Introduction to High-Throughput Sequencing for Variant Discovery and Analysis

Everything you need to know about HTS in order to use the GATK

Ultimate goal is to identify genomic variation in sequencing data

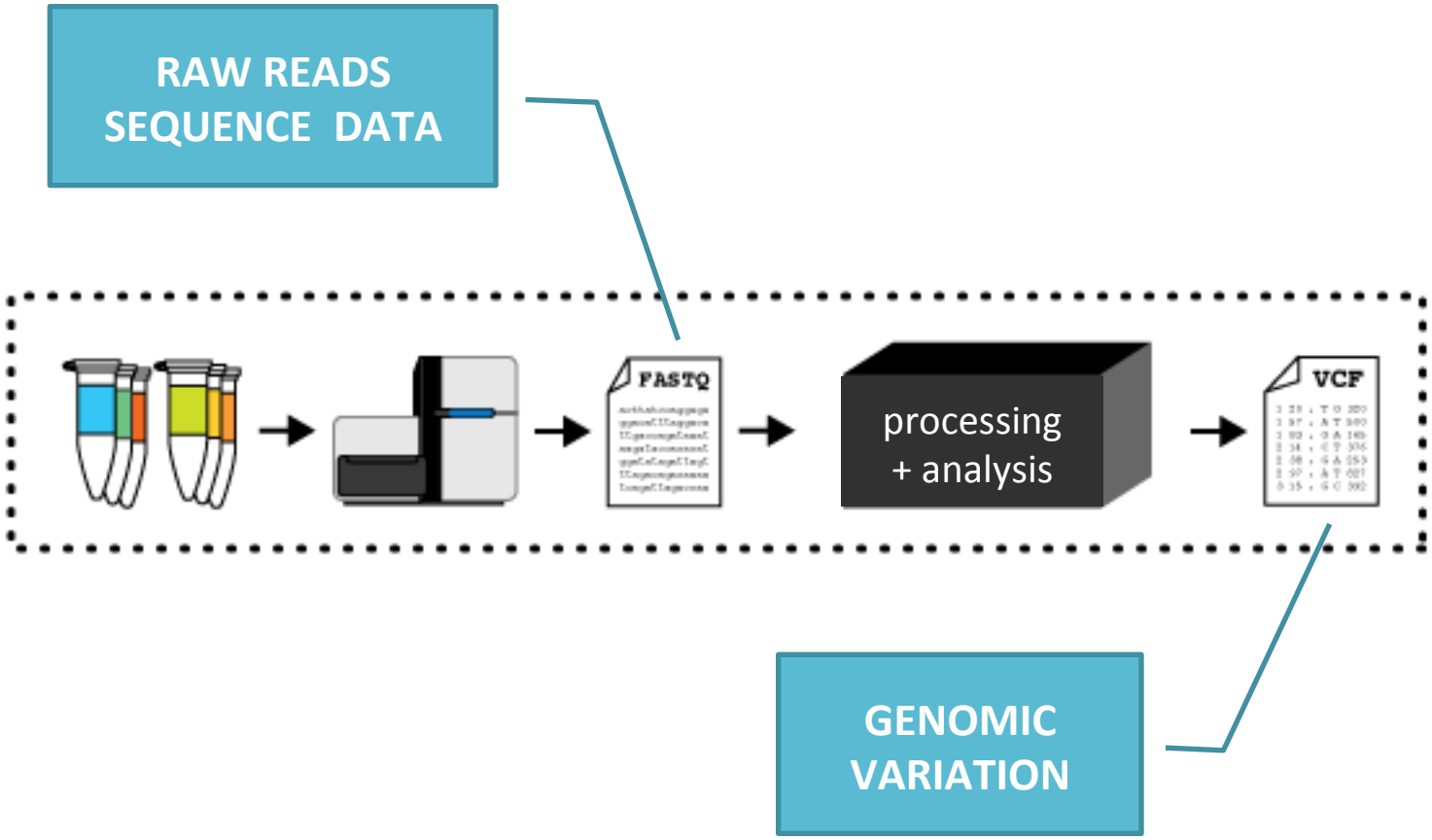


Table of Contents

PART 1:
HTS TECHNOLOGIES AND TERMINOLOGY

Today's sequencing machines produce a massive amount of data

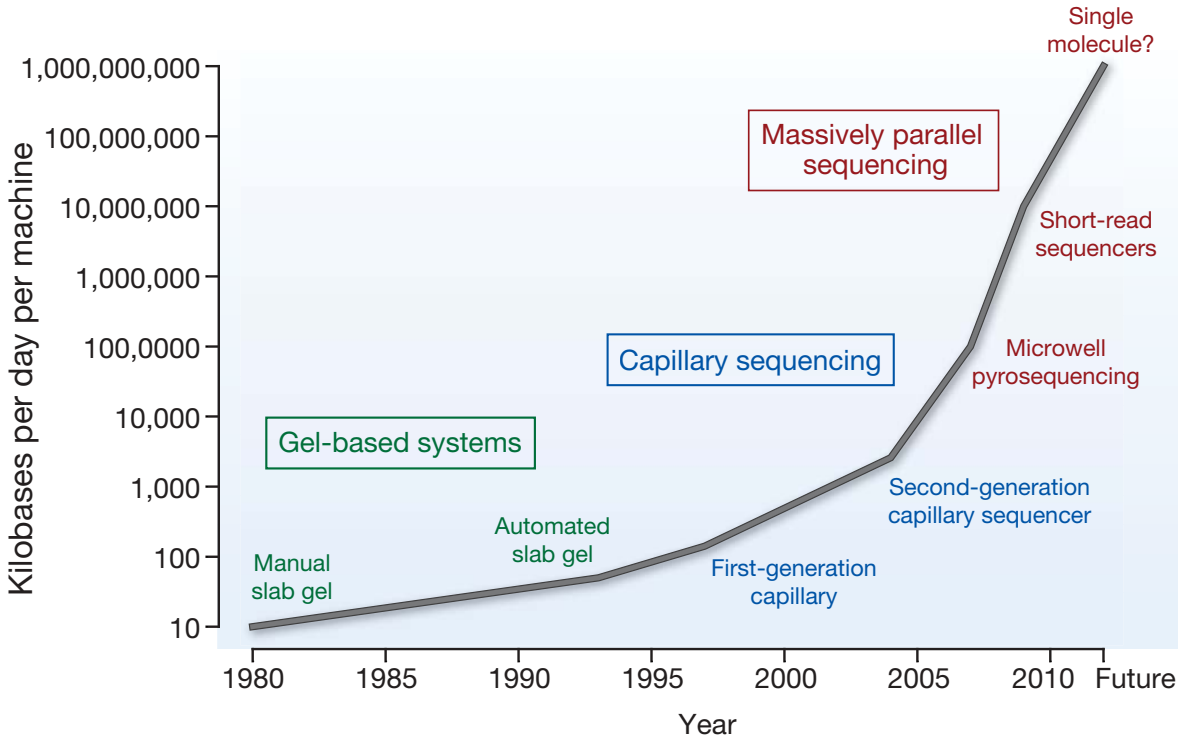
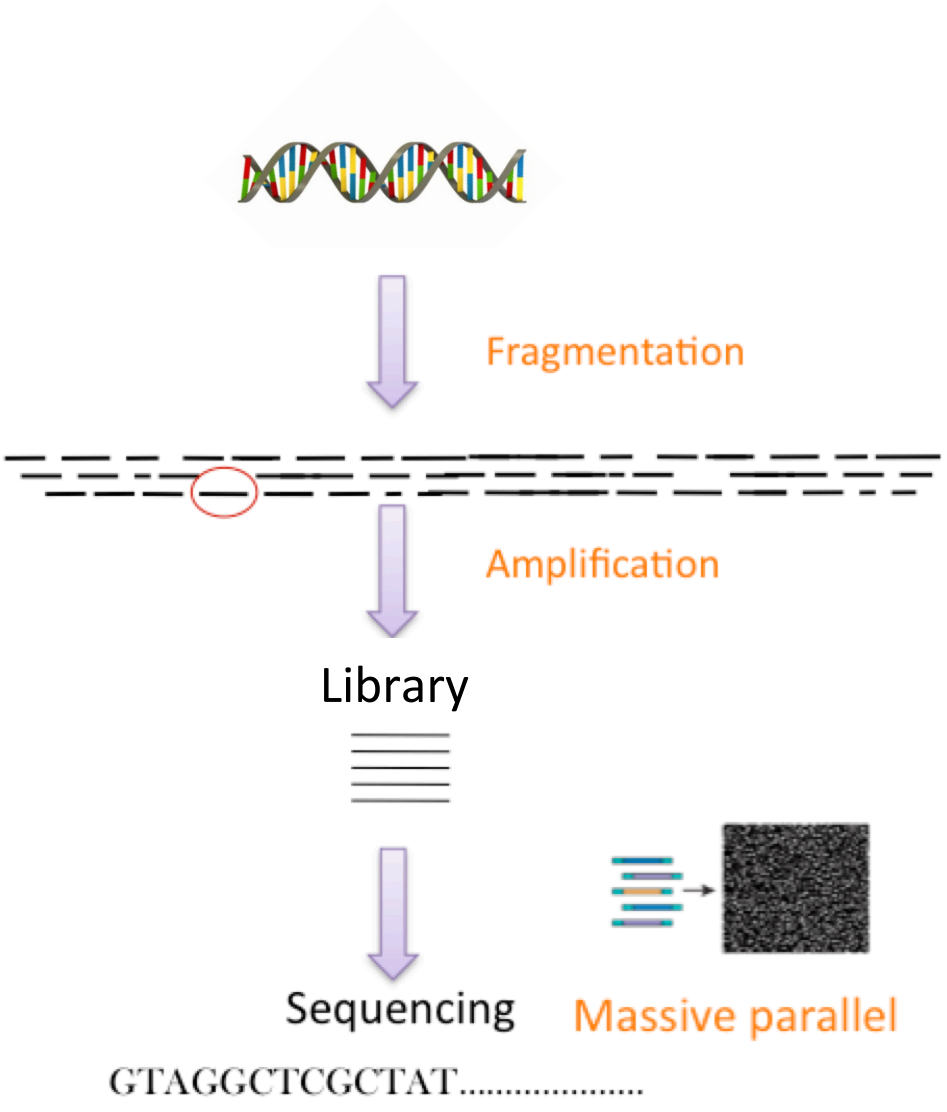
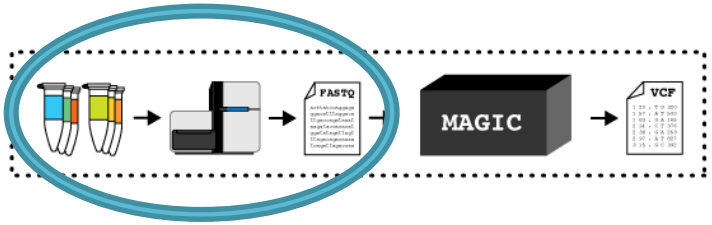
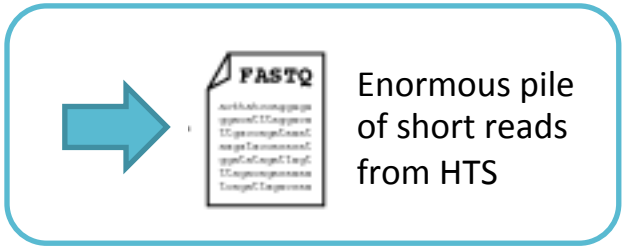


Figure 3 | Improvements in the rate of DNA sequencing over the past 30 years and into the future. From slab gels to capillary sequencing and second-generation sequencing technologies, there has been a more than a million-fold improvement in the rate of sequence generation over this time scale.

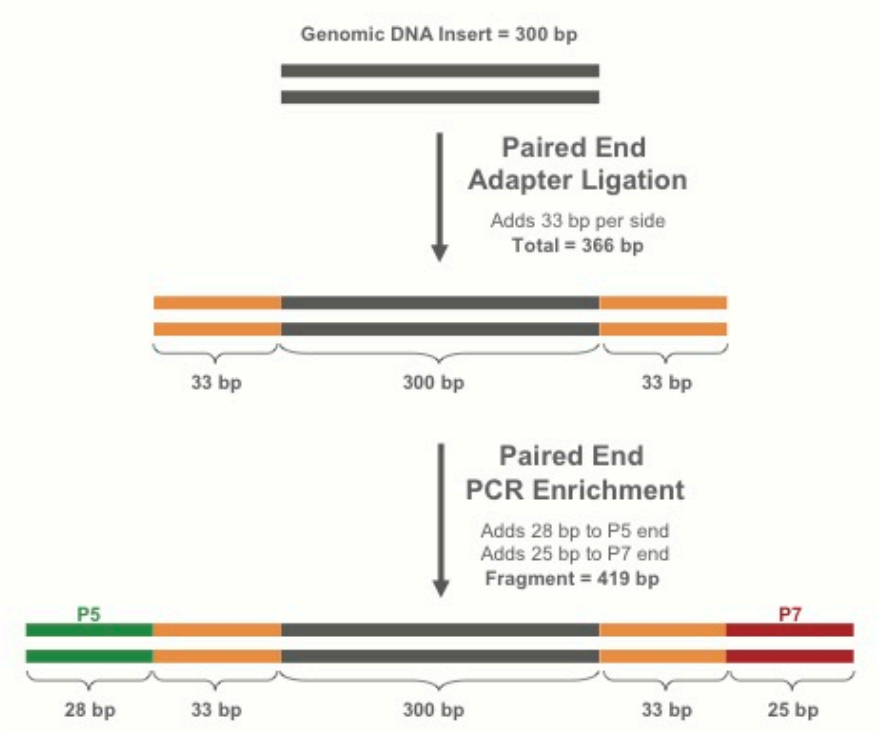
High-throughput sequencing yields a big pile of reads



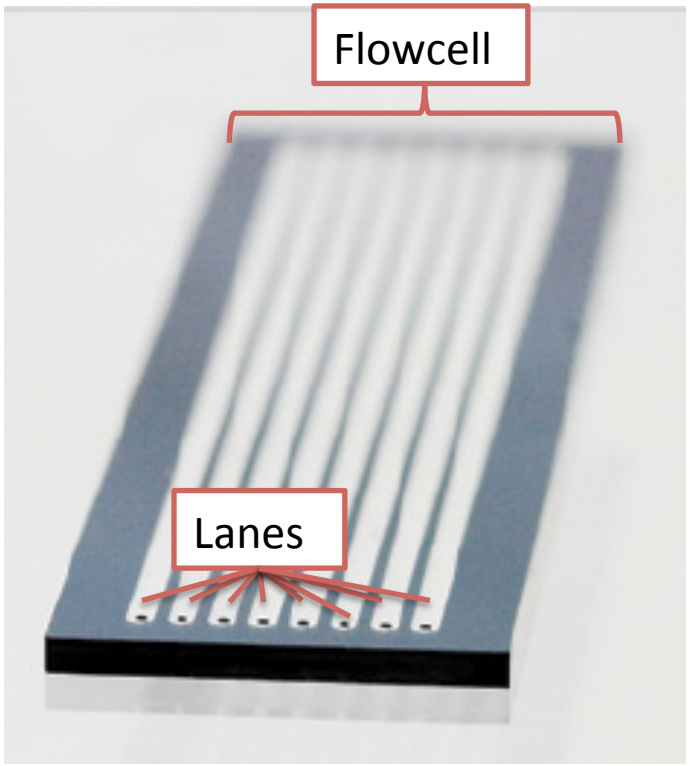




Terminology: libraries, lanes, and flowcells

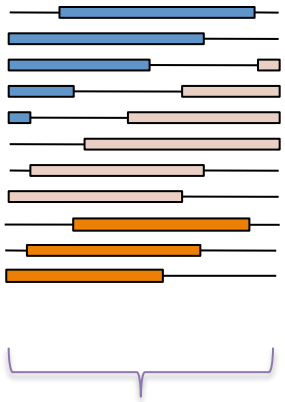


Each reaction produces a unique **library** of DNA fragments for sequencing.



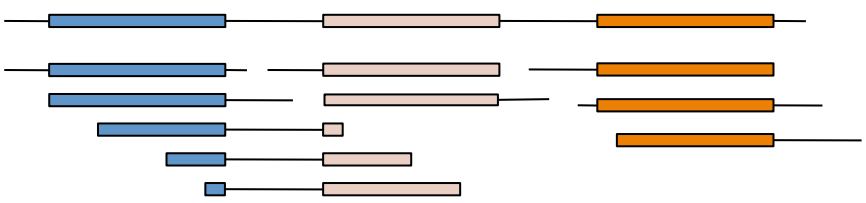
Each NGS machine processes a single **flowcell** containing several independent **lanes** during a single sequencing run

Instruments generate short reads that must be mapped to the reference



Enormous pile of short reads from HTS

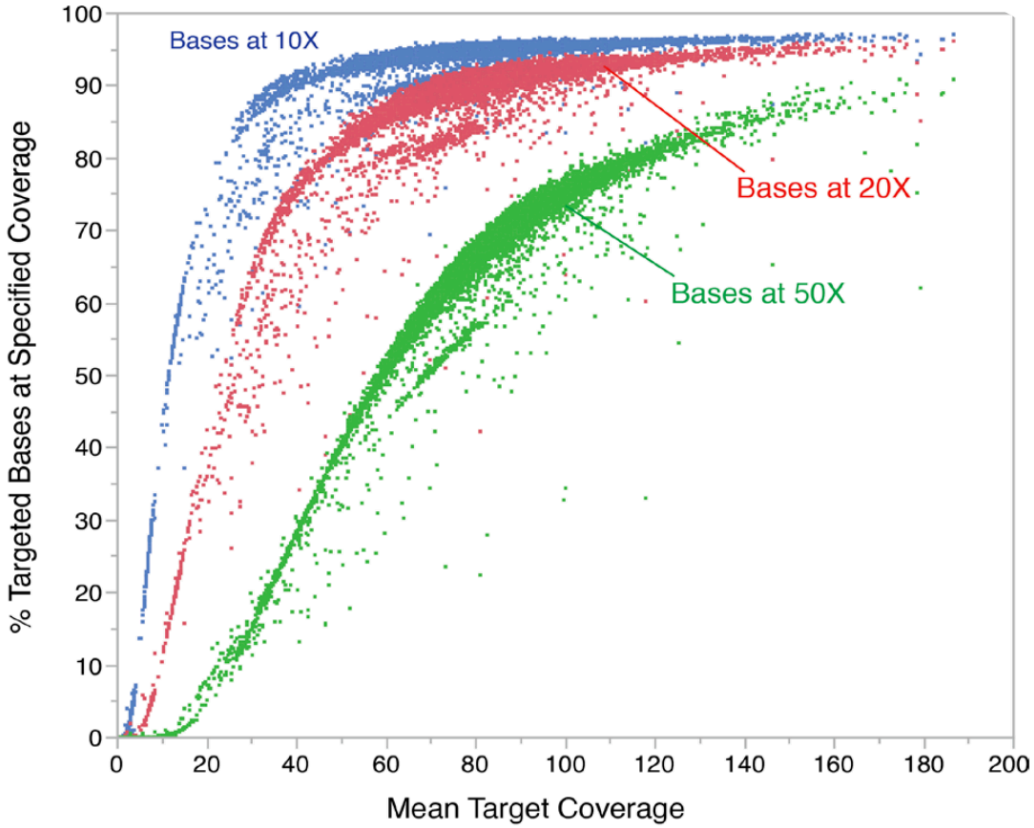
Mapping and alignment algorithms



Reference genome

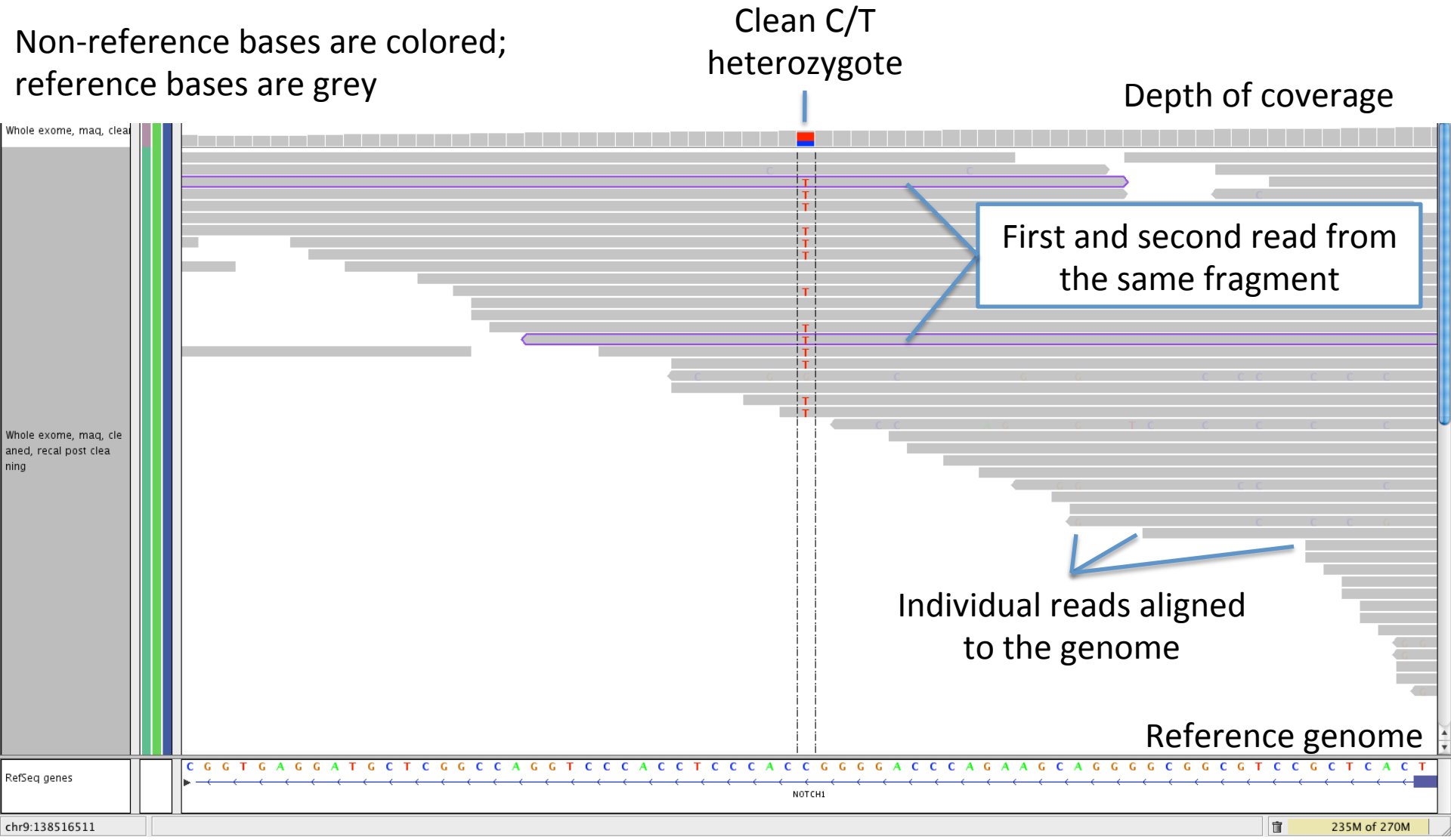
Reads mapped to reference

Terminology: coverage



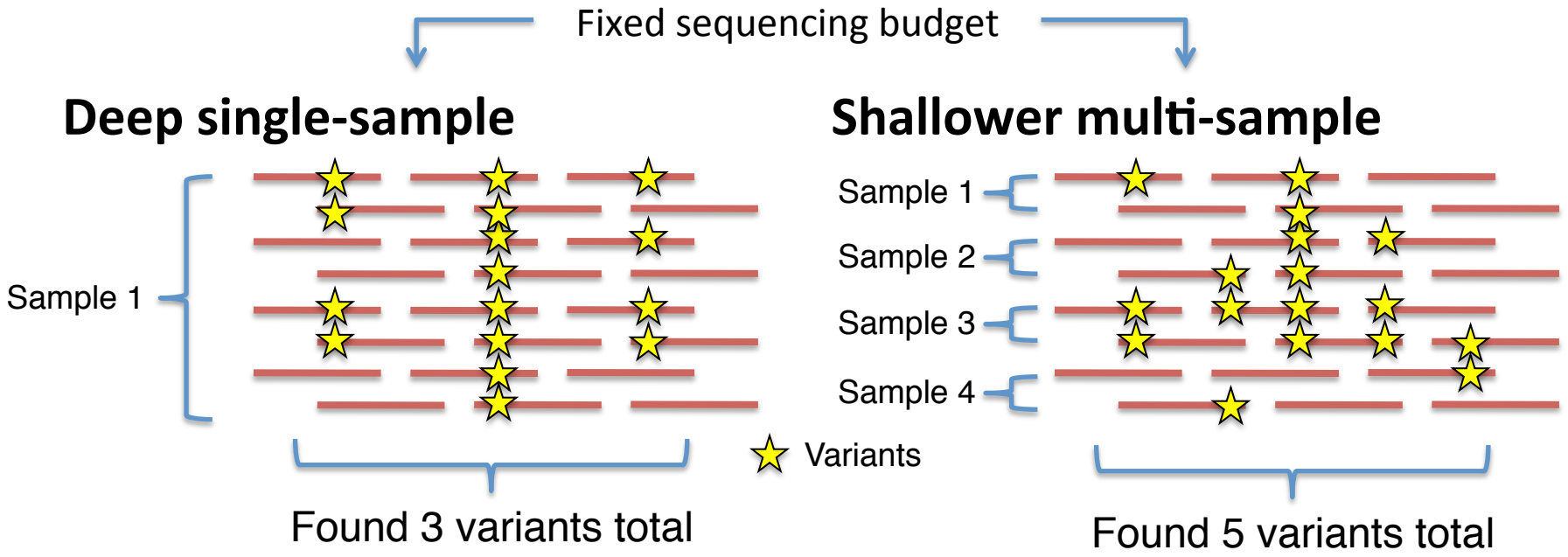
Distribution of coverage levels for targeted bases for representative samples sequenced to ~10X, ~20X, and ~50X mean target coverage

How we visualize aligned HTS reads (Integrated Genomics Viewer)



PART 2:
EXPERIMENTAL DESIGNS

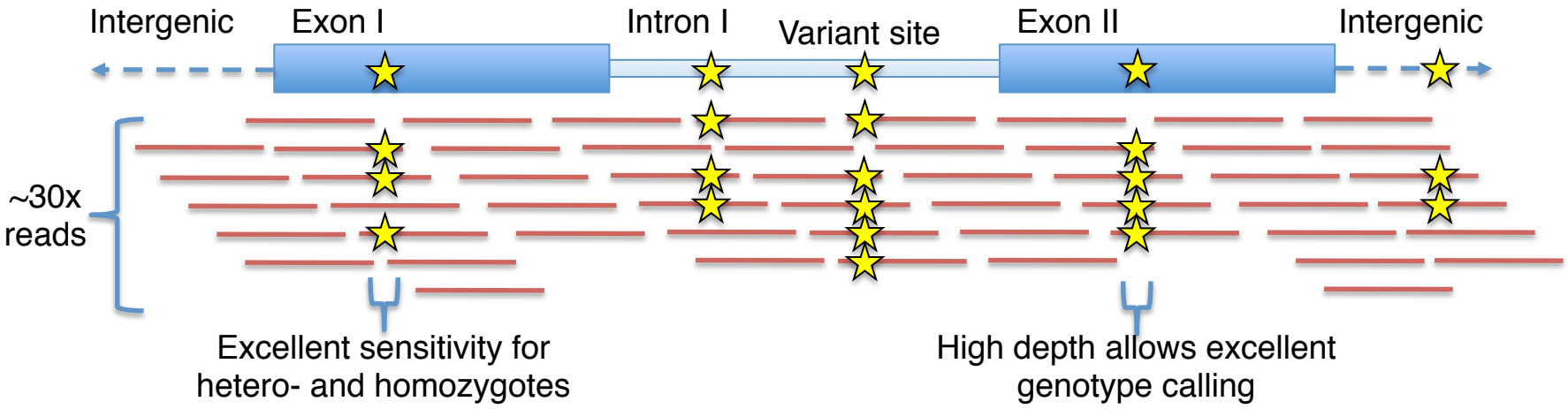
Single vs. multi-sample analysis



- Higher sensitivity for variants in the sample
- More accurate genotyping per sample
- Cost: no information about other samples

- Sensitivity dependent on frequency of variation
- Worse genotyping
- More total variants discovered

High-pass sequencing design



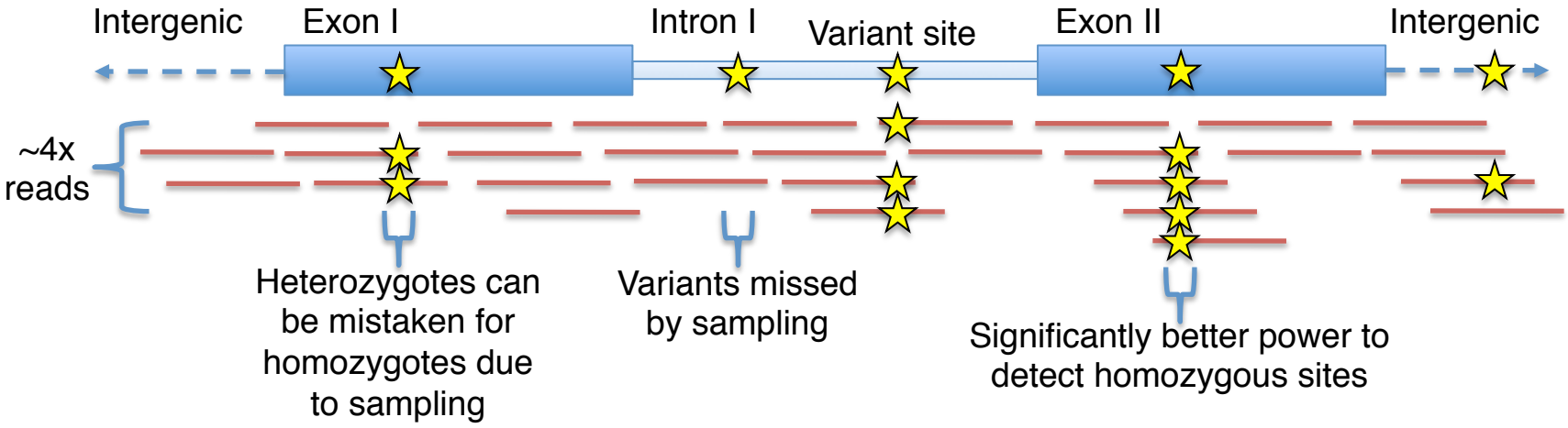
Data requirements per sample

Targeted bases	~3 Gb
Coverage	Avg. 30x
# sequenced bases	100 Gb
# lanes of HiSeq	~8 lanes

Variant detection among multiple samples

Variants found per sample	~3-5M
Percent of variation in genome	>99%
Pr{singleton discovery}	>99%
Pr{common allele discovery}	>99%

Low-pass sequencing design



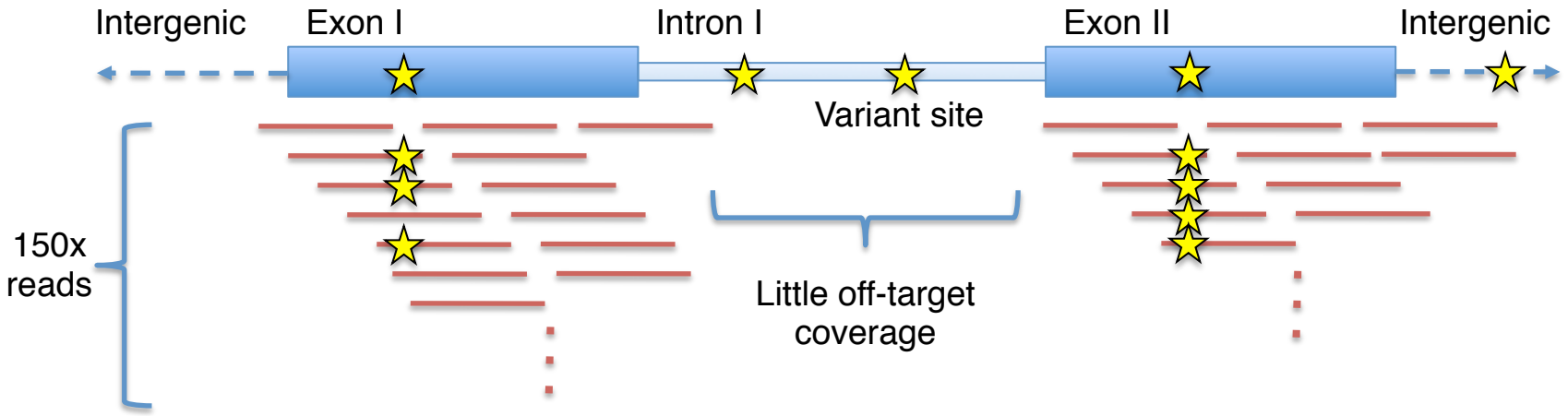
Data requirements per sample

Targeted bases	~3 Gb
Coverage	Avg. 4x
# sequenced bases	20 Gb
# lanes of HiSeq	~1.25

Variant detection among multiple samples

Variants found per sample	~3M
Percent of variation in genome	~90%
Pr{singleton discovery}	<50%
Pr{common allele discovery}	~99%

Exome capture sequencing design



Data requirements per sample

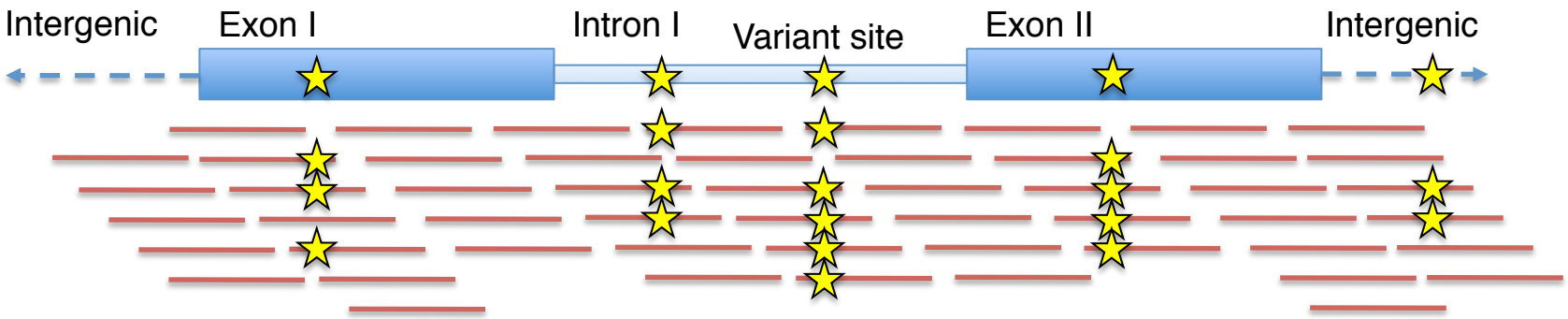
Targeted bases	~32Mb
Coverage	>80% 20x
# sequenced bases	5 Gb
# lanes of HiSeq	~0.33

Variant detection among multiple samples

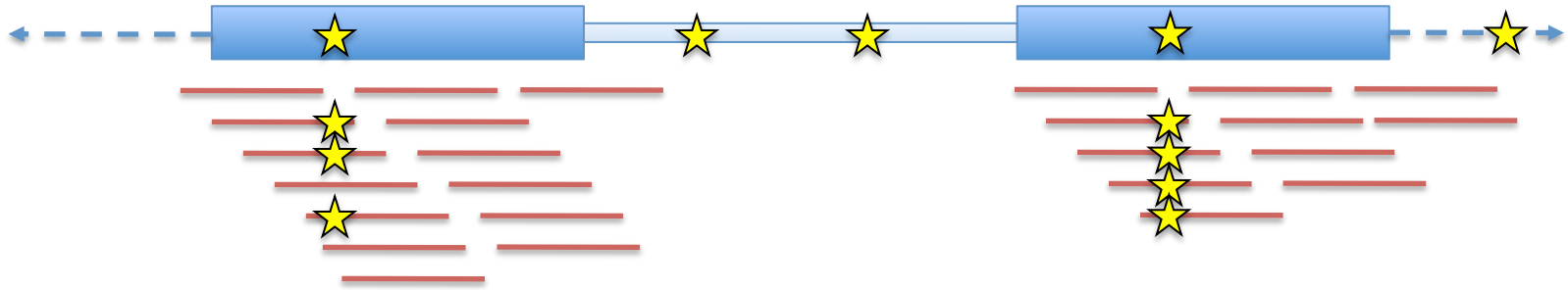
Variants found per sample	~20K
Percent of variation in genome	0.5%
Pr{singleton discovery}	~95%
Pr{common allele discovery}	~95%

Whole genome (WGS) vs. Exome (WEx)

Whole genome



Exome



Small targeted experiments, gene panels, RADseq

- Similar to exomes for most purposes

Key differences between whole genome and exome

Whole genome

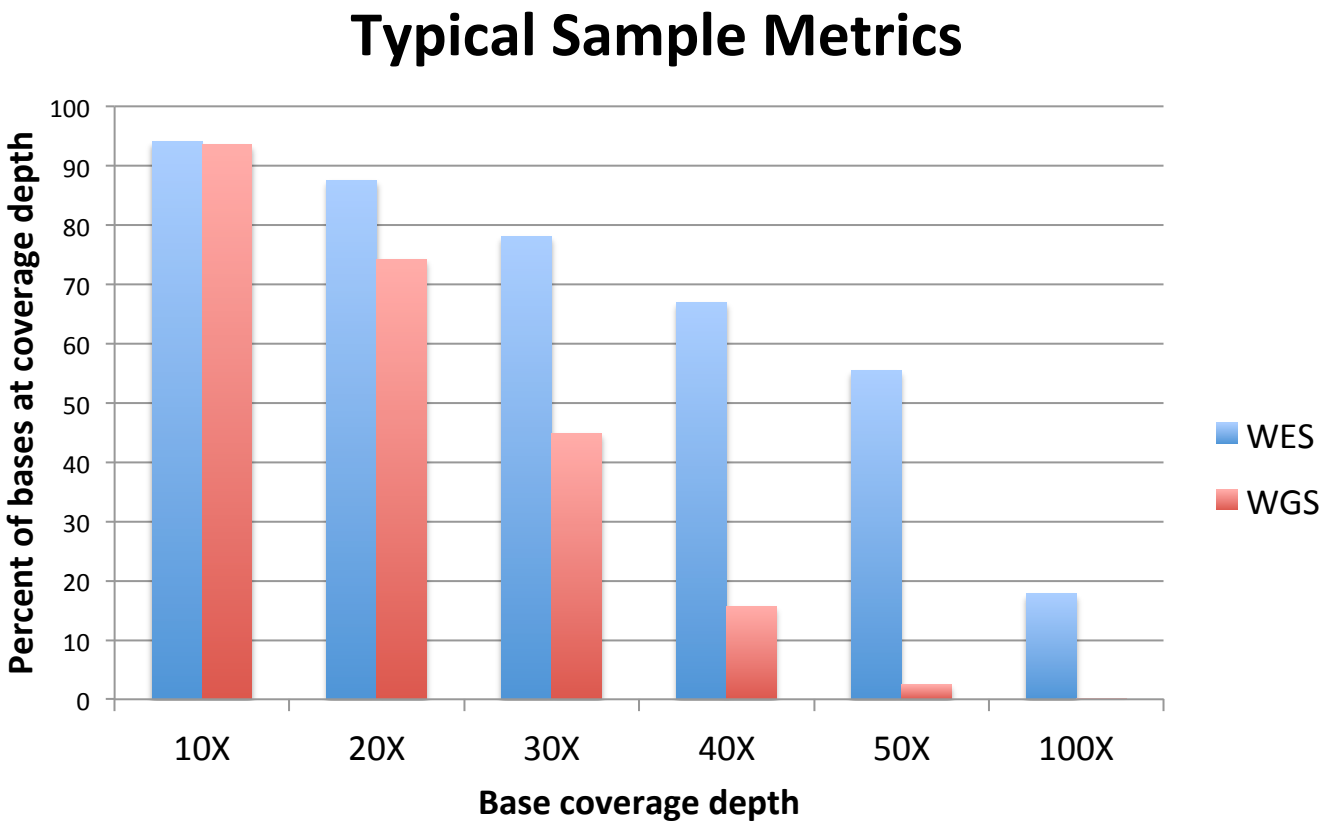
- Entire genome is prepped
- Possible to do PCR-free
- Higher price
- Exons + introns
- Coverage everywhere
  (well, almost – some missing)

Exome

- Capture of target regions
- PCR required
- Lower price
- Exons only
- Coverage only on targets
  (well, almost – some spillover)

- ➢ Use Best Practices as presented
- ➢ Must adapt Best Practices slightly

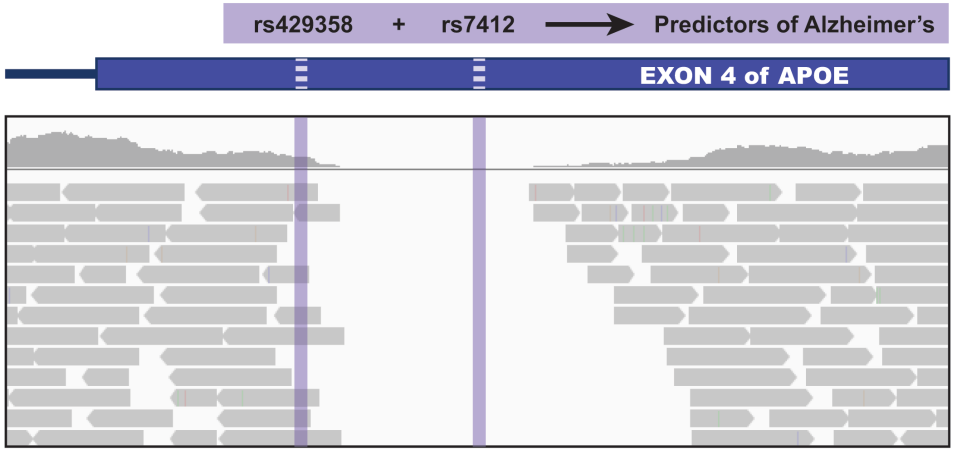
Key differences between whole genome and exome



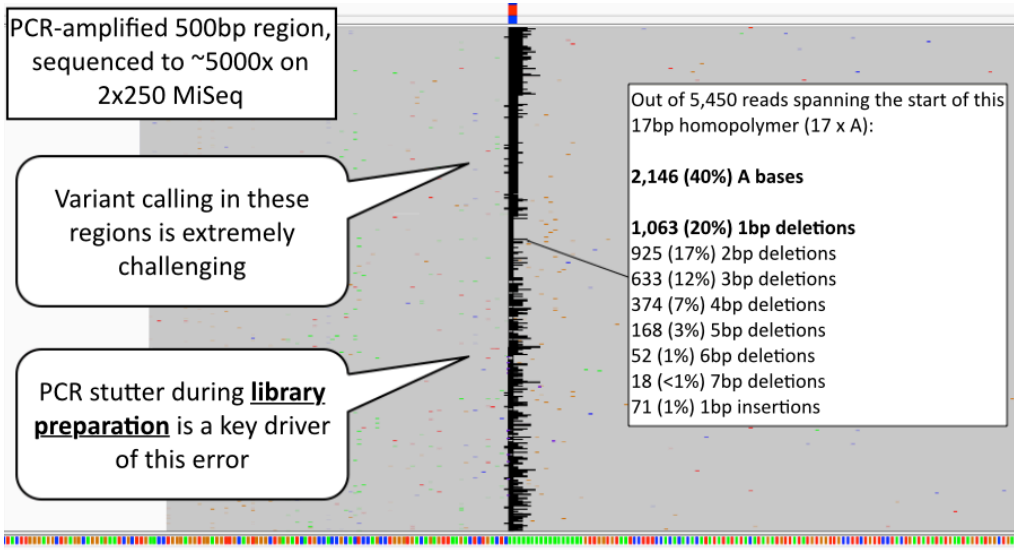
- To achieve 80% at 20X target for exomes, many bases are covered up to 100X
- WGS achieves mean coverage >20X with fewer high coverage bases

Drawbacks of PCR to keep in mind

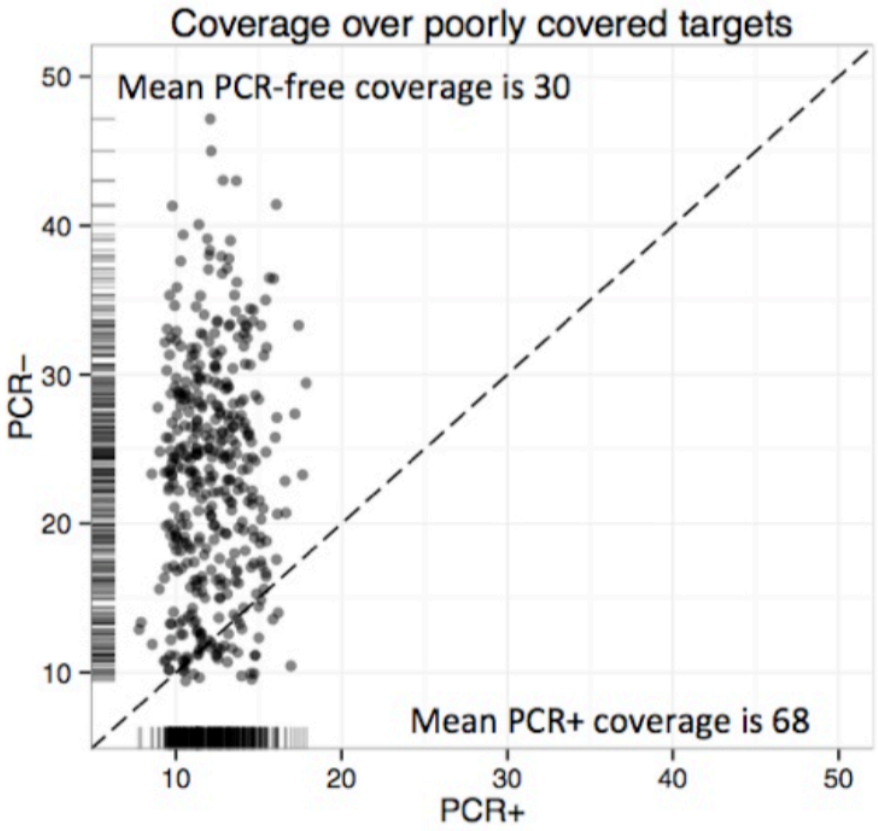
- Amplification rate depends on sequence GC content
  - Poor coverage of GC-rich regions
  - Skewed coverage profile

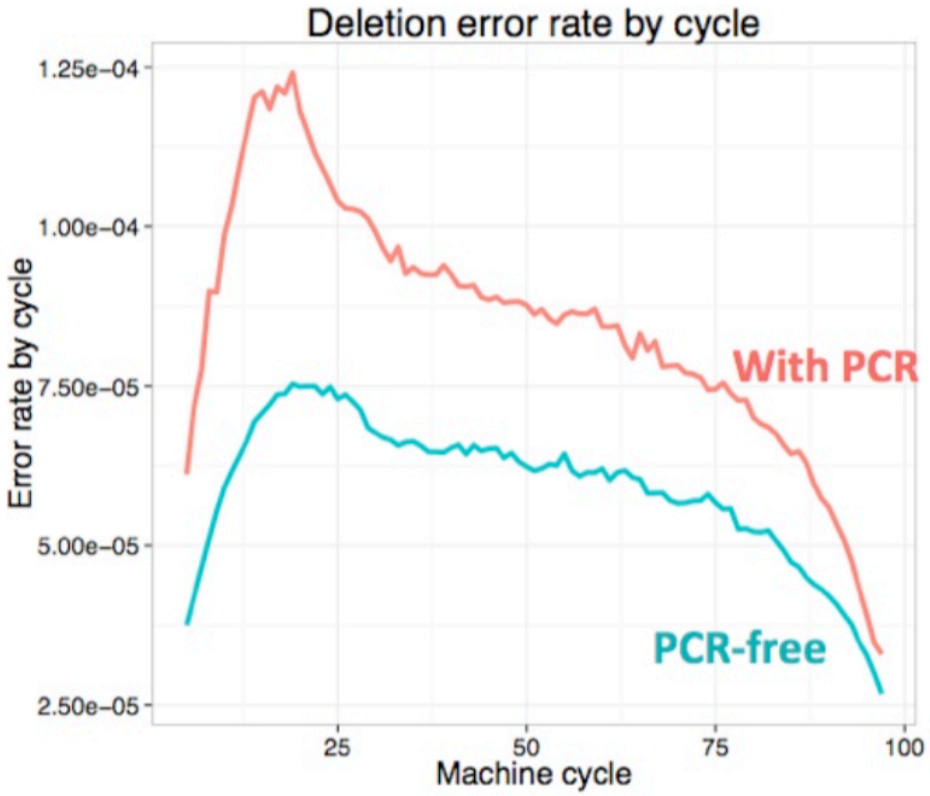


- Systematic PCR stutter in repetitive sequences
  - Complicates indel calling
  - **HaplotypeCaller includes --pcr_indel_model setting**



PCR-Free whole genomes are much better
(if you can afford them)



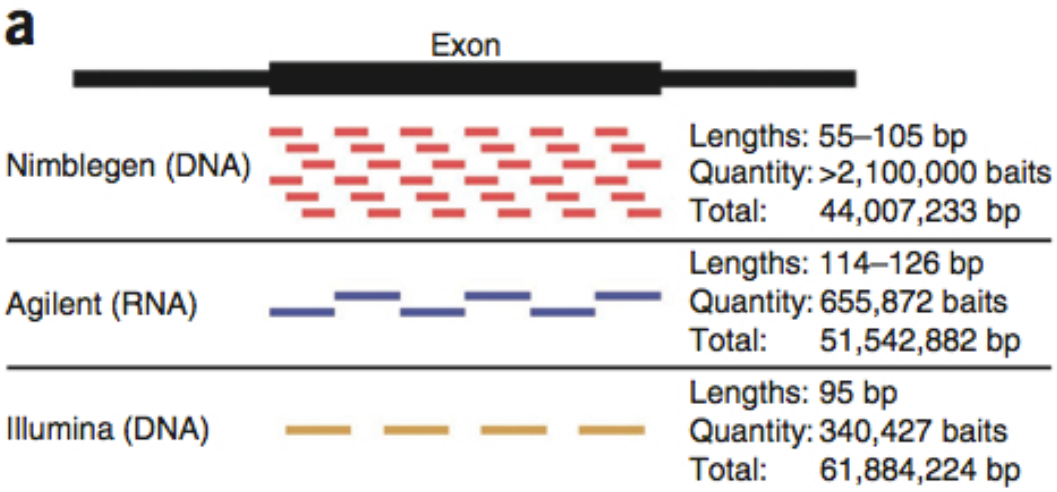


The regions with < 20 coverage in the PCR+ data receive only slightly less than average PCR-free coverage

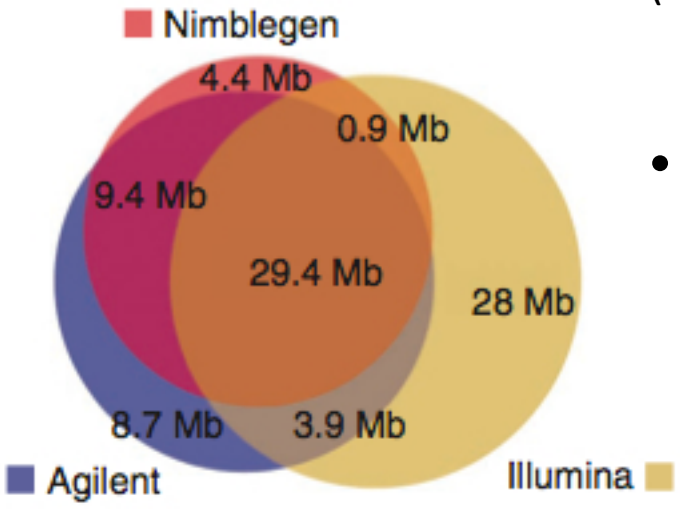
Average error rate is 2x lower. Errorful regions are improved even more

Exome capture target regions

- Involves **baits** (complementary sequences) tiled to capture segments of target regions



(Clark et al, Nat. Biotech., 2011)



(Clark et al, Nat. Biotech., 2011)

- Resulting covered intervals are specific to capture kit manufacturer

Broad uses a custom-designed bait set; the target intervals list is available in our resource bundle

Basic recommendation : restrict analyses to capture intervals

- Obtain the appropriate interval list from the capture kit manufacturer or sequence provider

➢ Use –L argument to restrict analyses at key steps*

```
java -jar GenomeAnalysisTK.jar
      [tool and other arguments] \
      -L intervals.list \
      --interval_padding 50
```

* RealignerTargetCreator, BaseRecalibrator, HaplotypeCaller

Can you include off-target regions in analysis?

- Some investigators argue **yes because =free data!**
  (e.g. Samuels *et al.*, Trends in Genetics 2013)

- **BUT** keep in mind adverse effects on
  - Runtime (lots of extra ground to cover)
  - Base recalibration (BQSR)
  - Variant calling metrics (FPs, Ti/Tv ratio)
  - Ability to do apples-to-apples comparisons on datasets from different origins

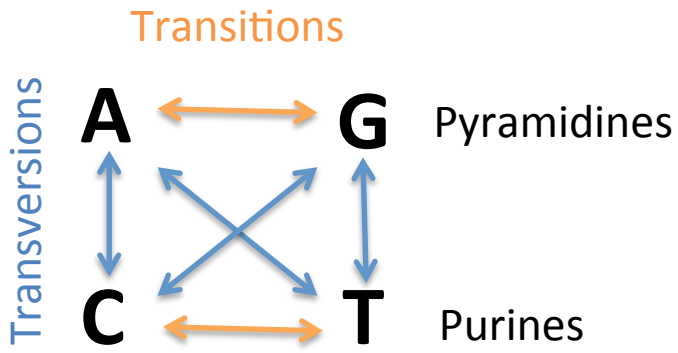
Effect on variant calling metrics

➢**If you do not use –L argument with HaplotypeCaller to restrict calling to capture targets :**

- Different expectations for total variant counts
- More low-confidence calls, possible false positives (FPs) likely in off-target regions due to lower coverage
- Different Ti/Tv expectations in exons (~3.0) vs. elsewhere (~2.0)

  – Evolution pressures are different
    - Different rates of mutation
    - Different types of mutation

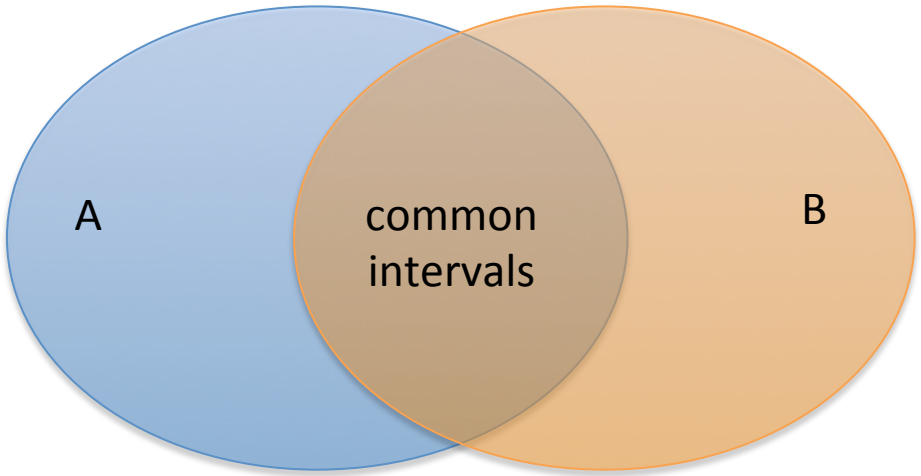
  ➢ Different Ti/Tv ratio



Effect on ability to compare datasets

- If samples were sequenced by different centers, with different capture or enrichment kits, covered (strictly usable) intervals are not equivalent.

- Must decide whether to use union or intersection of interval lists (we use intersection)



How does the Broad deal with intervals?

- Call bases and align reads over entire genome
- Call variants over “padded” intervals
  - Use “–ip 50” engine-level argument for HaplotypeCaller

- Perform QC evaluation over exact target intervals

What about gene panels, RADseq, ...?

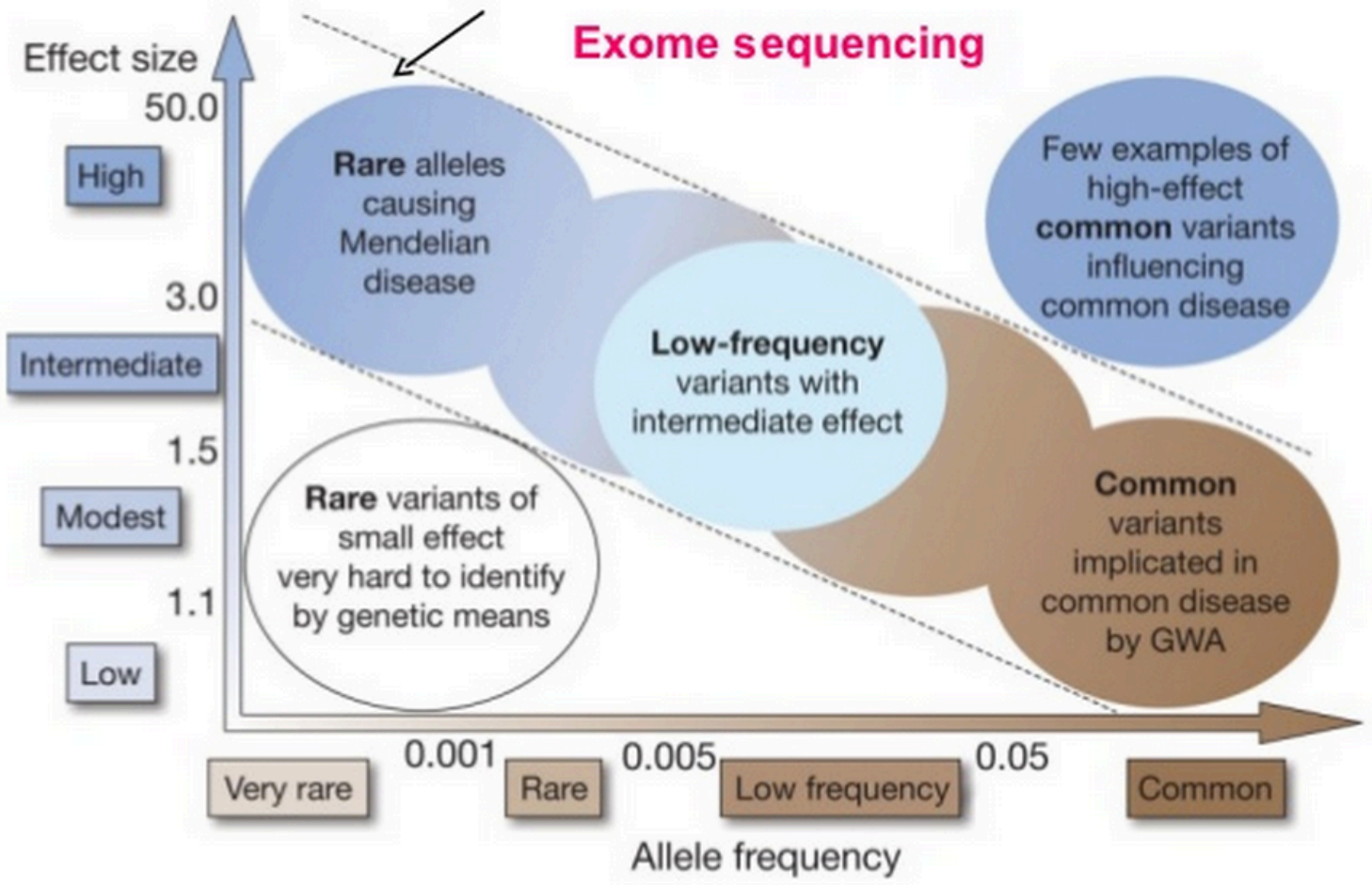
- Targeted gene panels
  - Functionally similar to exome with very small interval list

- RADseq
  - Uses **R**estriction **A**nalysis **D**igest i.e. chop up DNA at frequent restriction enzyme sites
  - Can be done without an assembled reference

"stacks" of sequence

- Same drawbacks as exomes PLUS some new ones
  - Produces many independent fragments with identical start and end, so MarkDuplicates should not be run
  - Too few variants for VQSR, must use hard-filtering

Application to clinical research



Manolio TA, et al (2009). Finding the missing heritability of complex diseases. Nature., 461(7265), 747

PART 3:
DATA FORMATS

From reads to variants :

Major steps involve transforming data and storing results in specific formats

Mapping + cleanup ⇨ **Variant Discovery** ⇨ **Variant Evaluation**

FASTQ -> BAM

from raw reads to aligned reads

BAM -> VCF

from aligned reads to genomic variation

Important file format #1: FASTQ (raw reads)

- Simple extension from traditional FASTA format.
- Each block has 4 elements (in 4 lines):
  - Sequence Name (read name, group, etc.)
  - Sequence
  - + (optional: Sequence name again)
  - Associated quality score.
- Example record:

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
```

Identifier

Sequence

Base Qualities
(ASCII 33 + Phred scaled Q)

Official specification in http://maq.sourceforge.net/fastq.shtml

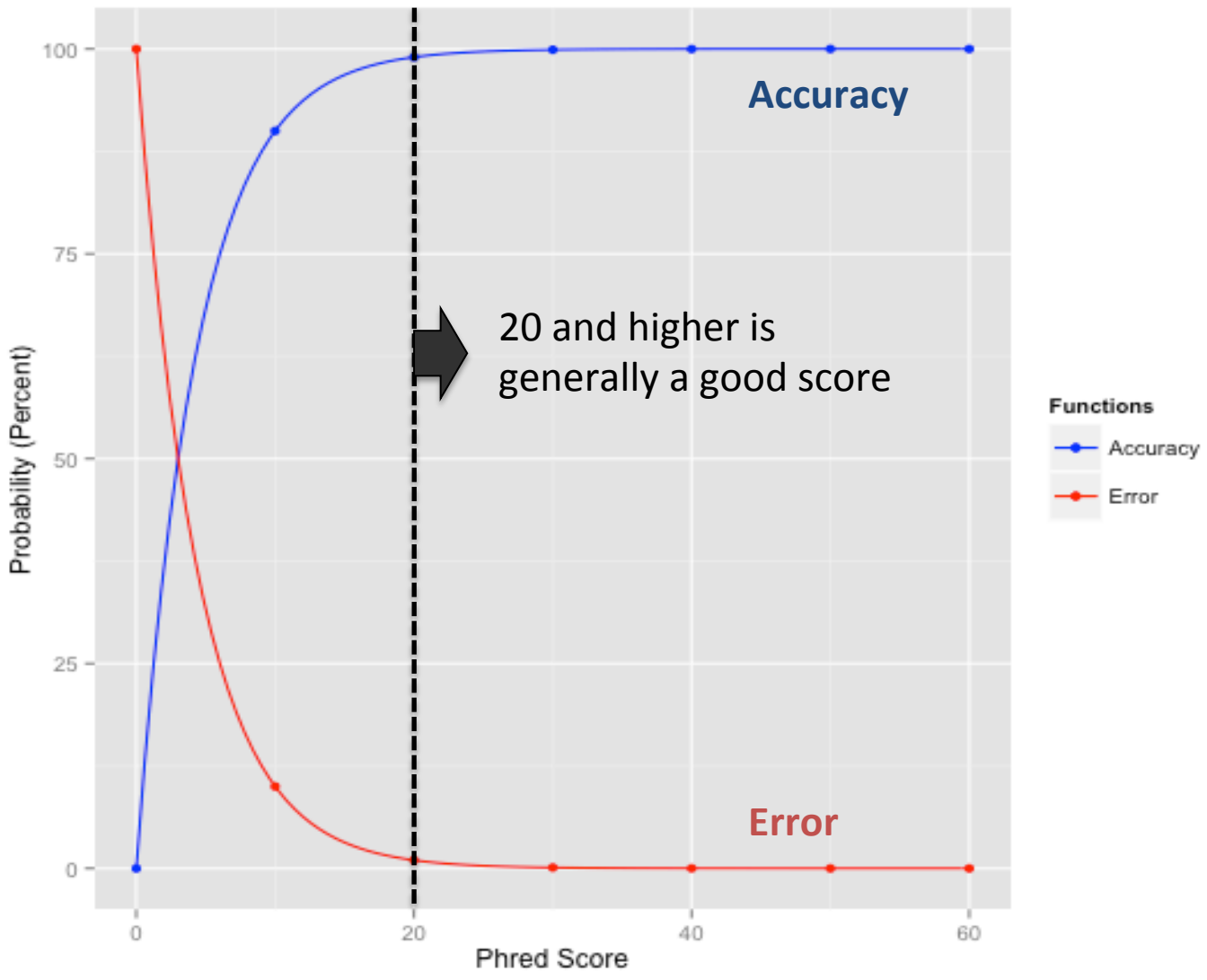
A quick guide to Phred scaling

- *Phred value* = −10 * *log*10(ε)

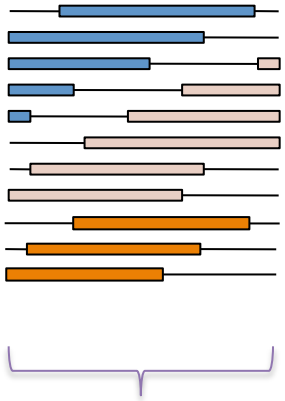
- Examples:
  - 90% confidence (10% error rate) = Q10
  - 99% confidence (1% error rate) = Q20
  - 99.9% confidence (.1% error rate) = Q30

- SAM encoding adds 33 to the value
  (because ASCII 33 is the first visible character)

Visualizing the meaning of Phred Scores

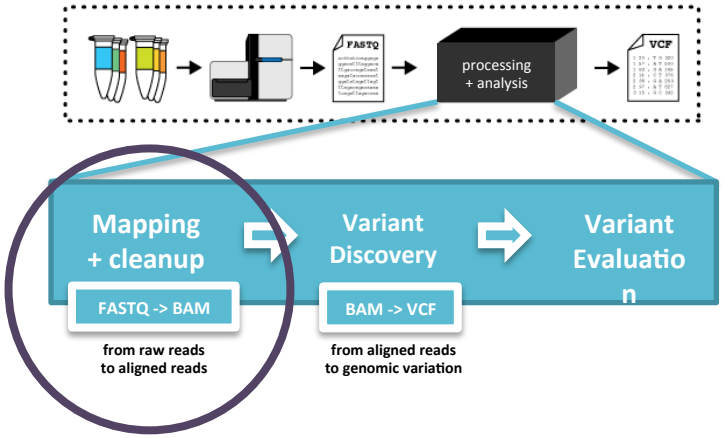


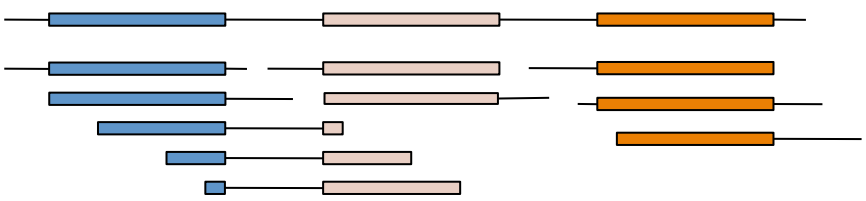
Mapping adds information : codified in SAM/BAM format



Enormous pile of short reads from HTS

Mapping and alignment algorithms





Reference genome

Reads mapped to reference

Important file format #2: SAM/BAM (aligned reads)

Sequence **A**lignment **M**ap / **B**inary **A**lignment **M**ap (compressed)

BAM file allows us to represent the data of any sequencer. Analyses can then be conducted largely agnostic to the particular sequencer used.

-> technology-independent

Data processing and analysis

A BAM file can contain data from a single or from several samples

BAM headers: an essential part of a BAM file

```
@HD    VN:1.0  GO:none SO:coordinate
@SQ    SN:chrM       LN:16571
@SQ    SN:chr1       LN:247249719
@SQ    SN:chr2       LN:242951149
[cut for clarity]
@SQ    SN:chr9       LN:140273252
@SQ    SN:chr10      LN:135374737
@SQ    SN:chr11      LN:134452384
[cut for clarity]
@SQ    SN:chr22   LN:49691432
@SQ    SN:chrX       LN:154913754
@SQ    SN:chrY       LN:57772954
@RG    ID:20FUK.1    PL:illumina    PU:20FUKAAXX100202.1    LB:Solexa-18483 SM:NA12878    CN:BI
@RG    ID:20FUK.2    PL:illumina    PU:20FUKAAXX100202.2    LB:Solexa-18484 SM:NA12878    CN:BI
@RG    ID:20FUK.3    PL:illumina    PU:20FUKAAXX100202.3    LB:Solexa-18483 SM:NA12878    CN:BI
@RG    ID:20FUK.4    PL:illumina    PU:20FUKAAXX100202.4    LB:Solexa-18484 SM:NA12878    CN:BI
@RG    ID:20FUK.5    PL:illumina    PU:20FUKAAXX100202.5    LB:Solexa-18483 SM:NA12878    CN:BI
@RG    ID:20FUK.6    PL:illumina    PU:20FUKAAXX100202.6    LB:Solexa-18484 SM:NA12878    CN:BI
@RG    ID:20FUK.7    PL:illumina    PU:20FUKAAXX100202.7    LB:Solexa-18483 SM:NA12878    CN:BI
@RG    ID:20FUK.8    PL:illumina    PU:20FUKAAXX100202.8    LB:Solexa-18484 SM:NA12878    CN:BI
@PG    ID:BWA  VN:0.5.7       CL:tk
@PG    ID:GATK PrintReads     VN:1.0.2864
20FUKAAXX100202:1:1:12730:189900      163    chrM  1     60    101M  =     282    381
        GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTA…[more bases]
        ?BA@A>BBBBACBBAC@BBCBBCBC@BC@CAC@:BBCBBCACAACBABCBCCAB…[more quals]
        RG:Z:20FUK.1    NM:i:1  AM:i:37 MD:Z:72G28   MQ:i:60 PG:Z:BWA       UQ:i:33
```

Required: Standard header

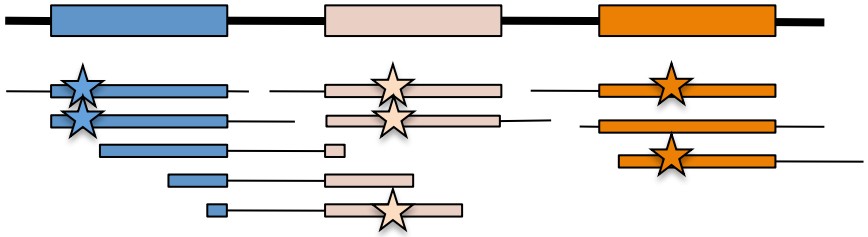
Essential: contigs of aligned reference sequence. Should be in karyotypic order.

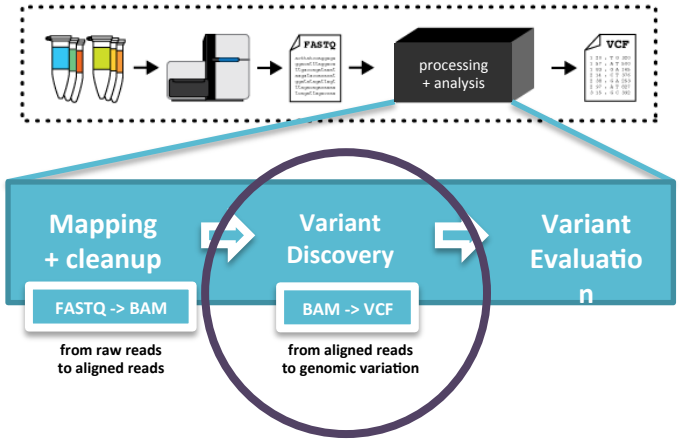
Essential: read groups. Carries platform (PL), library (LB), and sample (SM) information. Each read is associated with a read group

Useful: Data processing tools applied to the reads

Official specification in http://samtools.sourceforge.net/SAM1.pdf

Variant calls are a summarized representation of the original sequence data





Variant calling algorithms

- site 1 description + sample genotypes
- site 2 description + sample genotypes
- site 3 description + sample genotypes

Important file format #3: VCF (genomic variation)

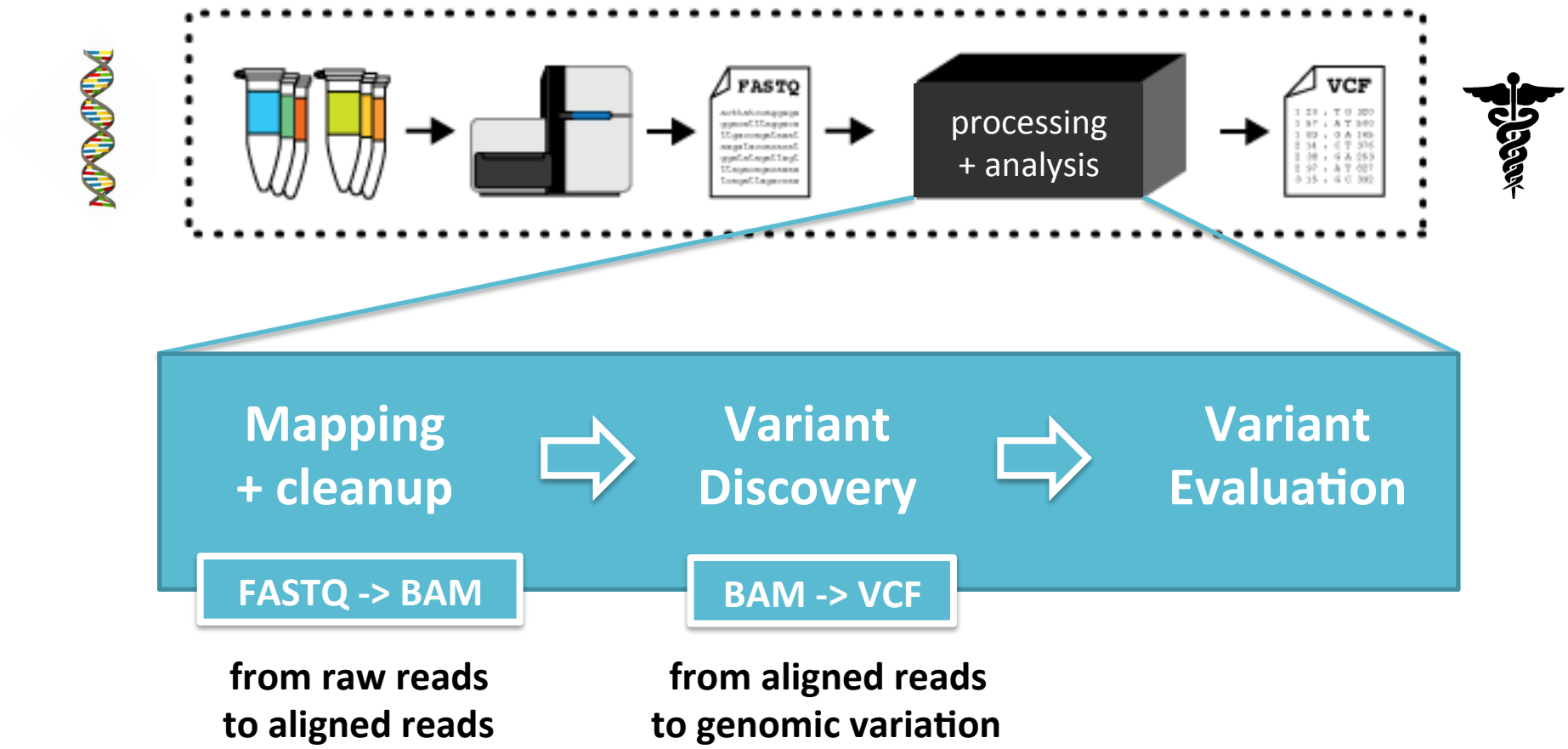
```
##fileformat=VCFv4.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership">
##FILTER=<ID=s50,Description="Less than 50% of 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">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO
    FORMAT      NA00001        NA00002        NA00003

20     14370   rs6054257 G       A       29   PASS   DP=14;AF=0.5;DB
    GT:GQ:DP 0/0:48:1 1/0:48:8 1/1:43:5
20     1110696 rs6040355 A       G,T     67   PASS   DP=10;AF=0.333,0.667;DB
    GT:GQ:DP 1/2:21:6 2/1:2:0   2/2:35:4
20     1230237 .         T       .       47   PASS   DP=13
    GT:GQ:DP 0/0:54:7 0/0:48:4 0/0:61:2
20     1234567 microsat1 GTCT    G,GTACT 50   PASS   DP=9
    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Header

Variant records

Official specification in
www.1000genomes.org/wiki/Analysis/Variant Call Format/vcf-variant-call-format-version-42

And that's all you need to know to get started



Further reading

http://www.broadinstitute.org/gatk/guide/best-practices