

# CS-E5875 High-Throughput Bioinformatics

## Variant calling and de novo assembly

Harri Lähdesmäki

Department of Computer Science  
Aalto University

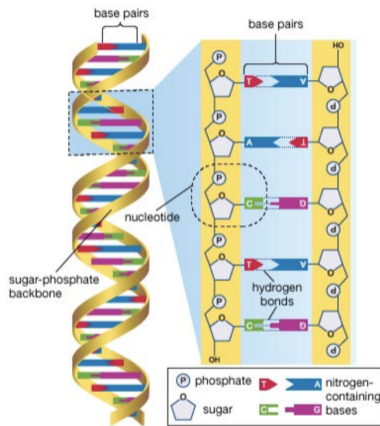
October 31, 2023

# Contents

- ▶ Genotype calling from HTS data
- ▶ Detecting somatic mutations from HTS data
- ▶ De novo assembly

# Human genome

- ▶ DNA is a double-stranded molecule with each strand being a linear sequence of nucleotides
- ▶ A nucleotide consists of a phosphate group, sugar, and nucleoside
- ▶ A nucleoside is a nitrogenous base connected to a deoxyribose sugar
- ▶ There are four different nucleotides (depending on the nucleoside): adenine (A), cytosine (C), guanine (G), thymine (T)
- ▶ The nucleotides have a specific base pairing in double-stranded DNA:
  - ▶ Adenine pairs w/ thymine
  - ▶ Cytosine pairs w/ guanine
- ▶ Total length: about 3 billion nucleotides



© 2007 Encyclopædia Britannica, Inc.

Figure from Wikipedia

# Types of human genome variation

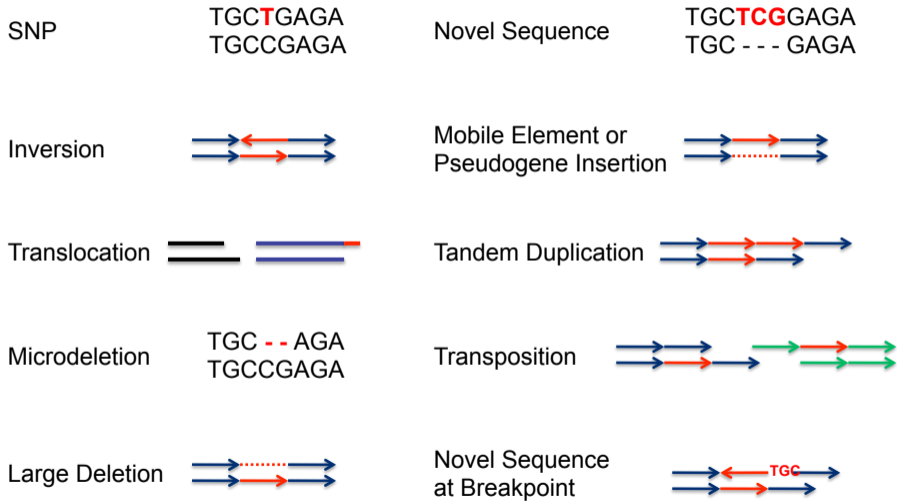


Figure from <https://web.stanford.edu/class/cs262/presentations/lecture4.pdf>

# Single-nucleotide polymorphism

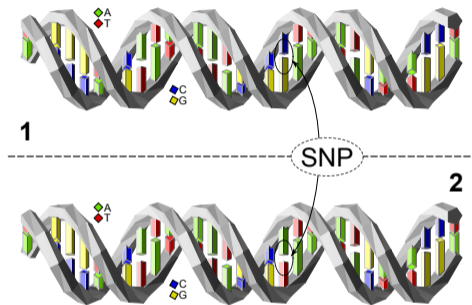
- ▶ Consider a specific nucleotide (chromosome and genomic coordinate) in human genome
- ▶ Most individuals have the same nucleotide at that position
- ▶ However, some individuals can have a different nucleotide at that position
- ▶ This nucleotide difference is called a genetic variant
- ▶ Different nucleotides at that variant position are called alleles
- ▶ There exist biallelic and multiallelic variants
  - ▶ Biallelic: a position in a genome can contain two different nucleotides
  - ▶ Multiallelic: a position in a genome can contain more than two different nucleotides
  - ▶ Much of the literature/published GWAS results focus on biallelic variants

# Single-nucleotide polymorphism

- ▶ Consider a specific nucleotide (chromosome and genomic coordinate) in human genome
- ▶ Most individuals have the same nucleotide at that position
- ▶ However, some individuals can have a different nucleotide at that position
- ▶ This nucleotide difference is called a genetic variant
- ▶ Different nucleotides at that variant position are called alleles
- ▶ There exist biallelic and multiallelic variants
  - ▶ Biallelic: a position in a genome can contain two different nucleotides
  - ▶ Multiallelic: a position in a genome can contain more than two different nucleotides
  - ▶ Much of the literature/published GWAS results focus on biallelic variants
- ▶ Minor allele is defined to be the allele that occurs with a lower frequency
- ▶ Variants with a minor allele frequency (MAF) of at least 5% are typically called common single-nucleotide polymorphisms (SNPs)
- ▶ Variants with MAF between 0.5% and 5% are called as low-frequency variants

# Single-nucleotide polymorphism

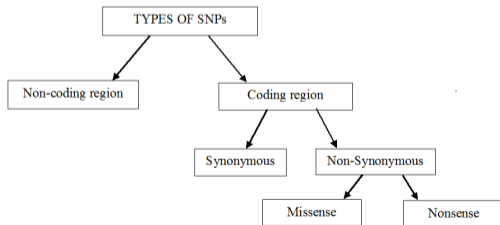
- ▶ An illustration of a SNP



Figures from [https://en.wikipedia.org/wiki/Single-nucleotide\\_polymorphism](https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism)

# Types of single-nucleotide polymorphism

- ▶ Non-coding SNP is located in a region of a genome that does not code for a protein
- ▶ Coding SNP is located in a region of a genome that codes for a protein
- ▶ Synonymous SNP does not change the amino acid sequence that is produced
- ▶ Nonsynonymous SNP changes the amino acid sequence that is produced
- ▶ Missense SNP causes a substitution of a different amino acid in the final amino acid sequence
- ▶ Nonsense SNP causes a premature stop codon / truncated protein amino acid sequence / non-functional protein



Figures from [https://en.wikipedia.org/wiki/Single-nucleotide\\_polymorphism](https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism)



# Genotype

- ▶ The genotype of a diploid individual at a single genomic variant position is the combination of the two alleles in the two chromosome copies
- ▶ Denote the two alleles of a biallelic variant by A and B
- ▶ Note that both A and B can take values in  $\{A,C,G,T\}$ 
  - ▶ The possible genotypes for the variant are then A/A, A/B and B/B
  - ▶ A/A: no mutation
  - ▶ A/B: heterozygous mutation
  - ▶ B/B: homozygous mutation
- ▶ For example: if the possible alleles of a biallelic SNP are  $A = C$  and  $B = T$ , then the possible genotypes are
  - ▶ A/A: C/C
  - ▶ A/B: C/T
  - ▶ B/B: T/T
- ▶ SNPs are the primary source of genetic differences between individuals

# Genotype calling

- ▶ Assume we have measured short DNA sequencing reads from a large number of cells for several individuals
- ▶ Having aligned the short sequencing reads of all individuals to a reference genome
  - ▶ SNP calling identifies variable sites (using sequencing reads from all individuals)
  - ▶ Genotype calling determines the genotype for each individual separately at each site (using sequencing reads from a single individual separately)
  - ▶ Genotype calling is typically only done for positions in which a SNP variant has already been called

Reference		AGTTTGACTCCAAACTGTAACGTAAGCTTAGCTACTACT
		TGACTCCAAACTGTAATGTA
		ACTCCAAACTCTAATGTAAG
		ACTCCAAACTGTAAGGTAAG
		CCAACTCTCTAATGTAAGCTT
		CCTAAGCTT
Aligned reads		AACTCTAATGTGAGCTTAGC
		ACTGTAATGTAAGCGTAGCT
		CTCTAATGTAAGCATAGCTA
		CTCTAATGTAAGCTTAGCTA
		GTAATGTGAGCTTAGCTACT
Genotype		-----C--T--A-----
		-----G--T--A-----

# Genotype calling

## Challenges in SNP and genotype calling

- ▶ A mismatch in an aligned read can be due to
  - ▶ A true SNP
  - ▶ An error while generating the sequencing library
  - ▶ Base calling error
  - ▶ Misalignment
  - ▶ Mistakes done earlier while building the reference sequence
- ▶ Many NGS studies rely on low-coverage sequencing (e.g.  $<5\times$ ), i.e., on average  $<5$  sequencing reads per site (nucleotide position) per individual
  - ▶ A high probability that only one of the two chromosomes of a diploid individual has been sampled / sequenced at a specified site

A probabilistic framework: so-called “genotype likelihoods” which incorporate errors that may have been introduced in base calling, alignment and assembly are coupled with prior information, such as allele frequencies and patterns of linkage disequilibrium (LD)

# GATK: a simple Bayesian genotyper

- ▶ Genotyping with GATK tool (McKenna et al, 2010)
- ▶ GATK computes the posterior probability of each genotype, given
  - ▶ the pileup of aligned reads that cover a given locus (i.e., a nucleotide location)
  - ▶ expected heterozygosity of the sample
- ▶ Define:
  - ▶  $G$  is the genotype
  - ▶  $D$  represents the data (pileup of the aligned reads at a given position)
  - ▶  $P(G)$  is a prior probability of seeing this genotype (in a given population)

## GATK: a simple Bayesian genotyper

- ▶ Genotyping with GATK tool (McKenna et al, 2010)
- ▶ GATK computes the posterior probability of each genotype, given
  - ▶ the pileup of aligned reads that cover a given locus (i.e., a nucleotide location)
  - ▶ expected heterozygosity of the sample
- ▶ Define:
  - ▶  $G$  is the genotype
  - ▶  $D$  represents the data (pileup of the aligned reads at a given position)
  - ▶  $P(G)$  is a prior probability of seeing this genotype (in a given population)
- ▶ The basic model is then (recall the Bayes' theorem)

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

# GATK: likelihood model

Reference		AGTTTGACTCCAAACTGTAACGTAAGCTTAGCTACTACT
		TGACTCCAAACTGTAA <b>T</b> GTA
		ACTCCAAACT <b>C</b> TAA <b>T</b> GTAAG
		ACTCCAAACTGTAA <b>G</b> GTAAG
		CCAA <b>T</b> CT <b>C</b> TAA <b>T</b> GTAAGCTT
		CCAAACTGTAA <b>T</b> GTAAGCTT
Aligned reads		AACT <b>C</b> TAA <b>T</b> GT <b>G</b> AGCTTAGC
		ACTGTAA <b>T</b> GTAAGC <b>G</b> TAGCT
		CT <b>C</b> TAA <b>T</b> GTAAGC <b>A</b> TAGCTA
		CT <b>C</b> TAA <b>T</b> GTAAGCTTAGCTA
		GTAA <b>T</b> GT <b>G</b> AGCTTAGCTACT
Genotype		----- <b>C</b> --- <b>T</b> --- <b>A</b> -----
		----- <b>G</b> --- <b>T</b> --- <b>A</b> -----

- ▶ The likelihood can be written as a product over the independent aligned reads

$$P(D|G) = \prod_{b_i \in \text{pileup}} P(b_i|G),$$

where  $b_i$  ( $i = 1, \dots, d$ ) represents the nucleotide base in the  $i$ th read covering the locus

# GATK: likelihood model

Reference		AGTTTGACTCCAAACTGTAACGTAAGCTTAGCTACTACT
		TGACTCCAAACTGTAA <b>T</b> GTA
		ACTCCAAACT <b>C</b> TAA <b>T</b> GTAAG
		ACTCCAAACTGTAA <b>G</b> GTAAG
		CCAA <b>T</b> CT <b>C</b> TAA <b>T</b> GTAAGCTT
		CCAAACTGTAA <b>T</b> GTAAGCTT
Aligned reads		AACT <b>C</b> TAA <b>T</b> GT <b>G</b> AGCTTAGC
		ACTGTAA <b>T</b> GTAAGC <b>G</b> TAGCT
		CT <b>C</b> TAA <b>T</b> GTAAGC <b>A</b> TAGCTA
		CT <b>C</b> TAA <b>T</b> GTAAGCTTAGCTA
		GTAA <b>T</b> GT <b>G</b> AGCTTAGCTACT
Genotype		----- <b>C</b> --- <b>T</b> --- <b>A</b> -----
		----- <b>G</b> --- <b>T</b> --- <b>A</b> -----

- ▶ For each position, decompose the genotype into its two alleles as  $G = (A_1, A_2)$
- ▶ The probability of a base given the genotype is defined as

$$P(b_i|G) = P(b_i|(A_1, A_2)) = \frac{1}{2}P(b_i|A_1) + \frac{1}{2}P(b_i|A_2),$$

because  $b_i$  can come from either of the chromosome copies with equal probability

# GATK: likelihood model

Reference		AGTTTGACTCCAAACTGTAACGTAAGCTTAGCTACTACT
		TGACTCCAAACTGTAA <b>T</b> GTA
		ACTCCAAACT <b>C</b> TAA <b>T</b> GTAAG
		ACTCCAAACTGTAA <b>G</b> GTAAG
		CCAA <b>T</b> CT <b>C</b> TAA <b>T</b> GTAAGCTT
		CCAAACTGTAA <b>T</b> GTAAGCTT
Aligned reads		AACT <b>C</b> TAA <b>T</b> GT <b>G</b> AGCTTAGC
		ACTGTAA <b>T</b> GTAAG <b>C</b> GTAGCT
		CT <b>C</b> TAA <b>T</b> GTAAG <b>C</b> ATAGCTA
		CT <b>C</b> TAA <b>T</b> GTAAGCTTAGCTA
		GTAA <b>T</b> GT <b>G</b> AGCTTAGCTACT
Genotype		----- <b>C</b> --- <b>T</b> --- <b>A</b> -----
		----- <b>G</b> --- <b>T</b> --- <b>A</b> -----

- ▶ Finally, the probability of seeing a base given an allele is

$$P(b_i|A) = \begin{cases} \frac{e_i}{3}, & \text{if } b_i \neq A \\ 1 - e_i, & \text{else} \end{cases},$$

where  $e_i = 10^{-\frac{q_i}{10}}$  is the reversed phred scaled quality score at the base in the  $i$ th read



# GATK: genotype selection

Reference		AGTTTGACTCCAAACTGTAACGTAAGCTTAGCTACTACT
		TGACTCCAAACTGTAA <b>T</b> GTA
		ACTCCAAACT <b>C</b> TAA <b>T</b> GTAAG
		ACTCCAAACTGTAA <b>G</b> GTAAG
		CCAA <b>T</b> CT <b>C</b> TAA <b>T</b> GTAAGCTT
		CCAAACTGTAA <b>T</b> GTAAGCTT
Aligned reads		AACT <b>C</b> TAA <b>T</b> GT <b>G</b> AGCTTAGC
		ACTGTAA <b>T</b> GTAAG <b>C</b> GTAGCT
		CT <b>C</b> TAA <b>T</b> GTAAG <b>C</b> ATAGCTA
		CT <b>C</b> TAA <b>T</b> GTAAGCTTAGCTA
		GTAA <b>T</b> GT <b>G</b> AGCTTAGCTACT
Genotype		----- <b>C</b> --- <b>T</b> --- <b>A</b> -----
		----- <b>G</b> --- <b>T</b> --- <b>A</b> -----

- ▶ The maximum a posteriori (MAP) estimate of the genotype is then

$$\hat{G} = \arg \max_G P(G|D)$$

- ▶ In other words, we choose  $G$  that gives the highest probability

$$P(G = (A, A) | D) \quad \text{or} \quad P(G = (A, B) | D) \quad \text{or} \quad P(G = (B, B) | D)$$

# Contents

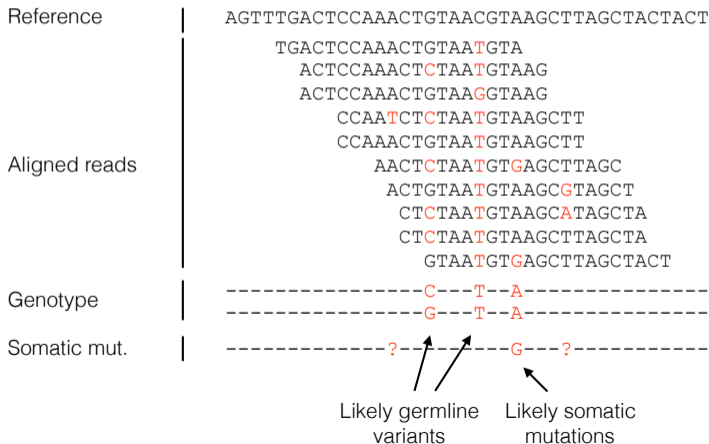
- ▶ Genotype calling from HTS data
- ▶ Detecting somatic mutations from HTS data
- ▶ De novo assembly

## Germline and somatic mutations

- ▶ Genetic variation represented by SNPs is inherited from parents and transmitted to offspring
  - ▶ SNPs are germline mutations
- ▶ In addition to SNPs, each individual will accumulate additional mutations during life time
  - ▶ These gained mutations are called somatic mutations (or gained mutations)
  - ▶ These mutations are generally not passed on to offspring through the germline
- ▶ Somatic mutations are involved in e.g. the development of many cancers

# Germline and somatic mutations

- ▶ Germline mutations that we have considered previously can be heterozygous or homozygous, i.e., appear with a frequency of 0%, 50% or 100%
- ▶ Somatic mutations can be present with any frequency (across a population of cells)



## Somatic mutation detection with Mutect

- ▶ Somatic mutation detection with Mutect (Cibulskis et al, 2013)
- ▶ Consider detecting a somatic mutation at a given position (chromosome and coordinate)
- ▶ Denote the reference allele as  $r \in \{A, C, G, T\}$
- ▶ Assume  $d$  aligned sequence reads overlap the position and denote
  - ▶  $b_i$  is the nucleobase called in the  $i$ th read ( $i \in \{1, \dots, d\}$ )
  - ▶  $e_i$  is the probability of error of the base called in the  $i$ th read

$$e_i = 10^{-\frac{q_i}{10}}$$

where  $q_i$  is the associated Phred quality score

## Mutect: two alternative models / hypotheses

- ▶ To detect a somatic mutation, try to explain the data at a specific position using two models:
  1. Model  $M_0$ : there is no somatic mutation at the given position and all non-reference nucleobases are explained by sequencing noise
  2. Model  $M_f^m$ : a variant allele  $m \neq r$  truly exists at the given position with an allele frequency  $f$  and reads are also subject to sequencing noise
- ▶ Note that  $M_0$  is equivalent to  $M_f^m$  with  $f = 0$  and  $m = r$

Reference		AGTTTGACTCCAAACTGTAACGTAAGCTTAGCTACTACT
		TGACTCCAAACTGTAA <b>T</b> GTA
		ACTCCAAACT <b>C</b> TAA <b>T</b> GTAAG
		ACTCCAAACTGTAAG <b>G</b> GTAAG
		CCAA <b>T</b> CT <b>C</b> TAA <b>T</b> GTAAGCTT
		CCAAACTGTAATGTAAGCTT
Aligned reads		AACT <b>C</b> TAA <b>T</b> GT <b>G</b> AGCTTAGC
		ACTGTAA <b>T</b> GTAAG <b>C</b> GTAGCT
		CT <b>C</b> TAA <b>T</b> GTAAG <b>C</b> ATAGCTA
		CT <b>C</b> TAA <b>T</b> GTAAGCTTAGCTA
		GTAA <b>T</b> GT <b>G</b> AGCTTAGCTACT
Genotype		----- <b>C</b> --- <b>T</b> --- <b>A</b> -----
		----- <b>G</b> --- <b>T</b> --- <b>A</b> -----

## Mutect: likelihood model

- ▶ The likelihood (or probability) for the model  $M_f^m$  is given by

$$P(D \mid M_f^m) = P(\{b_i\} \mid \{e_i\}, r, m, f) = \prod_{i=1}^d P(b_i \mid e_i, r, m, f)$$

assuming the sequencing errors are independent across reads

## Mutect: likelihood model

- ▶ If all substitution errors are equally likely and occur with probability  $e_i/3$ , then the likelihood is

$$P(b_i|e_i, r, m, f) = \begin{cases} f \cdot e_i/3 + (1 - f)(1 - e_i) & \text{if } b_i = r \\ f \cdot (1 - e_i) + (1 - f)e_i/3 & \text{if } b_i = m \\ e_i/3 & \text{otherwise} \end{cases}$$

- ▶ Recall the bulk sequencing experiment: a sequencing read  $b_i$  is measured from a DNA fragment that originates from a randomly selected cell in the biological input sample
  - ▶ Probability that  $b_i$  is from a cell that does not have the somatic mutation is  $1 - f$
  - ▶ Probability that  $b_i$  is from a cell that has the somatic mutation is  $f$
  - ▶ Probability that the nucleobase is measured correctly is  $1 - e_i$
  - ▶ Probability that  $b_i$  that is neither the reference nor the mutation (must be a sequencing error) is  $f \cdot e_i/3 + (1 - f) \cdot e_i/3 = e_i/3$



## Mutect: likelihood model

- ▶ The likelihood for the model  $M_0$ ,  $P(D | M_0)$ , can be computed similarly but by setting  $f = 0$  and  $m = r$ , where the likelihoods for the nucleobases are

$$P(b_i | e_i, m = r, f = 0) = \begin{cases} 1 - e_i & \text{if } b_i = m = r \\ e_i/3 & \text{otherwise} \end{cases}$$

Reference		AGTTTGACTCCAAACTGTAACGTAAGCTTAGCTACTACT
		TGACTCCAAACTGTAATGTA
		ACTCCAAACTCTAATGTAAG
		ACTCCAAACTGTAAGGTAAG
		CCAACTCTCTAATGTAAGCTT
		CCAAACTGTAATGTAAGCTT
Aligned reads		AACTCTAATGTGAGCTTAGC
		ACTGTAAATGTAAGCGTAGCT
		CTCTAATGTAAGCATAGCTA
		CTCTAATGTAAGCTTAGCTA
		GTAATGTGAGCTTAGCTACT
Genotype		-----C--T--A-----
		-----G--T--A-----

## Mutect: likelihood ratio

- ▶ Similarly as in genotyping, we can apply the Bayes rule to compute the probability of the models (or hypotheses)

$$P(M_f^m | D) \propto P(D | M_f^m)P(m, f)$$

$$P(M_0 | D) \propto P(D | M_0)P(m = r, f = 0)$$

where  $P(m, f)$  is a prior, i.e., the expected probability of a mutated nucleotide  $m$  and its frequency  $f$  for a given cancer type, and  $P(m = r, f = 0) = 1 - P(m, f \geq 0)$

## Mutect: likelihood ratio

- ▶ Similarly as in genotyping, we can apply the Bayes rule to compute the probability of the models (or hypotheses)

$$P(M_f^m | D) \propto P(D | M_f^m)P(m, f)$$

$$P(M_0 | D) \propto P(D | M_0)P(m = r, f = 0)$$

where  $P(m, f)$  is a prior, i.e., the expected probability of a mutated nucleotide  $m$  and its frequency  $f$  for a given cancer type, and  $P(m = r, f = 0) = 1 - P(m, f \geq 0)$

- ▶ Somatic mutation detection is performed by computing the likelihood ratio of the two models (or hypotheses),  $M_0$  and  $M_f^m$

$$\text{LOD}_T(m, f) = \log_{10} \frac{P(M_f^m | D)}{P(M_0 | D)} = \log_{10} \frac{P(D | M_f^m)P(m, f)}{P(D | M_0)(1 - P(m, f))}$$

## Mutect: likelihood ratio

- ▶ The key quantity  $P(M_f^m | D)$  involves two unknowns:
  - ▶ Mutated nucleotide  $m \neq r$
  - ▶ Variant frequency  $f$
- ▶ The unknown frequency  $f$  and mutated nucleotide  $m$  can be estimated using the maximum likelihood method to obtain  $\hat{f}_{ML}$  and  $\hat{m}_{ML}$

$$\hat{f}_{ML}, \hat{m}_{ML} = \arg \max_{f, m} P(M_f^m | D)$$

or set to plug-in estimates: e.g.  $m$  is chosen to be the most frequent non-reference allele and

$$\hat{f} = \frac{\text{the number of } m \text{ reads}}{\text{the total number of reads}}$$

## Mutect: likelihood ratio

- ▶ Somatic mutation is called if the above LOD score exceeds a certain significance level
- ▶ Note that the above LOD score corresponds to a likelihood ratio statistic

## Mutect: filtering false positives

The detected somatic mutations should be further filtered to avoid likely false positives

- ▶ Check that the detected variant is not a heterozygous germline SNP, i.e., test

$$\text{LOD}_N = \log_{10} \frac{P(M_0 | D)P(m, f)}{P(M_{0.5}^m | D)P(\text{"germline"})},$$

where frequency has been set to  $f = 0.5$  and terms have been reverted to avoid false positives

# Mutect: filtering false positives

- Filter other technical artifacts not accounted by the model

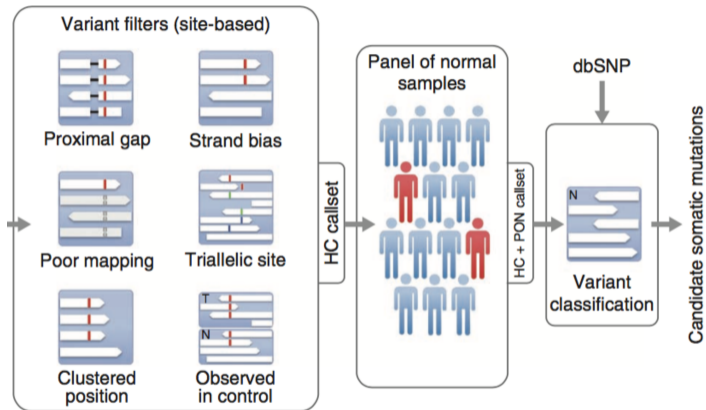


Figure from (Cibulskis et al, 2013)

# Mutect: performance

Sensitivity and specificity of detecting true somatic mutations depends on e.g.

- ▶ Sequencing depth, i.e., the number of reads overlapping a genomic position (on average)
- ▶ Variant frequency
- ▶ Abundance of sequencing errors
- ▶ etc.

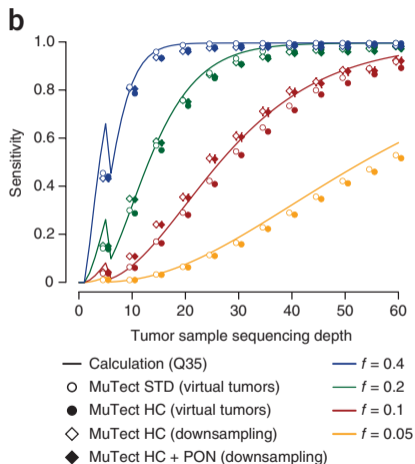


Figure from (Cibulskis et al, 2013)



# Types of human genome variation

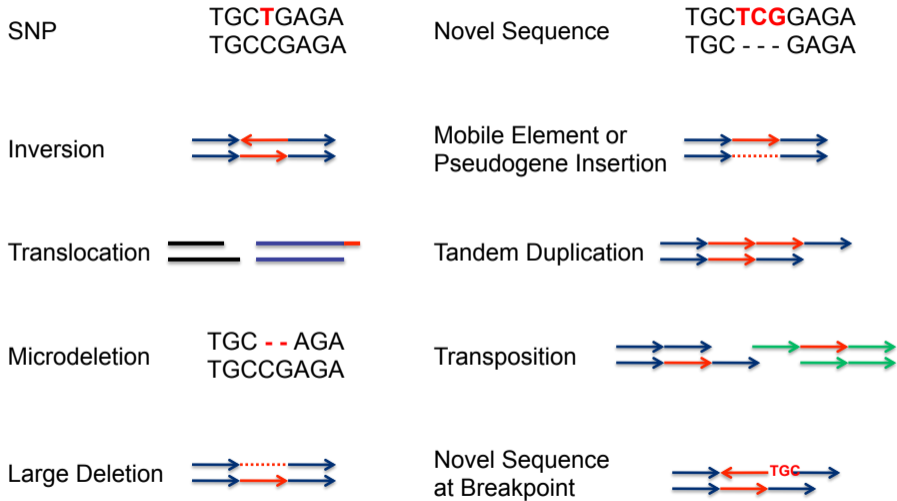


Figure from <https://web.stanford.edu/class/cs262/presentations/lecture4.pdf>

# Contents

- ▶ Genotype calling from HTS data
- ▶ Detecting somatic mutations from HTS data
- ▶ *De novo assembly*

## De novo genome assembly

- ▶ This section follows (Chaisson, et al, 2015)
- ▶ De novo: no reference genome available
- ▶ The goal of de novo genome assembly is to determine the sequence of a genome using only randomly sampled sequence fragments
  - ▶ Sequence fragments are typically less than one-millionth the size of a mammalian genome

## De novo genome assembly

- ▶ This section follows (Chaisson, et al, 2015)
- ▶ De novo: no reference genome available
- ▶ The goal of de novo genome assembly is to determine the sequence of a genome using only randomly sampled sequence fragments
  - ▶ Sequence fragments are typically less than one-millionth the size of a mammalian genome
- ▶ Most current approaches involve some aspect of a whole-genome shotgun sequencing and assembly (WGSA) strategy
  - ▶ Random fragments from a genome are sequenced and computationally stitched together to generate sequence contigs and scaffolds
- ▶ Under ideal conditions (i.e., uniformly high sequence coverage across the whole genome and a genome devoid of repetitive sequences), an assembly may be determined with the simple approach of merging reads with maximal overlap

## De novo genome assembly

- ▶ In practice such an approach does not work because:
  - ▶ Sequence coverage is almost never uniform
  - ▶ Genome contains repetitive sequences of varying length, and
  - ▶ Genome contains varying copy numbers (duplications)

# Types of genome assembly gaps

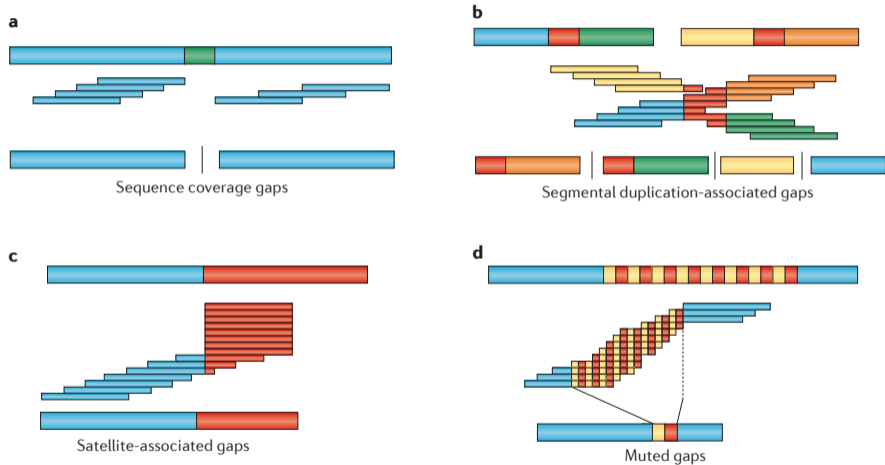


Figure from (Chaisson, et al, 2015)

## Early de novo assembly methods

- ▶ The most-widely used mammalian genomes, human and mouse, were not assembled using WGS
- ▶ Instead, human and mouse assemblies are relatively unique among mammalian genomes in that they were assembled almost entirely using clone-by-clone-based sequencing
  - ▶ Each genome/chromosome was divided into roughly 200-kb-long overlapping fragments that were cloned into bacterial artificial chromosomes (BACs) and individually assembled
  - ▶ These longer 200kb fragments were then connected

## Early de novo assembly methods

- ▶ The most-widely used mammalian genomes, human and mouse, were not assembled using WGS
- ▶ Instead, human and mouse assemblies are relatively unique among mammalian genomes in that they were assembled almost entirely using clone-by-clone-based sequencing
  - ▶ Each genome/chromosome was divided into roughly 200-kb-long overlapping fragments that were cloned into bacterial artificial chromosomes (BACs) and individually assembled
  - ▶ These longer 200kb fragments were then connected
- ▶ When the result of a de novo assembly is a sequence per chromosome without gaps and with 99.99% base-pair accuracy, the assembly is considered complete; otherwise, it is considered a draft.
  - ▶ Even a recent build of the human genome (GRCh38) contains gaps



## State-of-the-art assembly strategies

- ▶ Since 2013, de novo assembly of mammalian genomes has shifted from purely WGS data to assembly with longer sequence reads generated either synthetically or by single-molecule sequencing (SMS) (e.g. PacBio, Nanopore)
- ▶ The main algorithmic approaches to de novo assembly are
  - ▶ Overlap-layout-consensus (OLC)
  - ▶ de Bruijn
  - ▶ (The string graph)

# Overlap-layout-consensus (OLC)

- ▶ Contigs: Continuous (or 'contiguous') sequences produced in a de novo assembly, free of any gaps
- ▶ Basic steps of OLC algorithms:
  - ▶ Overlaps between all read pairs are first detected
  - ▶ Contigs are formed by iteratively merging overlapping reads until a read heuristically determined to be at the boundary of a repeat is reached

# Overlap-layout-consensus (OLC)

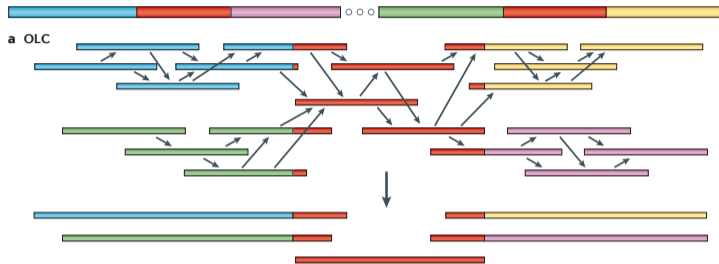


Figure from (Chaisson, et al, 2015)

- ▶ Some repeats can be resolved
- ▶ Imprecise read overlaps are allowed to account for sequencing errors

## de Bruin algorithms

- ▶ Basic steps of de Bruin algorithms:
  - ▶ Start by replacing each read with the set of all-overlapping sequences of a shorter, fixed length ( $k$  typically between 31 and 200)
  - ▶ Contigs are formed by merging  $k$ -mers appearing adjacently in reads stopping at  $k$ -mers from repeat boundaries
- ▶ Requires highly accurate reads
- ▶ Initially discards some of the ability for reads to resolve repeats longer than  $k$  bases

# de Bruin algorithms

**b** de Bruijn

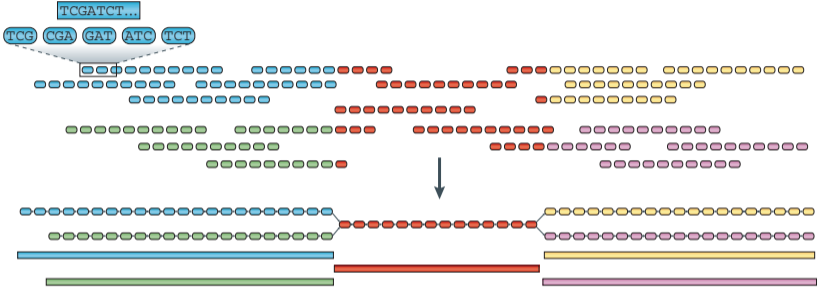


Figure from (Chaisson, et al, 2015)

# Genome annotation de novo

- ▶ Full genome assembly methods can be developed further
- ▶ Things to do next:
- ▶ Gene finding:
  - ▶ Ab initio prediction methods: based on statistical signals within the DNA
    - ▶ E.g.: hidden Markov model-based prediction of genes: Genscan, Augustus, HMMgene
  - ▶ Align known genes of model species against the new genome
  - ▶ If RNA-seq available from the same species, align RNA-seq data to the newly discovered genome
- ▶ Gene annotation:
  - ▶ Function of the genes that can be aligned to new genome give some hint about the newly sequenced organism

## References

- ▶ Chaisson MJP, et al, Genetic variation and the de novo assembly of human genomes, *Nature Reviews Genetics*, 16, 2015.
- ▶ Cibulskis K, et al., Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples, *Nature Biotechnology*, 31, 213–219, 2013.
- ▶ McKenna A, et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Res.*, 20(9):1297-303, 2010.
- ▶ Rasmus Nielsen, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song, Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet.* 12(6): 443-451, 2011.