HG471/BIOS21216: NGS Technology and Variant Calling

Prof. Xin He

March 19, 2016

Lecture Overview

Sequencing technologies: Sanger sequencing and next generation sequencing (NGS).

NGS workflow: alignment, variant calling, postprocessing (filtering/QC).

Applications of NGS in sequencing studies.

Reasons of Doing Sequencing Studies

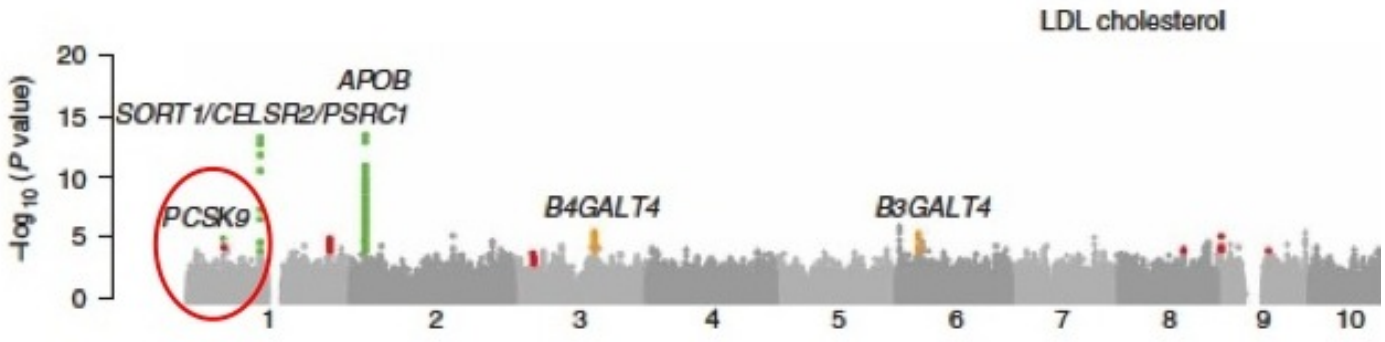
Comparing with GWAS, sequencing studies can directly identify causal variants.

Rare variants make significant contribution to genetics of Mendelian and complex diseases.

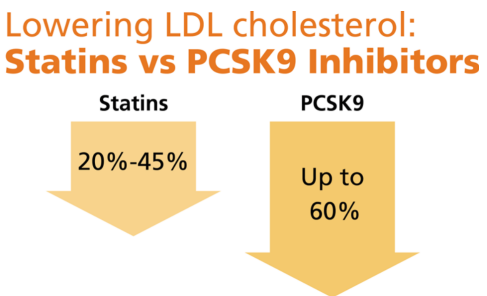
- Variants with large effects tend to be under purifying selection, thus are present in low frequencies.
- A special case: *de novo* mutations - impossible to study with GWAS.

Detection of **structural variations** would be much more complete with sequencing data.

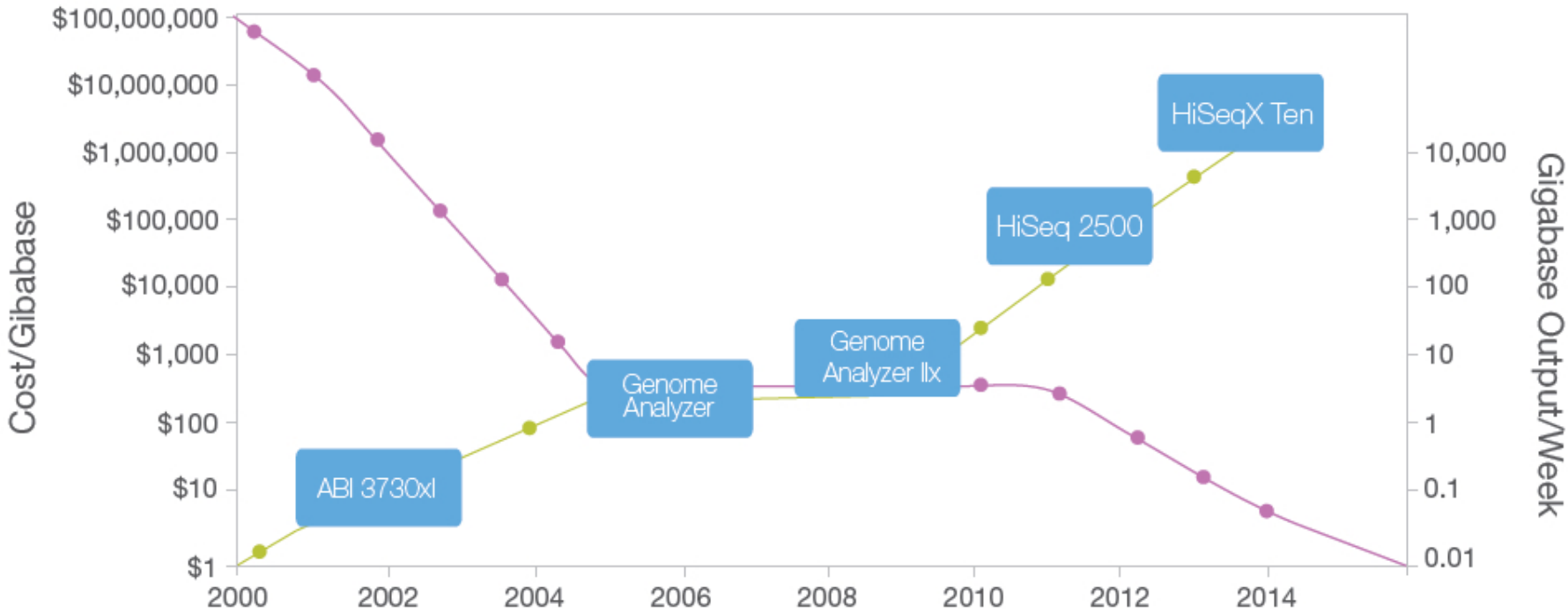
Multiple Variants in PCSK9 Reduce LDL



- Common variants (MAF 20%) in PCSK9 change LDL by ∼3 mg/dl (Willer et al, 2008)
- Rare variants (MAF 1%) in PCSK9 can change LDL by ∼16 mg/dl (Cohen et al, 2005)
- Private mutations in PCSK9 change LDL by $>$ 100 mg/dl (Abifadel et al, 2003)



DNA Sequencing Revolution

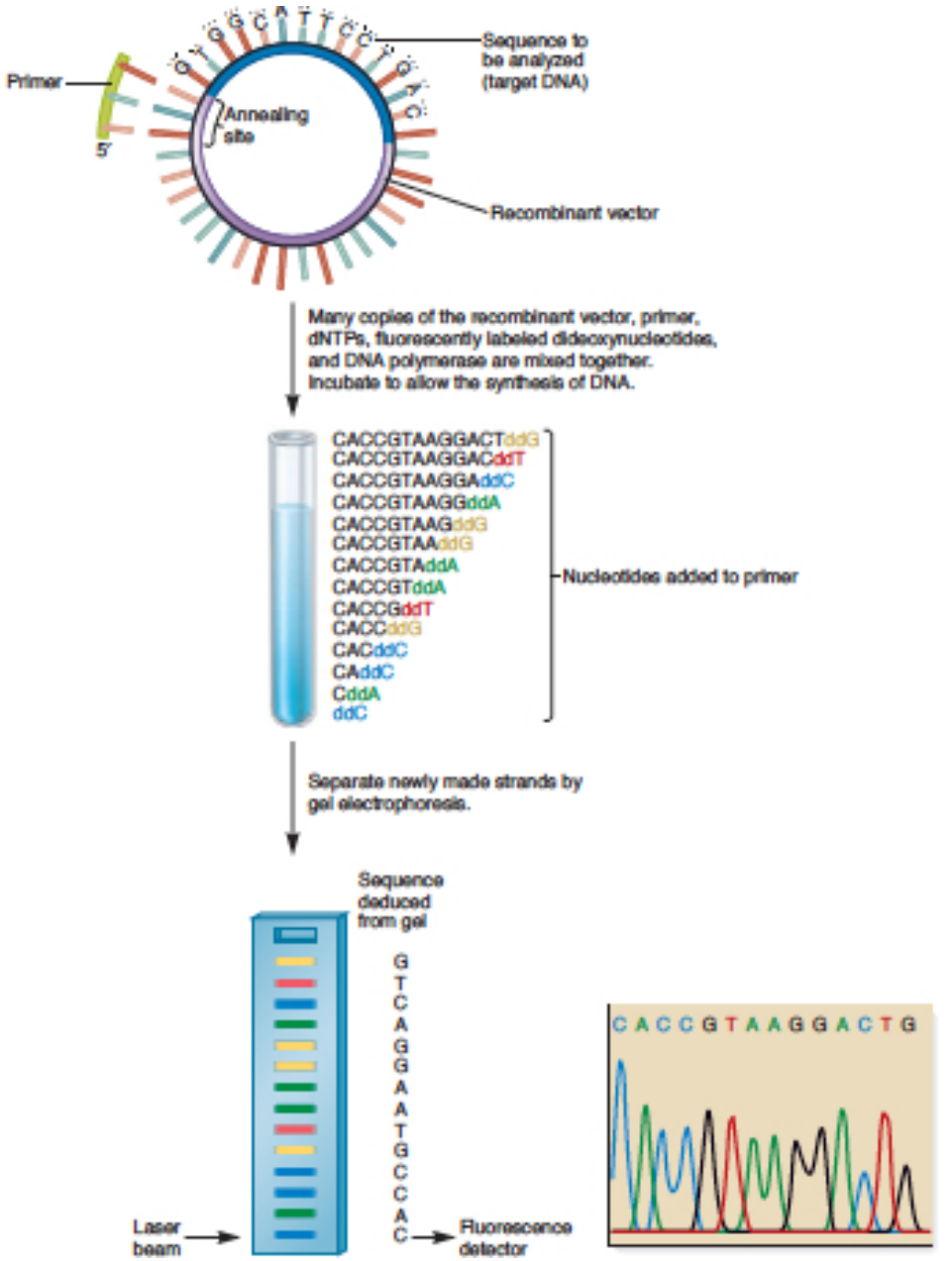


Background: Sanger Sequencing

Use synthetic nucleotides, dideoxyribonucleotides (ddNTP): if a ddNTP is added to a growing strand of DNA, DNA can no longer be replicated. This is called **chain termination**.

To sequence a DNA, add mixture of normal nucleotides and low concentration of ddNTPs. The ddNTPs are **fluorescently labeled**: each type of ddNTP with a different color.

Sanger sequencing: Figure 18.19 of Brooker, Genetics: Analysis & Principles, 4ed.


Primer
5′
Annealing site
Sequence to be analyzed (target DNA)
Recombinant vector
Many copies of the recombinant vector, primer, dNTPs, fluorescently labeled dideoxynucleotides, and DNA polymerase are mixed together. Incubate to allow the synthesis of DNA.
CACCGTAAGGACTddG
CACCGTAAGGACddT
CACCGTAAGGAddC
CACCGTAAGGddA
CACCGTAAGddG
CACCGTAAddG
CACCGTAddA
CACCGTddA
CACCGddT
CACCddG
CACddC
CAddC
CddA
ddC
Nucleotides added to primer
Separate newly made strands by gel electrophoresis.
Sequence deduced from gel
G T C A G G A A T G C C A C
Laser beam
Fluorescence detector
C A C C G T A A G G A C T G

Next Generation Sequencing by Illumina

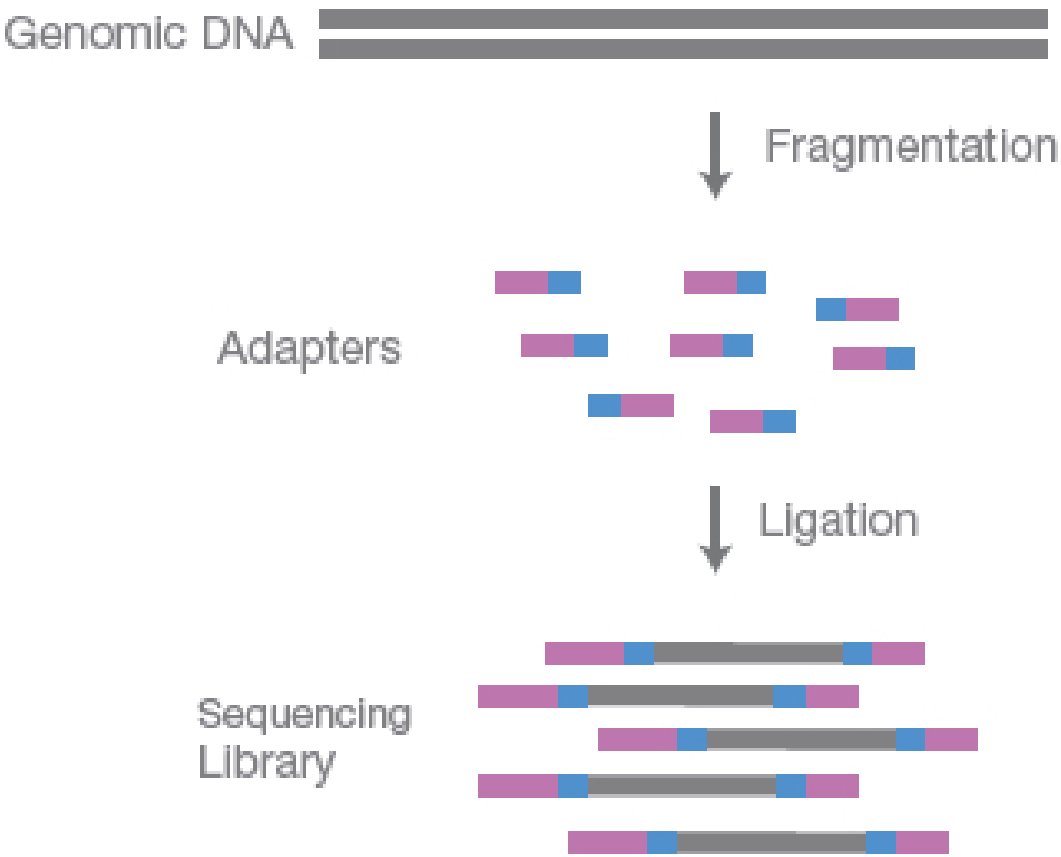
Multiple sequencing platforms are available such as Illumina, Ion Torrent, Complete Genomics, PacBio, and so on. Illumina is currently the dominant one: low sequencing error (about 1% error) and high throughput ($>$ 100Gb data per run).

Key concepts of Illumina sequencing:

- **Fragmentation** of a genome: into 200-300 bp fragments.
- **Spatially separate** these fragments at a slide.
- **Sequence by synthesis**: read each fragment in a way similar to Sanger sequencing.

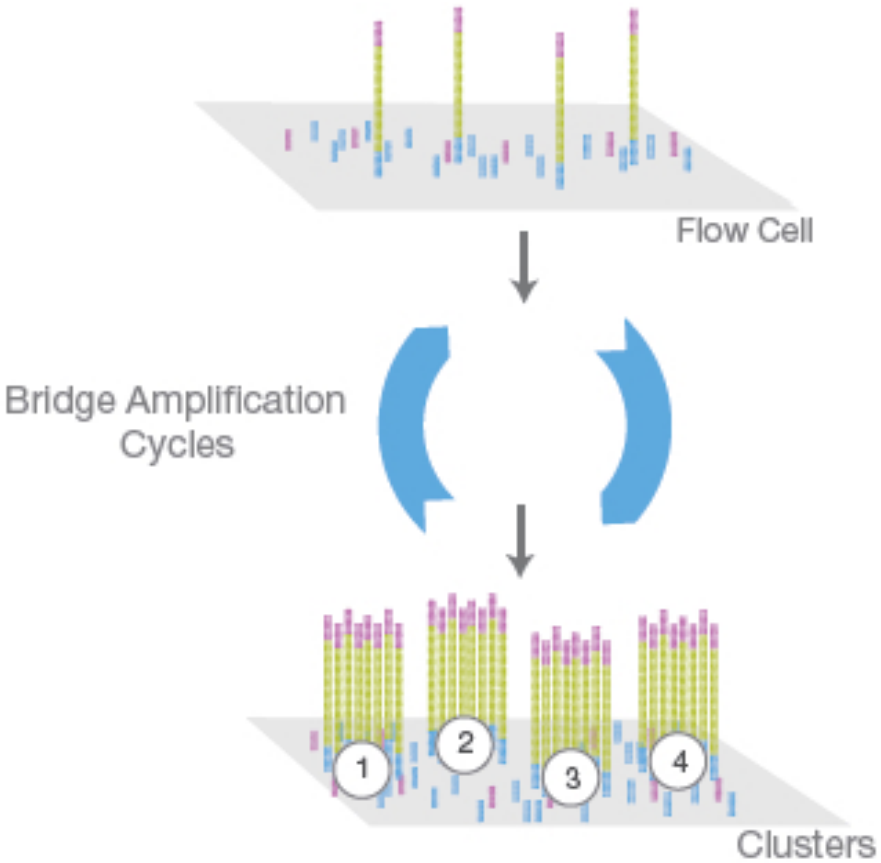
Reference: Illumina, An Introduction to Next-Generation Sequencing Technology.

Step 1. Library Preparation



NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

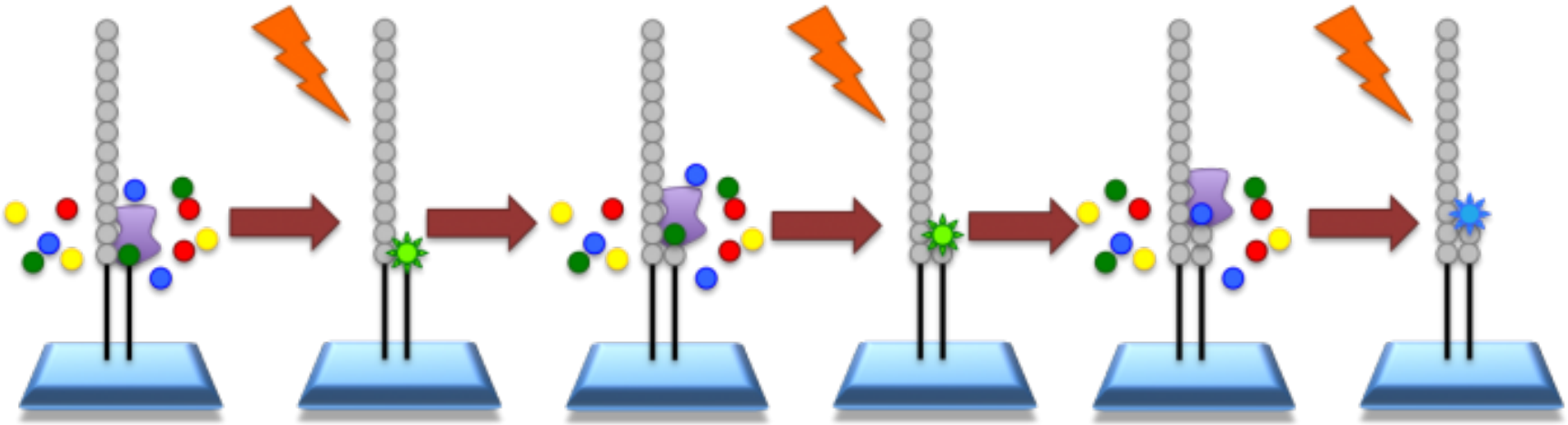
Step 2. Cluster Amplification



Library is loaded into a flow cell and the fragments hybridize to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

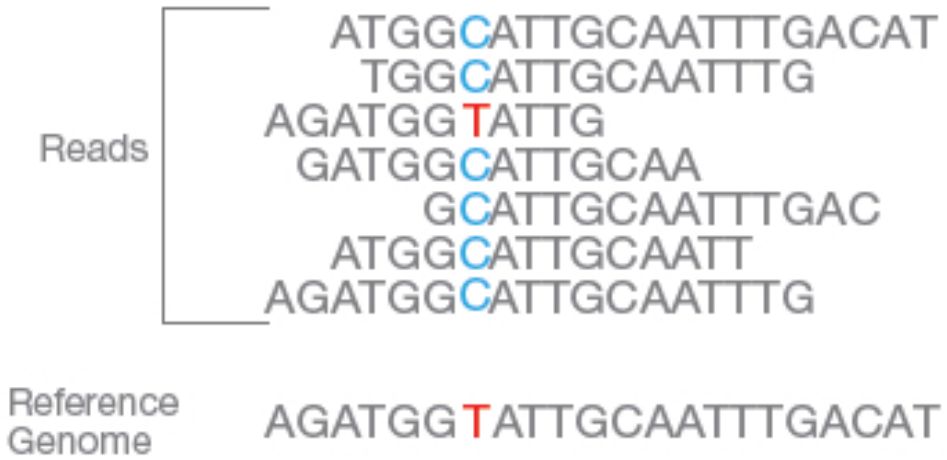
See video: `https://www.youtube.com/watch?v=HMyCqWhwB8E`

Step 3. Sequencing by Synthesis



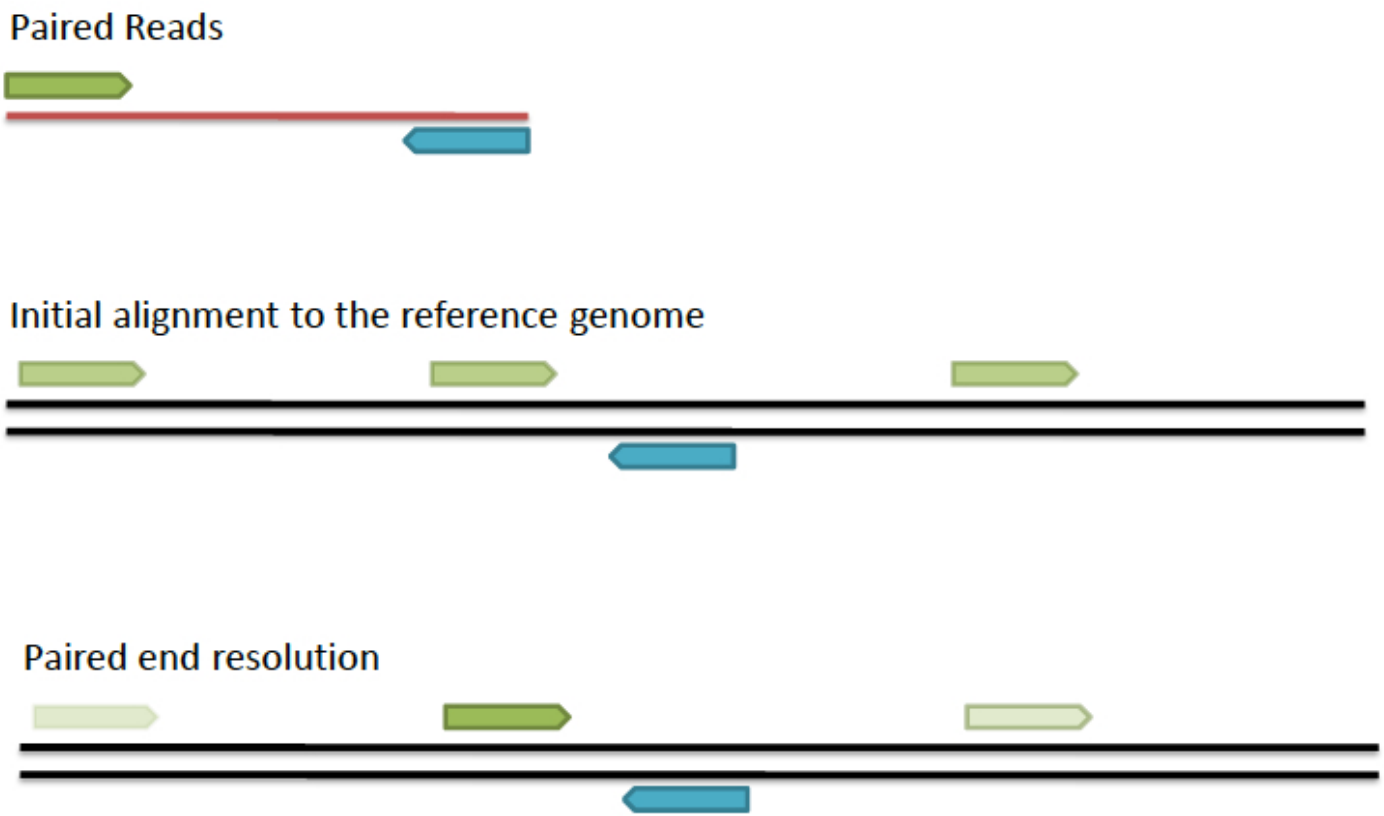
In each cycle, only one base (ddNTP with a distinct color) is added, then an image is taken. In next cycle, the terminators are removed, and the same process repeats.

Step 4. Alignment and Data Analysis



Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

Paired End Sequencing



Paired end sequencing enables both ends of a DNA fragment to be sequenced. The distance between each paired read is known, facilitating read alignment.

NGS Applications

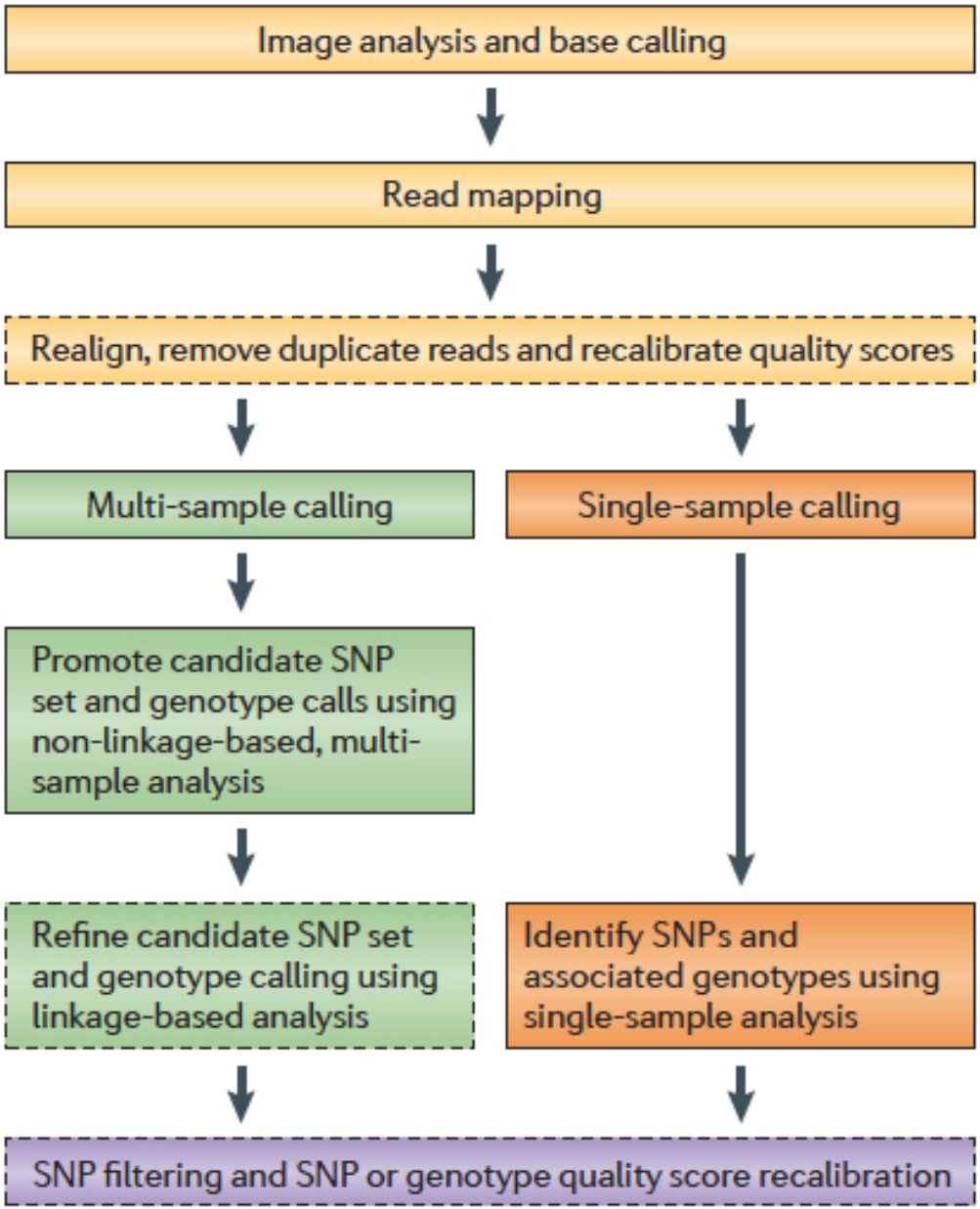
Genomics:

- Whole genome sequencing.
- Whole exome sequencing and targeted sequencing.

Transcriptomics and epigenomics:

- RNA-seq: gene expression.
- Methylation sequencing.
- ChIP-seq: transcription factor binding sites and histone modification.

From Sequencing Data to Variant Calls



Raw reads: FASTQ files

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTT
+ OPTIONAL
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC6
```

- Sequence identifier
- Sequence of raw bases
- Optional comment line (often repetition of sequence identifier)
- Sequence of ASCII characters representing Phred-scaled base quality scores

Phred Scores

Base quality score: estimate the error via analysis of raw sequence images and measure the quality using Phred quality score:

$$Q_{\text{base}} = -10 \log_{10}(P(\text{error}))$$

A base quality score of 30 encodes an error probability of 0.001. $Q > 20$ is often used in filters for considering a base.

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
|                         |    |        |
0.2.....................26...31.....41

L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

ASCII encoding of Phred scores.

Read Mapping to a Reference Genome

Sequence alignment: arranging the sequences to identify regions of similarity.

```
      |      *
AGGC  AAGGGTGCAGT  sample DNA
AGGCAAAGGTTGCAGT   reference
```

Read Mapping to a Reference Genome

Sequence alignment: arranging the sequences to identify regions of similarity.

```
      |     *
AGGC  AAGGGTGCAGT  sample DNA
AGGCAAAGGTTGCAGT   reference
```

Alignment programs:

- Hasing based: MAQ
- Burroughs-Wheeler Transform: Bowtie, BWA.

Alignment quality score: similar to base quality score, Phred-scale.

$$Q_{\text{map}} = -10 \log_{10}(P(\text{read is wrongly mapped}))$$

Output: BAM/SAM files.

Errors in Read Alignment

Common source of errors for variant analysis.

Repeat regions are often difficult to align.

Polymorphism, especially indels, in the sequenced DNA (relative to reference genome), can lead to alignment errors.

Data Preprocessing for Variant Calling

Realignment: initial alignment is based on the reference genome. We can infer the haplotype of the sample genome, then realign reads into the sample genome.

Discard **duplicate reads**: often due to amplification errors.

Base quality score recalibration: the base quality scores do not reflect true error rates, so need recalibration.

Base Quality Score Recalibration

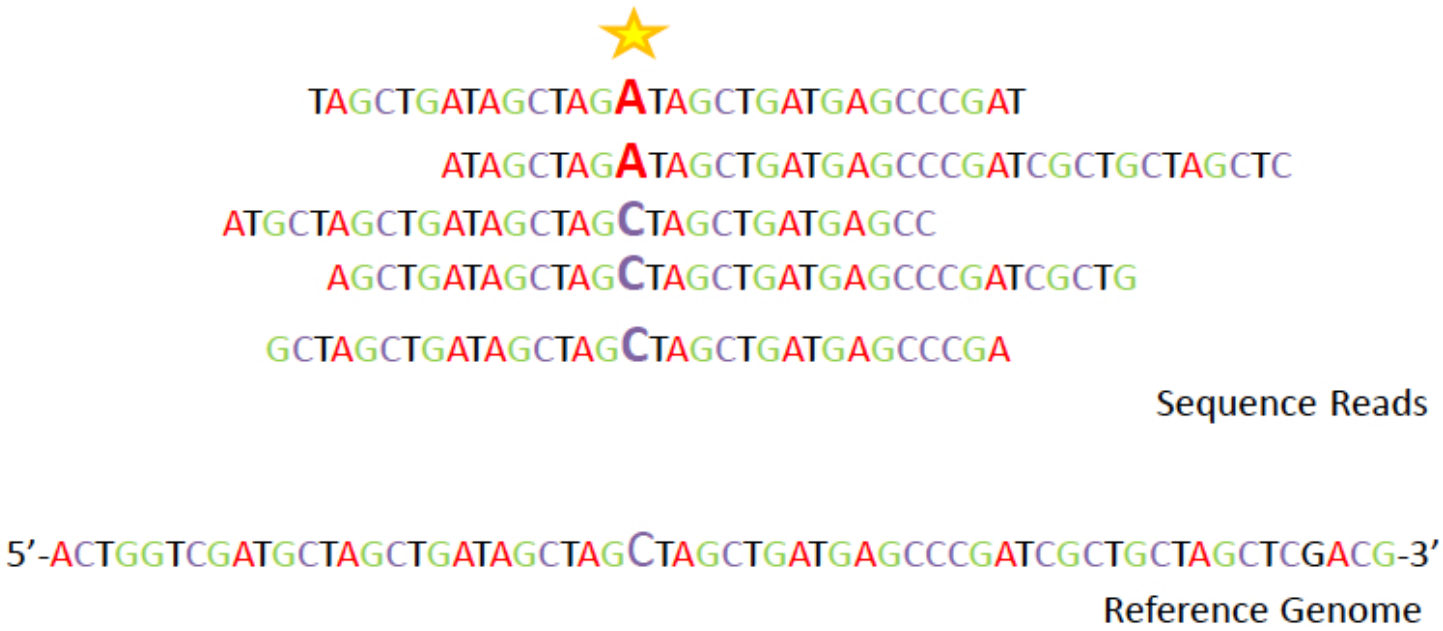
Use training data where no variants should be found. Any mismatches in these regions are due to base calling error.

Procedure for quality score recalibraton (GATK):

- **Categorize** all bases by their quality score (R), machine cycle in the read (C) and dinucleotide context (D).
- For each category, compute the **empirical error rate**: number of mismatches / number of bases.
- **Adjust** quality scores to match the actual error rates.

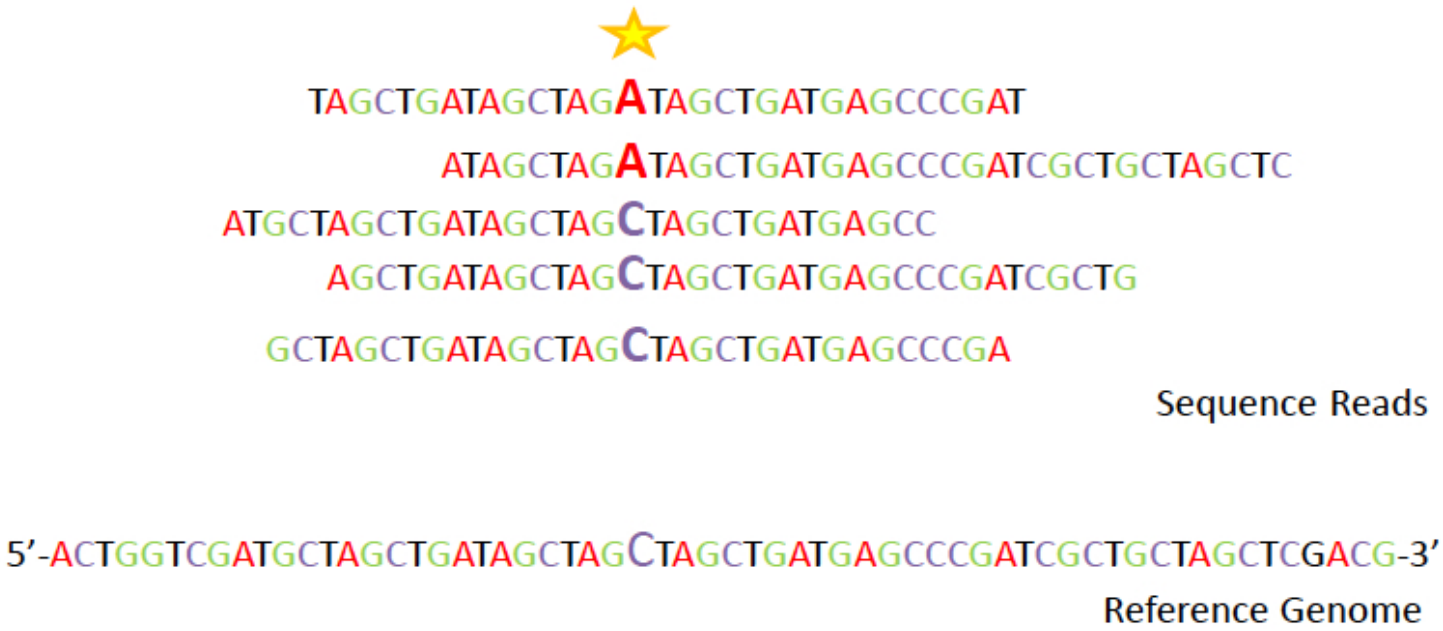
Genotype Calling and Variant Calling

Genotype calling: determine the genotype of an individual.



Genotype Calling and Variant Calling

Genotype calling: determine the genotype of an individual.



Variant calling: determine if any position differs from the reference allele in at least one individual.

Probabilistic Model for Genotype Calling

Let D be the reads of an individual and G be the genotype (to be determined), use Bayes Theorem, the **posterior probability** of G:

$$P(G|D) = \frac{P(D|G)P(G)}{\sum_G P(D|G)P(G)} \propto P(D|G)P(G)$$

Probabilistic Model for Genotype Calling

Let D be the reads of an individual and G be the genotype (to be determined), use Bayes Theorem, the **posterior probability** of G:

$$P(G|D) = \frac{P(D|G)P(G)}{\sum_G P(D|G)P(G)} \propto P(D|G)P(G)$$

Genotype likelihood $P(D|G)$: let D_j be data of read j of an indivdiaul, assuming independence of reads

$$P(D|G) = \prod_j P(D_j|G)$$

Prior of genotype $P(G)$: incorporate information of allele frequencies and LD.

Genotype Likelihood

For any read j of an individual: suppose $G = H_1 H_2$ (two haplotypes)

$$P(D_j|G) = \frac{1}{2}\left[P(D_j|H_1) + P(D_j|H_2)\right]$$

Suppose B is the base at H_1 or H_2, and ϵ_j the error rate (computed from minimum of base and read map quality scores)

$$P(D_j|B) = \begin{cases} 1-\epsilon_j & \text{if } D_j = B \\ \epsilon_j & \text{otherwise} \end{cases}$$

Example of Genotype Likelihood

P(reads|A/A, read mapped)= 0.00000098

P(reads|A/C, read mapped)= 0.03125

P(reads|C/C, read mapped)= 0.000097

Individual Prior of Genotypes

Most sites do not vary: $P(\text{non-reference base}) \sim 0.001$.

When a site does vary, it is usually heterozygous

- $P(\text{non-reference heterozygote}) \sim 0.001 * 2/3$.
- $P(\text{non-reference homozygote}) \sim 0.001 * 1/3$.

Mutation model: transitions account for most variants (C $\leftrightarrow$ T or A $\leftrightarrow$ G).

Single Sample Genotype Calling

P(reads|A/A)= 0.00000098	**Prior(A/A) =** 0.00034	**Posterior(A/A) =** <.001
P(reads|A/C)= 0.03125	**Prior(A/C) =** 0.00066	**Posterior(A/C) =** 0.175
P(reads|C/C)= 0.000097	**Prior(C/C) =** 0.99900	**Posterior(C/C) =** 0.825

Population Prior of Genotypes

Use allele frequency information of the population or other samples. Ex. frequency of A is 0.2.

P(reads|A/A)= 0.00000098	**Prior(A/A)** = 0.04	**Posterior(A/A)** = <.001
P(reads|A/C)= 0.03125	**Prior(A/C)** = 0.32	**Posterior(A/C)** = 0.999
P(reads|C/C)= 0.000097	**Prior(C/C)** = 0.64	**Posterior(C/C)** = <.001

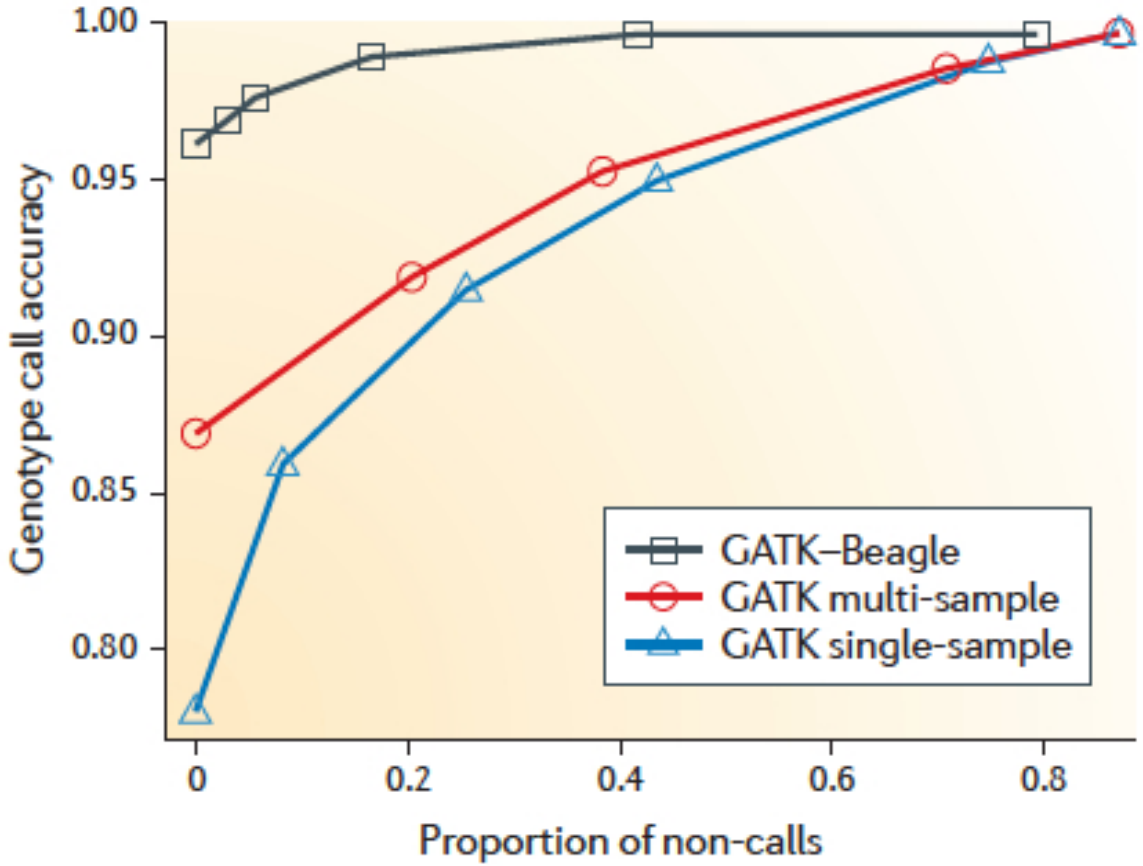
Haplotype-based Prior of Genotypes

Use imputation analysis: compares individuals with similar flanking haplotypes to infer genotypes.

Can make accurate genotype calls with 2-4x coverage of the genome.

Works most effectively at high or moderate-frequency variants.

Performance of Genotype Calling Methods



Variant Calling (GATK)

A position contains a variant if at least one alternative allele is present across all samples.

Let $q_i = \{0, 1, 2\}$ be the number of alternative alleles in sample i, and $q = \sum_i q_i$. Our goal is to infer $P(q|D)$, where D is data of all samples: a position contains a variant if $q > 0$.

$$P(q|D) = \frac{P(q = X)P(D|q = X)}{\sum_Y P(q = Y)P(D|q = Y)}.$$

Variant Call Quality Score: $Q_{\text{variant}} = -10 \log_{10}(P(q = 0|D))$

Prior distribution $P(q)$: this is the site frequency spectrum (SFS), and GATK uses a prior based on infinite-site model of population genetics. Let $\theta = 4N_e\mu$ (a constant about 6×10^{-4}):

$$P(q = X) = \theta/X \qquad \text{for } X > 0$$

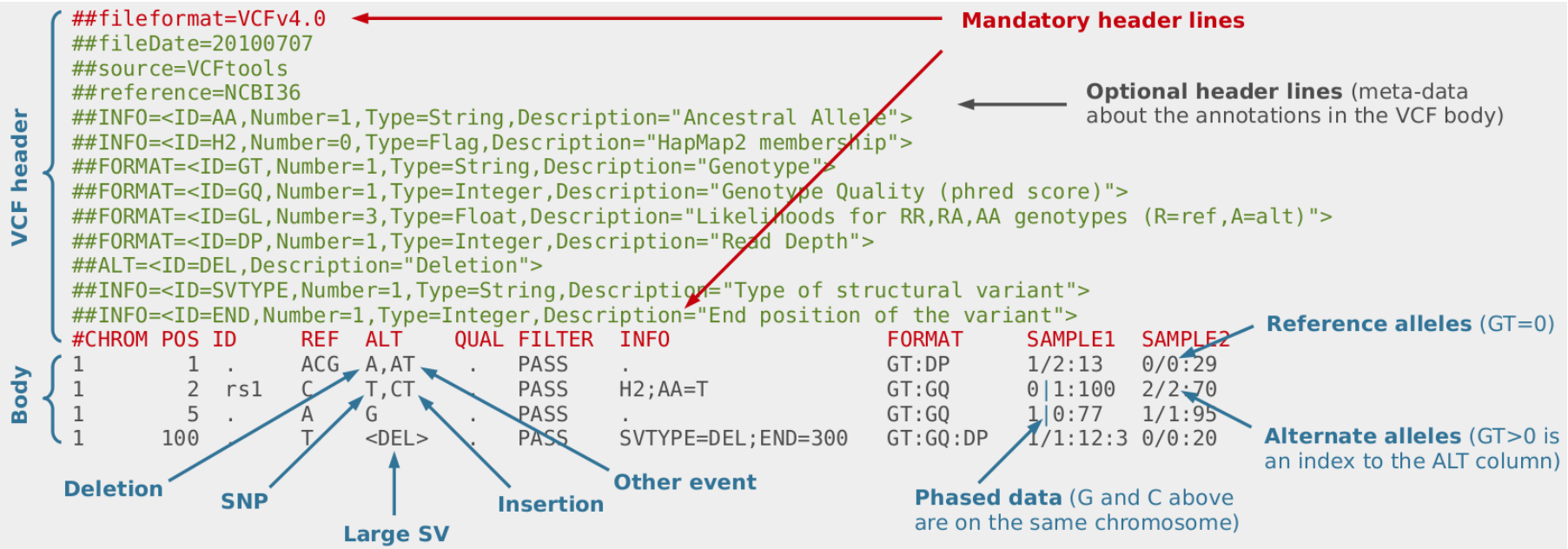
Likelihood function of q:

$$P(D|q = X) = \sum_{G \in \Gamma} \prod_i P(D_i|G_i)$$

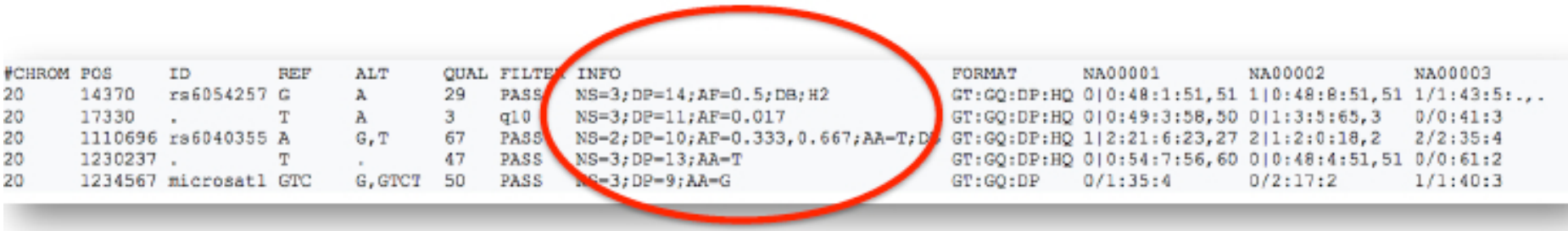
where G_i is the genotype of sample i and Γ is the space of all genotypes such that $\sum_i G_i = X$.

Results of Variant and Genotype Calling: VCF

Variant Call Format (VCF): standard format for storing sequence variations.



VCF Format



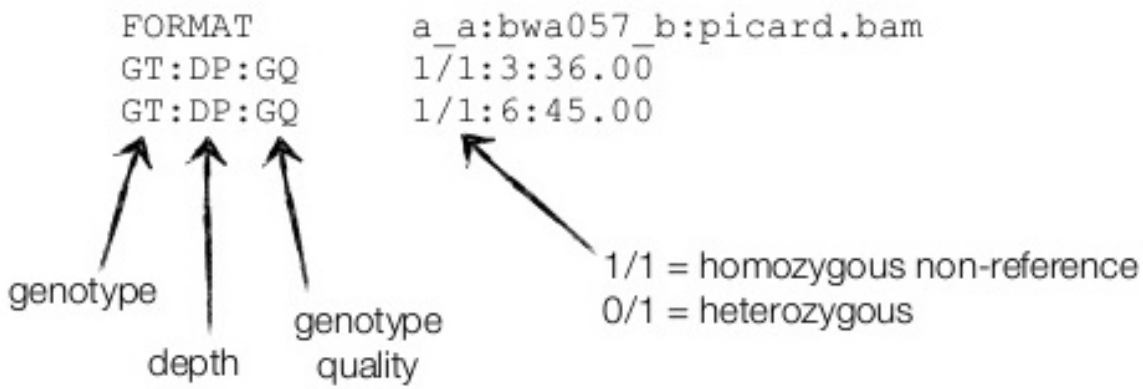
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	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

INFO: additional information of variants. Optional. Some predefined options:

- DP: combined depth across all the samples
- NS: number of samples with data
- AF: estimated allele frequency
- SB: strand bias at this position

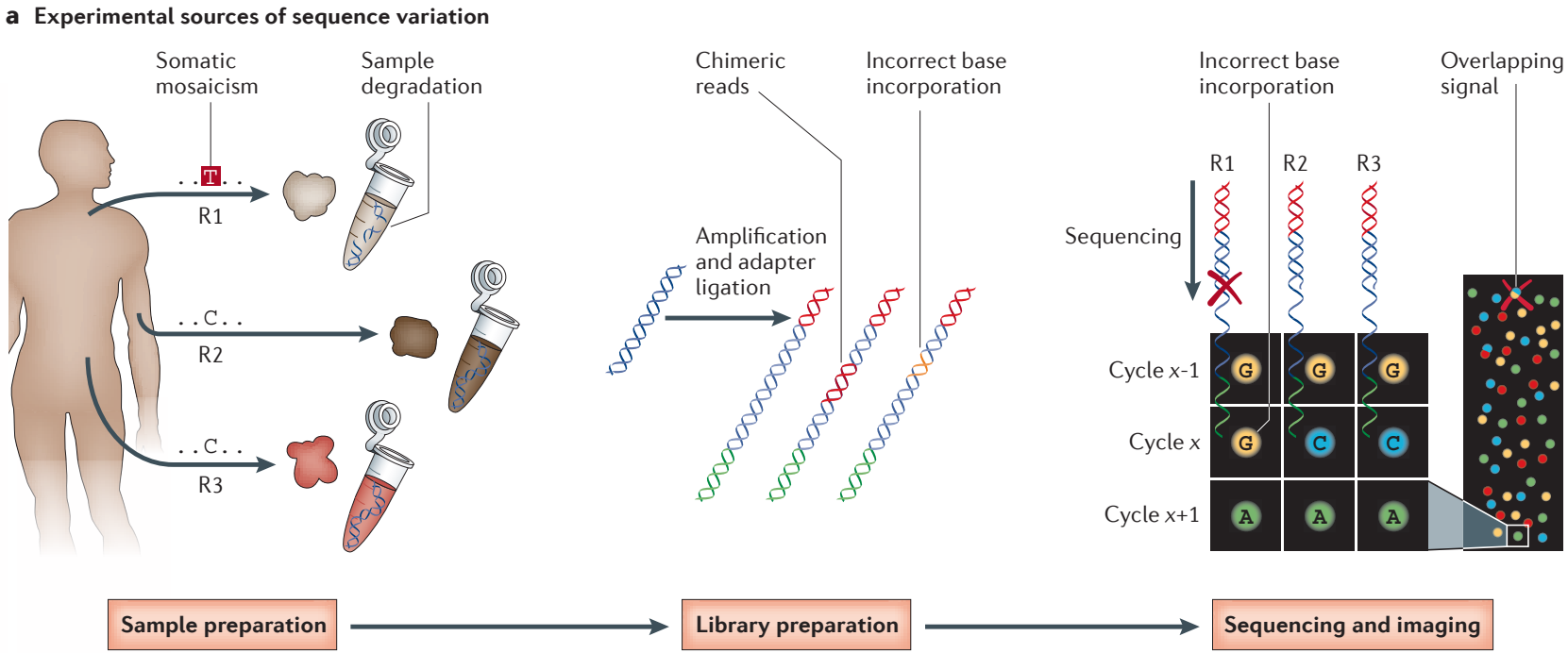
VCF Format

Genotype data: format defined in the FORMAT field.



Experimental Sources of Sequence Errors

Robasky et al, The role of replicates for error mitigation in next-generation sequencing, *Nature Reviews Genetics*, 2014



Filtering and Quality Control

Remove reads with low base calling or mapping quality scores.

Remove genotypes or variants with low read depth, strand bias, allelic imbalance, or low quality scores.

Remove samples with unusually high heterozygosity (excess of variants).

Other QC metrics:

- Concordance among replicates: biological, technical and cross-platform replicates.
- Mendelian incompatibilities (e.g. in trio data).
- Transition to transversion ratio: close to 2:1.

Main Applications of NGS in Genetics

Discover rare variants of Mendelian Diseases (in particular, those that are refractory to linkage).

Causal variants of undiagonised childhood diseases.

Rare variants in common complex diseases, often in case-control settings.

Considerations of Designing Sequencing Studies

Determine sequence coverage (average no. of reads mapped to a base): optimize specificity-sensitivity trade off by the type of study. Ex. given 125Gb of 2×100 data,

- Single sample (e.g. clinic diagnosis): high coverage, 40x recommended.
- Low frequency variants in common diseases: prefer to have low coverage in many samples. 3-5x recommended.
- *De novo* mutations: 60x of all samples in a family.

Considerations of Designing Sequencing Studies

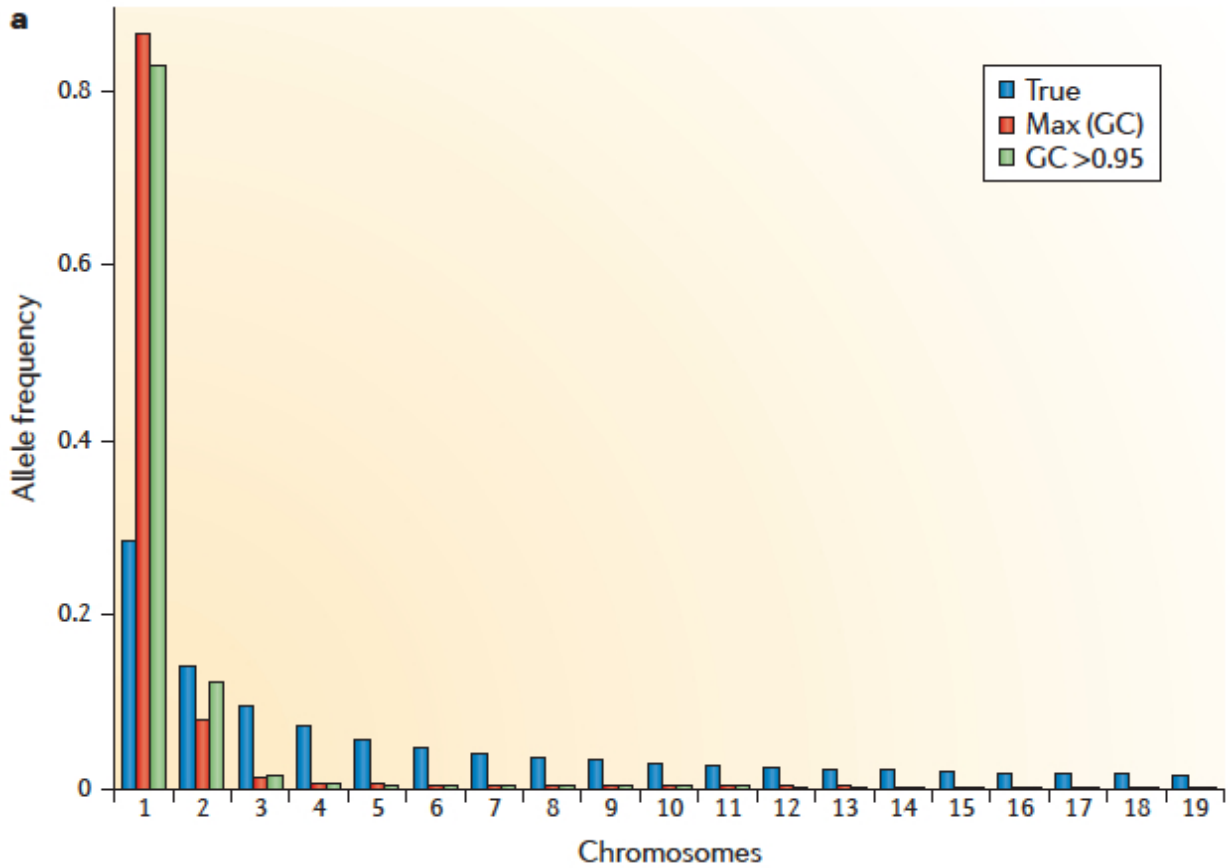
Determine sequence coverage (average no. of reads mapped to a base): optimize specificity-sensitivity trade off by the type of study. Ex. given 125Gb of 2×100 data,

- Single sample (e.g. clinic diagnosis): high coverage, 40x recommended.
- Low frequency variants in common diseases: prefer to have low coverage in many samples. 3-5x recommended.
- *De novo* mutations: 60x of all samples in a family.

Be aware of batch effects when combining data from multiple sources (especially for case-control studies): e.g. data from different platforms have different error profiles.

Ignoring Genotype Uncertainty May Introduce Bias

Estimating SFS in 50 individuals: using best genotype calls (GC) or only genotypes with posterior probability > 0.95 leads to an excess of singletons.



Summary

NGS: Illumina Sequencing by Synthesis (SBS) technology.

Base calling, read mapping, preprocessing (realignment, duplicate reads, base score recalibration).

Genotype and variant calling: Bayesian approach with population (AF) prior and LD prior.

Filtering and QC important for NGS data.

Applications of NGS: practical considerations in coverage, possible bias and batch effect.

Recommended Readings

- Nielsen et al (2011) Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*. 12(6):443
- DePristo et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nat Genet.* 43(5):491

Optional Readings

- Goldstein et al (2013) Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet.* 14(7):460
- An Introduction to Next-Generation Sequencing Technology: `http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf`