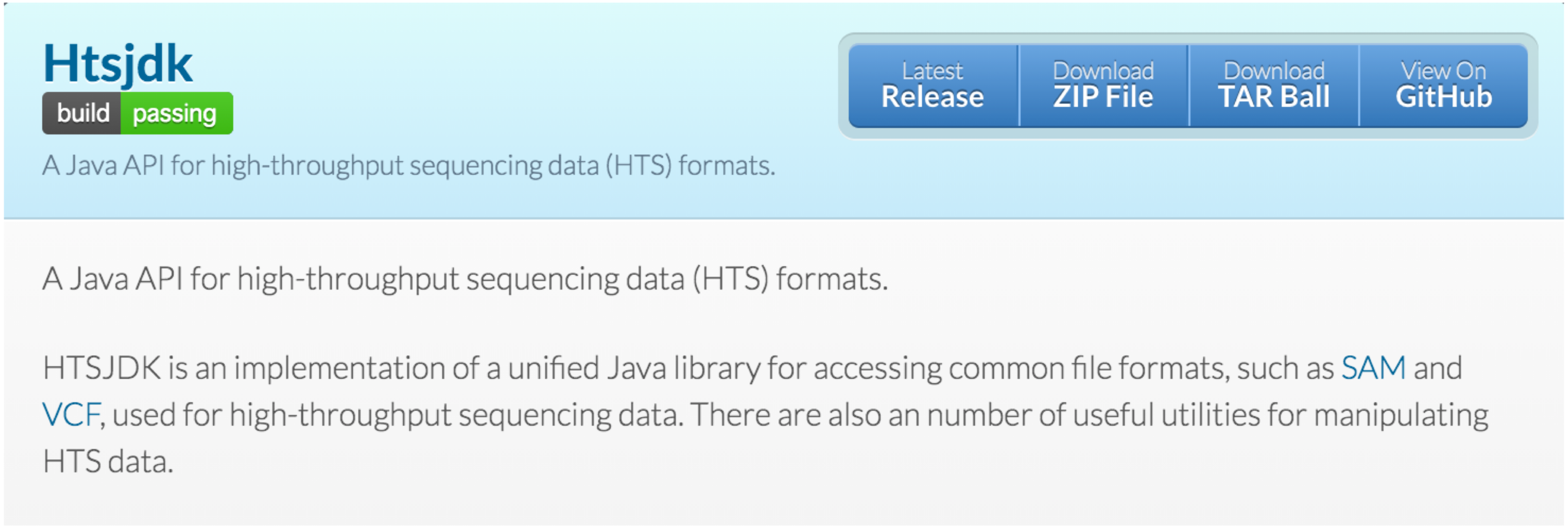
HTSJDK Project

CZI EOSS Conference December 2020

Chris Norman

BROAD INSTITUTE

High Throughput Sequencing Java Development Kit



Htsjdk

A Java API for high-throughput sequencing data (HTS) formats.

A Java API for high-throughput sequencing data (HTS) formats.

HTSJDK is an implementation of a unified Java library for accessing common file formats, such as SAM and VCF, used for high-throughput sequencing data. There are also an number of useful utilities for manipulating HTS data.

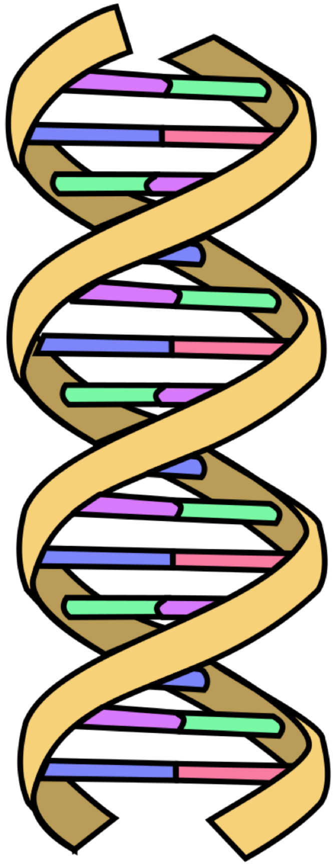
HTSJDK Features/APIs

- Read and write for common genomics file formats
  - SAM / BAM / CRAM / VCF / BED / FASTA / FASTQ
  - companion file formats (indices, dictionaries, checksum, etc.)
- Reference implementations for GA4GH formats
- Random access queries
- Remote access (i.e., SRA, htsGET)
- Compression/Decompression
- Ancillary structures (genomic coordinates, intervals)
- Format Validation/Spec Compliance

HTSJDK Consumers

- GATK (Genome Analysis Toolkit)
- Picard
- IGV (Integrated Genomics Viewer)
- Disq
- GenomicsDB
- Hadoop-BAM
- VariantQC
- fgbio

DNA Molecule Structure



DNA

🟩 = Adenine

🟪 = Thymine

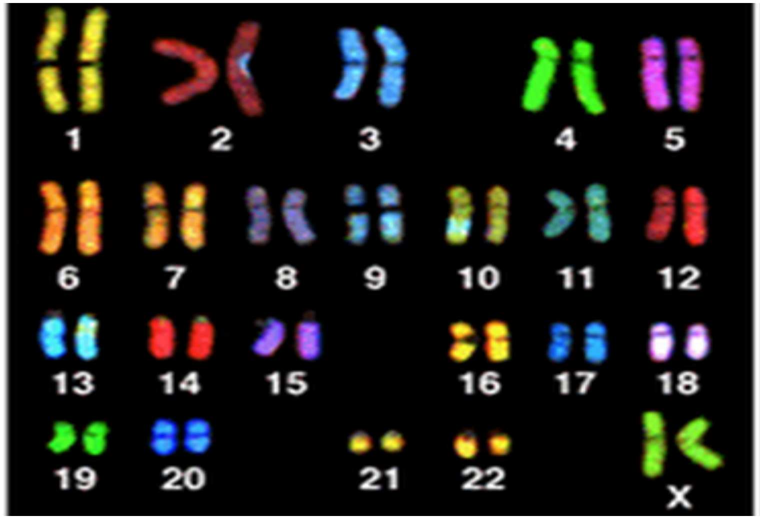
🟥 = Cytosine

🟦 = Guanine

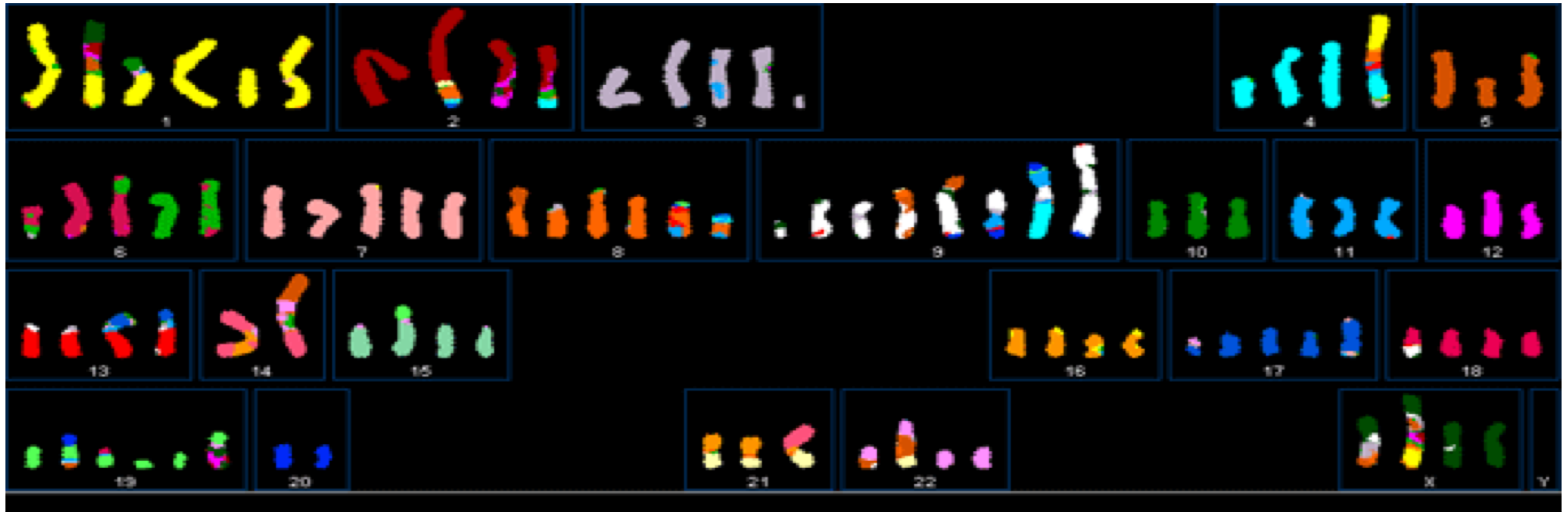
🟨 = Phosphate backbone

High Level Structure of the Genome

Normal Cell

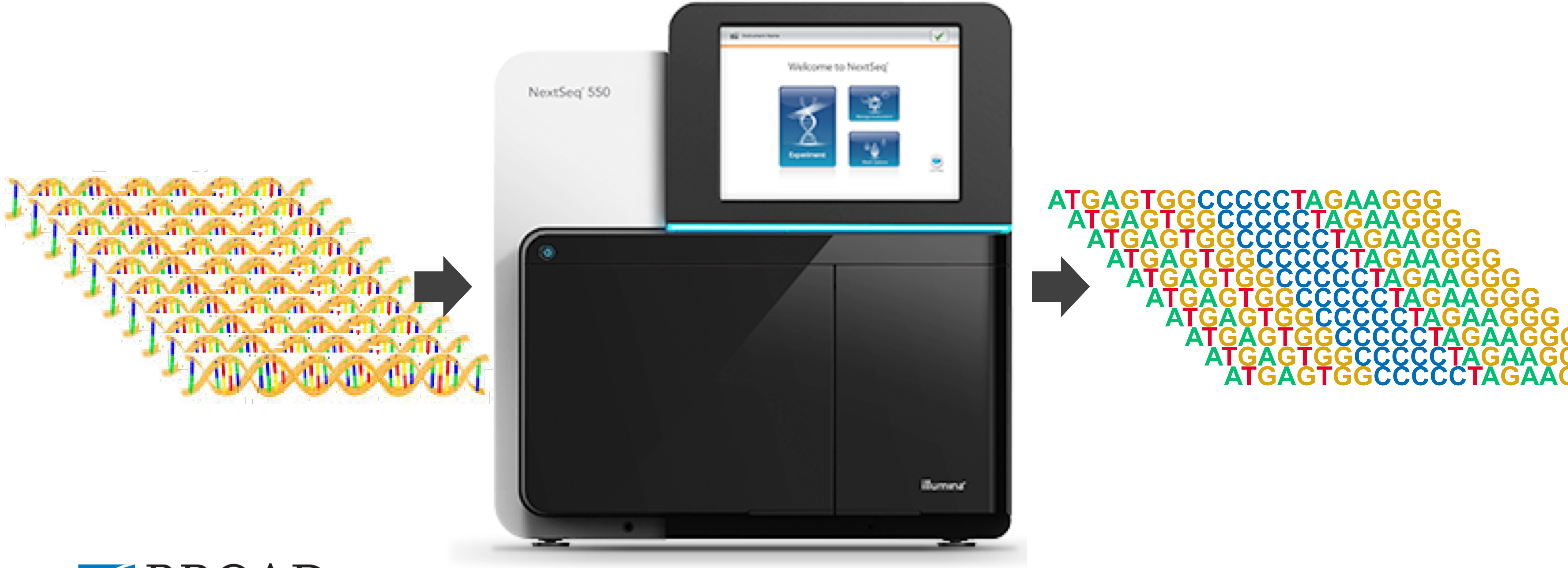


Cancer Cell Line HCC1954



- Spectral karyotyping paints each chromosome pair with a color
- Alterations can vary dramatically between cancers and within cancers

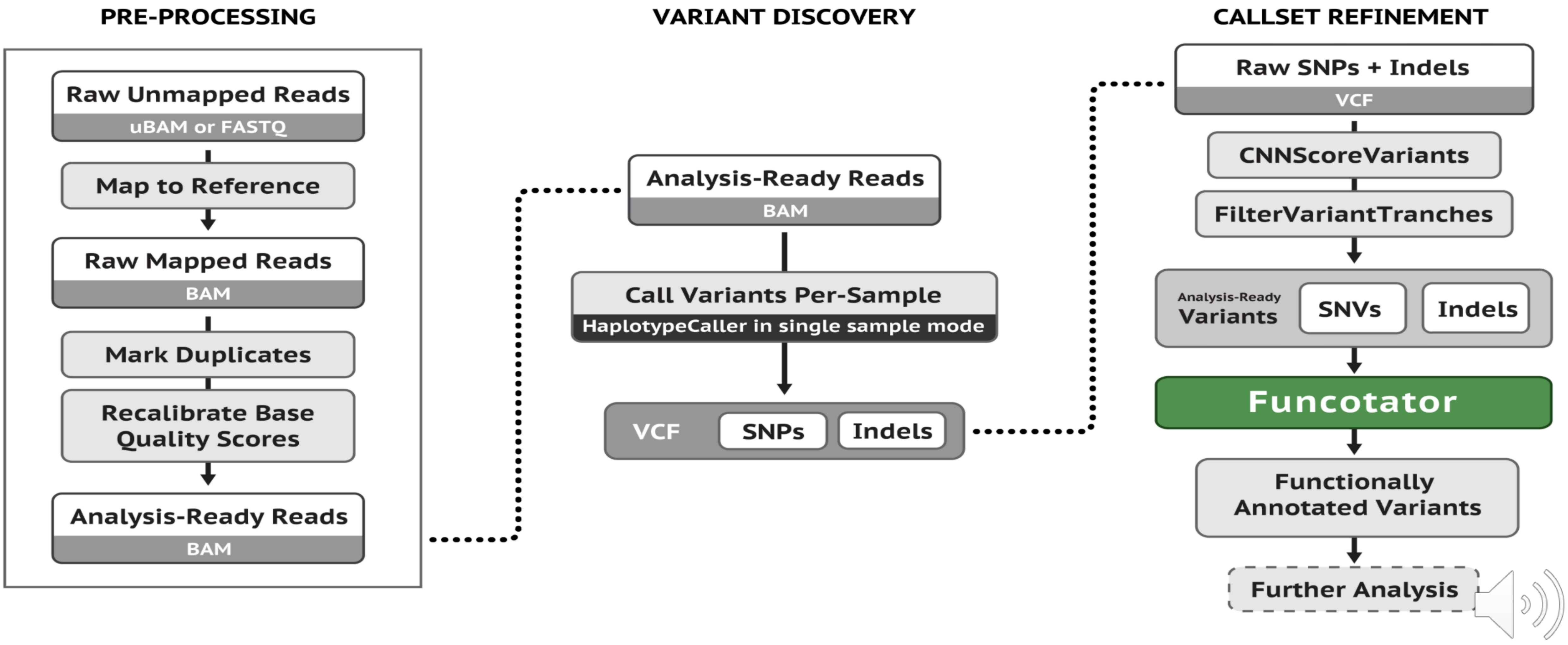
DNA Sequencing



ATGAGTGGCCCCCTAGAAGGG
ATGAGTGGCCCCCTAGAAGGG
ATGAGTGGCCCCCTAGAAGGG
ATGAGTGGCCCCCTAGAAGGG
ATGAGTGGCCCCCTAGAAGGG
ATGAGTGGCCCCCTAGAAGGG
ATGAGTGGCCCCCTAGAAGGG
ATGAGTGGCCCCCTAGAAGG
ATGAGTGGCCCCCTAGAAG
ATGAGTGGCCCCCTAGAA

BROAD
INSTITUTE

Germline Single-Sample Pipeline



Data Standardization

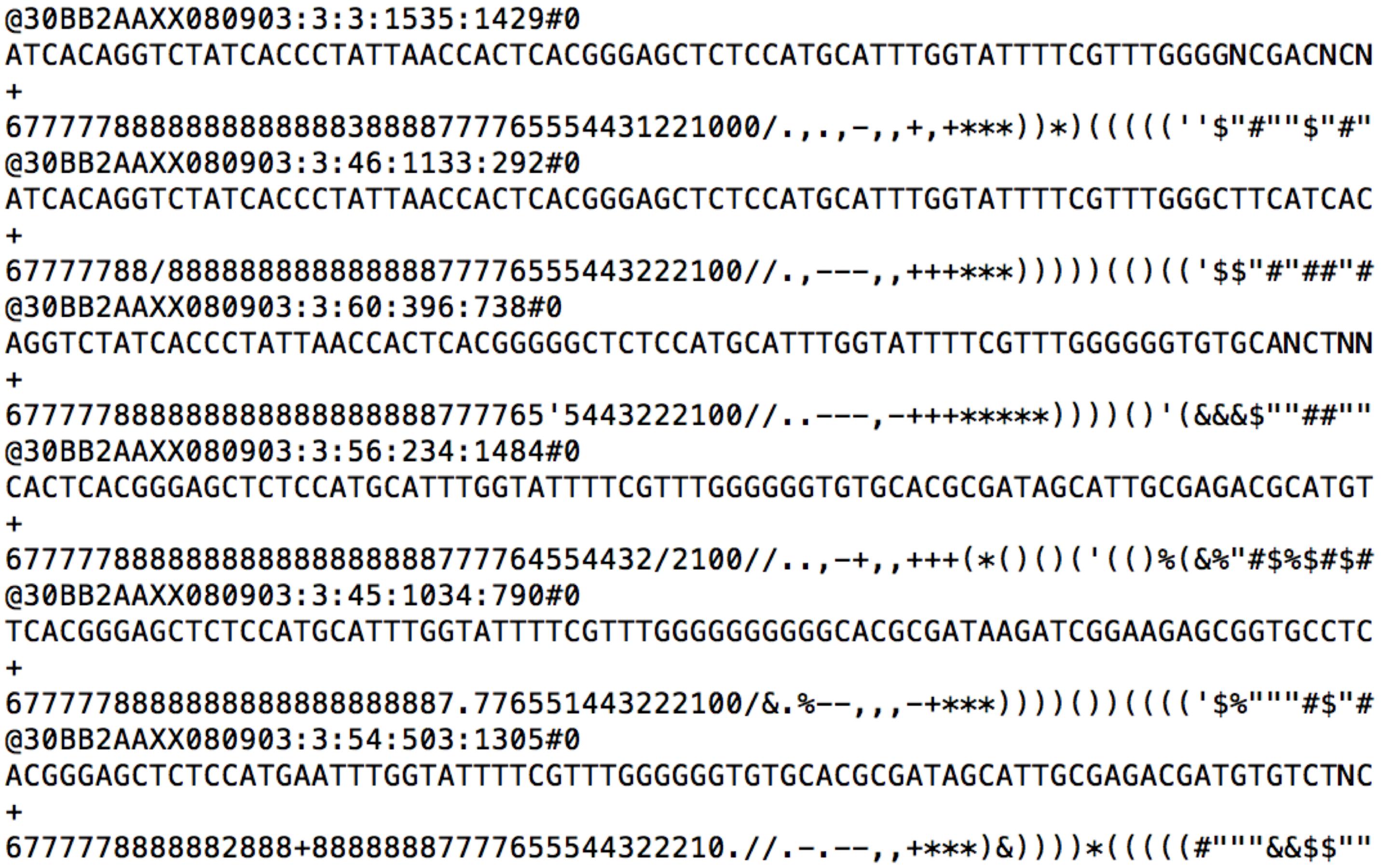
- Need for solid software engineering practices
  - Lack of standardization results in unusable data
  - Thousands of *Terabytes* of data created annually
- HTSJDK provides and API over standard formats
  - Data ingested and generated can be used with any standard genomics software

File Format	Data
BAM, CRAM	Reads / Genomes
FASTA	Reference Genome
VCF	Variants

Global Alliance
for Genomics & Health

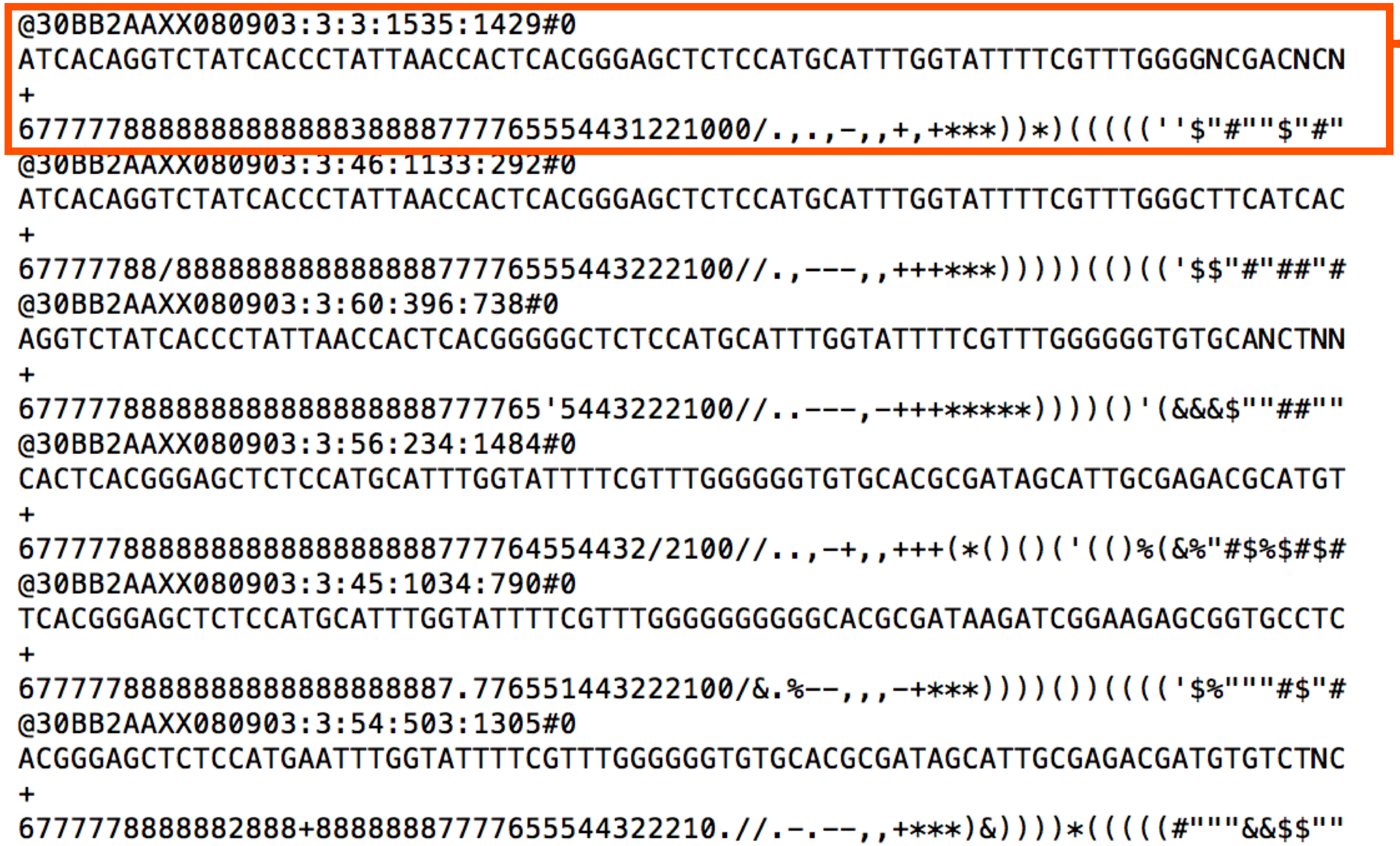
Collaborate. Innovate. Accelerate.

FASTQ Format Reads w/Quality Scores



```
@30BB2AAXX080903:3:3:1535:1429#0
ATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGNCGACNCN
+
6777778888888888888838888777765554431221000/.,.,-,,+,+***))*)(((((''$"#""$"#"
@30BB2AAXX080903:3:46:1133:292#0
ATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGCTTCATCAC
+
67777788/88888888888888877776555443222100//.,---,,+++***)))))(()(('$$"#"##"#
@30BB2AAXX080903:3:60:396:738#0
AGGTCTATCACCCTATTAACCACTCACGGGGGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGTGTGCANCTNN
+
677777888888888888888888777765'5443222100//..---,-+++*****))))()'(&&&$""##""
@30BB2AAXX080903:3:56:234:1484#0
CACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCATGT
+
677777888888888888888888777764554432/2100//..,-+,,+++(*()()('(()%(&%"#$%$#$#
@30BB2AAXX080903:3:45:1034:790#0
TCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGGGGCACGCGATAAGATCGGAAGAGCGGTGCCTC
+
6777778888888888888888887.776551443222100/&.%--,,,-+***))))())(((('$%"""#$"#
@30BB2AAXX080903:3:54:503:1305#0
ACGGGAGCTCTCCATGAATTTGGTATTTTCGTTTGGGGGGTGTGCACGCGATAGCATTGCGAGACGATGTGTCTNC
+
6777778888882888+88888887777655544322210.//.-.--,,+***)&))))*(((((#"""&&$$""
```

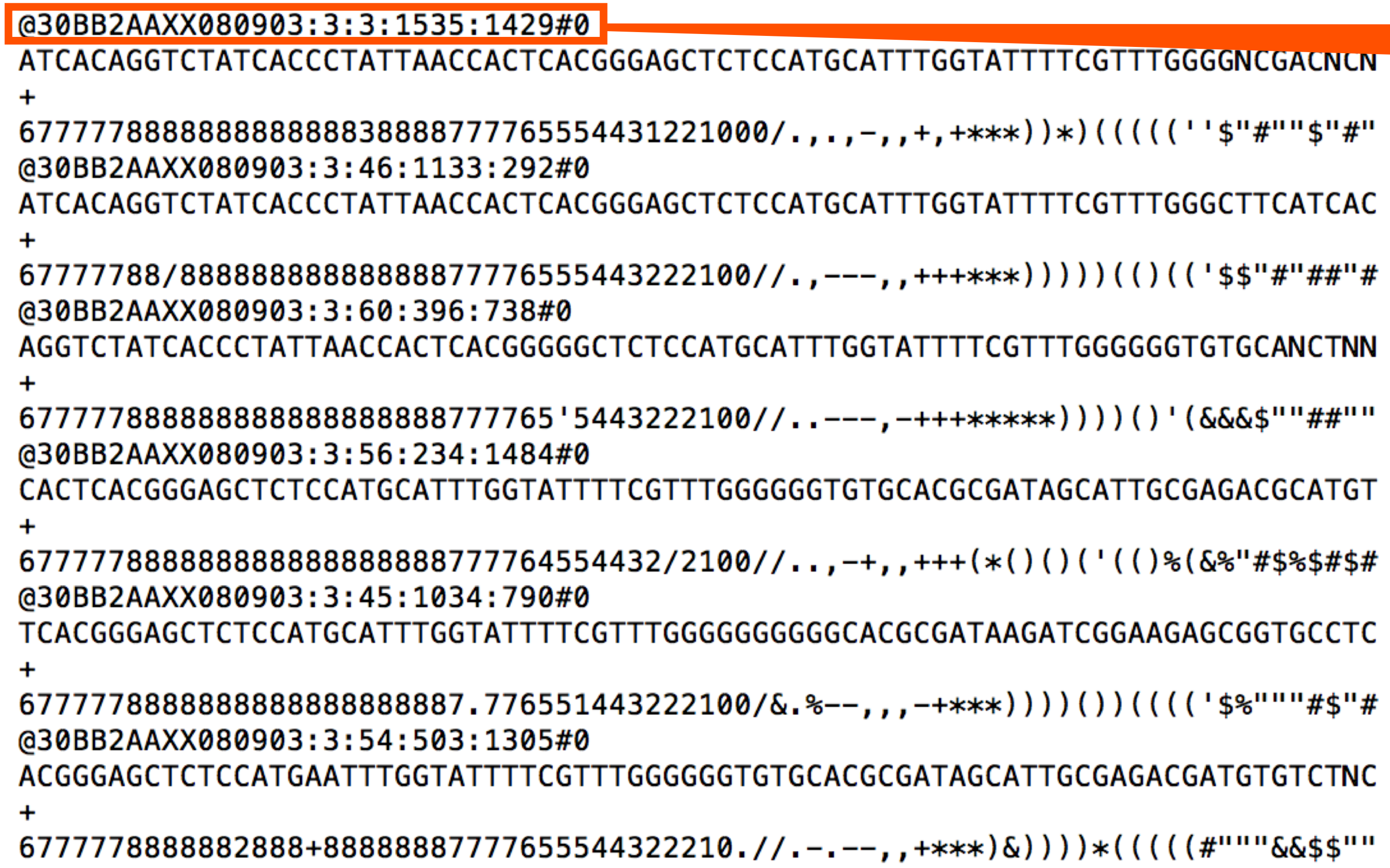
FASTQ Format Reads w/Quality Scores



```
@30BB2AAXX080903:3:3:1535:1429#0
ATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGNCGACNCN
+
6777778888888888888838888777765554431221000/.,.,-,,+,+***))*)(((((''$"#""$"#"
@30BB2AAXX080903:3:46:1133:292#0
ATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGCTTCATCAC
+
67777788/88888888888888877776555443222100//.,---,,+++***)))))(()(('$$"#"##"#
@30BB2AAXX080903:3:60:396:738#0
AGGTCTATCACCCTATTAACCACTCACGGGGGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGTGTGCANCTNN
+
677777888888888888888888777765'5443222100//..---,-+++*****))))()'(&&&$""##""
@30BB2AAXX080903:3:56:234:1484#0
CACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCATGT
+
677777888888888888888888777764554432/2100//..,-+,,+++(*()()('(()%(&%"#$%$#$#
@30BB2AAXX080903:3:45:1034:790#0
TCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGGGGGCACGCGATAAGATCGGAAGAGCGGTGCCTC
+
6777778888888888888888887.776551443222100/&.%--,,,-+***))))())(((('$%"""#$"#
@30BB2AAXX080903:3:54:503:1305#0
ACGGGAGCTCTCCATGAATTTGGTATTTTCGTTTGGGGGGTGTGCACGCGATAGCATTGCGAGACGATGTGTCTNC
+
6777778888882888+88888887777655544322210.//.-.--,,+***)&))))*(((((#"""&&$$""
```

Single Read

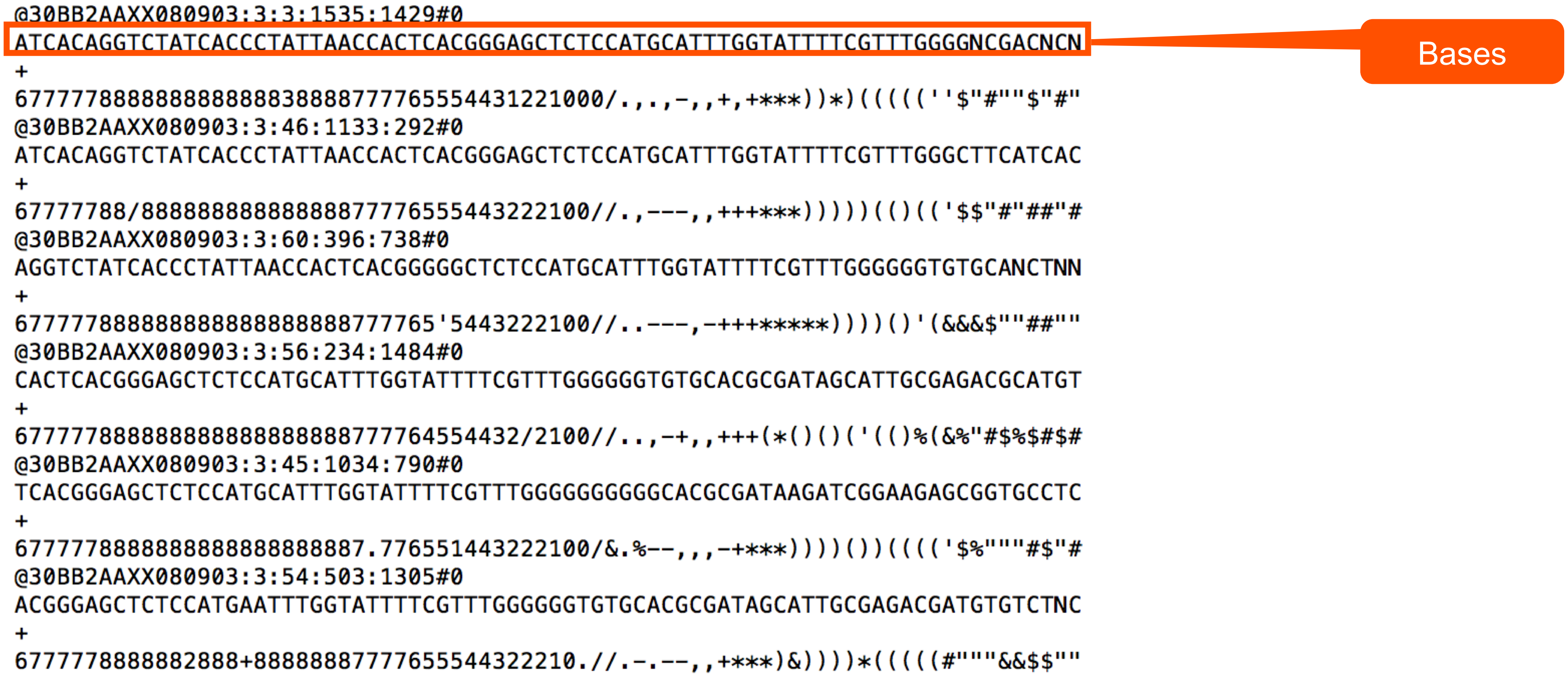
FASTQ



```
@30BB2AAXX080903:3:3:1535:1429#0
ATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGNCGACNCN
+
6777778888888888888838888777765554431221000/.,.,-,,+,+***))*)(((((''$"#""$"#"
@30BB2AAXX080903:3:46:1133:292#0
ATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGCTTCATCAC
+
67777788/888888888888888877776555443222100//.,---,,+++***)))))(()(('$$"#"##"#
@30BB2AAXX080903:3:60:396:738#0
AGGTCTATCACCCTATTAACCACTCACGGGGGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGTGTGCANCTNN
+
67777788888888888888888887777765'5443222100//..---,-+++*****))))()'(&&&$""##""
@30BB2AAXX080903:3:56:234:1484#0
CACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCATGT
+
6777778888888888888888888777764554432/2100//..,-+,,+++(*()()('(()%(&%"#$%$#$#
@30BB2AAXX080903:3:45:1034:790#0
TCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGGGGCACGCGATAAGATCGGAAGAGCGGTGCCTC
+
6777778888888888888888887.776551443222100/&.%--,,,-+***))))())(((('$%"""#$"#
@30BB2AAXX080903:3:54:503:1305#0
ACGGGAGCTCTCCATGAATTTGGTATTTTCGTTTGGGGGGTGTGCACGCGATAGCATTGCGAGACGATGTGTCTNC
+
677777888882888+888888877776555443222 10.//.-.--,,+***)&))))*(((((#"""&&$$""
```

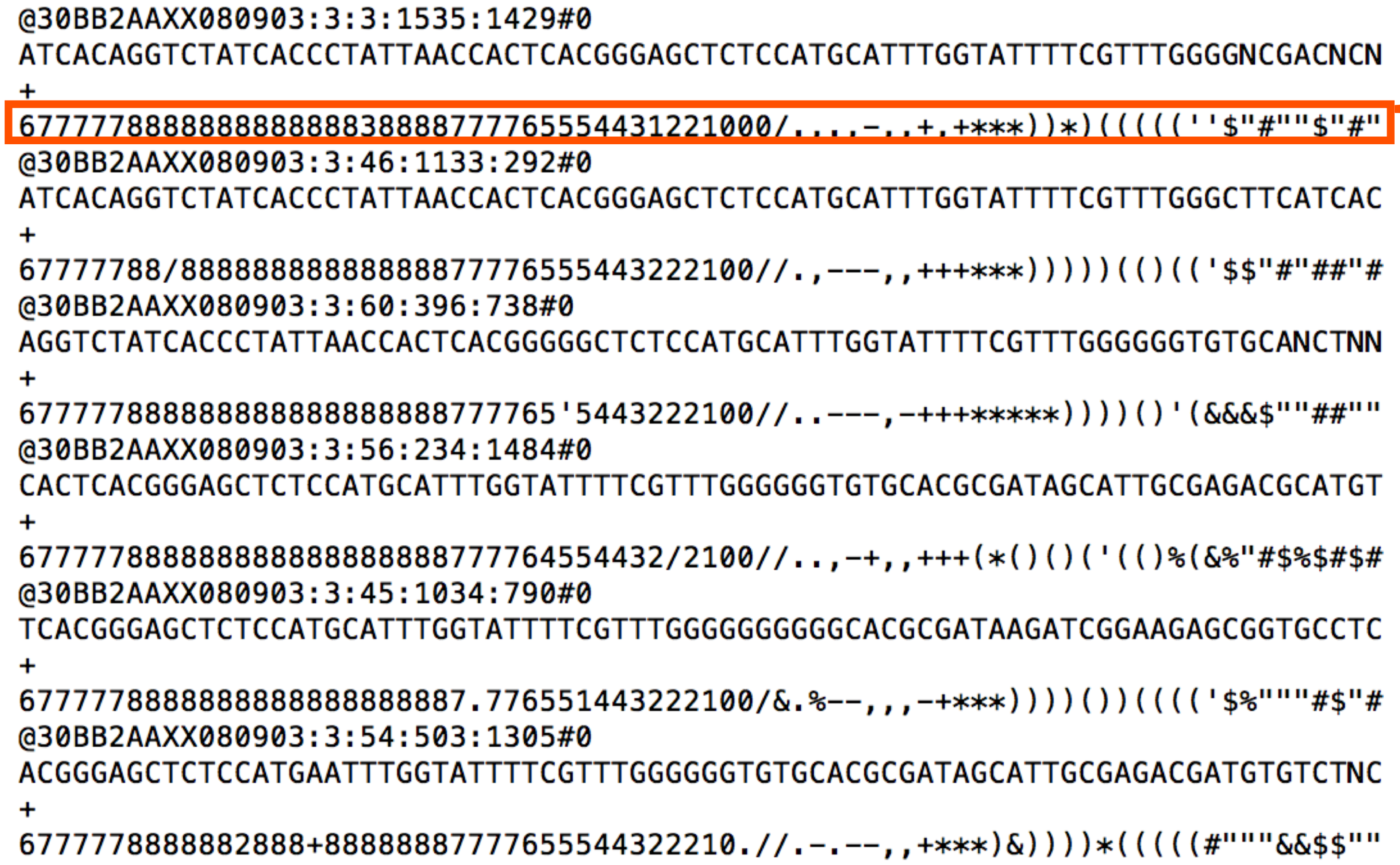
Sequence Identifier

FASTQ



```
@30BB2AAXX080903:3:3:1535:1429#0
ATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGNCGACNCN
+
6777778888888888888838888777765554431221000/.,.,-,,+,+***))*)(((((''$"#""$"#"
@30BB2AAXX080903:3:46:1133:292#0
ATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGCTTCATCAC
+
67777788/88888888888888877776555443222100//.,---,,+++***)))))(()(('$$"#"##"#
@30BB2AAXX080903:3:60:396:738#0
AGGTCTATCACCCTATTAACCACTCACGGGGGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGTGTGCANCTNN
+
677777888888888888888888777765'5443222100//..---,-+++*****))))()'(&&&$""##""
@30BB2AAXX080903:3:56:234:1484#0
CACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCATGT
+
677777888888888888888888777764554432/2100//..,-+,,+++(*()()('(()%(&%"#$%$#$#
@30BB2AAXX080903:3:45:1034:790#0
TCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGGGGGCACGCGATAAGATCGGAAGAGCGGTGCCTC
+
6777778888888888888888887.776551443222100/&.%--,,,-+***))))())(((('$%"""#$"#
@30BB2AAXX080903:3:54:503:1305#0
ACGGGAGCTCTCCATGAATTTGGTATTTTCGTTTGGGGGGTGTGCACGCGATAGCATTGCGAGACGATGTGTCTNC
+
6777778888882888+888888877776555443222210.//.-.--,,+***)&))))*(((((#"""&&$$""
```

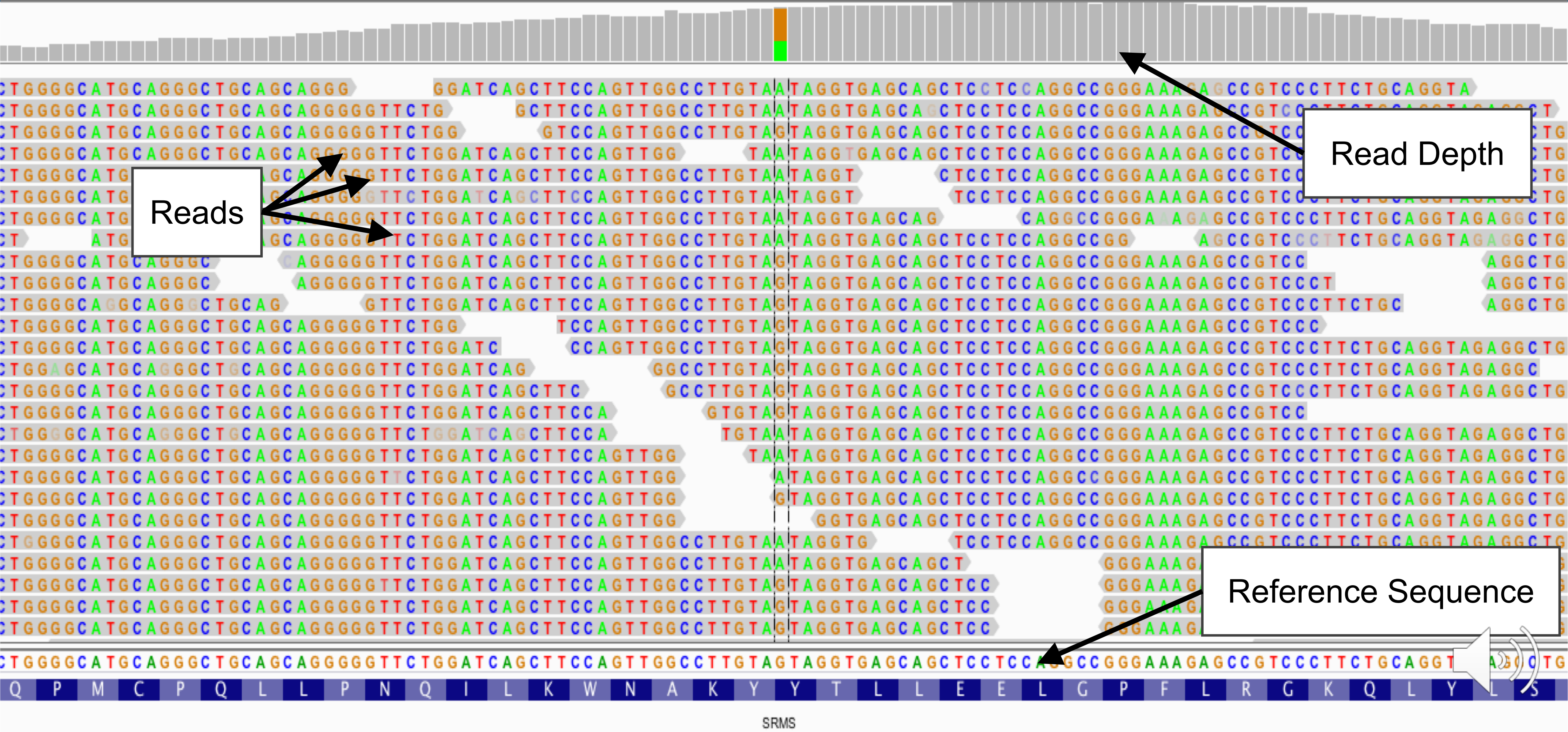
FASTQ



```
@30BB2AAXX080903:3:3:1535:1429#0
ATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGNCGACNCN
+
677777888888888888838888777765554431221000/....-..+.+***))*)(((((''$"#""$"#"
@30BB2AAXX080903:3:46:1133:292#0
ATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGCTTCATCAC
+
67777788/88888888888888877776555443222100//.,---,,+++***)))))(()(('$$"#"##"#
@30BB2AAXX080903:3:60:396:738#0
AGGTCTATCACCCTATTAACCACTCACGGGGGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGTGTGCANCTNN
+
677777888888888888888888777765'5443222100//..---,-+++*****))))()'(&&&$""##""
@30BB2AAXX080903:3:56:234:1484#0
CACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCATGT
+
67777788888888888888888877776455443 2/2100//..,-+,,+++(*()()('(()%(&%"#$%$#$#
@30BB2AAXX080903:3:45:1034:790#0
TCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGGGGGGGCACGCGATAAGATCGGAAGAGCGGTGCCTC
+
6777778888888888888888887.776551443222100/&.%--,,,-+***))))())(((('$%"""#$"#
@30BB2AAXX080903:3:54:503:1305#0
ACGGGAGCTCTCCATGAATTTGGTATTTTCGTTTGGGGGGTGTGCACGCGATAGCATTGCGAGACGATGTGTCTNC
+
6777778888882888+88888887777655544322210.//.-.--,,+***)&))))*(((((#"""&&$$""
```

Base Call Quality Scores

Huge Pile of Short Reads



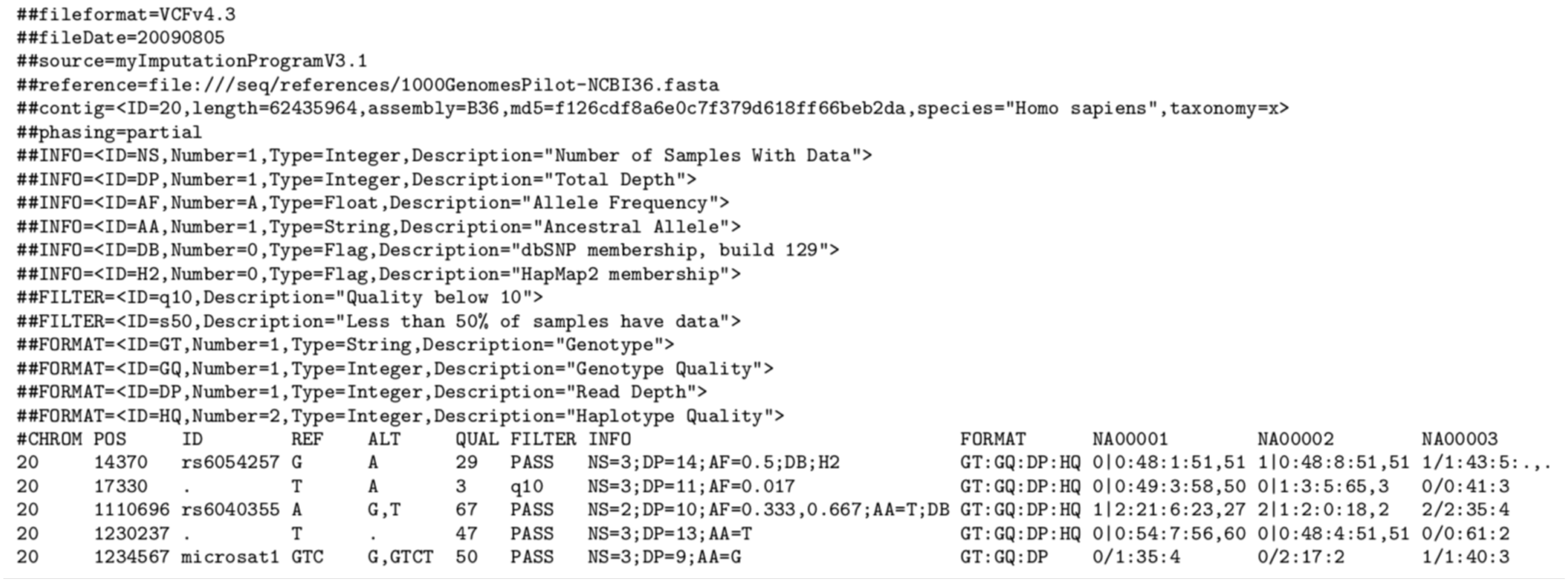
Read Depth

Reads

Reference Sequence

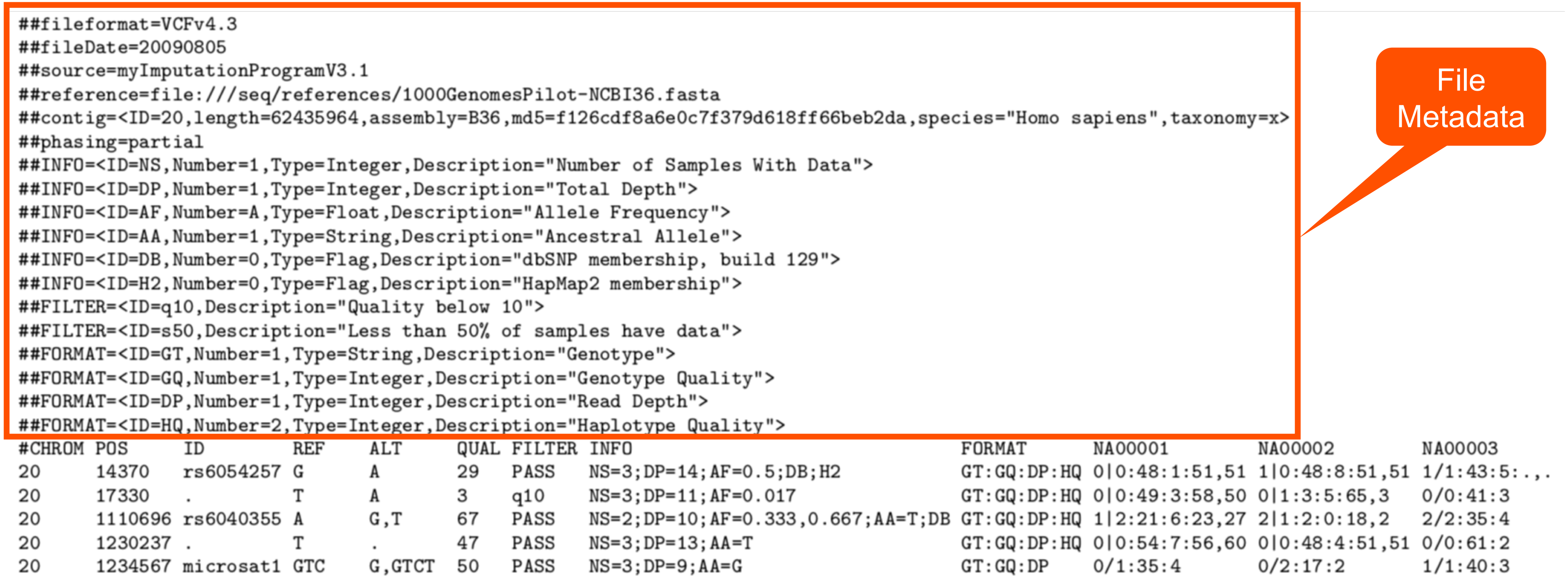
CTGGGGCATGCAGGGCTGCAGCAGGGGGTTCTGGATCAGCTTCCAGTTGGCCTTGTAGTAGGTGAGCAGCTCCTCCAGGCCGGGAAAGAGCCGTCCCTTCTGCAGGTAGGCTG
Q P M C P Q L L P N Q I L K W N A K Y Y T L L E E L G P F L R G K Q L Y L S
SRMS

VCF (Variant Call Format)



```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

VCF (Variant Call Format)

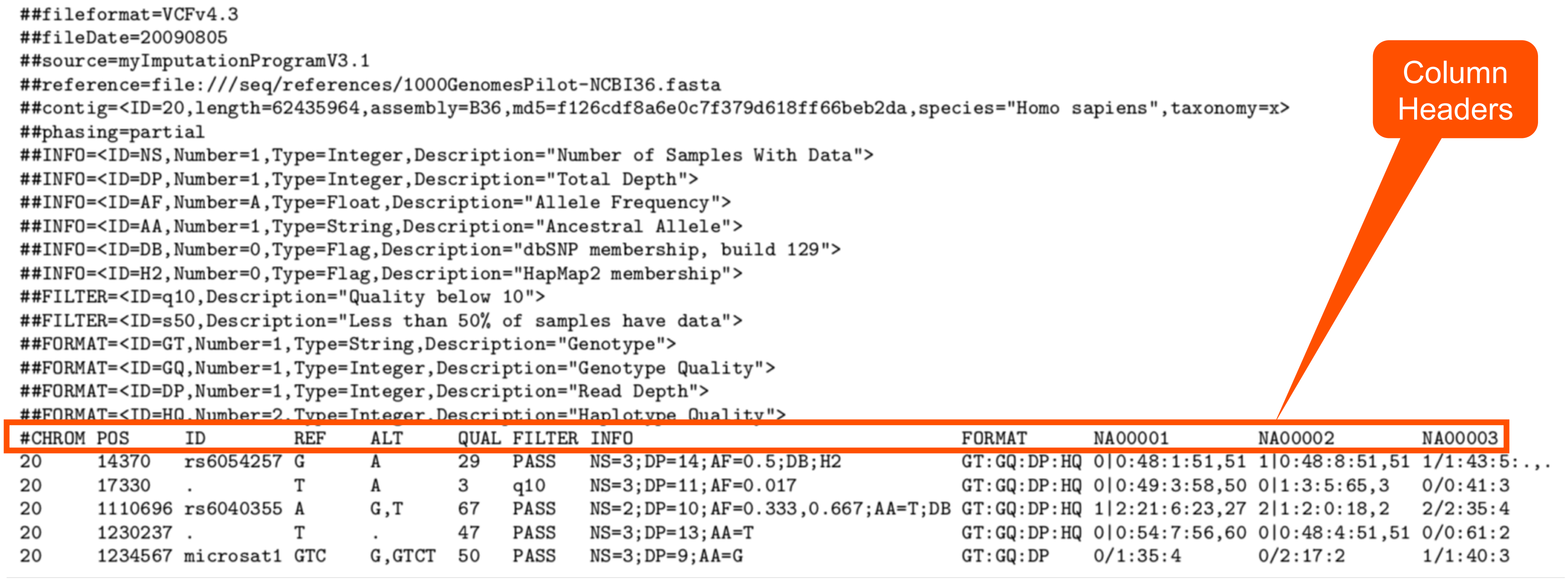


```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

File Metadata

BROAD
INSTITUTE

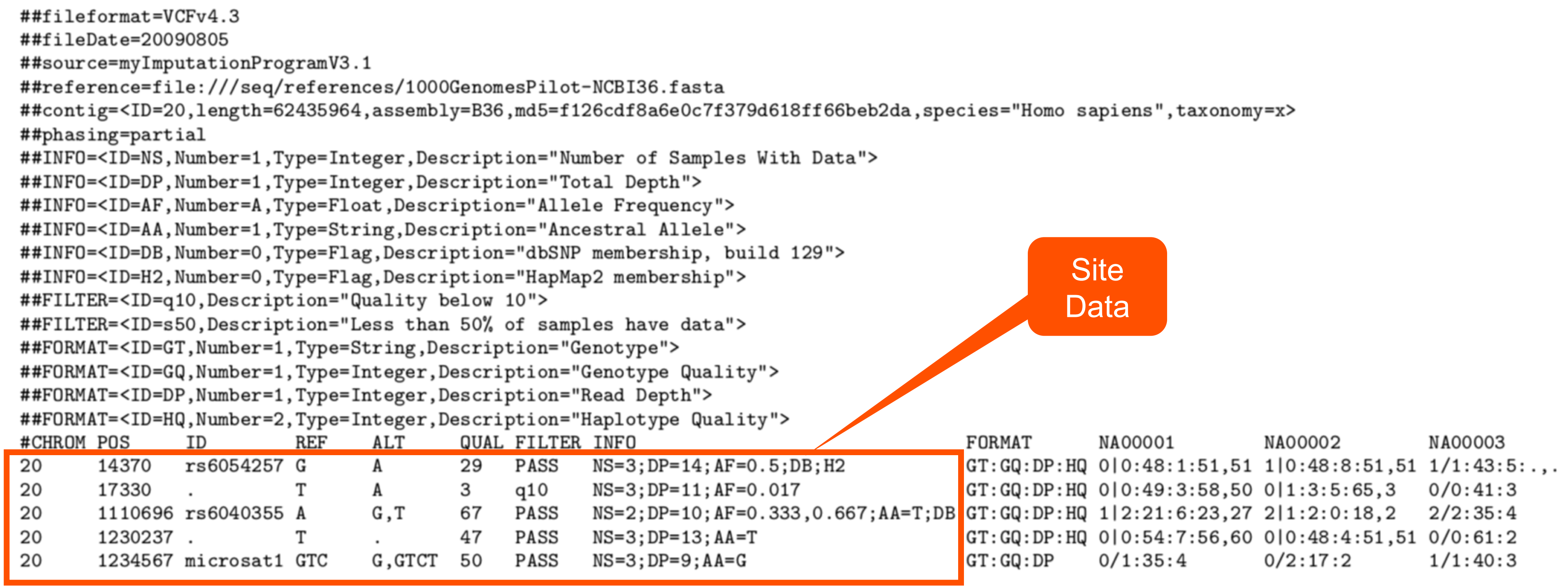
VCF (Variant Call Format)



```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Column Headers

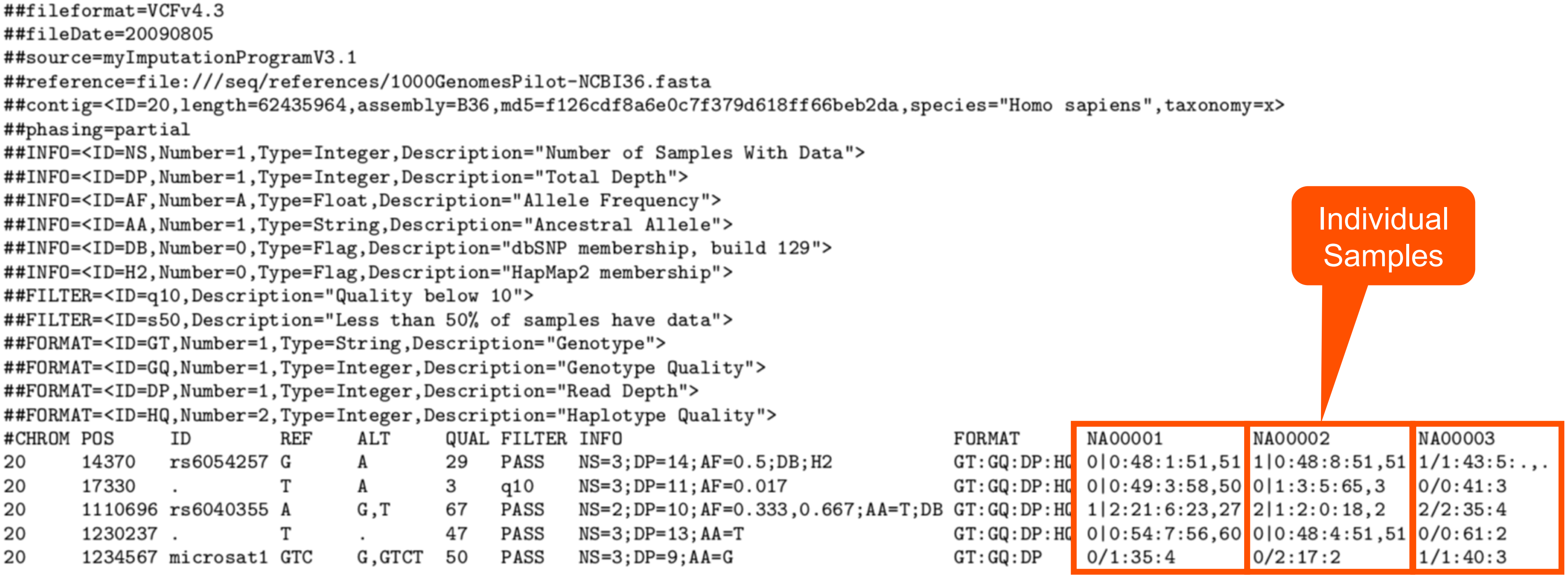
VCF (Variant Call Format)



```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Site Data

VCF (Variant Call Format)



```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT    QUAL FILTER INFO                             FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A      29   PASS   NS=3;DP=14;AF=0.5;DB;H2          GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A      3    q10    NS=3;DP=11;AF=0.017              GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T    67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .      47   PASS   NS=3;DP=13;AA=T                  GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT 50   PASS   NS=3;DP=9;AA=G                   GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Individual Samples

The technology and specs evolve…

- Long reads technology
- Long references
- Circular genomes
- Alternative reference formats
- Better cloud storage/streaming support
- Transparent support for encrypted formats such as crypt4gh
- …

HTSJDK History

- Much of the code is 10+ years old
- No dedicated development team
- Code originated in several disparate projects
- Lack of well defined versioning strategy
- Lack of well defined interfaces
- Much of the codebase predates language support for generic/functional programming

Goals for the CZI EOSS Project

- Define & publish a versioning scheme
- Implement a plugin system for file format codecs
- Enable extensibility via dynamic codec discovery
- Explicit support for side-by-side file format versions
- Code Discipline and Modernization
  - Interface-driven!
  - Refactoring
  - 80%+ test coverage
- And ultimately...increase the rate at which we can deliver new features!