Marine Genomics

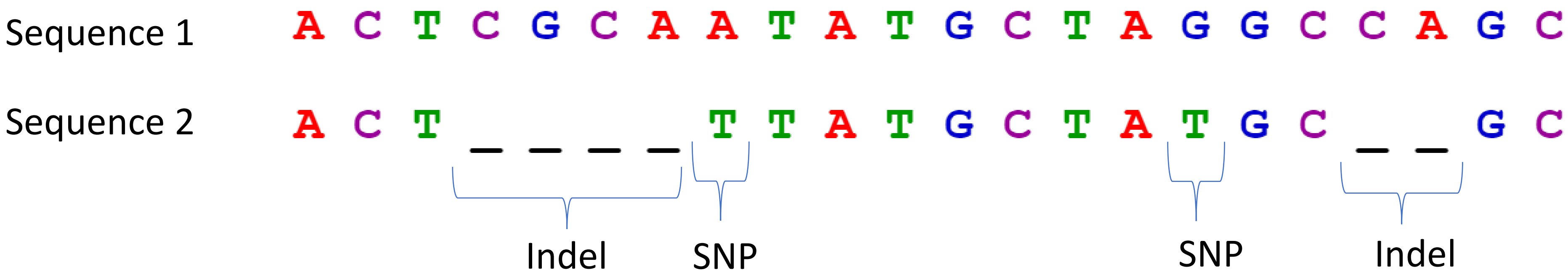
April 13th 2021

Mapping and calling variants

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

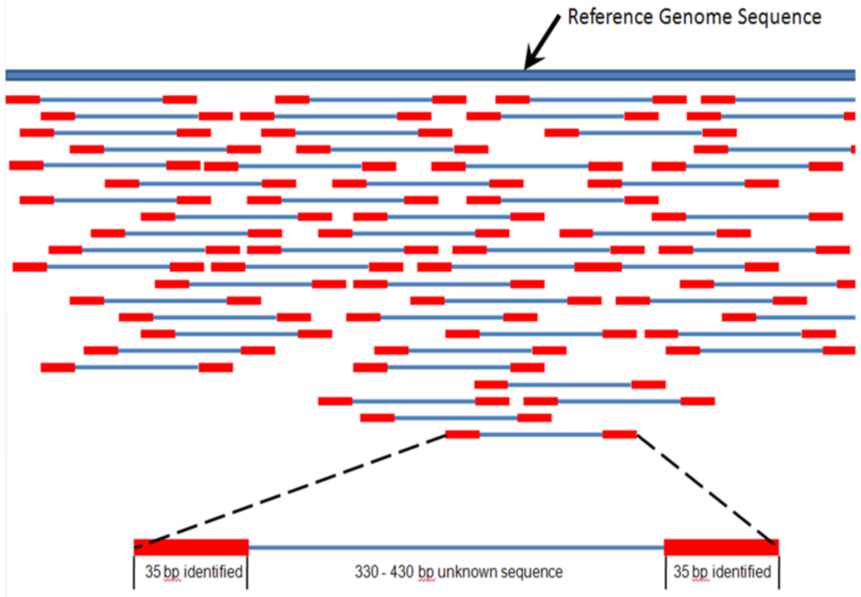
Indel SNP SNP Indel

How do we find a variant?

Map and align sequences from other individuals to a reference genome

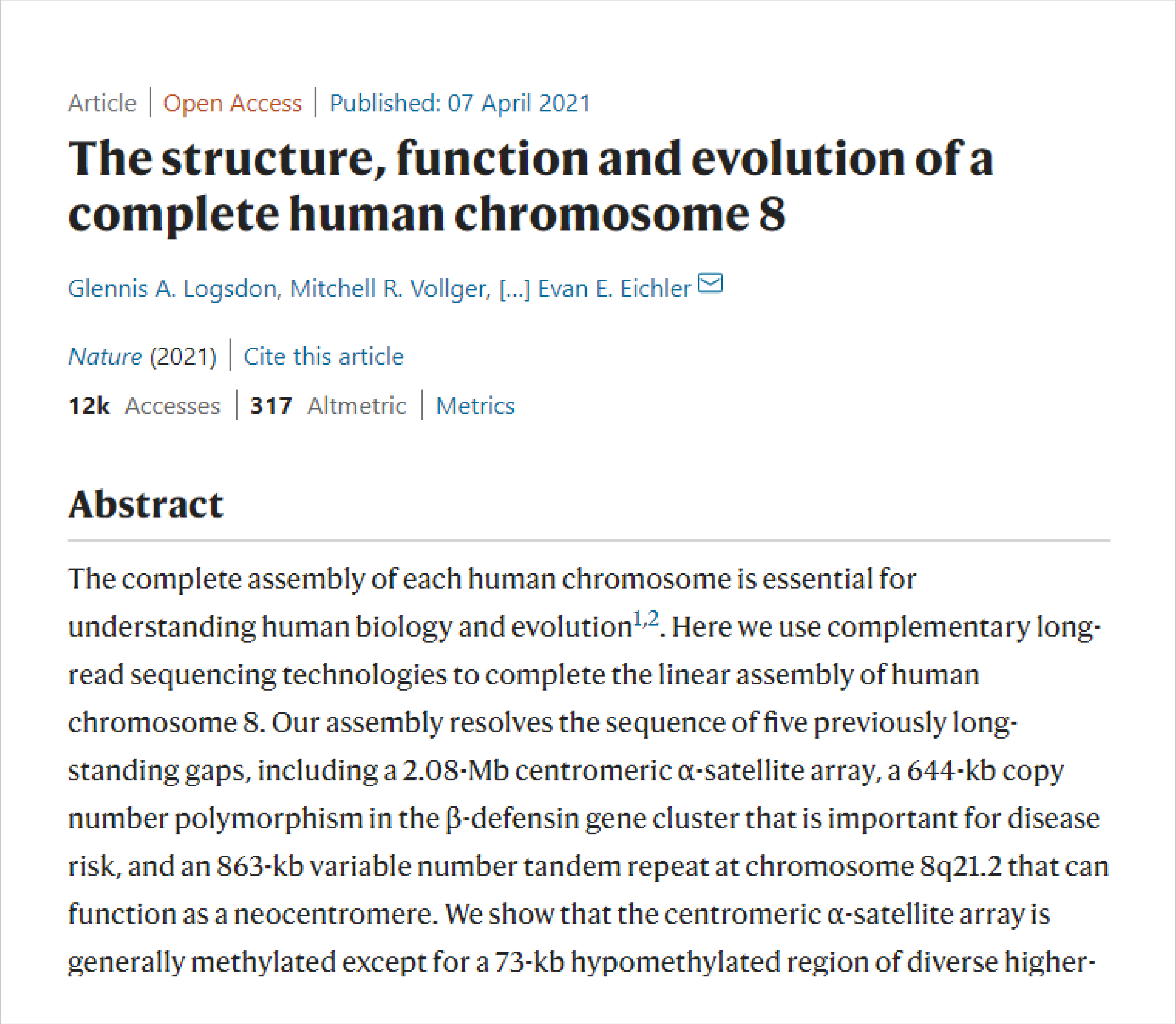
- Does It matter what your reference genome is?
  - Is it the same or different species?
  - Is it from the same population?

- Short answer: Yes, it matters!



Genomes are continually being improved

- More genomes are being sequenced all the time
- Many marine organisms don't yet have a genome sequence available



Article | Open Access | Published: 07 April 2021

The structure, function and evolution of a complete human chromosome 8

Glennis A. Logsdon, Mitchell R. Vollger, [...] Evan E. Eichler

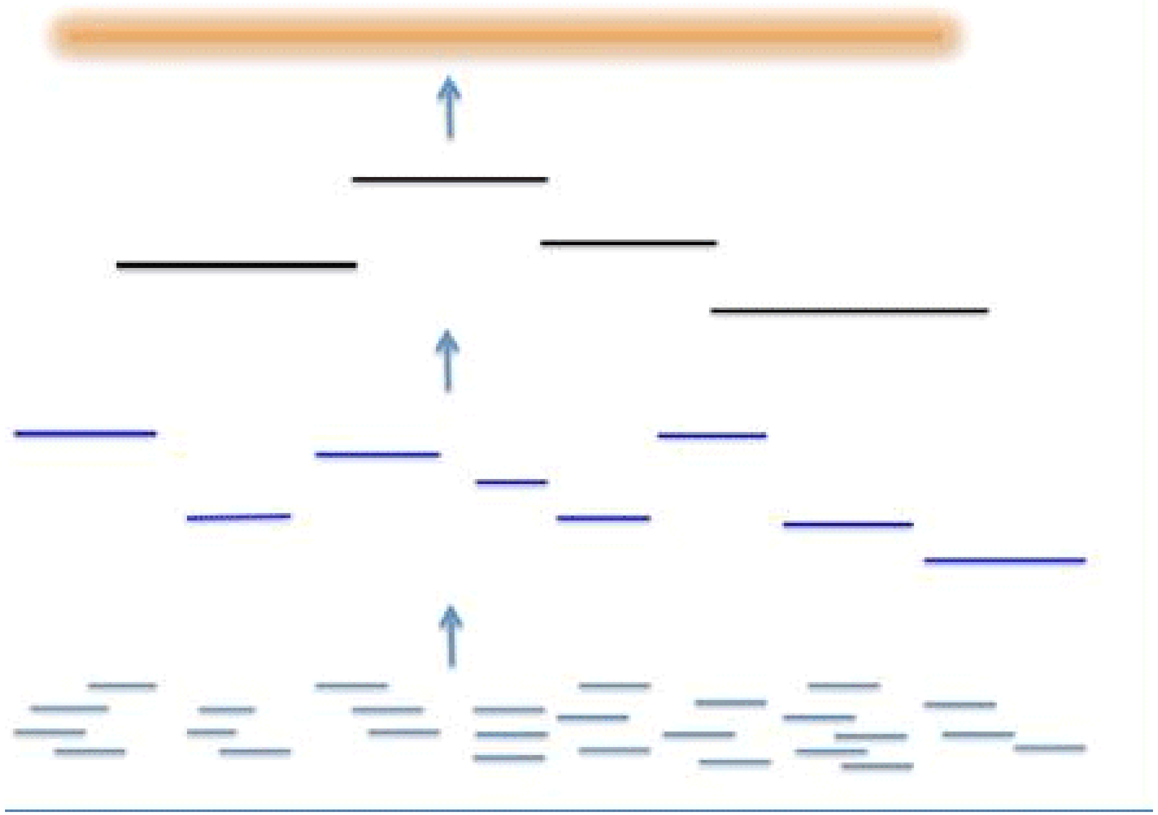
Nature (2021) | Cite this article

12k Accesses | **317** Altmetric | Metrics

Abstract

The complete assembly of each human chromosome is essential for understanding human biology and evolution[1,2]. Here we use complementary long-read sequencing technologies to complete the linear assembly of human chromosome 8. Our assembly resolves the sequence of five previously long-standing gaps, including a 2.08-Mb centromeric α-satellite array, a 644-kb copy number polymorphism in the β-defensin gene cluster that is important for disease risk, and an 863-kb variable number tandem repeat at chromosome 8q21.2 that can function as a neocentromere. We show that the centromeric α-satellite array is generally methylated except for a 73-kb hypomethylated region of diverse higher-

Finding variants – some terminology



Chromosome

Scaffolds

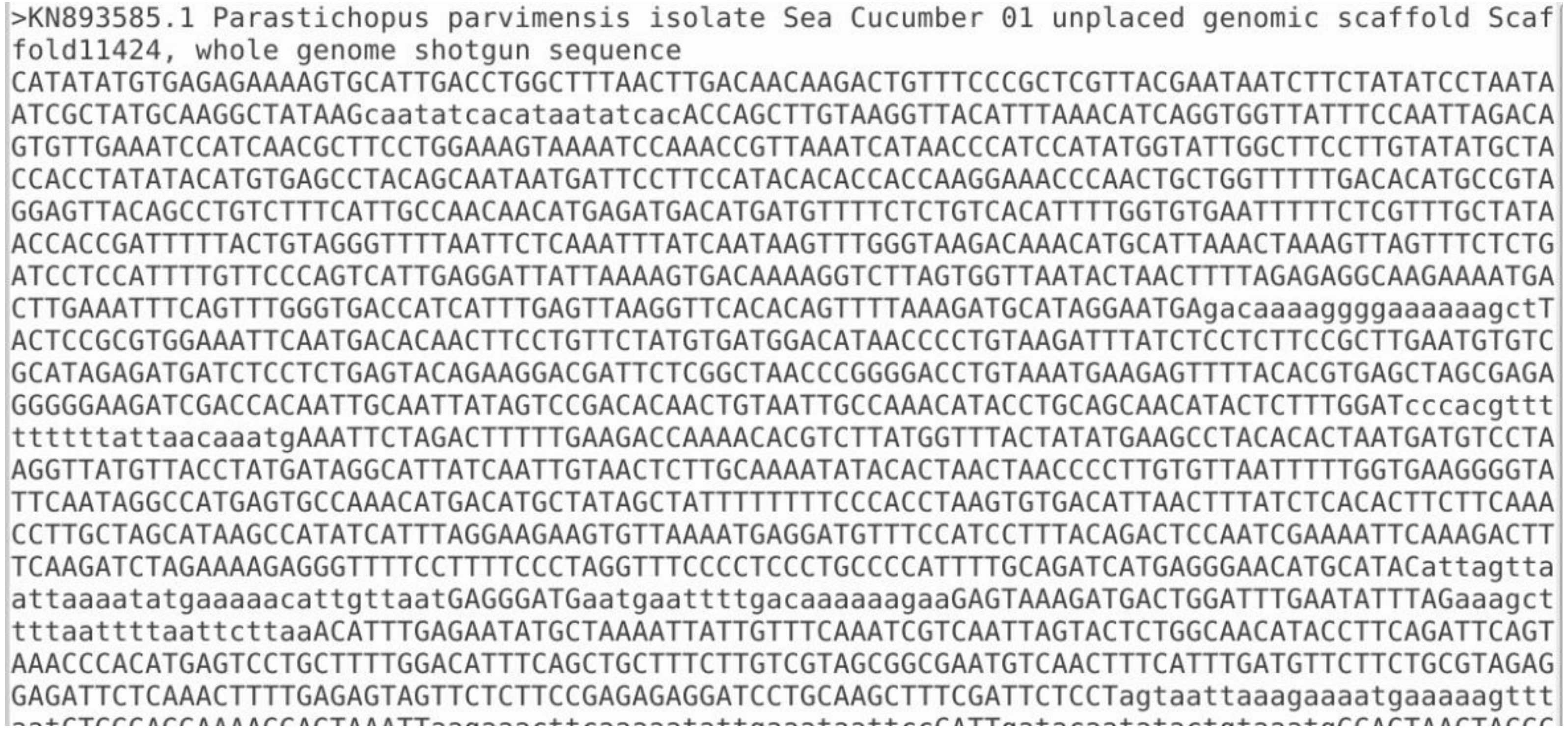
Contigs

Reads

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

A reference genome is a collection of contigs

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
```

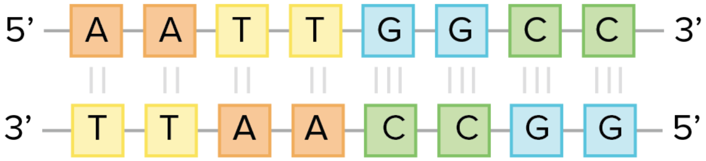
A reference genome is a collection of contigs

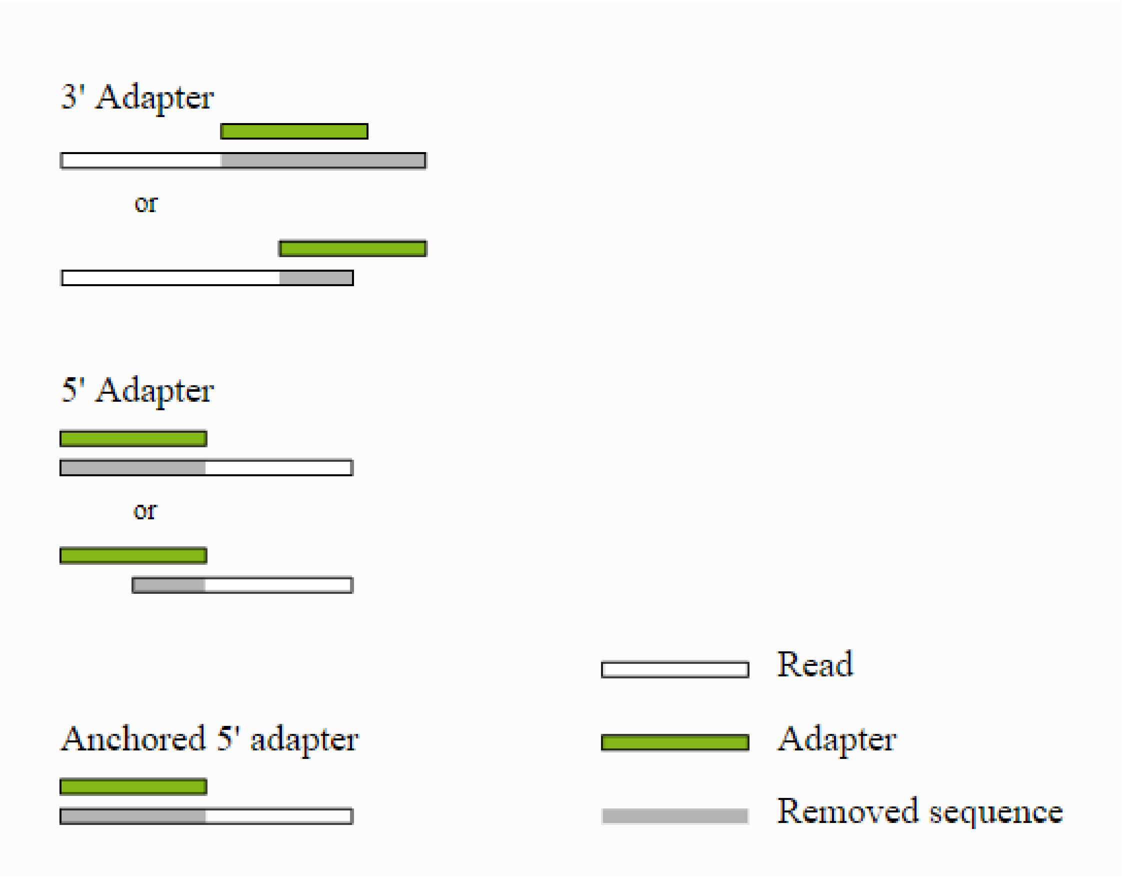
Typically in fasta format

Finding variants - pipeline

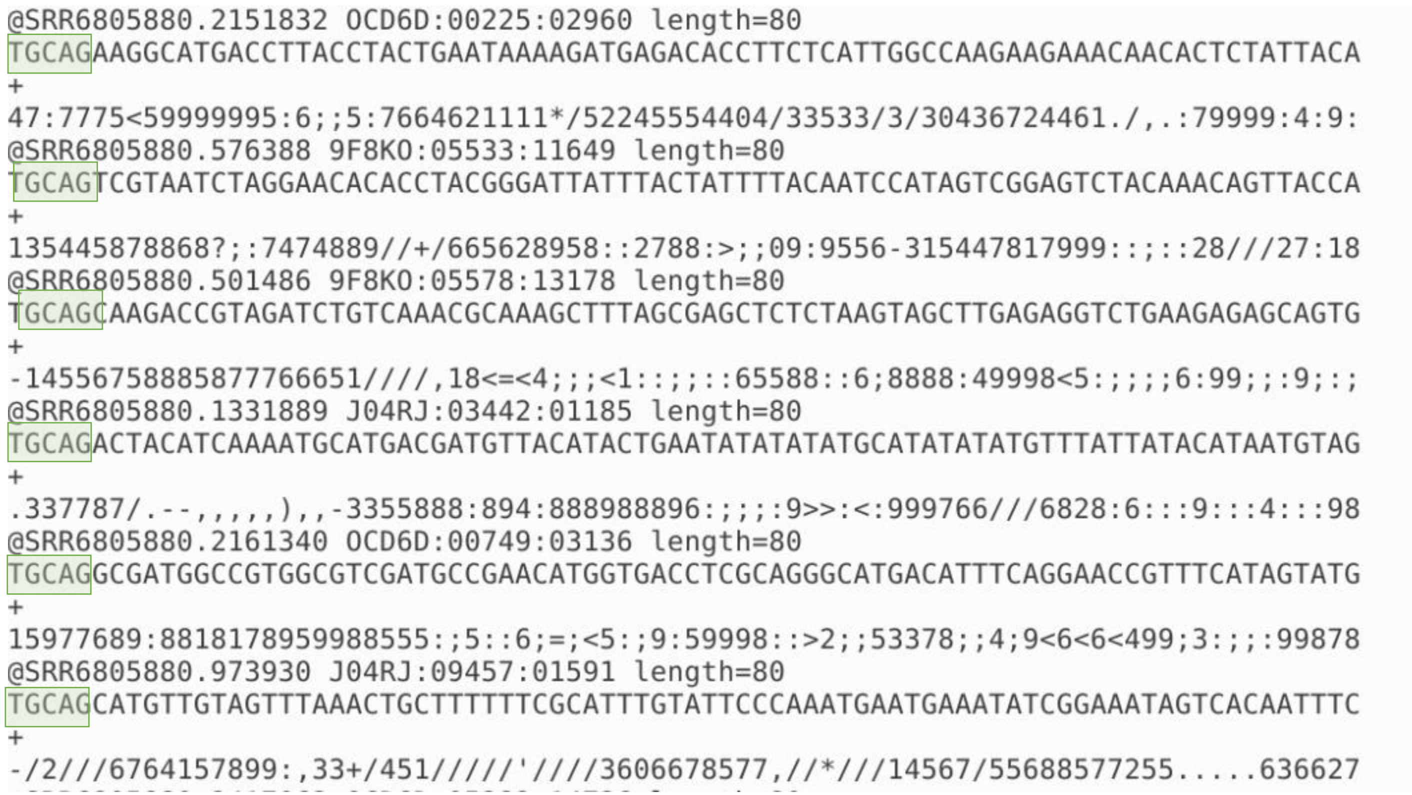
- Get reads and genome (download from git hub)
- Trim adapters off of reads (cutadapt)
- Index genome (bowtie2)
- Map reads to genome -> generate a sam file -> convert to bam file (bowtie2, samtools)
- Calculate genotype likelihoods (angsd, samtools)
- Happy dance

Trimming adaptors from reads

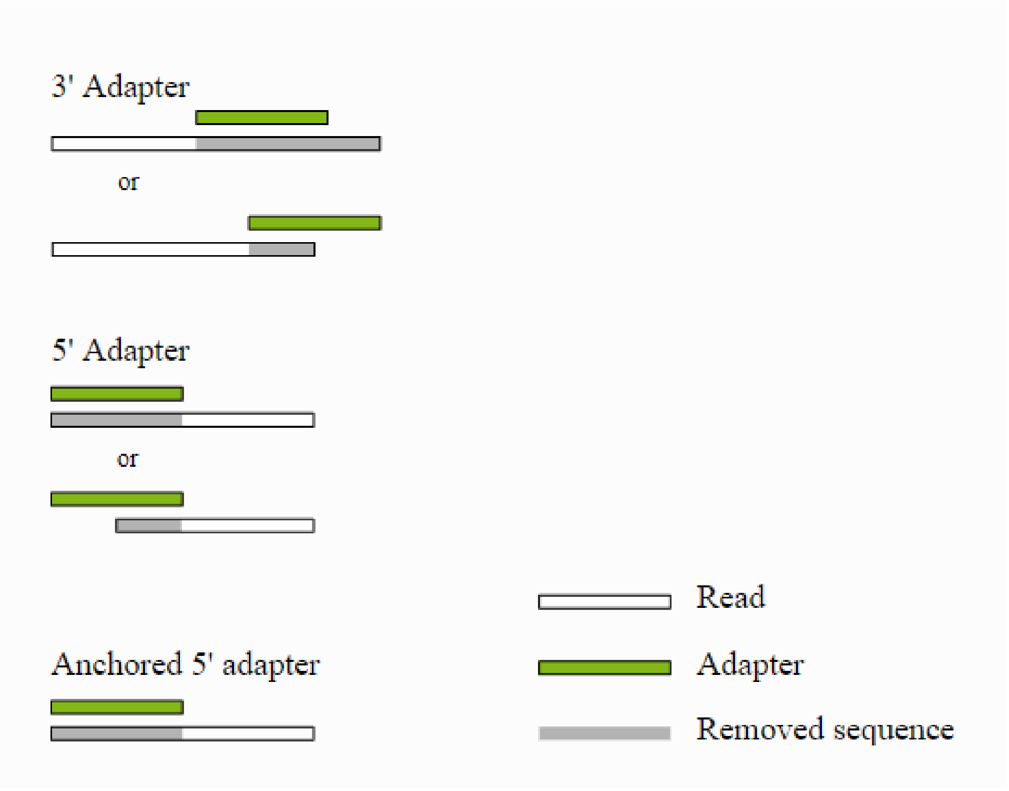




Trimming adaptors from reads

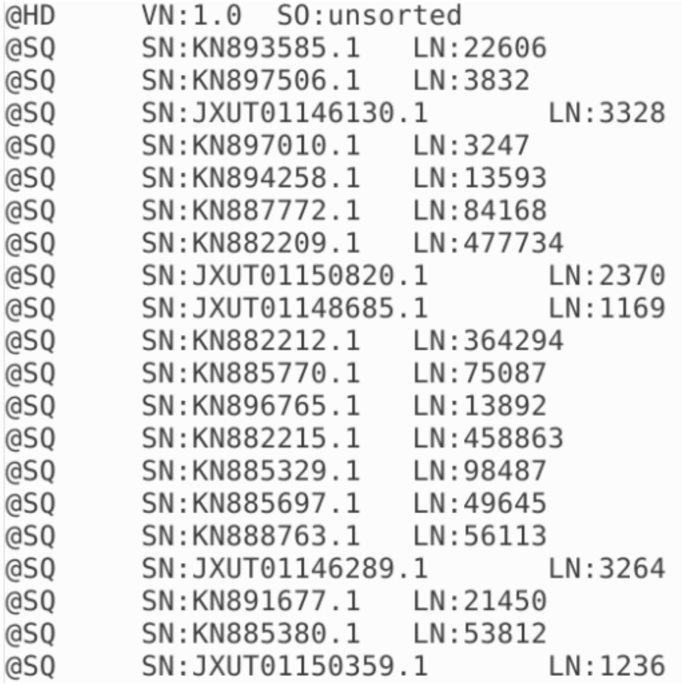


```
@SRR6805880.2151832 OCD6D:00225:02960 length=80
TGCAGAAGGCATGACCTTACCTACTGAATAAAAGATGAGACACCTTCTCATTGGCCAAGAAGAAACAACACTCTATTACA
+
47:7775<59999995:6;;5:7664621111*/52245554404/33533/3/30436724461./,.:79999:4:9:
@SRR6805880.576388 9F8KO:05533:11649 length=80
TGCAGTCGTAATCTAGGAACACACCTACGGGATTATTTACTATTTTACAATCCATAGTCGGAGTCTACAAACAGTTACCA
+
135445878868?;:7474889//+/665628958::2788:>;;09:9556-315447817999::;::28///27:18
@SRR6805880.501486 9F8KO:05578:13178 length=80
TGCAGCAAGACCGTAGATCTGTCAAACGCAAAGCTTTAGCGAGCTCTCTAAGTAGCTTGAGAGGTCTGAAGAGAGCAGTG
+
-1455675888587776665l////,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
```



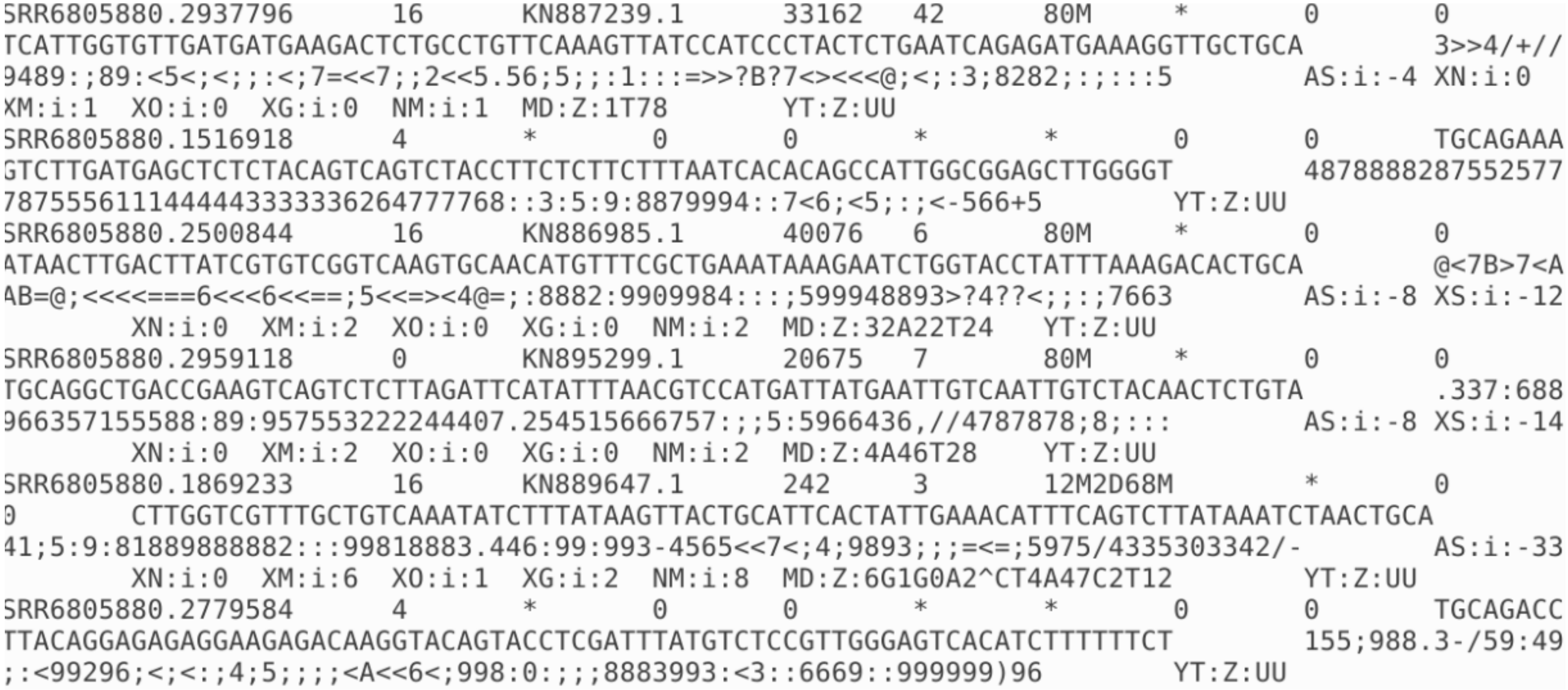
SAM and BAM file formats

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
78755561114444433333362647777768::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

SAM and BAM file formats

Sequence Alignment Map, Binary Alignment Map

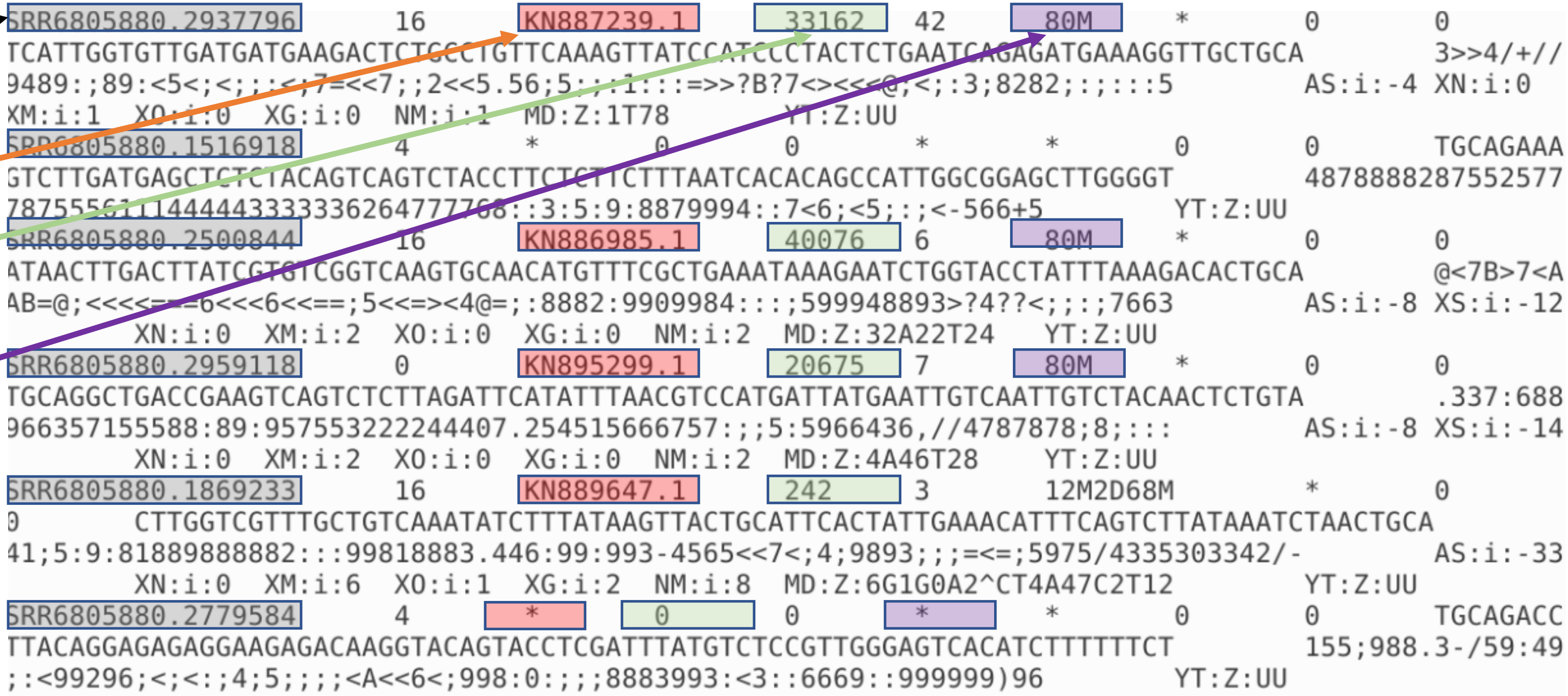
Name of read

Name of contig where read aligns

Position on contig where 5’ end starts

Alignment information “cigar string”

80M = contiguous match of 80bp



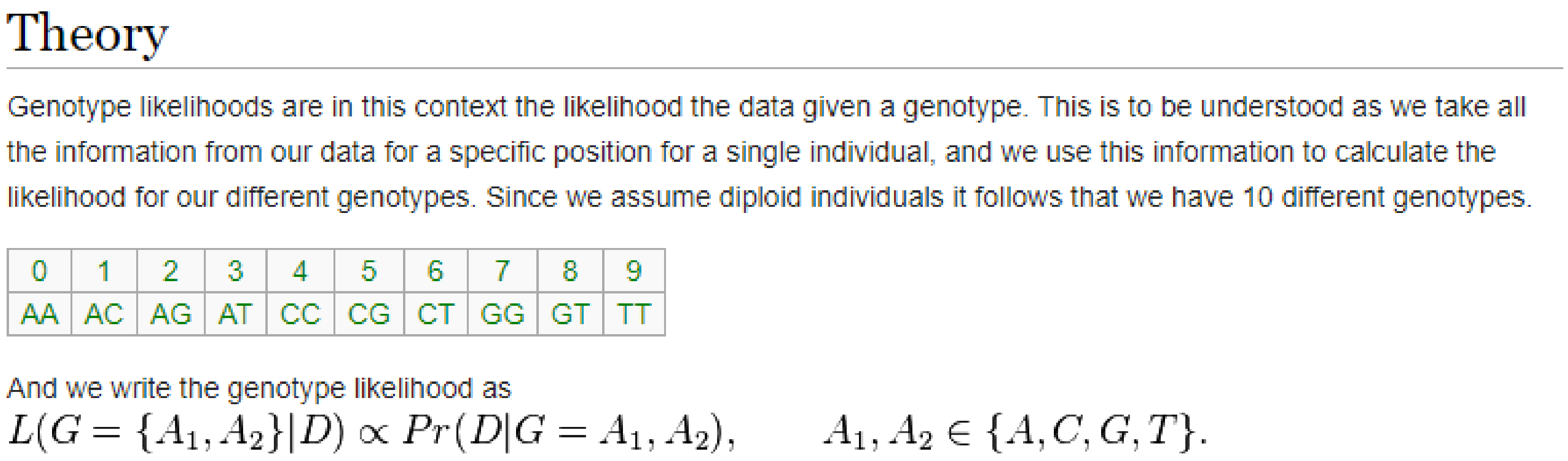
```
SRR6805880.2937796      16      KN887239.1      33162   42      80M     *       0       0
TCATTGGTGTTGATGATGAAGACTCTGCCTGTTCAAAGTTATCCATCCCTACTCTGAATCAGAGATGAAAGGTTGCTGCA          3>>4/+//
9489:;89:<5<;<;;:<;7=<<7;;2<<5.56;5:;:1:::=>>?B?7<><<<@;<;:3;8282:;;:::5          AS:i:-4 XN:i:0
XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:1T78       YT:Z:UU
SRR6805880.1516918      4       *       0       0       *       *       0       0       TGCAGAAA
GTCTTGATGAGCTCTCTACAGTCAGTCTACCTTCTCTTCTTTAATCACACAGCCATTGGCGGAGCTTGGGGT          4878888287552577
78755561114444433333362647777768::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

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

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