# NGS Bioinformatics Overview

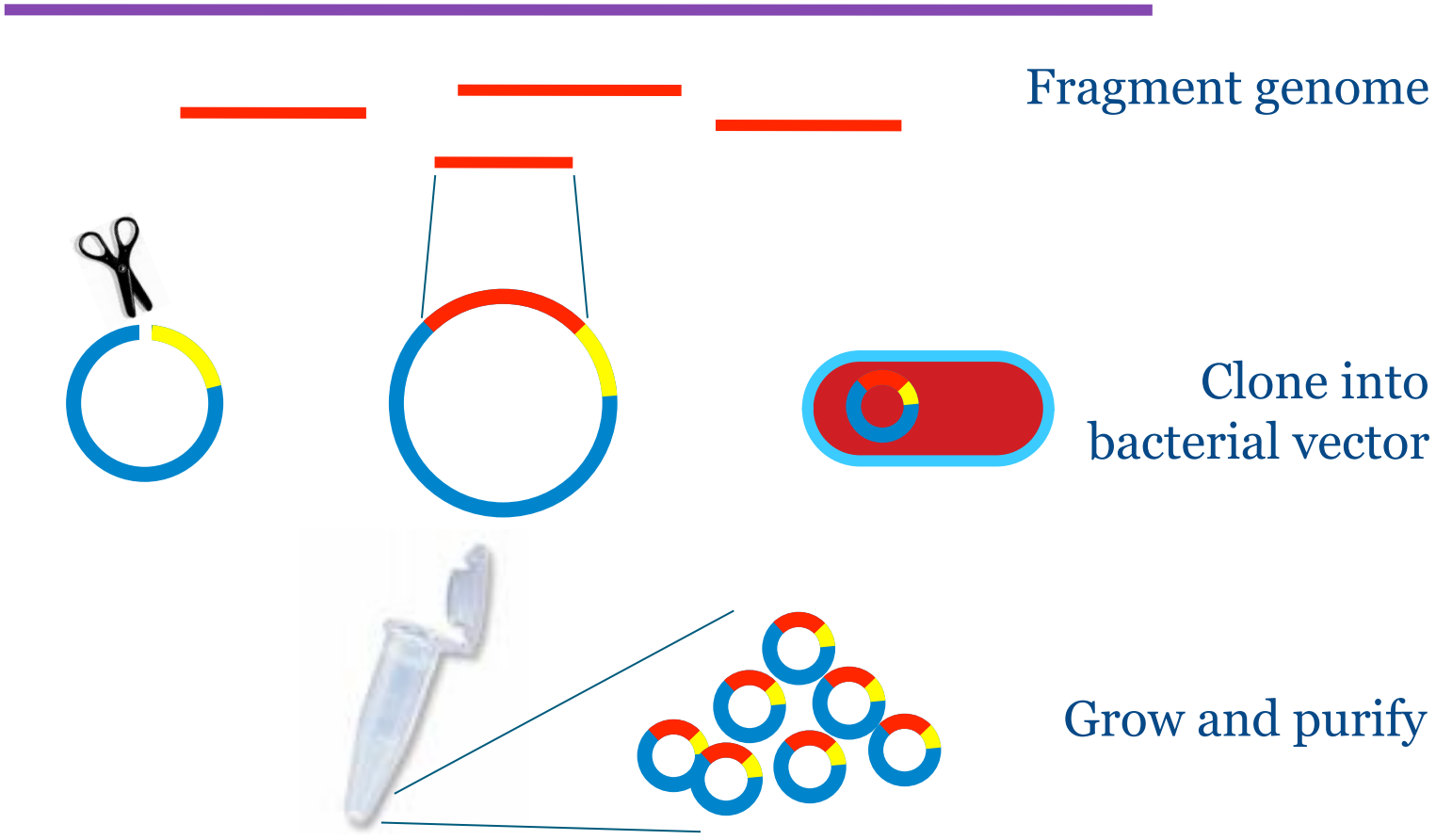
Joshua Randall
Human Genetics Informatics
Wellcome Trust Sanger Institute

Thanks to Thomas Keane and Harold Swerdlow for some slide content

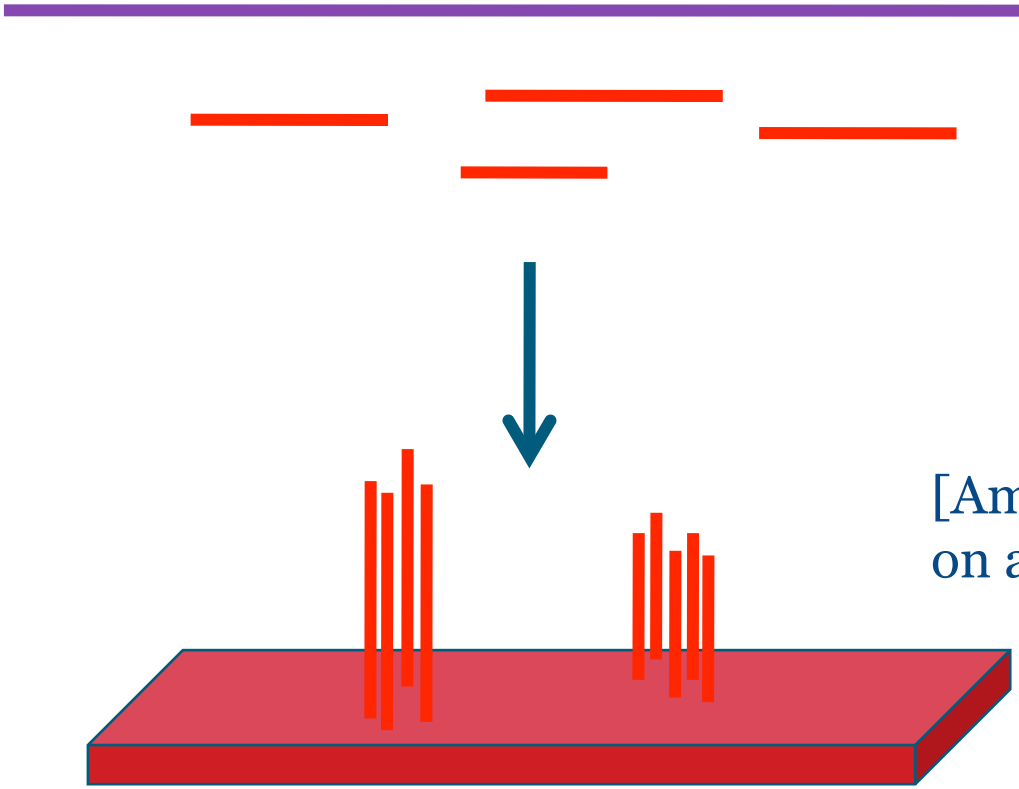
# NGS Bioinformatics Overview

- NGS Overview
- NGS Analysis Workflows
- NGS Data Formats
- Lab Exercises

# Capillary Sample Prep

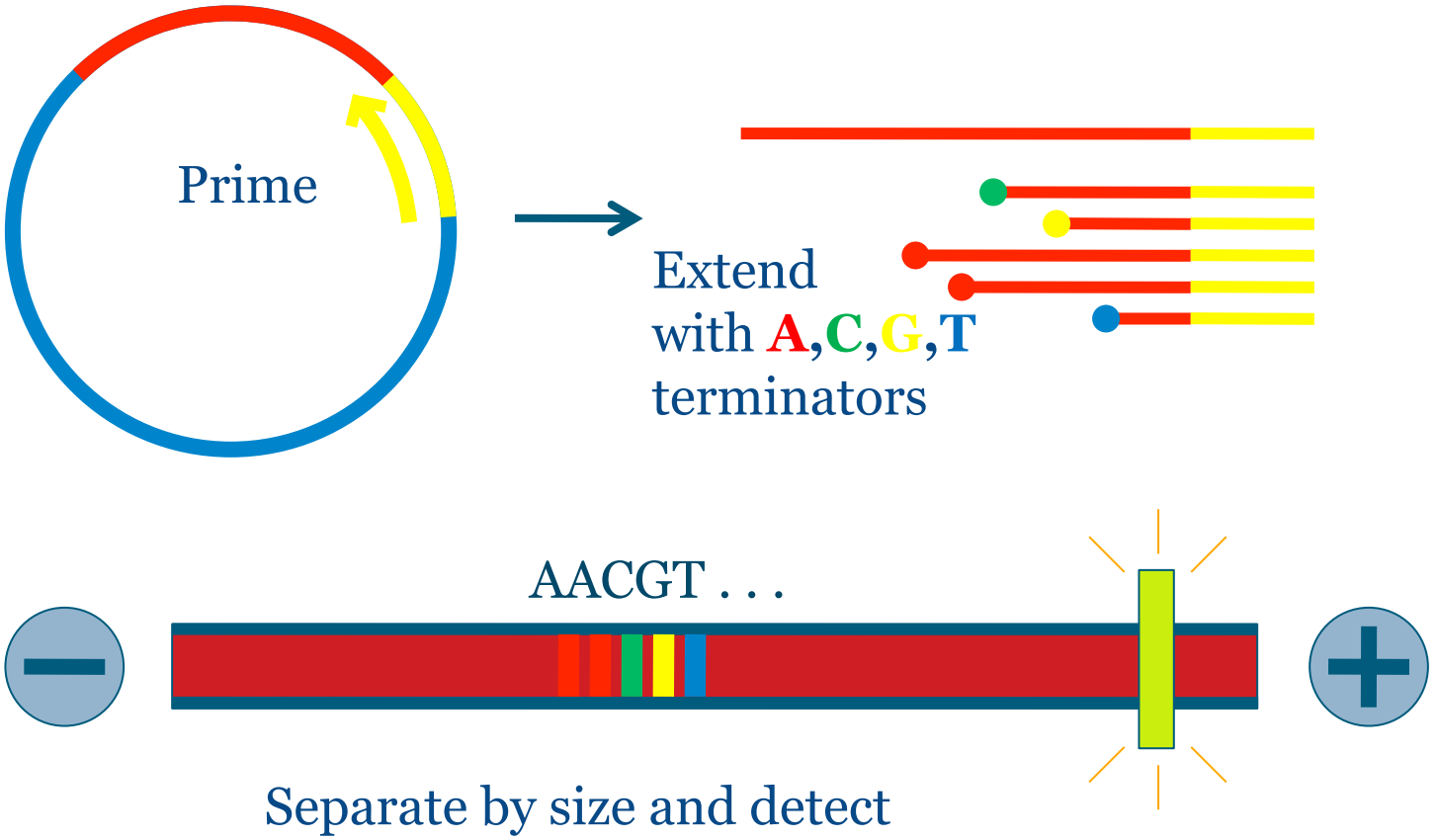


# Next-Generation Sample Prep

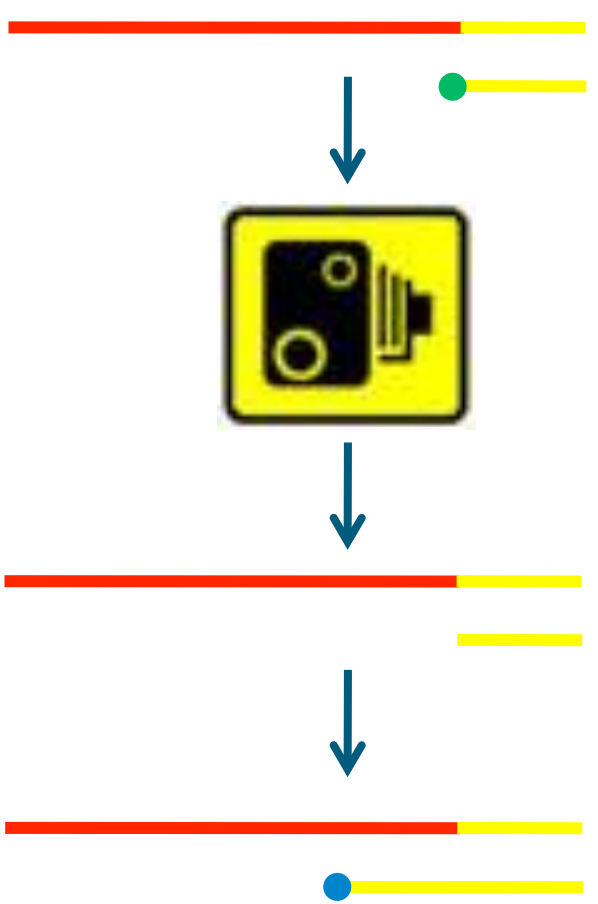


[Amplify] fragments directly on a surface (bead, chip, etc.)

# Capillary Sequencing



# Sequencing by Synthesis



Extend by 1 base

Image

Reverse termination

Repeat

# Capillary Reactions

1 tube
1 template

→

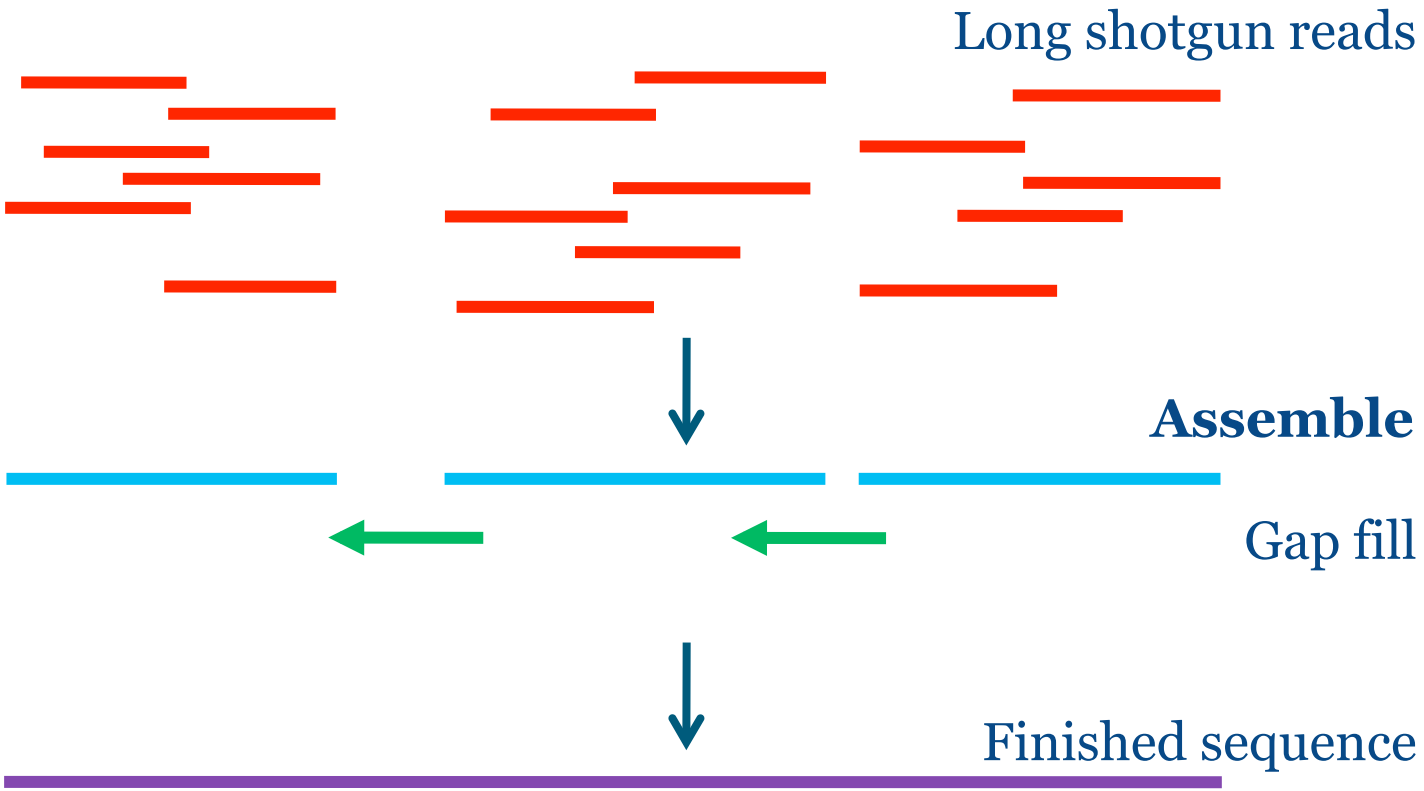
1 capillary
1000 bases

# Next-Generation Reactions

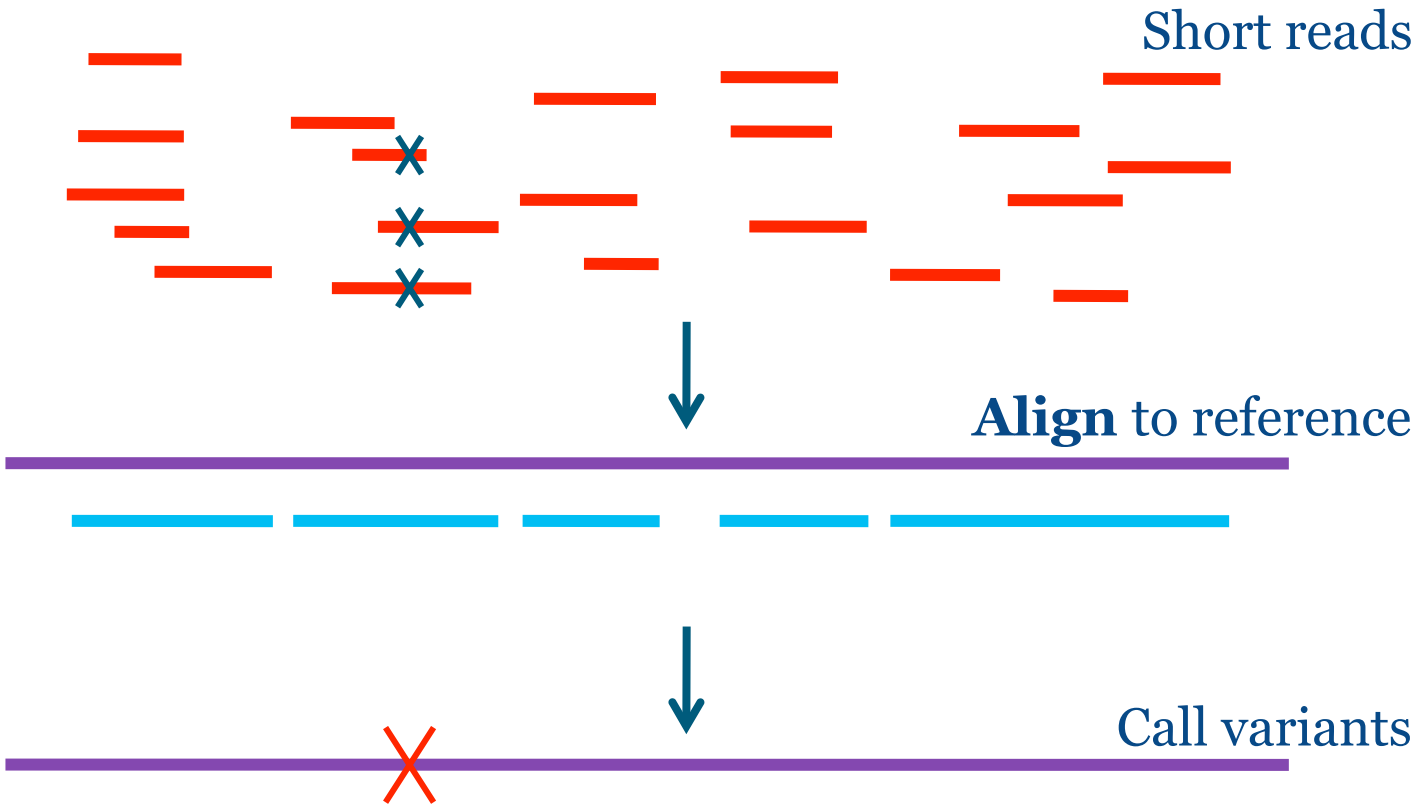
1 feature
1 template

1 chip
gigabases

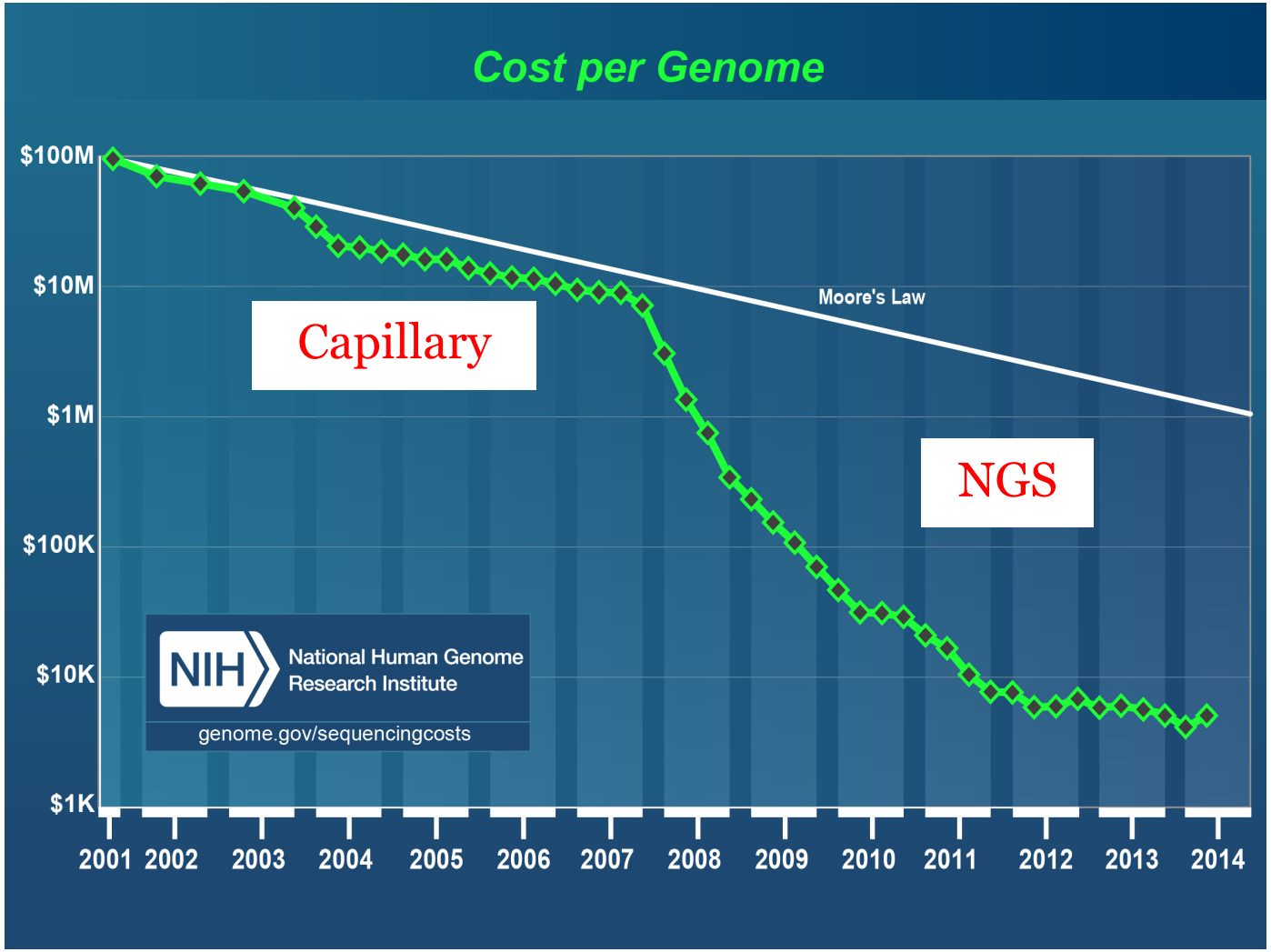
# *De-novo* Sequencing



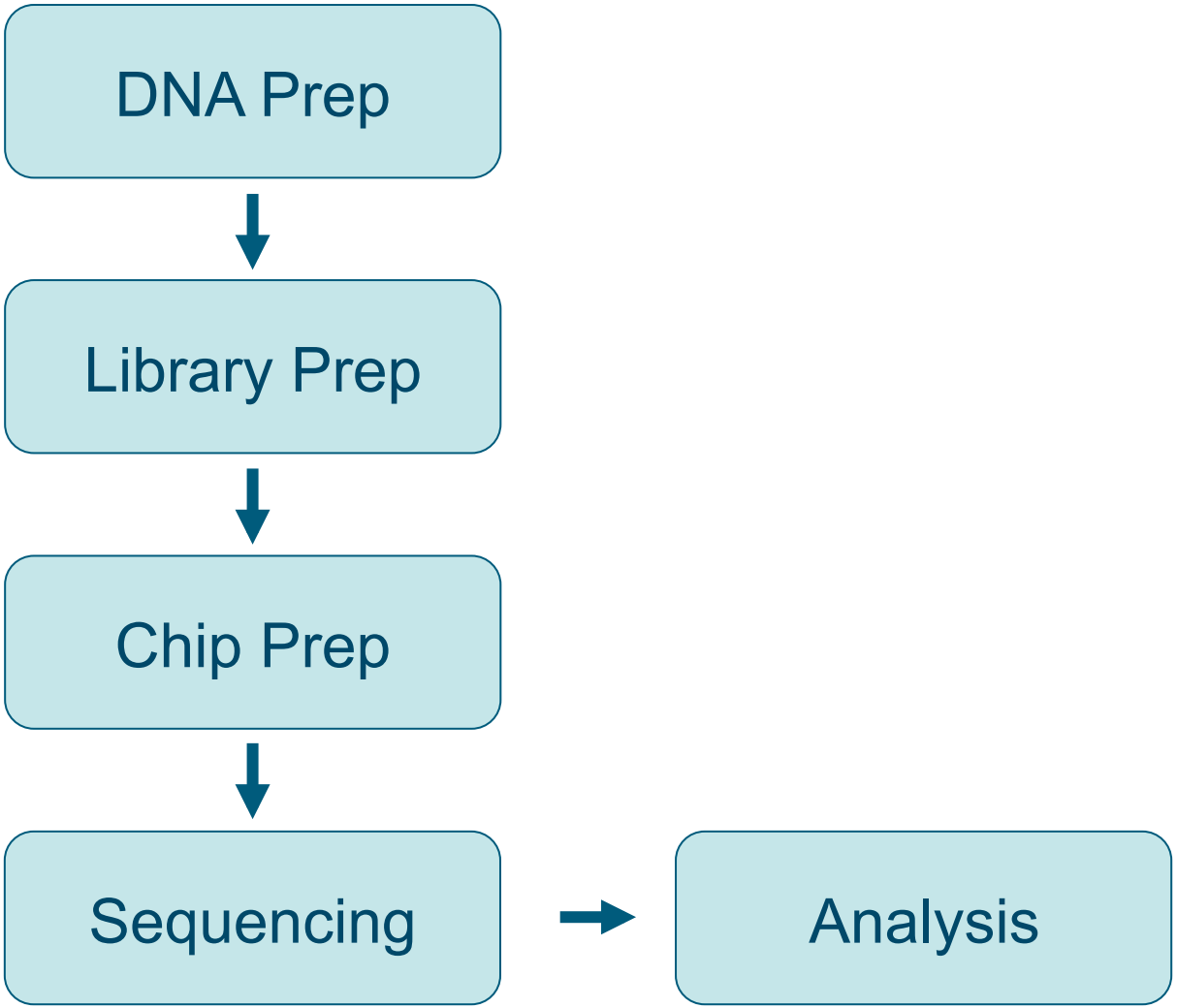
# Re-sequencing



# Global Sequencing Progress



# NGS Process



# NGS Bioinformatics Overview

- NGS Overview
- **NGS Analysis Workflows**
- NGS Data Formats
- Lab Exercises

# NGS Workflows

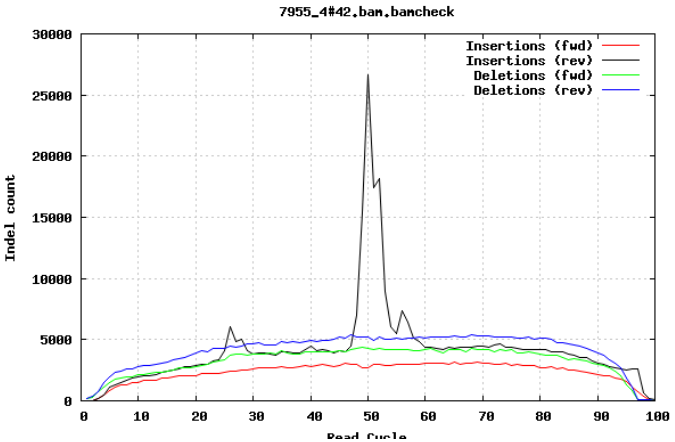
Next-gen sequencing experiments:

- Vary from one or two up to thousands of samples
- Can generate one or more sequencing libraries per sample
- Multiple libraries can be sequenced per lane (multiplex)
  - We call data from one index within a lane a ‘lanelet’
- Data can be quite large
  - can be difficult to do even ‘easy’ things like splitting and merging

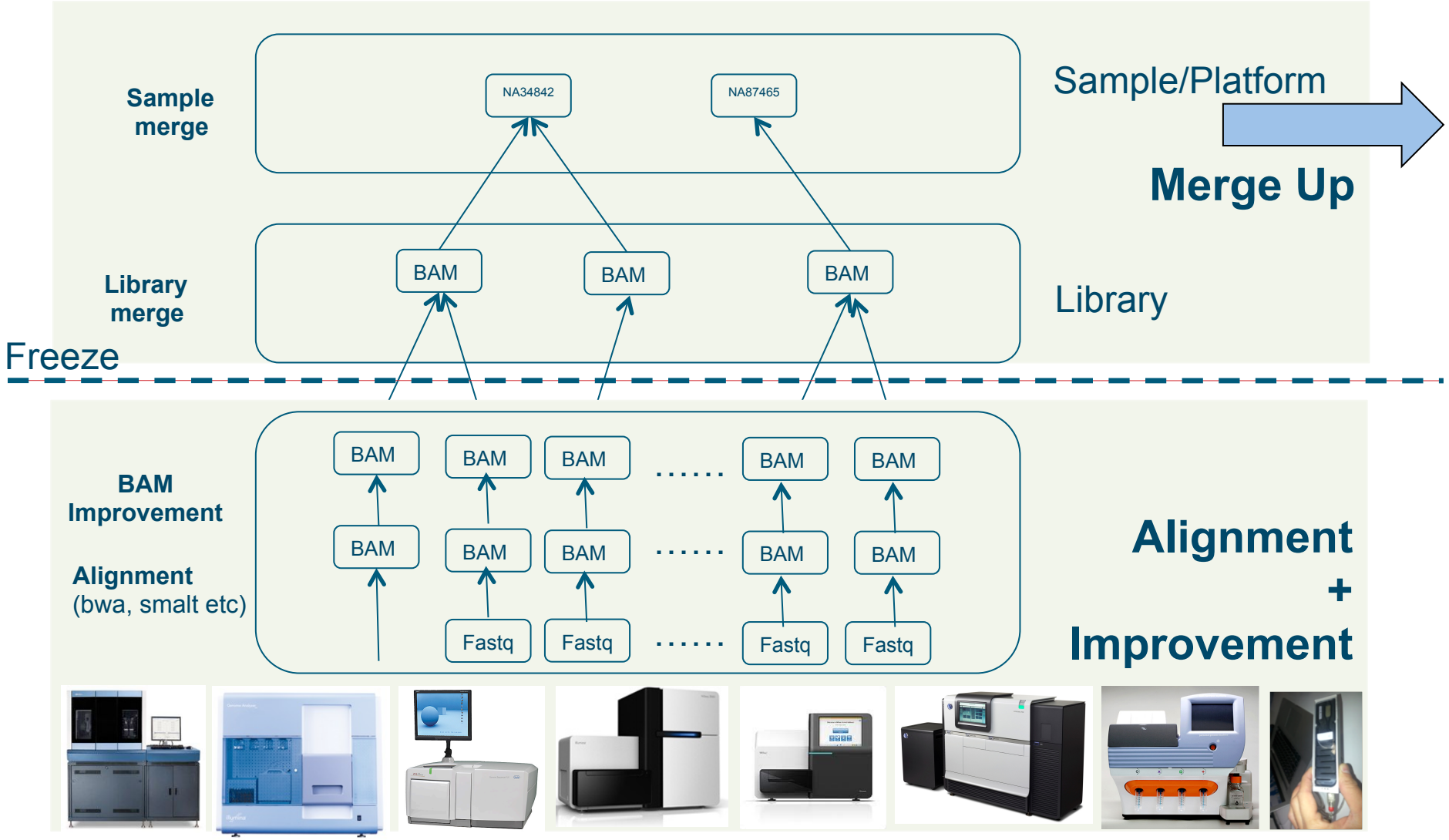
Major modes: de-novo assembly or reference-based analysis

In resequencing, alignment of reads onto reference is just the 1st step

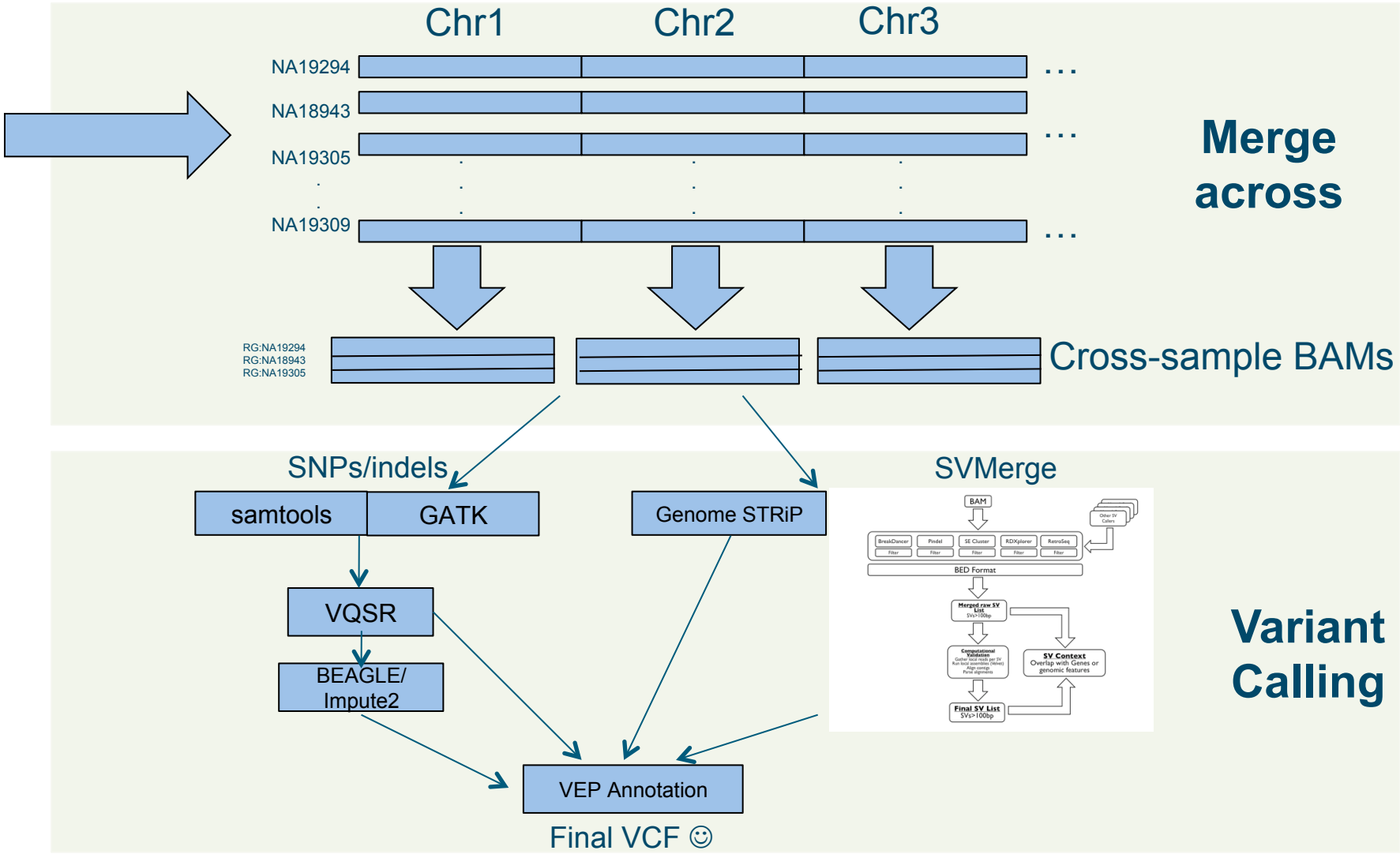
- QC of data is very important for good calls
  - Biases in the library or sequence data will produce unexpected results or missed variant calls (e.g. GC or context bias)
  - False indels due to ‘bubbles’ in sequencing run



# Resequencing Data Production Workflow



# Resequencing Data Production Workflow

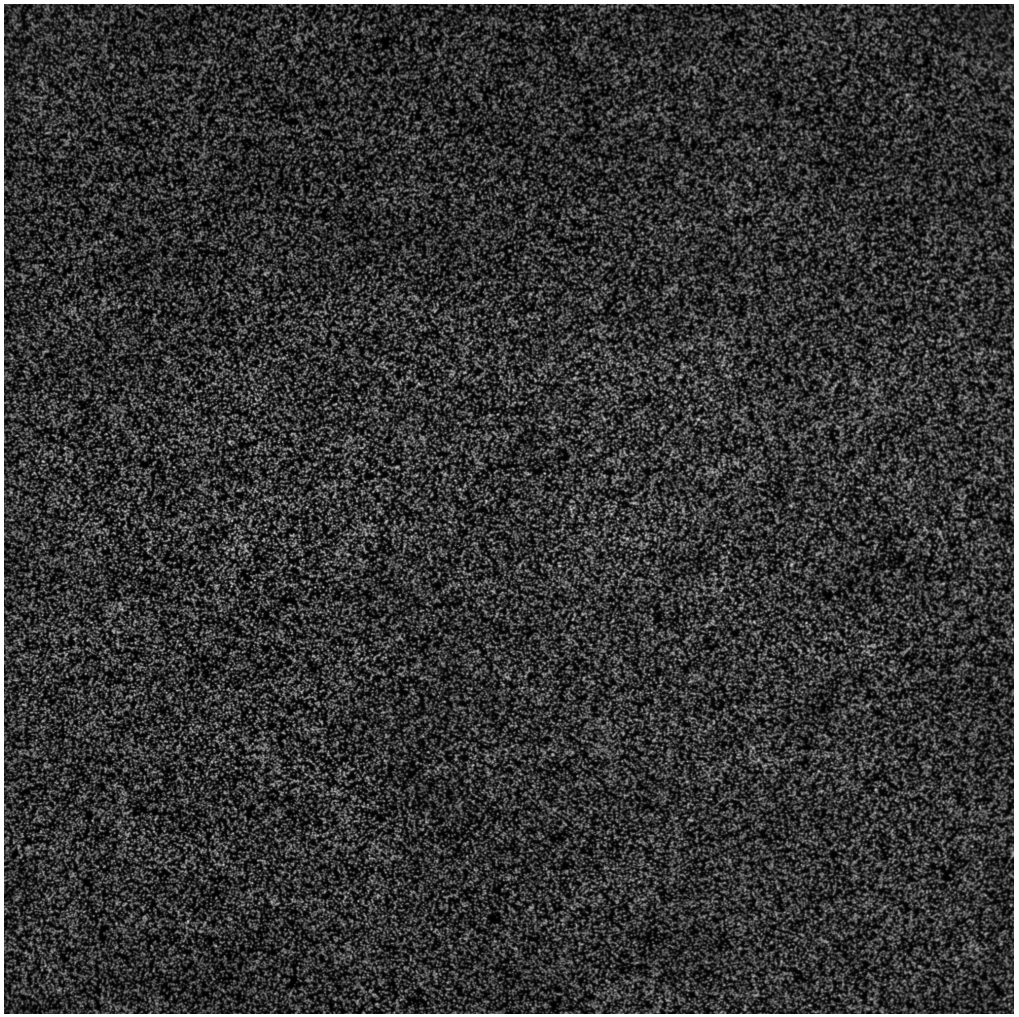


# NGS Bioinformatics Overview

- NGS Overview
- NGS Analysis Workflows
- **NGS Data Formats**
- Lab Exercises

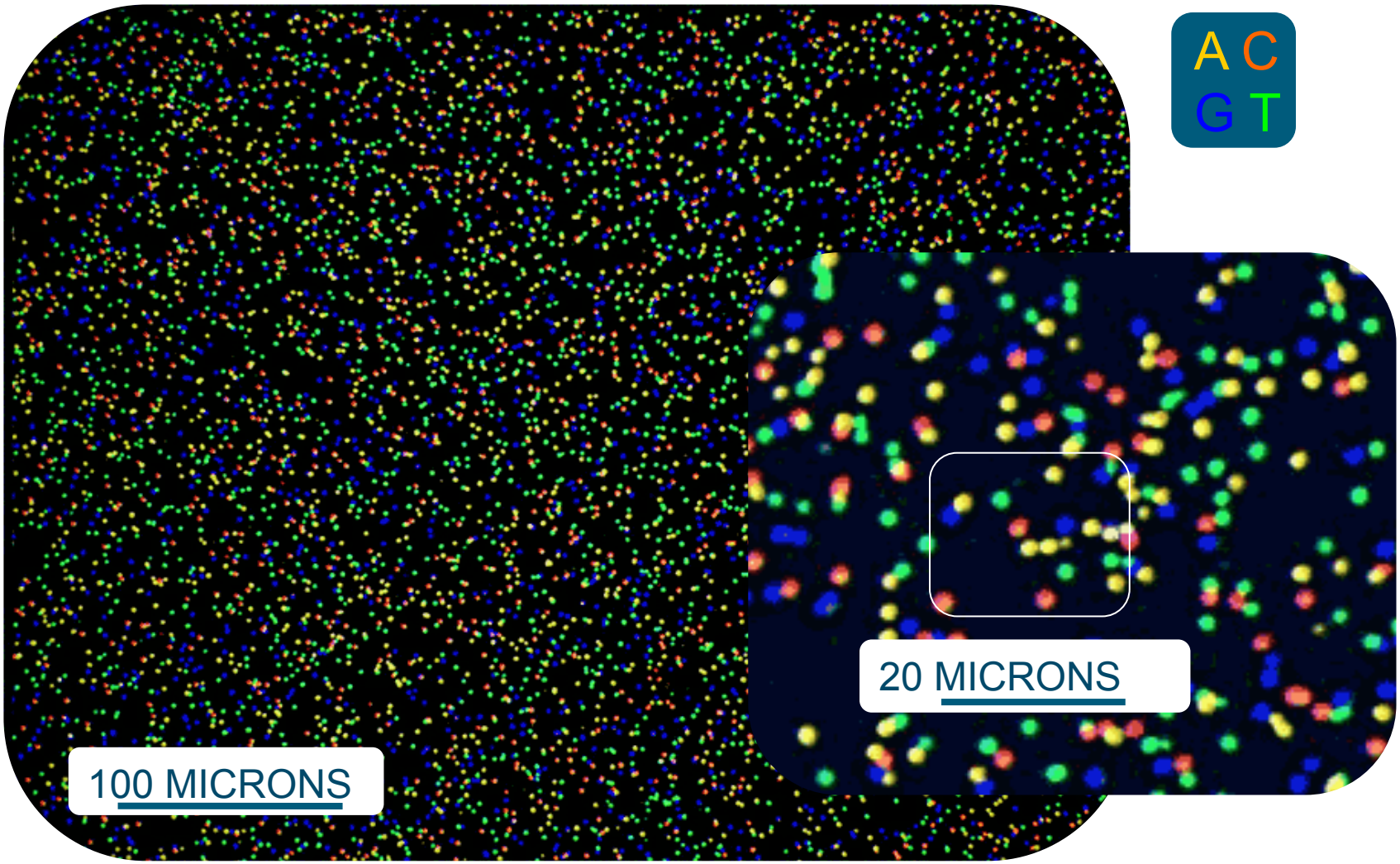
# Raw NGS Data

Small region of an Illumina image file (one per read cycle)

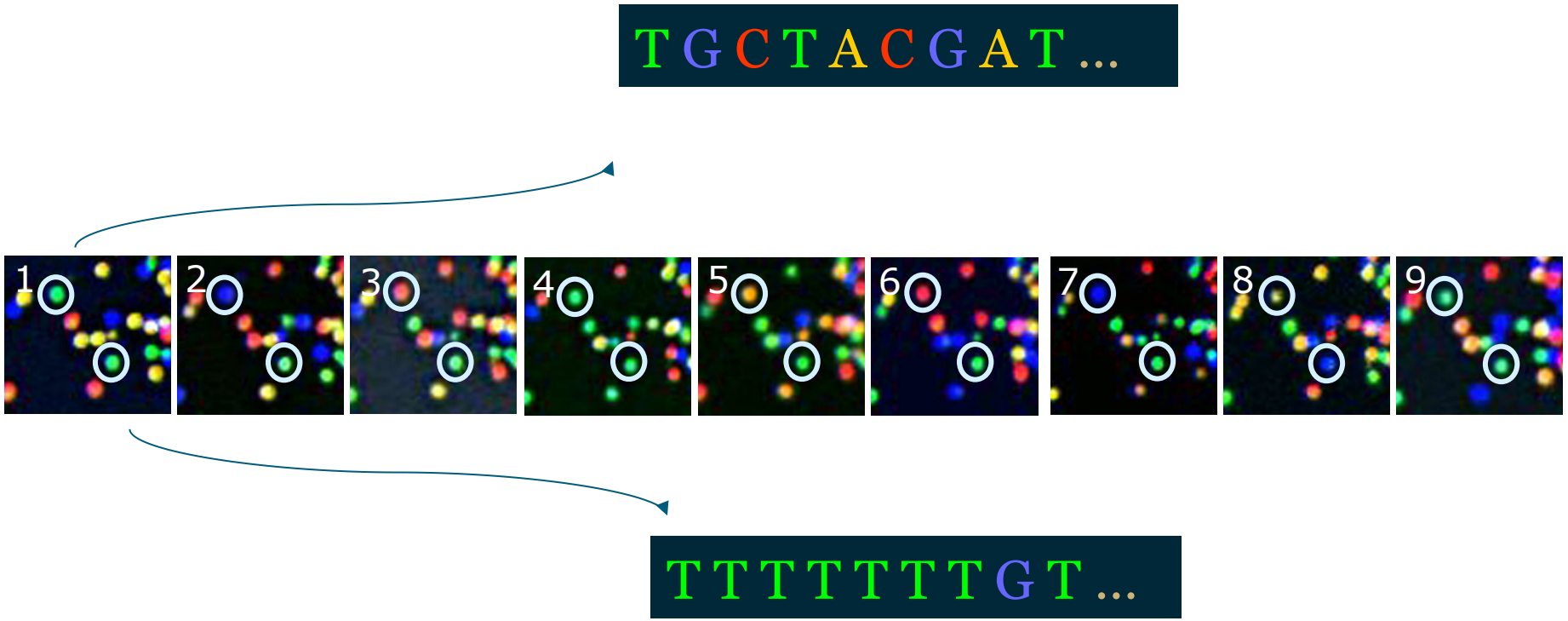


# Raw NGS Data

Illumina image data: zoomed in to reveal individual clusters

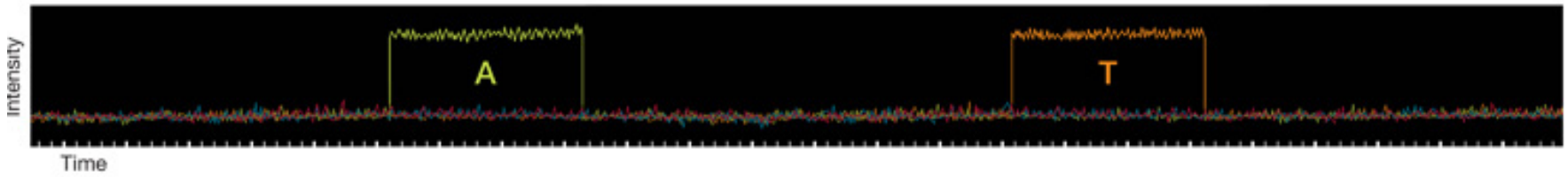


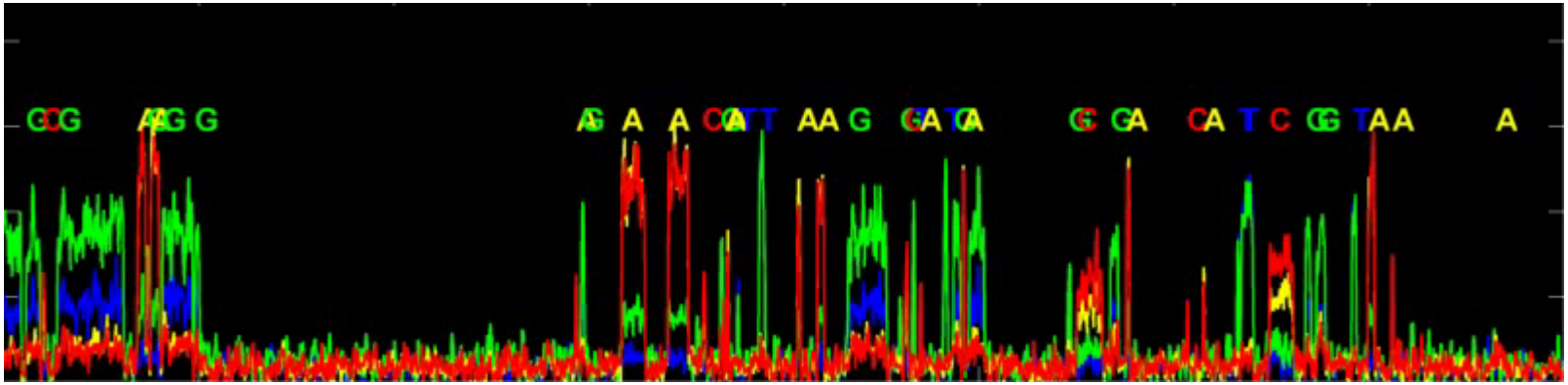
# Base calling from raw Illumina data




Base calling is now done in real-time (Illumina RTA) and raw images are typically discarded once calling is complete. Illumina RTA output includes a sequence of bases for each read, along with quality scores for each base. It can be exported to FASTQ or converted directly to unaligned BAM.

# Raw PacBio data

# Primary NGS Data Formats

## FASTA

- Unaligned sequence data (in NGS context, used only for reference data)

## FASTQ

- Unaligned read sequences with base qualities

## BAM

- Aligned or unaligned reads
- Text and binary formats

## CRAM

- Aligned or unaligned reads
- Advanced compression models

## VCF

- Flexible variant call format
- Arbitrary types of sequence variation

# FASTA

FASTA is a simple format for raw sequencing data

- In NGS workflows, typically only used for reference data
- File extension is normally .fasta or .fa

Format

- File can contain any number of sequence sections
- Each sequence starts with a single-line description beginning with the '>' character
- Following this come any number of lines consisting of sequence data which when concatenated together gives the full sequence

Cock et al. (2009) *NAR*

# FASTQ

FASTQ is a simple format for raw unaligned sequencing reads

- Simple addition to the FASTA format
- Sequence and an associated per base quality score

Originally standard for storing capillary data

Format

- Subset of the ASCII printable characters
- ASCII 33–126 inclusive with a simple offset mapping
- perl -w -e "print ( unpack( 'C', '%' ) - 33 );"

| | Range | Offset | Type | Range |
|---|---|---|---|---|
| Sanger standard | | | | |
| `fastq-sanger` | 33–126 | 33 | PHRED | 0 to 93 |
| Solexa/early Illumina | | | | |
| `fastq-solexa` | 59–126 | 64 | Solexa | −5 to 62 |
| Illumina 1.3+ | | | | |
| `fastq-illumina` | 64–126 | 64 | PHRED | 0 to 62 |

```
@HS16_08055:2:2103:5574:3860#1/1
CTGGCTGGATCCACTCGAGGTATGCAACAAGCAAACTGA
+
A@C@EEFD0GFFFGFGGFHFEFEFE7EGGIDEGEFGGG
```

Cock et al. (2009) *NAR*

# SAM/BAM Format

SAM (Sequence Alignment/Map) format

- Single unified format for storing read alignments to a reference genome

BAM (Binary Alignment/Map) format

- Binary equivalent of SAM
- Developed for fast processing/indexing

Advantages

- Can store alignments from most aligners
- Supports multiple sequencing technologies
- Supports indexing for quick retrieval/viewing
- Compact size (e.g. 112Gbp Illumina = 116Gbytes disk space)
- Reads can be grouped into logical groups e.g. lanes, libraries, samples
- Widely support by variant calling software packages

Replacement for SRF & FASTQ

# Read Entries in SAM

| No. | Name | Description |
|---|---|---|
| 1 | QNAME | Query NAME of the read or the read pair |
| 2 | FLAG | Bitwise FLAG (pairing, strand, mate strand, etc.) |
| 3 | RNAME | Reference sequence NAME |
| 4 | POS | 1-Based leftmost POSition of clipped alignment |
| 5 | MAPQ | MAPping Quality (Phred-scaled) |
| 6 | CIGAR | Extended CIGAR string (operations: MIDNSHP) |
| 7 | MRNM | Mate Reference NaMe ('=' if same as RNAME) |
| 8 | MPOS | 1-Based leftmost Mate POSition |
| 9 | ISIZE | Inferred Insert SIZE |
| 10 | SEQ | Query SEQuence on the same strand as the reference |
| 11 | QUAL | Query QUALity (ASCII-33=Phred base quality) |

```
HS18_07983:1:2203:5095:109107#36    163    ENA|AJ011856|AJ011856.1 412    60    100M    =    471    159
ATAAAATTATTAATAATAATCAATATGAAATTAATAAAAATCTTATAAAAAAGTAATGAATACTCCTTTTTAAAAATAAAAAGGGGTTCGGTCCCCCCCC
9BCDGDEHGEHFHHGFHHJGHFHIGHFIGHFHGGGHGHGHGHJGHHGHHHGGHHHIGGGGGFGGDGGHFHFIGEGHGFGGHFEDGG4GHGGGFHGFHIE
  X0:i:1  X1:i:0  MD:Z:100       RG:Z:1#36.1
```

Heng Li et al (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics,* 25:2078-2079

# Cigar Format

Cigar has been traditionally used as a compact way to represent a sequence alignment

Operations include

- M - match or mismatch
- I - insertion
- D - deletion

SAM extends these to include

- S - soft clip (ignore these bases)
- H - hard clip (ignore and remove these bases)
- E.g. Read: `ACGCA-TGCAGTtagacgt`

  Ref: `ACTCAGTG--GT`

  Cigar: 5M1D2M2I2M7S

# What is the cigar line?

E.g. Read: `tgtcgtcACGCATG---CAGTtagacgt`

Ref: `ACGCATGCGGCAGT`

Cigar:

# Read Group Tag

Each lane has a unique RG tag that contains meta-data for the lane

RG tags

- ID: SRR/ERR number
- PL: Sequencing platform
- PU: Run name
- LB: Library name
- PI: Insert fragment size
- SM: Individual
- CN: Sequencing center

# 1000 Genomes BAM File

```
@HD     VN:1.0  GO:none SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3    LN:198022430    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:fdfd811849cc2fadebc929bb925902e5
@SQ     SN:4    LN:191154276    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:23dccd106897542ad87d2765d28a19a1
@SQ     SN:5    LN:180915260    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:0740173db9ffd264d728f32784845cd7
@SQ     SN:6    LN:171115067    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1d3a93a248d92a729ee764823acbbc6b
@SQ     SN:7    LN:159138663    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:618366e953d6aaad97dbe4777c29375e
@SQ     SN:8    LN:146364022    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:96f514a9929e410c6651697bded59aec
@SQ     SN:9    LN:141213431    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:3e273117f15e0a400f01055d9f393768
@SQ     SN:10   LN:135534747    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:988c28e000e84c26d552359af1ea2e1d
@SQ     SN:11   LN:135006516    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:98c59049a2df285c76ffb1c6db8f8b96
@SQ     SN:12   LN:133851895    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:51851ac0e1a115847ad36449b0015864
@SQ     SN:13   LN:115169878    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:283f8d7892baa81b510a015719ca7b0b
@SQ     SN:14   LN:107349540    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:98f3cae32b2a2e9524bc19813927542e
@SQ     SN:15   LN:102531392    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:e5645a794a8238215b2cd77acb95a078
@SQ     SN:16   LN:90354753     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:fc9b1a7b42b97a864f56b348b06095e6
@SQ     SN:17   LN:81195210     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:351f64d4f4f9ddd45b35336ad97aa6de
@SQ     SN:18   LN:78077248     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:b15d4b2d29dde9d3e4f93d1d0f2cbc9c
@SQ     SN:19   LN:59128983     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1aacd71f30db8e561810913e0b72636d
@SQ     SN:20   LN:63025520     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:0dec9660ec1efaaf33281c0d5ea2560f
@SQ     SN:21   LN:48129895     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:2979a6085bfe28e3ad6f552f361ed74d
@SQ     SN:22   LN:51304566     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:a718acaa6135fdca8357d5bfe94211dd
@SQ     SN:X    LN:155270560    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:7e0e2e580297b7764e31dbc80c2540dd
@SQ     SN:Y    LN:59373566     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1fa3474750af0948bdf97d5a0ee52e51
@SQ     SN:MT   LN:16569        AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:c68f52674c9fb33aef52dcf399755519
@SQ     SN:GL000207.1   LN:4262 AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:f3814841f1939d3ca19072d9e89f3fd7
@RG     ID:ERR000047    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR000048    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR000071    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR000091    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR000094    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR000105    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR000377    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR001126    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR001127    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR001128    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR001180    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR001181    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR005185    PL:ILLUMINA     LB:HUMsgR2ABDDCAPE      PI:148  DS:SRP000031    SM:NA18582      CN:BGI
@RG     ID:ERR008994    PL:ILLUMINA     LB:HUMsgR22AFDFAAPE     PI:333  DS:SRP000546    SM:NA18582      CN:BGI
@RG     ID:ERR009030    PL:ILLUMINA     LB:HUMsgR22AFDFAAPE     PI:333  DS:SRP000546    SM:NA18582      CN:BGI
@PG     ID:bwa  VN:0.5.5
ERR001127.3207020       163     5       9998    0       45M     =       10089   136     AACTAACCCTAACCCTAACCCTAACCATAACCCTAACCCTAACAC    /3/@<></>+<?=??A=?3@@A>??@9A>?11A9=@%%@A=?:$8
XT:A:R  XN:i:3  SM:i:0  AM:i:0  X0:i:2  X1:i:4    XM:i:2  XO:i:0  XG:i:0  RG:Z:ERR001127  NM:i:5  MD:Z:0N0N0N23C16C1      OQ:Z:=>7IIII1I0IIIIIIII0IIIIEII1IIH))?0=I%%II=?-$4
```

```
samtools view –h mybam.bam | less -S
```

# 1000 Genomes BAM File

```
@HD     VN:1.0  GO:none SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3    LN:198022430    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:fdfd811849cc2fadebc929bb925902e5
@SQ     SN:4    LN:191154276    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:23dccd106897542ad87d2765d28a19a1
@SQ     SN:5    LN:180915260    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:0740173db9ffd264d728f32784845cd7
@SQ     SN:6    LN:171115067    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1d3a93a248d92a729ee764823acbbc6b
@SQ     SN:7    LN:159138663    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:618366e953d6aaad97dbe4777c29375e
@SQ     SN:8    LN:146364022    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:96f514a9929e410c6651697bded59aec
@SQ     SN:9    LN:141213431    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:3e273117f15e0a400f01055d9f393768
@SQ     SN:10   LN:135534747    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:988c28e000e84c26d552359af1ea2e1d
@SQ     SN:11   LN:135006516    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:98c59049a2df285c76ffb1c6db8f8b96
@SQ     SN:12   LN:133851895    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:51851ac0e1a115847ad36449b0015864
@SQ     SN:13   LN:115169878    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:283f8d7892baa81b510a015719ca7b0b
@SQ     SN:14   LN:107349540    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:98f3cae32b2a2e9524bc19813927542e
@SQ     SN:15   LN:102531392    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:e5645a794a8238215b2cd77acb95a078
@SQ     SN:16   LN:90354753     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:fc9b1a7b42b97a864f56b348b06095e6
@SQ     SN:17   LN:81195210     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:351f64d4f4f9ddd45b35336ad97aa6de
@SQ     SN:18   LN:78077248     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:b15d4b2d29dde9d3e4f93d1d0f2cbc9c
@SQ     SN:19   LN:59128983     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1aacd71f30db8e561810913e0b72636d
@SQ     SN:20   LN:63025520     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:0dec9660ec1efaaf33281c0d5ea2560f
@SQ     SN:21   LN:48129895     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:2979a6085bfe28e3ad6f552f361ed74d
@SQ     SN:22   LN:51304566     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:a718acaa6135fdca8357d5bfe94211dd
@SQ     SN:X    LN:155270560    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:7e0e2e580297b7764e31dbc80c2540dd
@SQ     SN:Y    LN:59373566     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1fa3474750af0948bdf97d5a0ee52e51
@SQ     SN:MT   LN:16569        AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:c68f52674c9fb33aef52dcf399755519
@SQ     SN:GL000207.1   LN:4262 AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:f3814841f1939d3ca19072d9e89f3fd7
@RG     ID:ERR001268    PL:ILLUMINA     LB:NA12878.1    PI:200  DS:SRP000032    SM:NA12878      CN:MPIMG
@RG     ID:ERR001269    PL:ILLUMINA     LB:NA12878.1    PI:200  DS:SRP000032    SM:NA12878      CN:MPIMG
@RG     ID:ERR001698    PL:ILLUMINA     LB:g1k-sc-NA12878-CEU-1 PI:200  DS:SRP000032    SM:NA12878      CN:SC
@RG     ID:SRR001114    PL:ILLUMINA     LB:Solexa-3620  PI:0    DS:SRP000032    SM:NA12878      CN:BI
@RG     ID:SRR001115    PL:ILLUMINA     LB:Solexa-3623  PI:0    DS:SRP000032    SM:NA12878      CN:BI
@PG     ID:GATK TableRecalibration.4    VN:v2.2.16      CL:Covariates=[ReadGroupCovariate, QualityScoreCovariate, DinucCovariate, CycleCovariate], use_original_quals=true,
default_read_group=DefaultReadGroup, default_platform=ILLUMINA, force_read_group=null, force_platform=null, solid_recal_mode=SET_Q_ZERO, window_size_nqs=5, homopolymer_nback=7,
exception_if_no_tile=false, pQ=5, maxQ=40, smoothing=1
@PG     ID:bwa  VN:0.5.5
```

samtools view –H my.bam | less -S

How is the BAM file sorted?
How many different sequencing centres contributed lanes to this BAM file?
What is the alignment tool used to create this BAM file?
How many different sequencing libraries are there in this BAM? Hint: RG tag

# SAM/BAM Tools

Several tools and programming APIs for interacting with SAM/BAM files

- **Samtools** - Sanger/C (http://htslib.org/)
  - Convert SAM <-> BAM (samtools view)
  - Sort and index BAM files (samtools sort, samtools index)
  - Give summary of the mapping flags (samtools flagstat)
  - Merge multiple sorted BAM files (samtools merge)
  - Remove PCR duplicates from the library preparation (samtools rmdup)
- **Picard** - Broad Institute/Java (http://picard.sourceforge.net)
  - MarkDuplicates, CollectAlignmentSummaryMetrics, CreateSequenceDictionary, SamToFastq, MeanQualityByCycle, FixMateInformation, … (and many others)
- Pysam – Python library based on htslib (http://code.google.com/p/pysam/)

BAM Visualisation

- Samtools (samtools tview): http://htslib.org/
- BamView, LookSeq, Gap5: http://www.sanger.ac.uk/resources/software/
- IGV: http://www.broadinstitute.org/igv/
- Tablet: http://bioinf.scri.ac.uk/tablet/
- PyBamView: http://melissagymrek.com/pybamview/

# CRAM

BAM files are too large

- ~1.5-2 bytes per base pair (12-16 bits/bp)

Increases in disk capacity are being far outstripped by sequencing technologies (per unit cost of each)

BAM stored all of the data

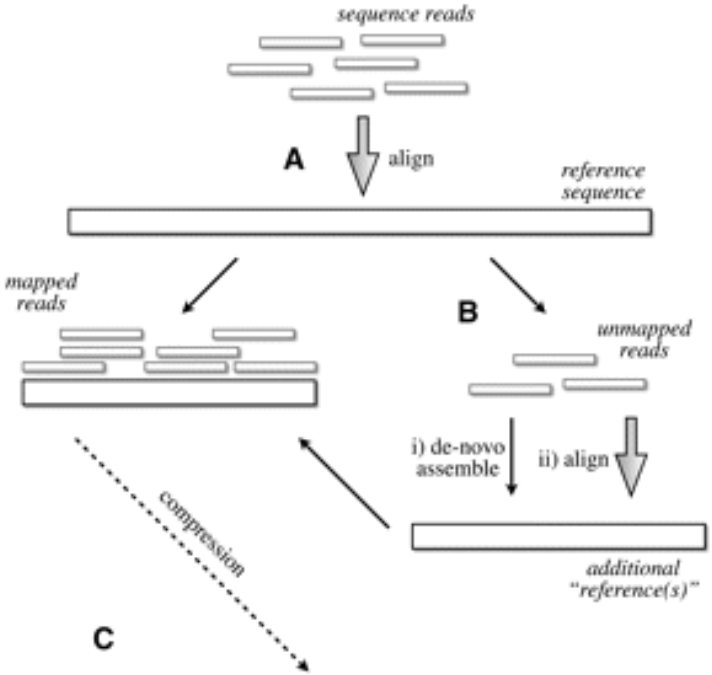
- every read base
- every base quality
- using conventional compression techniques

CRAM: Two important concepts

- Reference based compression
- Controlled loss of quality information

Supported by samtools (and htslib) – htsjdk support in development

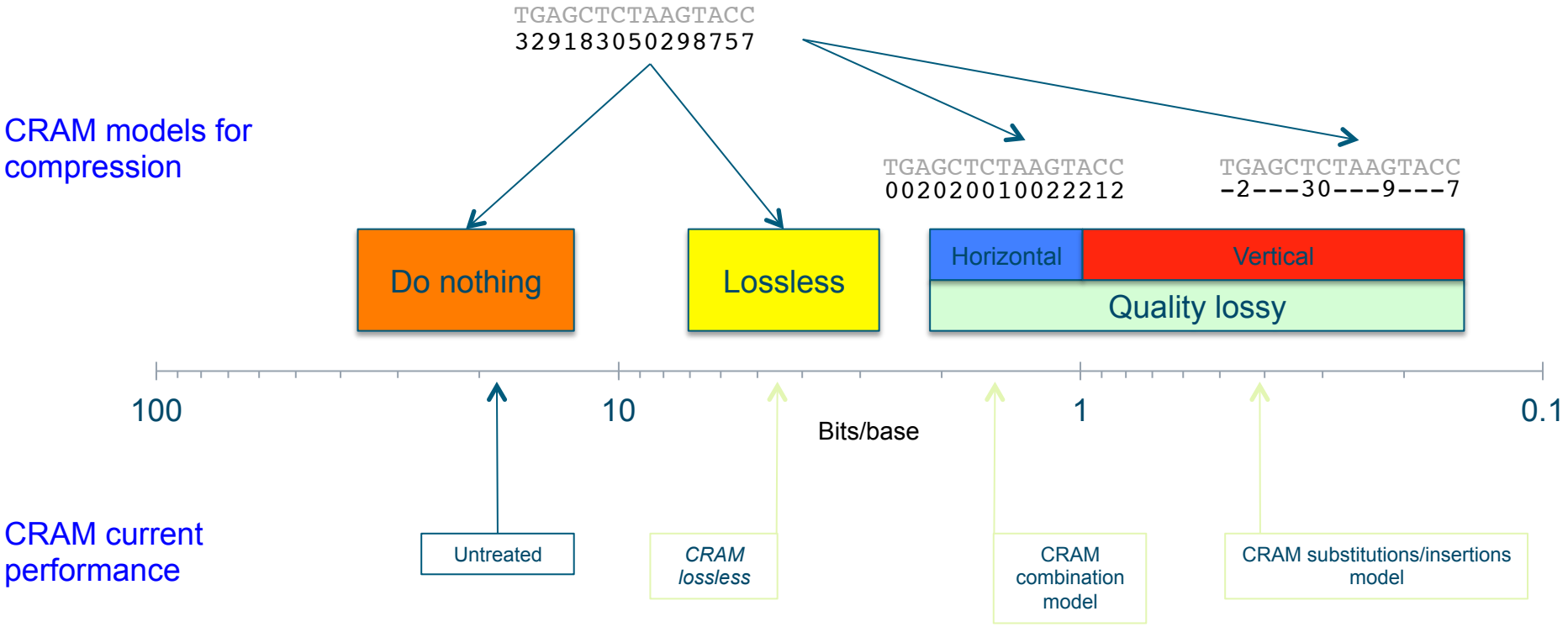
# Reference Based Compression




| Position | Strand | Substitutions | Insertions | Deletions |
|---|---|---|---|---|
| 4 | + | 4-G | *none* | 5-3 |
| 6 | + | *none* | *none* | *none* |
| 7 | + | *none* | 8-AT 4-C | *none* |

Fritz et al. (2012) *Gen Res*

# CRAM: Reference-based sequence data compression




- CRAM v0.9 released 03.08.12:
- Pairing information preservation regardless of distance
- Revised and improved lossless mode
- Option to preserve all unmapped reads
- Performance and bug fixes
- Arbitrary tags

http://www.ebi.ac.uk/ena/about/cram_toolkit
http://listserver.ebi.ac.uk/mailman/listinfo/cram-dev

# Variant Call Format (VCF)

VCF is a standardised format for storing DNA polymorphism data

- SNPs, insertions, deletions and structural variants
- With rich annotations

Can be indexed for fast data retrieval of variants from a range of positions

Store variant information across many samples

Record meta-data about the site

- dbSNP accession, filter status, validation status

Very flexible format

- Arbitrary tags can be introduced to describe new types of variants
- No two VCF files are necessarily the same
  - User extensible annotation fields supported
- Same event can be expressed in multiple ways by including different numbers
  - Recommendation on VCF format website to ensure consistency

# VCF file structure

Header section and a data section

Header

- Arbitrary number of meta-information lines
- Starting with characters '##'
- Column definition line starts with single '#'

Mandatory columns

- Chromosome (CHROM)
- Position of the start of the variant (POS)
- Unique identifiers of the variant (ID)
- Reference allele (REF)
- Comma separated list of alternate non-reference alleles (ALT)
- Phred-scaled quality score (QUAL)
- Site filtering information (FILTER)
- User extensible annotation (INFO)

# Example VCF – SNPs/indels

```
##fileformat=VCFv4.1
##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
```

# Indexing and compression

Text formats (VCF, SAM, FASTQ) can be block compressed

- Using `bgzip` or built-in capability of other software (e.g. `samtools`)

Sorted, bgzip-compressed files can then be indexed with `tabix`

- Using the index, software can access specific regions without having to stream through the full file

Other formats such as BAM, CRAM have their own indexing system (e.g. '.bai', '.crai', '.csi') but work similarly

# NGS Bioinformatics Overview

- NGS Overview
- NGS Analysis Workflows
- NGS Data Formats
- **Lab Exercises**