTGGAAATTCAATGACACAACTTCCTGTTCTATGTGATGGACATAACCCCTGTAAGATTTATCTCCTCTTCCGCTTGAATGTGTC
GCATAGAGATGATCTCCTCTGAGTACAGAAGGACGATTCTCGGCTAACCCGGGGACCTGTAAATGAAGAGTTTTACACGTGAGCTAGCGAGA
GGGGGAAGATCGACCACAATTGCAATTATAGTCCGACACAACTGTAATTGCCAAACATACCTGCAGCAACATACTCTTTGGATcccacgttt
ttttttattaacaaatgAAATTCTAGACTTTTTGAAGACCAAAACACGTCTTATGGTTTACTATATGAAGCCTACACACTAATGATGTCCTA
AGGTTATGTTACCTATGATAGGCATTATCAATTGTAACTCTTGCAAAATATACACTAACTAACCCCTTGTGTTAATTTTTGGTGAAGGGGTA
TTCAATAGGCCATGAGTGCCAAACATGACATGCTATAGCTATTTTTTTTCCCACCTAAGTGTGACATTAACTTTATCTCACACTTCTTCAAA
CCTTGCTAGCATAAGCCATATCATTTAGGAAGAAGTGTTAAAATGAGGATGTTTCCATCCTTTACAGACTCCAATCGAAAATTCAAAGACTT
TCAAGATCTAGAAAAGAGGGTTTTCCTTTTCCCTAGGTTTCCCCTCCCTGCCCCATTTTGCAGATCATGAGGGAACATGCATACattagtta
attaaaatatgaaaaacattgttaatGAGGGATGaatgaattttgacaaaaaagaaGAGTAAAGATGACTGGATTTGAATATTTAGaaagct
tttaattttaattcttaaACATTTGAGAATATGCTAAAATTATTGTTTCAAATCGTCAATTAGTACTCTGGCAACATACCTTCAGATTCAGT
AAACCCACATGAGTCCTGCTTTTGGACATTTCAGCTGCTTTCTTGTCGTAGCGGCGAATGTCAACTTTCATTTGATGTTCTTCTGCGTAGAG
GAGATTCTCAAACTTTTGAGAGTAGTTCTCTTCCGAGAGAGGATCCTGCAAGCTTTCGATTCTCCTagtaattaaagaaaatgaaaaagttt
```

A reference genome is a collection of contigs

Typically, in fasta format

Finding variants - pipeline



Download reference genome and sample reads

Trim adaptors (cutadapt)

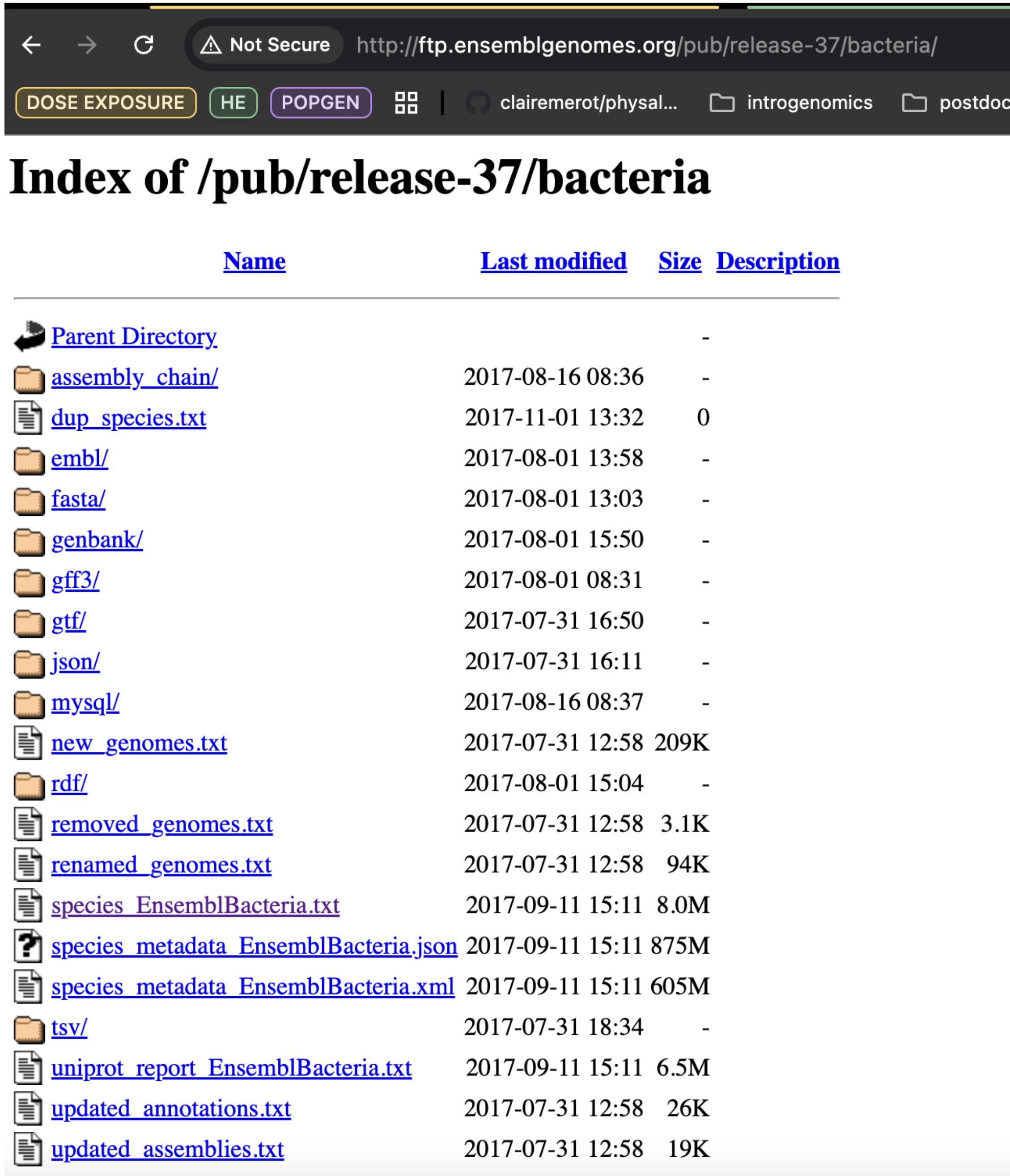
Index the reference genome (bowtie2)

Map the reads to the genome (bowtie2)

Convert sam file to bam file (samtools)

Calculate genotype likelihoods (angsd)

Step 1: Download the data!

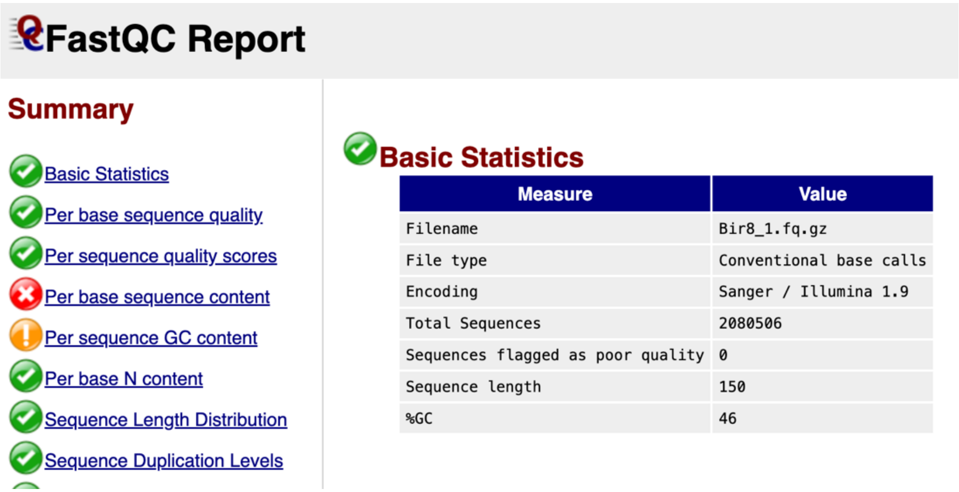


Index of /pub/release-37/bacteria

Name	Last modified	Size	Description
Parent Directory		-	
assembly_chain/	2017-08-16 08:36	-	
dup_species.txt	2017-11-01 13:32	0	
embl/	2017-08-01 13:58	-	
fasta/	2017-08-01 13:03	-	
genbank/	2017-08-01 15:50	-	
gff3/	2017-08-01 08:31	-	
gtf/	2017-07-31 16:50	-	
json/	2017-07-31 16:11	-	
mysql/	2017-08-16 08:37	-	
new_genomes.txt	2017-07-31 12:58	209K	
rdf/	2017-08-01 15:04	-	
removed_genomes.txt	2017-07-31 12:58	3.1K	
renamed_genomes.txt	2017-07-31 12:58	94K	
species_EnsemblBacteria.txt	2017-09-11 15:11	8.0M	
species_metadata_EnsemblBacteria.json	2017-09-11 15:11	875M	
species_metadata_EnsemblBacteria.xml	2017-09-11 15:11	605M	
tsv/	2017-07-31 18:34	-	
uniprot_report_EnsemblBacteria.txt	2017-09-11 15:11	6.5M	
updated_annotations.txt	2017-07-31 12:58	26K	
updated_assemblies.txt	2017-07-31 12:58	19K	

```
wget ftp://ftp.ensemblgenomes.org/pub/release-37/bacteria/species_EnsemblBacteria.txt
```

Step 1.2: Quality Control with fastqc



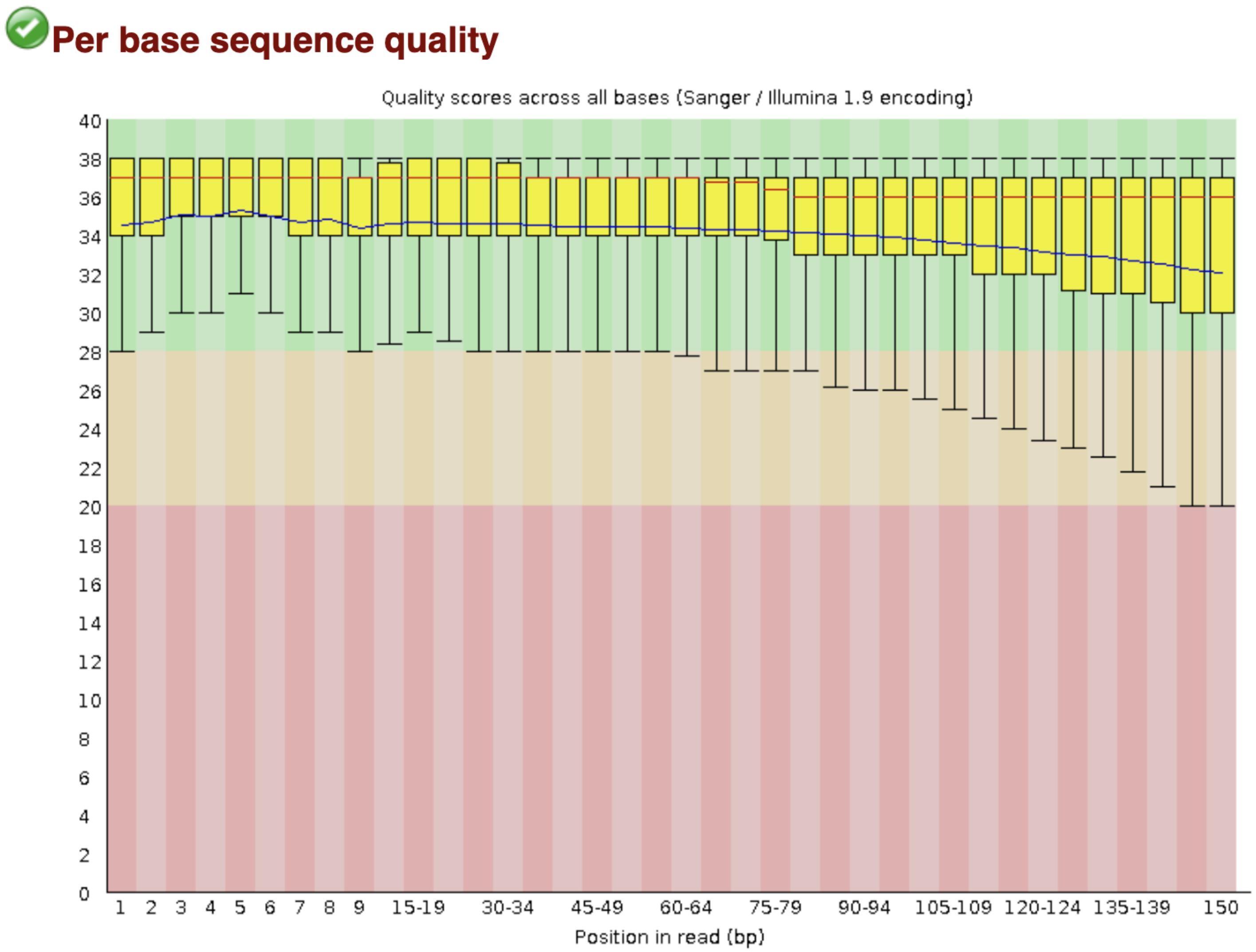
FastQC Report

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels

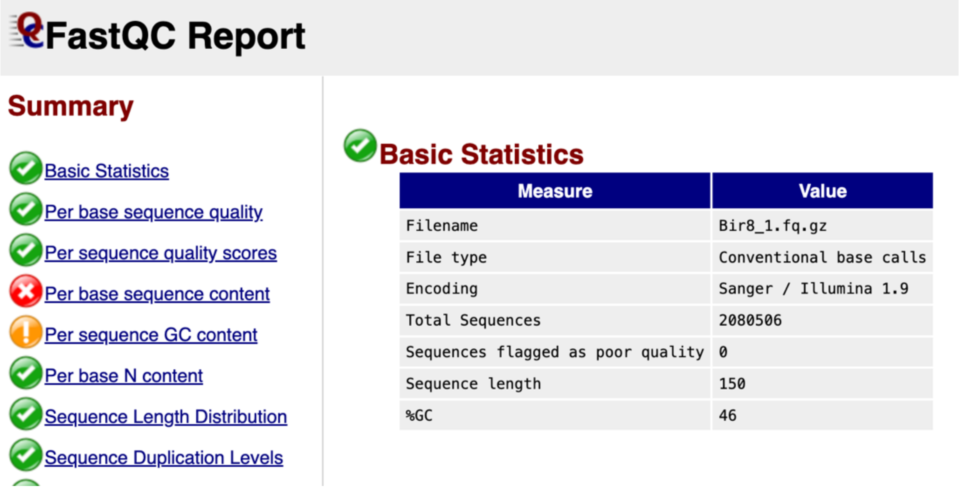
✓ Basic Statistics

Measure	Value
Filename	Bir8_1.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2080506
Sequences flagged as poor quality	0
Sequence length	150
%GC	46



✓ Per base sequence quality

Step 1.2: Quality Control with fastqc



FastQC Report

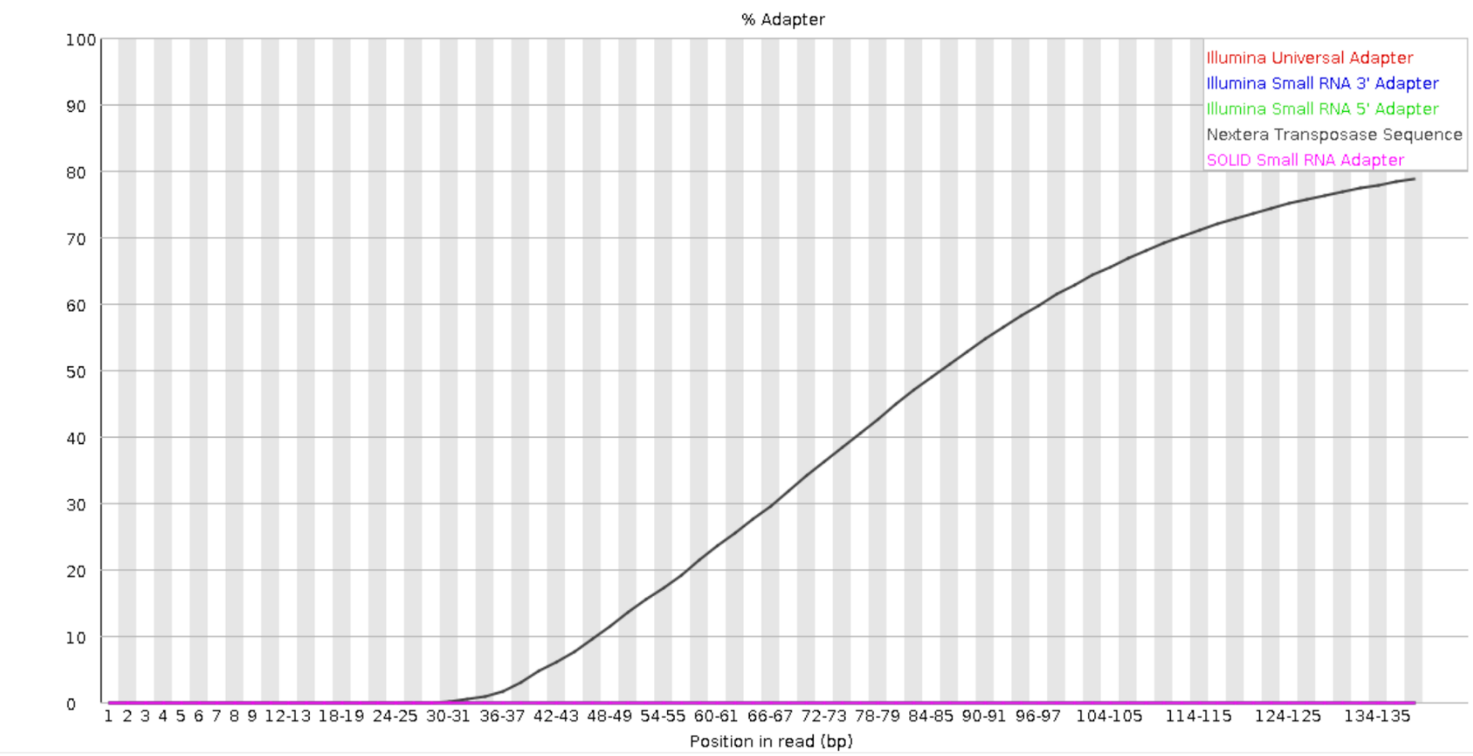
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels

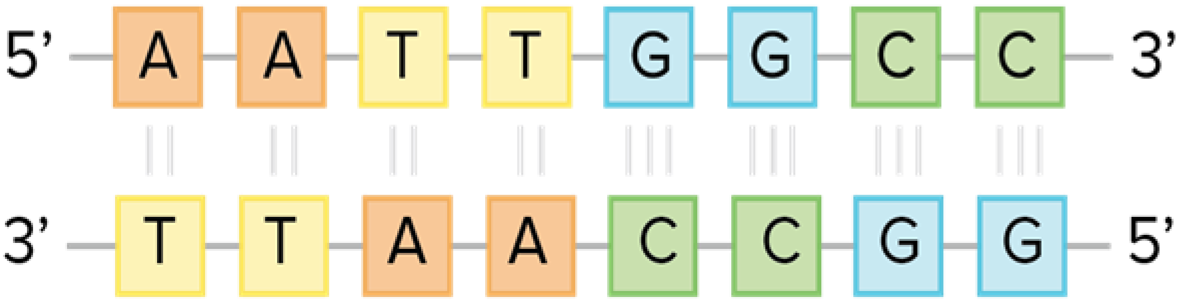
✓ Basic Statistics

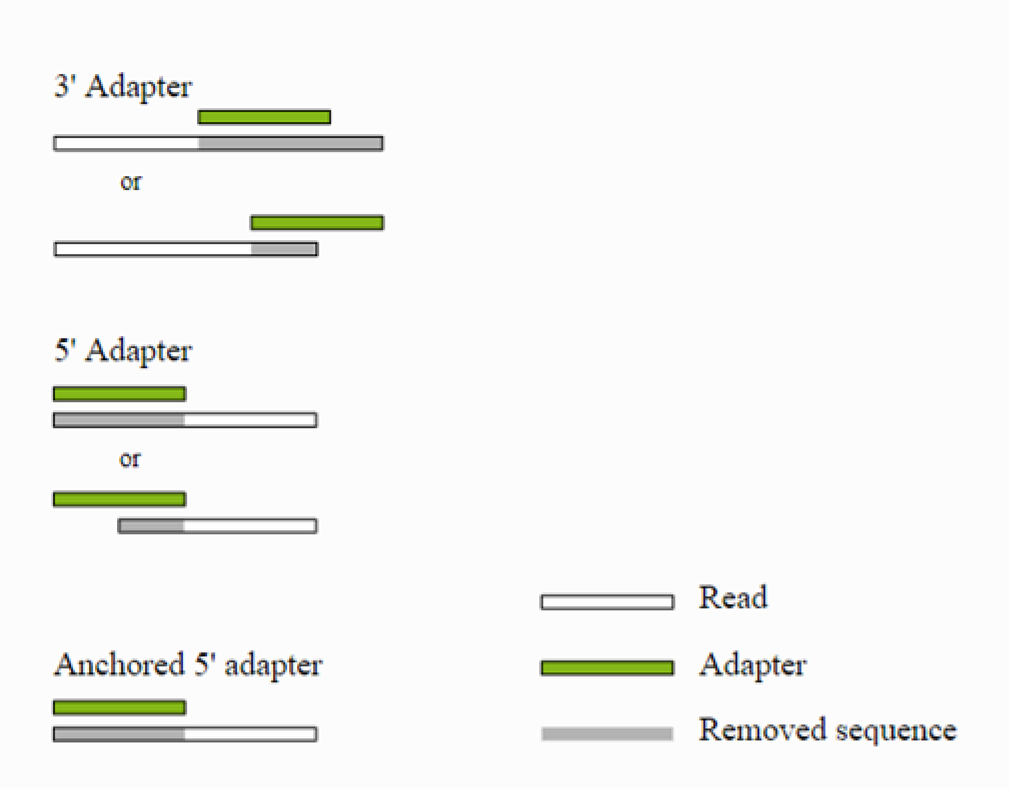
Measure	Value
Filename	Bir8_1.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2080506
Sequences flagged as poor quality	0
Sequence length	150
%GC	46

✗ Adapter Content

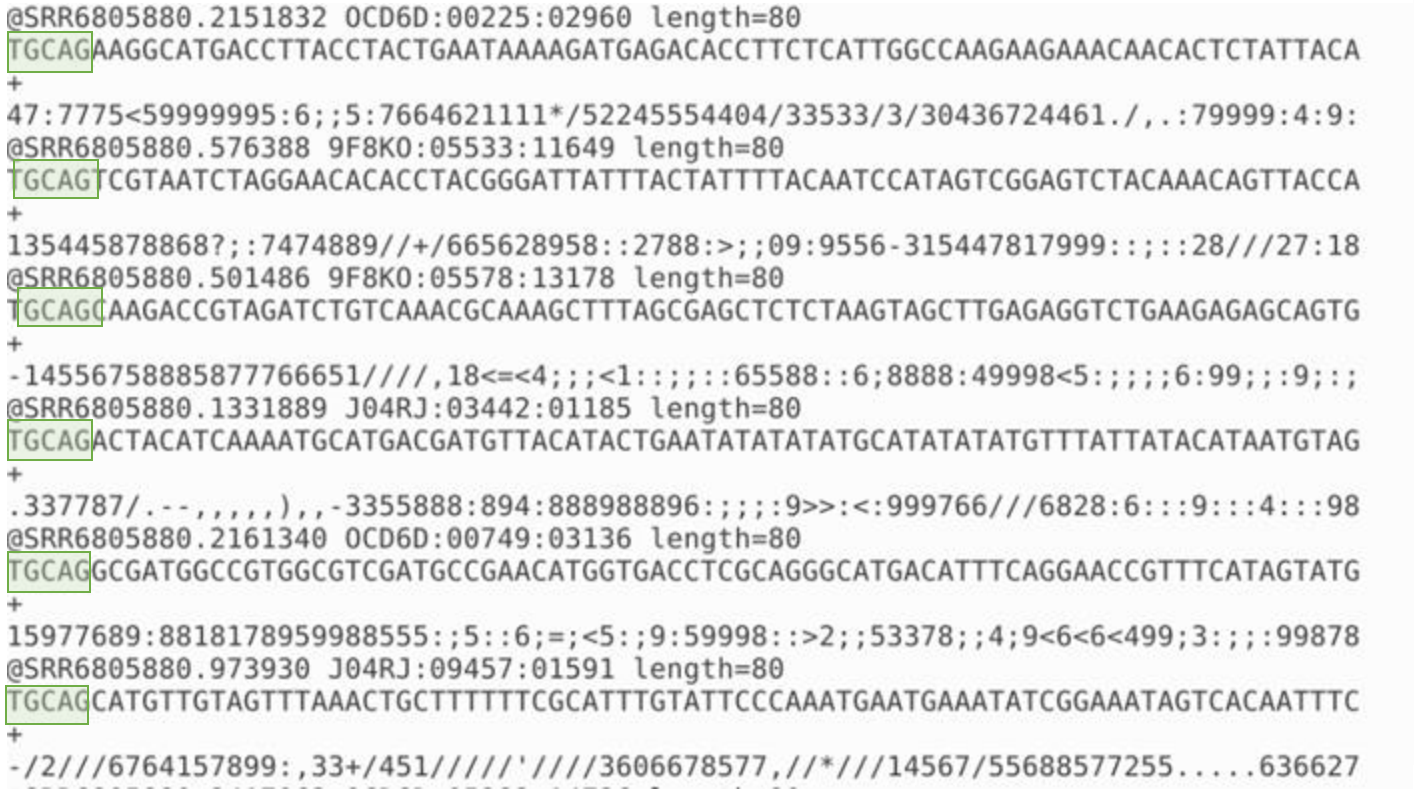


Step 2: Trimming adaptors from reads

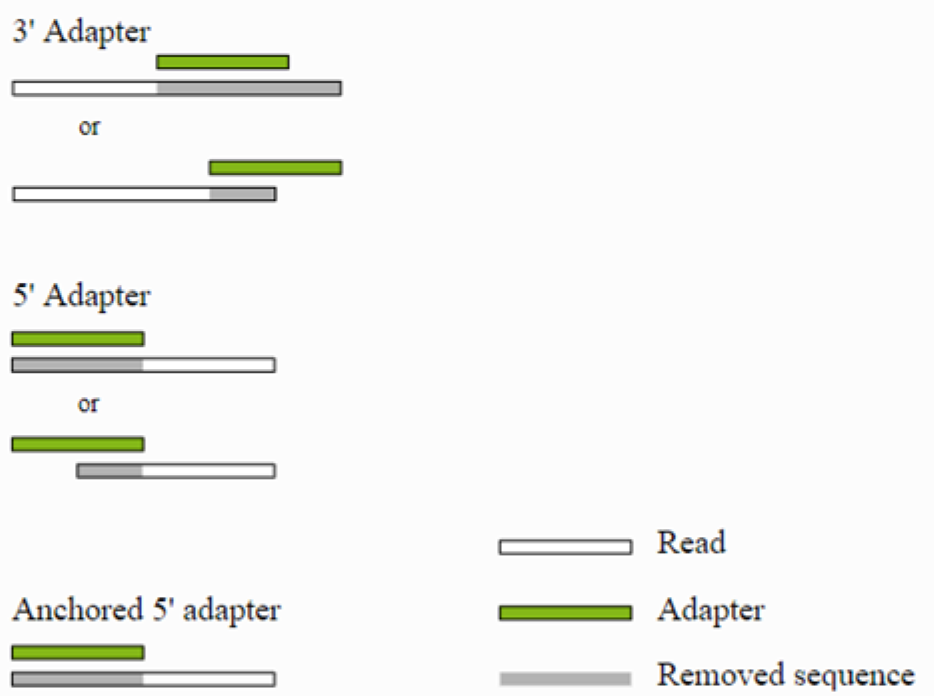




Step 2: Trimming adaptors from reads



```
@SRR6805880.2151832 OCD6D:00225:02960 length=80
TGCAGAAGGCATGACCTTACCTACTGAATAAAAGATGAGACACCTTCTCATTGGCCAAGAAGAAACAACACTCTATTACA
+
47:7775<59999995:6;;5:7664621111*/52245554404/33533/3/30436724461./,.:79999:4:9:
@SRR6805880.576388 9F8KO:05533:11649 length=80
TGCAGTCGTAATCTAGGAACACACCTACGGGATTATTTACTATTTTACAATCCATAGTCGGAGTCTACAAACAGTTACCA
+
135445878868?;:7474889//+/665628958::2788:>;;09:9556-315447817999::;::28///27:18
@SRR6805880.501486 9F8KO:05578:13178 length=80
TGCAGCAAGACCGTAGATCTGTCAAACGCAAAGCTTTAGCGAGCTCTCTAAGTAGCTTGAGAGGTCTGAAGAGAGCAGTG
+
-14556758885877766651////,18<=<4;;;<1::;;::65588::6;8888:49998<5:;;;;6:99;;:9;:;
@SRR6805880.1331889 J04RJ:03442:01185 length=80
TGCAGACTACATCAAAATGCATGACGATGTTACATACTGAATATATATATGCATATATATGTTTATTATACATAATGTAG
+
.337787/.--,,,,,),,-3355888:894:888988896:;;;:9>>:<:999766///6828:6:::9:::4:::98
@SRR6805880.2161340 OCD6D:00749:03136 length=80
TGCAGGCGATGGCCGTGGCGTCGATGCCGAACATGGTGACCTCGCAGGGCATGACATTTCAGGAACCGTTTCATAGTATG
+
15977689:8818178959988555:;5::6;=;<5:;9:59998::>2;;53378;;4;9<6<6<499;3:;;:99878
@SRR6805880.973930 J04RJ:09457:01591 length=80
TGCAGCATGTTGTAGTTTAAACTGCTTTTTTCGCATTTGTATTCCCAAATGAATGAAATATCGGAAATAGTCACAATTTC
+
-/2///6764157899:,33+/451/////'////3606678577,//*///14567/55688577255.....636627
```



Step 3: Index to Reference Genome

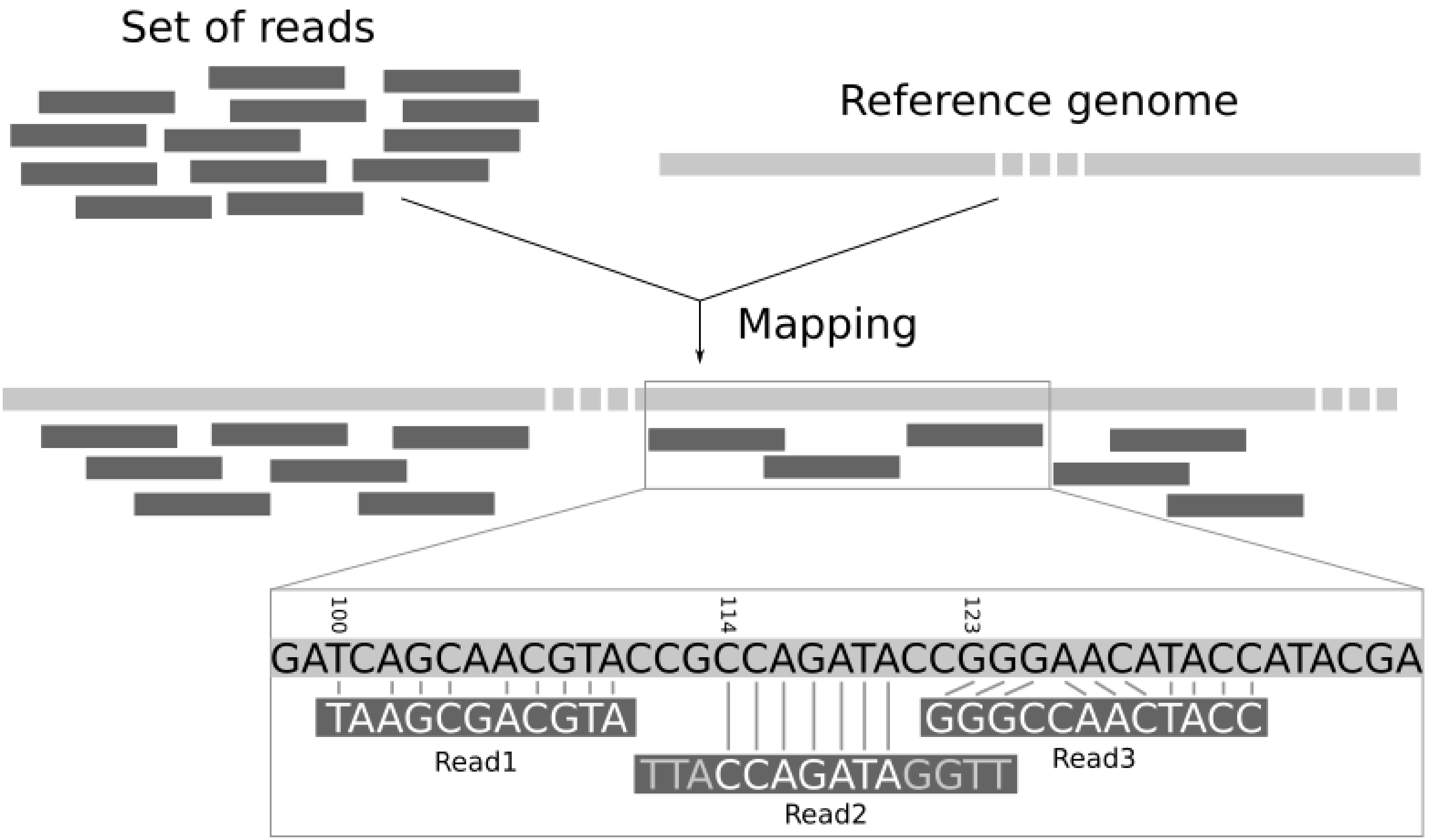
Bowtie 2

Fast and sensitive read alignment

Step 3: Index to Reference Genome

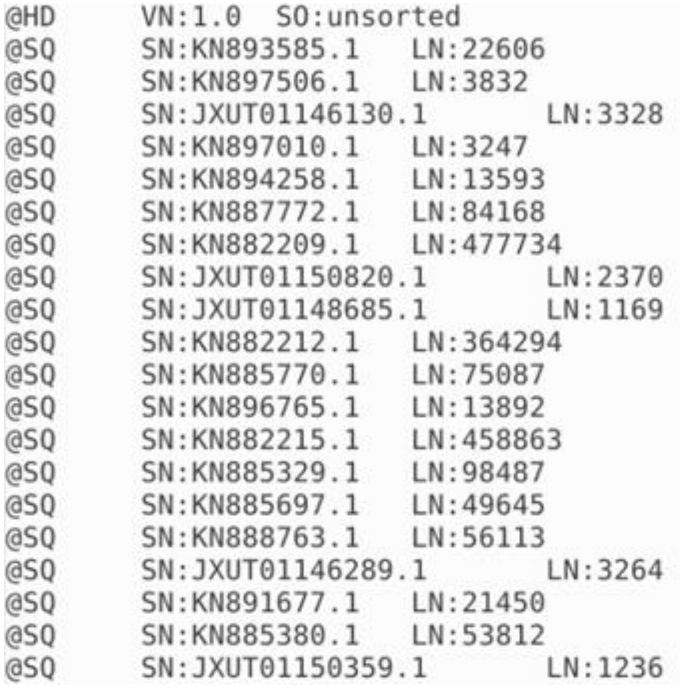
Bowtie 2
Fast and sensitive read alignment

Step 4: Mapping to the Reference Genome



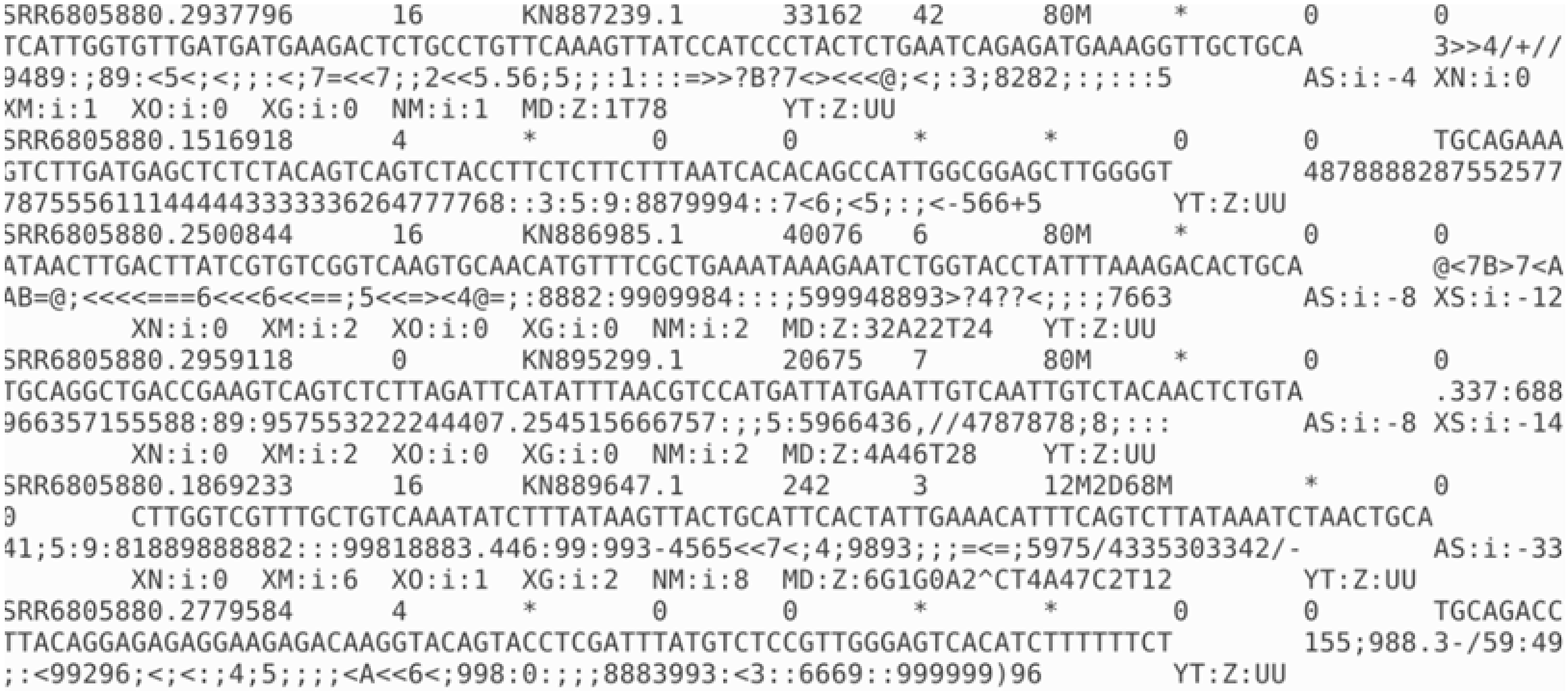
Step 5: SAM and BAM file formats

Sequence Alignment Map, Binary Alignment Map



```
@HD     VN:1.0  SO:unsorted
@SQ     SN:KN893585.1   LN:22606
@SQ     SN:KN897506.1   LN:3832
@SQ     SN:JXUT01146130.1       LN:3328
@SQ     SN:KN897010.1   LN:3247
@SQ     SN:KN894258.1   LN:13593
@SQ     SN:KN887772.1   LN:84168
@SQ     SN:KN882209.1   LN:477734
@SQ     SN:JXUT01150820.1       LN:2370
@SQ     SN:JXUT01148685.1       LN:1169
@SQ     SN:KN882212.1   LN:364294
@SQ     SN:KN885770.1   LN:75087
@SQ     SN:KN896765.1   LN:13892
@SQ     SN:KN882215.1   LN:458863
@SQ     SN:KN885329.1   LN:98487
@SQ     SN:KN885697.1   LN:49645
@SQ     SN:KN888763.1   LN:56113
@SQ     SN:JXUT01146289.1       LN:3264
@SQ     SN:KN891677.1   LN:21450
@SQ     SN:KN885380.1   LN:53812
@SQ     SN:JXUT01150359.1       LN:1236
```

Head of .sam file



```
SRR6805880.2937796      16      KN887239.1      33162   42      80M     *       0       0
TCATTGGTGTTGATGATGAAGACTCTGCCTGTTCAAAGTTATCCATCCCTACTCTGAATCAGAGATGAAAGGTTGCTGCA        3>>4/+//
9489:;89:<5<;<;;:<;7=<<7;;2<<5.56;5;;:1:::=>>?B?7<><<<@;<;:3;8282:;;:::5        AS:i:-4 XN:i:0
XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:1T78       YT:Z:UU
SRR6805880.1516918      4       *       0       0       *       *       0       0       TGCAGAAA
GTCTTGATGAGCTCTCTACAGTCAGTCTACCTTCTCTTCTTTAATCACACAGCCATTGGCGGAGCTTGGGGT        4878888287552577
78755561114444433333362647777 68::3:5:9:8879994::7<6;<5;:;<-566+5        YT:Z:UU
SRR6805880.2500844      16      KN886985.1      40076   6       80M     *       0       0
ATAACTTGACTTATCGTGTCGGTCAAGTGCAACATGTTTCGCTGAAATAAAGAATCTGGTACCTATTTAAAGACACTGCA        @<7B>7<A
AB=@;<<<<===6<<<6<<==;5<<=><4@=;:8882:9909984:::;599948893>?4??<;;:;7663        AS:i:-8 XS:i:-12
        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:32A22T24   YT:Z:UU
SRR6805880.2959118      0       KN895299.1      20675   7       80M     *       0       0
TGCAGGCTGACCGAAGTCAGTCTCTTAGATTCATATTTAACGTCCATGATTATGAATTGTCAATTGTCTACAACTCTGTA        .337:688
966357155588:89:957553222244407.254515666757:;;5:5966436,//4787878;8;:::        AS:i:-8 XS:i:-14
        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:4A46T28    YT:Z:UU
SRR6805880.1869233      16      KN889647.1      242     3       12M2D68M        *       0
0       CTTGGTCGTTTGCTGTCAAATATCTTTATAAGTTACTGCATTCACTATTGAAACATTTCAGTCTTATAAATCTAACTGCA
41;5:9:81889888882:::99818883.446:99:993-4565<<7<;4;9893;;;=<=;5975/4335303342/-        AS:i:-33
        XN:i:0  XM:i:6  XO:i:1  XG:i:2  NM:i:8  MD:Z:6G1G0A2^CT4A47C2T12        YT:Z:UU
SRR6805880.2779584      4       *       0       0       *       *       0       0       TGCAGACC
TTACAGGAGAGAGGAAGAGACAAGGTACAGTACCTCGATTTATGTCTCCGTTGGGAGTCACATCTTTTTTCT        155;988.3-/59:49
;:<99296;<;<:;4;5;;;;<A<<6<;998:0:;;;8883993:<3::6669::999999)96        YT:Z:UU
```

Tail of .sam file

Step 5: SAM and BAM file formats

Sequence Alignment Map, Binary Alignment Map

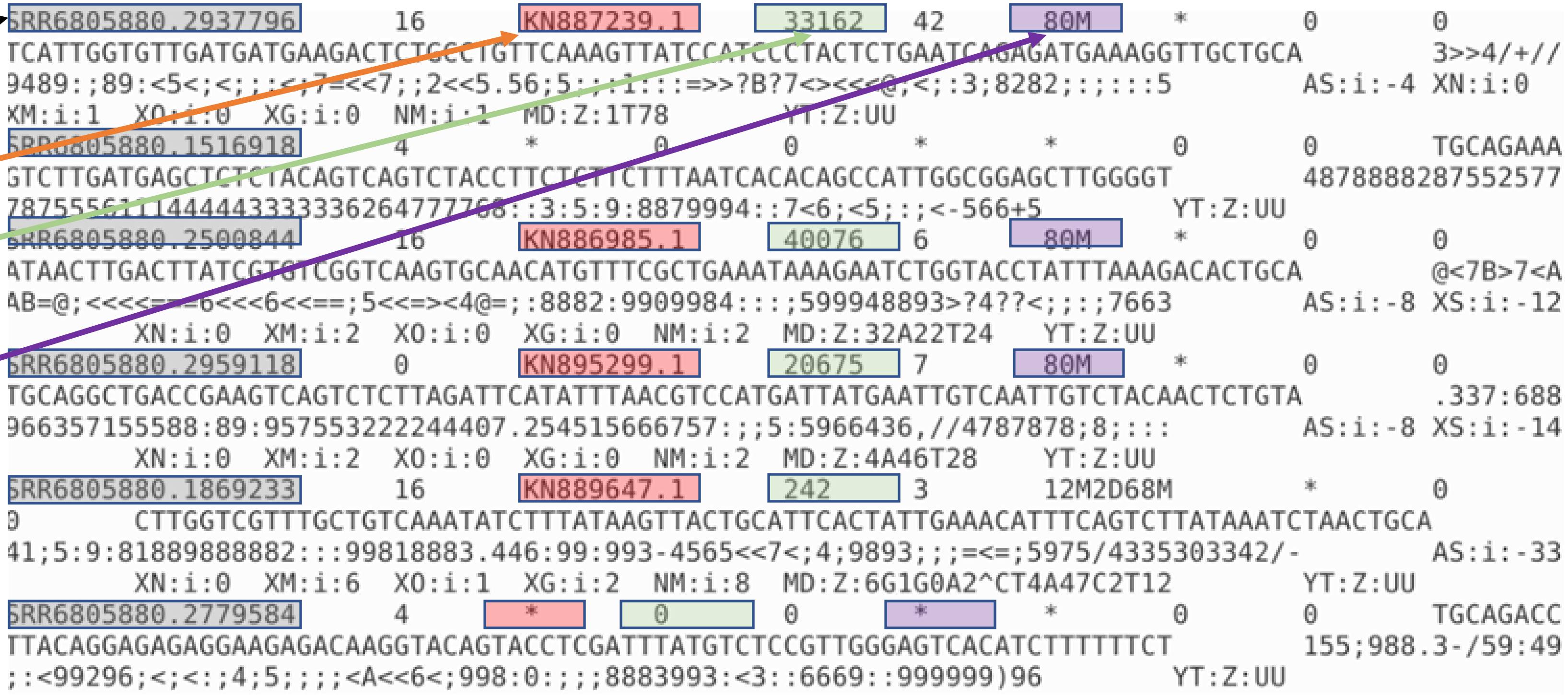
Name of read

Name of contig where read aligns

Position on contig where 5’ end starts

Alignment information “cigar string”

80M = contiguous match of 80bp



```
SRR6805880.2937796      16      KN887239.1      33162   42      80M     *       0       0
TCATTGGTGTTGATGATGAAGACTCTGCCTGTTCAAAGTTATCCATCCCTACTCTGAATCAGAGATGAAAGGTTGCTGCA          3>>4/+//
9489:;89:<5<;<;;;<;7=<<7;;2<<5.56;5;;:1:::=>>?B?7<><<<@;<;:3;8282;:;:::5          AS:i:-4 XN:i:0
XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:1T78       YT:Z:UU
SRR6805880.1516918      4       *       0       0       *       *       0       0       TGCAGAAA
GTCTTGATGAGCTCTCTACAGTCAGTCTACCTTCTCTTCTTTAATCACACAGCCATTGGCGGAGCTTGGGGT          4878888287552577
7875556111444443333336264777768::3:5:9:8879994::7<6;<5;:;<-566+5          YT:Z:UU
SRR6805880.2500844      16      KN886985.1      40076   6       80M     *       0       0
ATAACTTGACTTATCGTGTCGGTCAAGTGCAACATGTTTCGCTGAAATAAAGAATCTGGTACCTATTTAAAGACACTGCA          @<7B>7<A
AB=@;<<<<===6<<<6<<==;5<<=><4@=;:8882:9909984:::;599948893>?4??<;;:;7663          AS:i:-8 XS:i:-12
        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:32A22T24   YT:Z:UU
SRR6805880.2959118      0       KN895299.1      20675   7       80M     *       0       0
TGCAGGCTGACCGAAGTCAGTCTCTTAGATTCATATTTAACGTCCATGATTATGAATTGTCAATTGTCTACAACTCTGTA          .337:688
966357155588:89:957553222244407.254515666757:;;5:5966436,//4787878;8;:::          AS:i:-8 XS:i:-14
        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:4A46T28    YT:Z:UU
SRR6805880.1869233      16      KN889647.1      242     3       12M2D68M        *       0
0       CTTGGTCGTTTGCTGTCAAATATCTTTATAAGTTACTGCATTCACTATTGAAACATTTCAGTCTTATAAATCTAACTGCA
41;5:9:81889888882:::99818883.446:99:993-4565<<7<;4;9893;;;=<=;5975/4335303342/-          AS:i:-33
        XN:i:0  XM:i:6  XO:i:1  XG:i:2  NM:i:8  MD:Z:6G1G0A2^CT4A47C2T12        YT:Z:UU
SRR6805880.2779584      4       *       0       0       *       *       0       0       TGCAGACC
TTACAGGAGAGAGGAAGAGACAAGGTACAGTACCTCGATTTATGTCTCCGTTGGGAGTCACATCTTTTTTCT          155;988.3-/59:49
;:<99296;<;<:;4;5;;;;<A<<6<;998:0:;;;8883993:<3::6669::999999)96          YT:Z:UU
```

Tail of .sam file

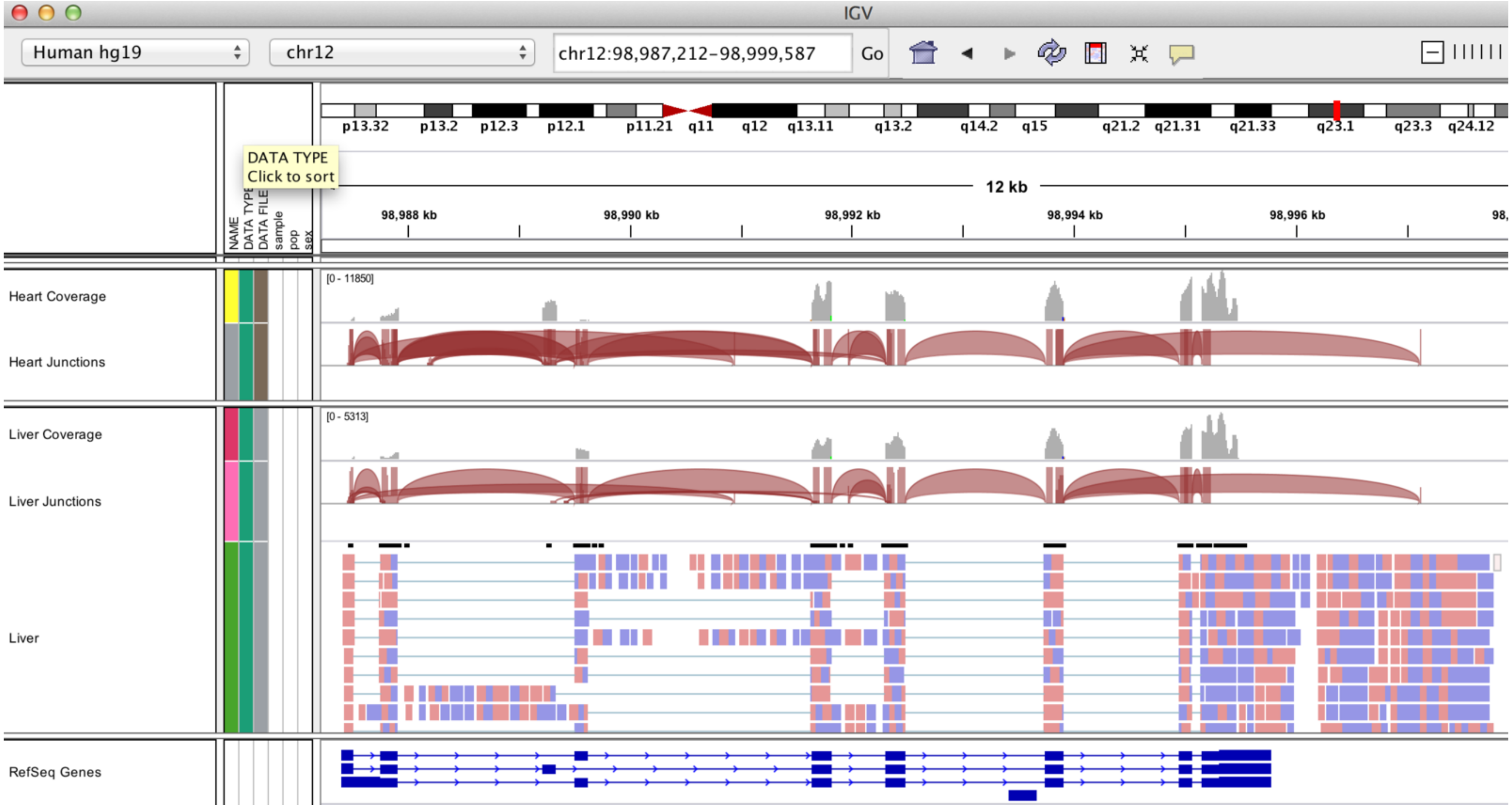
Step 6: “Calling” a genotype

Two main ways to estimate a genotype

1. Hard genotype calling
2. Genotype likelihoods

Depends on the type of data that you have. If you have reads with a high degree of coverage (many copies of the same read) you can do hard genotype calling. If you have variable coverage or low coverage you can do genotype likelihoods, to account for some uncertainty in the genotype.

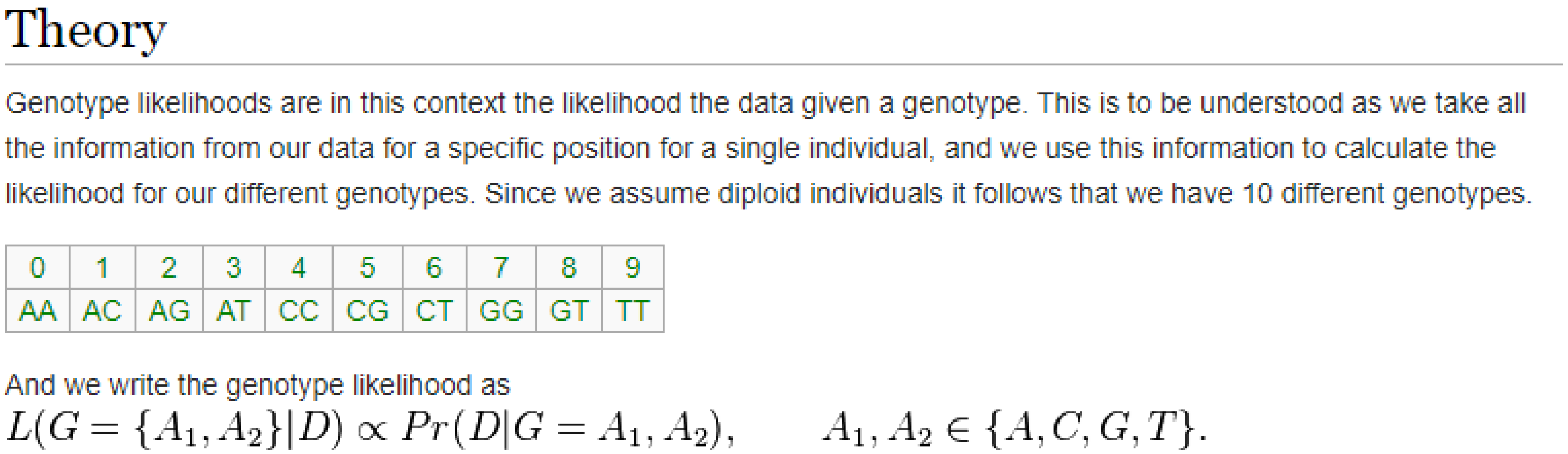
Step 6: “Calling” a genotype: alignment



Step 6: Genotype likelihoods

In ANGSD http://www.popgen.dk/angsd/index.php/Genotype_Likelihoods

Accounts for some uncertainty in the genotype estimation



Theory

Genotype likelihoods are in this context the likelihood the data given a genotype. This is to be understood as we take all the information from our data for a specific position for a single individual, and we use this information to calculate the likelihood for our different genotypes. Since we assume diploid individuals it follows that we have 10 different genotypes.

0	1	2	3	4	5	6	7	8	9
AA	AC	AG	AT	CC	CG	CT	GG	GT	TT

And we write the genotype likelihood as

$$L(G=\{A_1,A_2\}|D) \propto Pr(D|G=A_1,A_2), \qquad A_1,A_2 \in \{A,C,G,T\}.$$