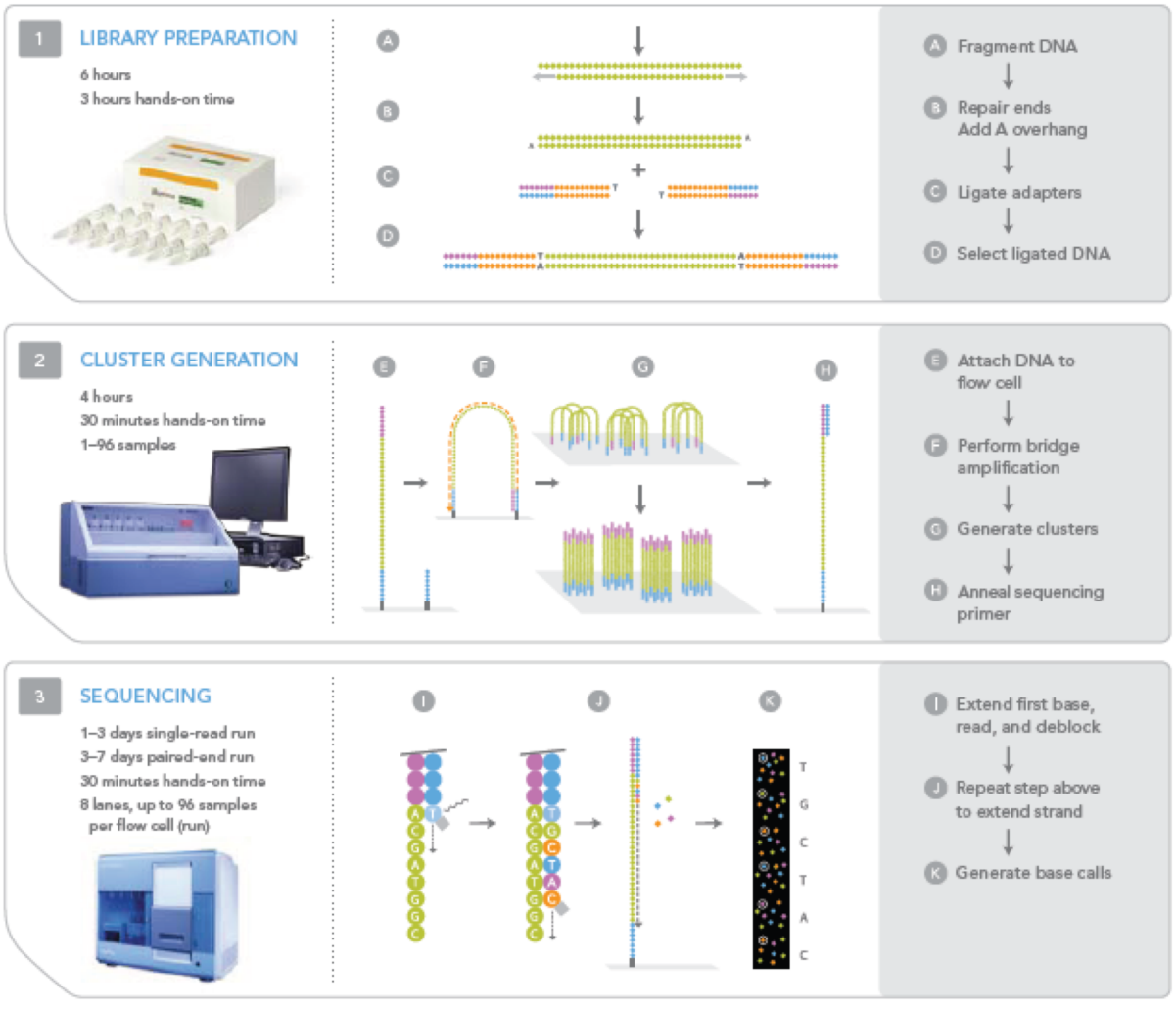
Sequencing data formats

Stephanie Gogarten

Sequencing process



Blood samples (or other tissue)

↓

Fragment DNA

↓

Replicate DNA fragments

↓

Read sequence

Sequence data

- Sequencing center produces “reads” - strings of base pairs
- Align reads to a reference genome
  - note this can have different “builds”, or versions
- Reads that have been aligned to the reference are stored in SAM files

Types of variants - SNP/SNV, INDEL

SNP = Single Nucleotide Polymorphism (usually common)

SNV = Single Nucleotide Variant (can be rare!)

```
Reference      ACTGACGCATGCATCATGCATGC
SNP            ACTGACGCATGCATCATTCATGC
```

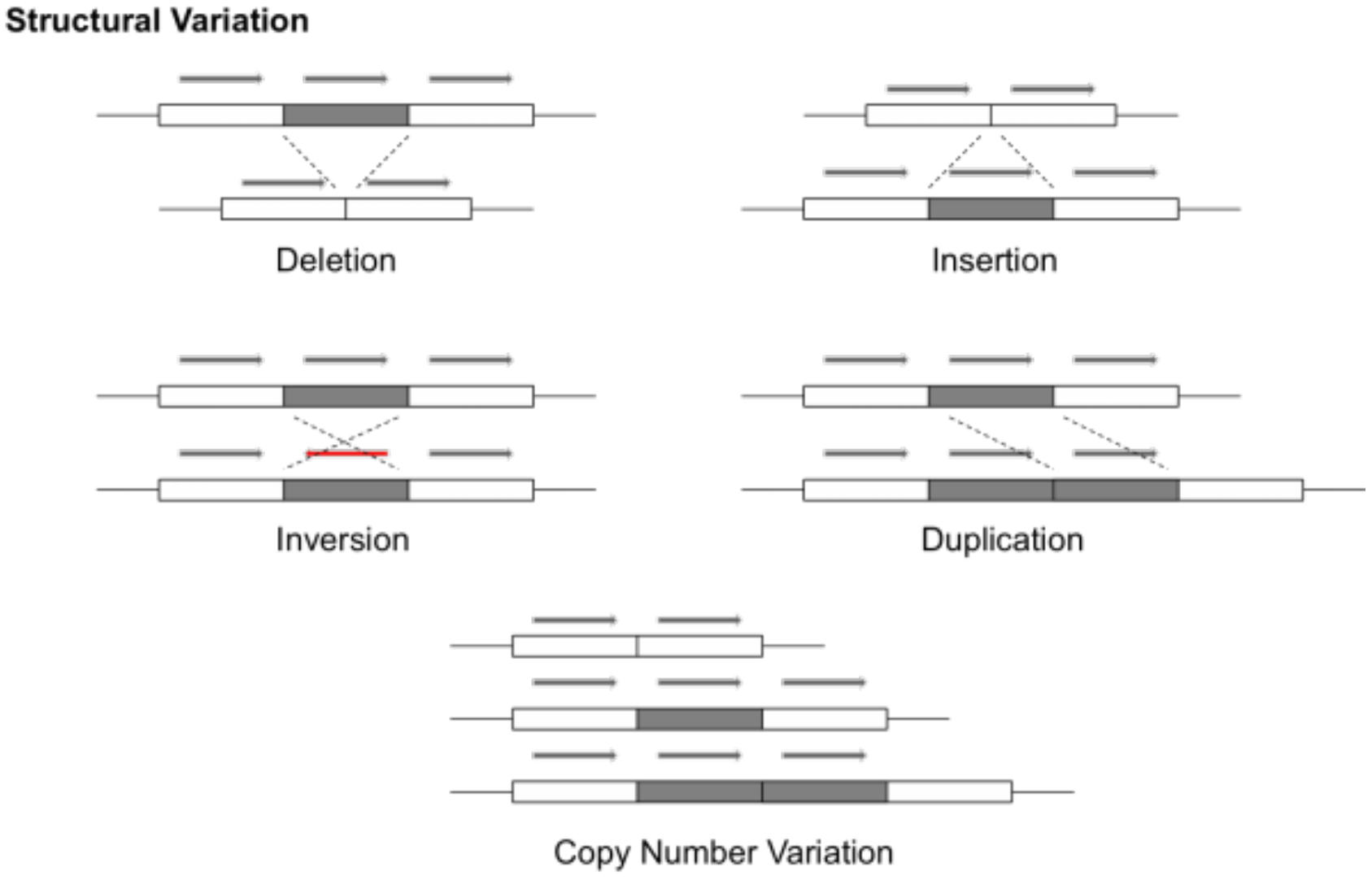
INDEL = INsertion or DELetion

```
Reference      ACTGACGCATGCATCATGCATGC
Insertion      ACTGACGCATGGTACATCATGCATGC
Deletion       ACTGACG--TGCATCATGCATGC
```

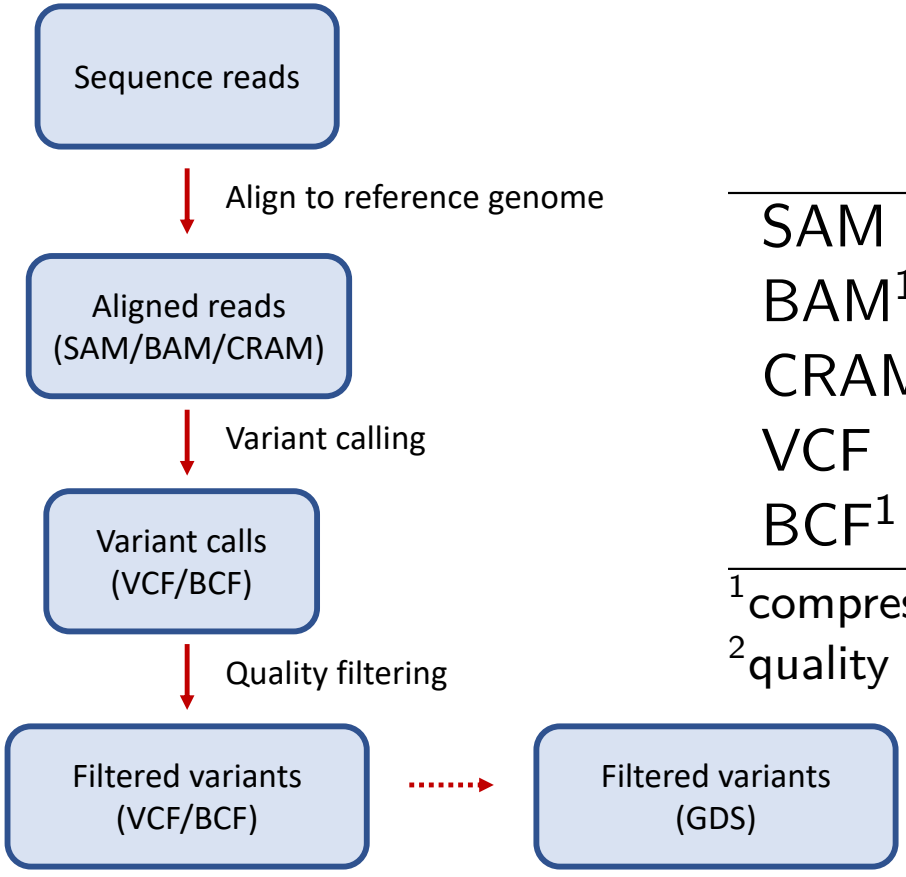
Indel

Types of variants - structural

Structural variants can be complex, and are much harder to call!



File formats



Same data, different format

SAM	Sequence Alignment Map
BAM[1]	Binary Alignment Map
CRAM[2]	
VCF	Variant Call Format
BCF[1]	Binary Call Format

[1] compressed with bgzip

[2] quality scores binned

GDS: Genomic Data Storage

Variant Call Format

VCF text file:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample1	Sample2
20	14370	rs6054257	G	A	29	PASS	NS=3	GT:GQ	0|0: 48	1|0: 48
20	17330	.	T	A	3	q10	NS=3	GT:GQ	0|0: 49	0|1: 03
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;AA=T	GT:GQ	1|2: 21	2|1: 02
20	1230237	.	T	.	47	PASS	NS=3;AA=T	GT:GQ	0|0: 54	0|0: 48
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;AA=G	GT:GQ	0/1: 35	0/2: 17

- VCF consists of a header section and a data section
- each variant is stored in a line
- genotypes
  - GT: alleles + phasing states
- annotations in INFO and FORMAT fields
  - NS (number of samples)
  - AA (ancestral allele)
  - GQ (genotype quality)
- Tabix indexing allows faster access to subsets of variants

Further resources

- Work directly with SAM/BAM files:
  - samtools: http://www.htslib.org/doc/samtools.html
- Work directly with VCF/BCF files:
  - bcftools: http://www.htslib.org/doc/bcftools.html
  - vcftools: https://vcftools.github.io/
- Compression and indexing:
  - bgzip: http://www.htslib.org/doc/bgzip.html
  - tabix: http://www.htslib.org/doc/tabix.html
- More details
  - https://genome.sph.umich.edu/wiki/SAM
  - https://en.wikipedia.org/wiki/SAM_(file_format)
  - https://en.wikipedia.org/wiki/Variant_Call_Format
  - https://samtools.github.io/hts-specs/VCFv4.2.pdf