Genome-scale technologies 2 / Algorithmic and statistical aspects of DNA sequencing

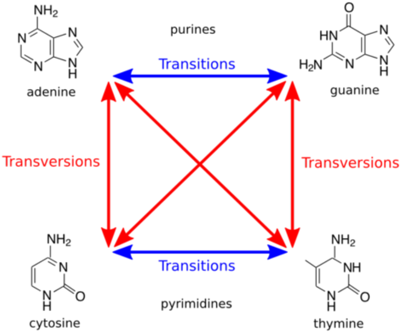
Single nucleotide polymorphism discovery

Ewa Szczurek
szczurek@mimuw.edu.pl

Instytut Informatyki
Uniwersytet Warszawski

Short biological prerequisites

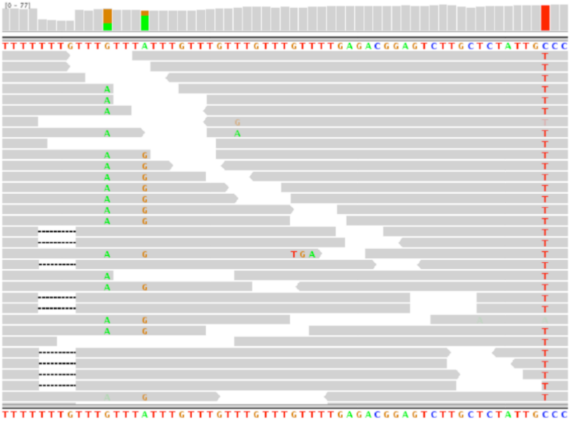
- Haploid genome: each chromosome has a single copy in the nucleus
- Diploid genome: each chromosome has two copies in the nucleus (one from father, one from mother)
- Mutation types: transitions and transversions



Short biological prerequisites

- Allele: one of a number of alternative forms of the same genetic locus.
- Haplotype:
  1. a specific group of genes that a progeny inherits from one parent
  2. a collection of specific alleles (that is, specific DNA sequences) in a cluster of tightly-linked genes on a chromosome that are likely to be inherited together
  3. a set of (several) single-nucleotide polymorphisms (SNPs, śnips")—also known as DNA sequence variations at specific nucleotide sites, or as polymorphic sites—on a single chromosome that are associated statistically.

Single nucleotide polymorphism calling


[0 - 77]
TTTTTTTGTTTGTTTATTTGTTTGTTTGTTTGTTTTGAGACGGAGTCTTGCTCTATTGCCC
TGA

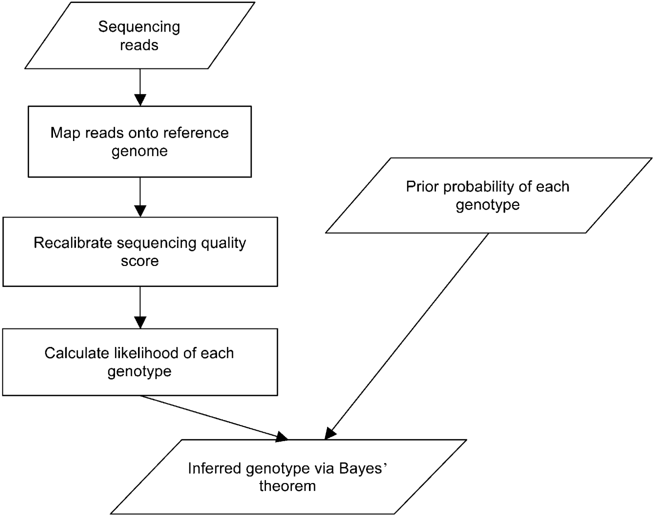
IGV visualization

- Reads: arrows oriented by increasing machine cycle;
- Highlighted bases: mismatches to the reference
- coverage histogram per base above the reads.

Single nucleotide polymorphism calling: challenges

- Reads need to be accurately mapped to the reference
- Each read aligned independently ⇒ many reads spanning indels will be misaligned
- Per base Phred scores are inacurate, co-vary with machine cycle, or sequence context.
- Separating true variation from machine artifacts due to the high rate and context-specific nature of sequencing errors requires sensitive and specific statistical models.

SNP-calling probabilistic model in SOAPsnp


Sequencing reads
Map reads onto reference genome
Recalibrate sequencing quality score
Calculate likelihood of each genotype
Prior probability of each genotype
Inferred genotype via Bayes' theorem

SNP-calling probabilistic model in SOAPsnp

- The probability of genotype T_i given observed data D (reads)s at a locus i is given by
  - $$P(T_i|D) = \frac{P(T_i)P(D|T_i}{\sum_{x=1}^{S} P(T_x)P(D|T_x)}, \tag{1}$$
    where S is the total number of genotypes.
- For a haploid genome H_m there are four types of genotypes:
  - $T_i = H_m \in \{A, C, G, T\}$, $S = 4$.
- For a diploid genome $H_m H_n$ there are four types of genotypes:
  - $T_i = H_m H_n \in \{AA, CC, GG, TT, AC, AG, AT, CG, CT, GT\}$, $S = 10$.

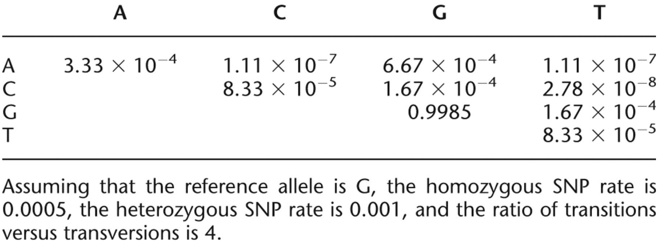
The genotype with the highest posterior $P(T_i|D)$ is chosen as the consensus, with a Phred-like score $-10\log_{10}[1 - P(T_i|D)]$.

SNP-calling model in SOAPsnp: Prior probability of genotypes

- $P(T_i)$ for a haploid genotype
    - Assumptions: SNP rate is 0s.001, transitions are 4× more frequent than transversions.
    - Given the reference allele G, the prior probabilities for relevant allele in reads are: 6.67×10^{-4} for A, 1.67×10^{-4} for C and T and 0.999 for G.

SNP-calling model in SOAPsnp: Prior probability of genotypes

- $P(T_i)$ for a haploid genotype
  - Assumptions: SNP rate is 0s.001, transitions are 4× more frequent than transversions.
  - Given the reference allele G, the prior probabilities for relevant allele in reads are: 6.67×10^{-4} for A, 1.67×10^{-4} for C and T and 0.999 for G.
- $P(T_i)$ for a diploid genotype



	A	C	G	T
A	3.33×10^{-4}	1.11×10^{-7}	6.67×10^{-4}	1.11×10^{-7}
C		8.33×10^{-5}	1.67×10^{-4}	2.78×10^{-8}
G			0.9985	1.67×10^{-4}
T				8.33×10^{-5}

Assuming that the reference allele is G, the homozygous SNP rate is 0.0005, the heterozygous SNP rate is 0.001, and the ratio of transitions versus transversions is 4.

SNP-calling model in SOAPsnp: Observation likelihood

- $P(D|T)$ calculated from observed allele types in the sequencing reads.
- Let $P(d_k|H)$ be the likelihood of observing allele d_k for a possible haploid genotype H.
- For a diplold genome with the assumption that the two copies are independent
  - $$P(d_k|T) = \frac{P(d_k|H_m) + P(d_k|H_n)}{2},$$
- For a set of n observed alleles at a locus i, $D = \{d_1, \ldots, d_n\}$,
  - $P(D|T) = \prod_{k=1}^{n} P(d_k|T)$.

SNP-calling model in SOAPsnp: Recalibration of base calling quality

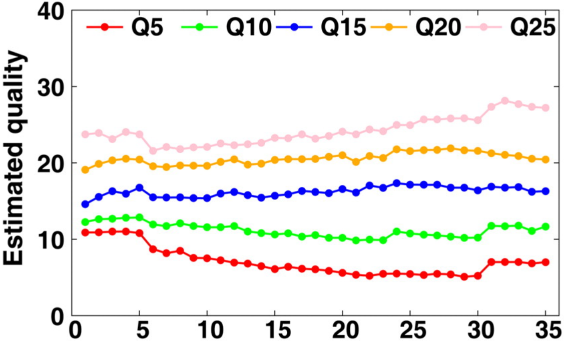
- For each allele d_k observed for an assumed genotype T we define
  1. o_k, observed allele type
  2. q_k, quality score
  3. c_k, sequencing cycle (coordinate on read).
- Then the likelihood $P(d_k|T)$ becomes
  - $$P(d_k|T) = P((o_k, q_k, c_k)|T) = P((o_k, c_k)|(T, q_k))P(q_k|T).$$

SNP-calling model in SOAPsnp: Recalibration of base calling quality

- Four-dimensional matrix built to store $P((o_k, c_k)|(T, q_k))$.
- Precalculated from unique alignments, counting the number of substitutions, and estimating the mismatch rate for each combination of q_k, c_k and substitution type.
- Effectively, each quality score rescaled by each sequencing cycle and for each substitution combination.

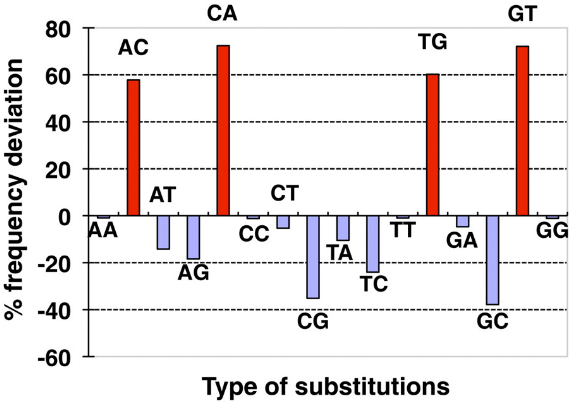
- $P(q_k|T)$ is the probability of a genotype T to have an observation with quality q_K
  - Assumed that for $T = A, C, G, T$ these distributions are the same
  - $P(q_k|T)$ becomes a function of q_k only, $P(q_k|T) = f(q_k)$,
  - $f(q_k)$ reduces in the Bayesian formula (1).

SNP-calling model in SOAPsnp: Recalibration of base calling quality



- Extracted bases with each quality value from raw aligned reads
- Estimated quality $= -10\log_1 0(\mathit{mismatchrate})$.

SNP-calling model in SOAPsnp: Estimated vs quality-based mismatch rate



- % frequency deviation = [(Error rate by alignment mismatch rate) - (error rate by quality value)]/ (Error rate by quality value).

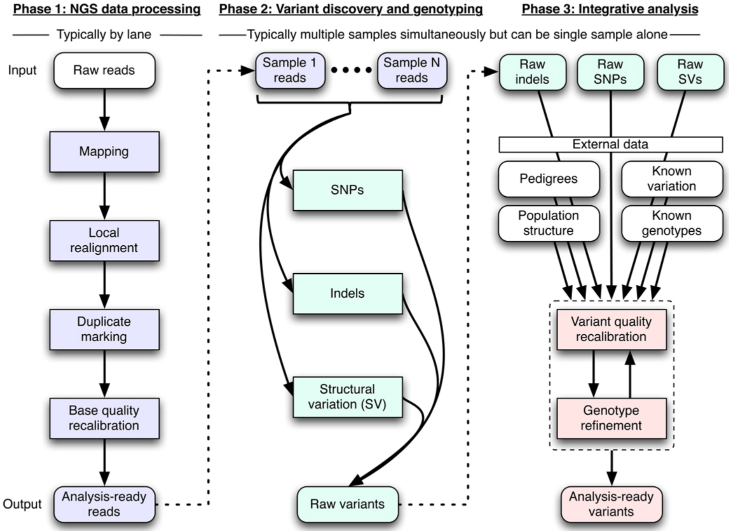
SNP-calling model in SOAPsnp: Dealing with dependent errors

- The same alleles from reads mapped at the same location ranked by sequencing quality (low to high).
- Reduction of quality based on dependence for t_k-th observation

  $$q'_k = \theta^{t_k} q_k \,,$$

  where the dependency coefficient $0 < \theta < 1$. $\theta = 0$ means the completely dependent model, $\theta = 1$ means completely independent model.
- The reduced qualities q_k used instead of the original q_k in the likelihood matrix.

Genome Analysis Toolkit (GATK)



Genome Analysis Toolkit (GATK): Steps of local realignment of reads spanning an indel.

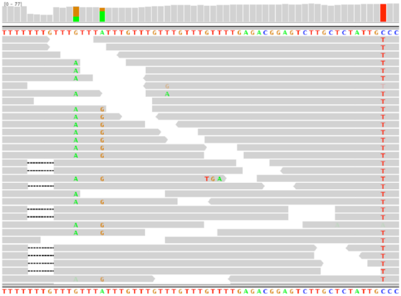
1. Identify regions for realignment where
   - at least one read contains an indel,
   - there exists a cluster of mismatching bases or
   - an already known indel segregates at the site (e.g., dbSNP).
2. At each region, construct haplotypes by incorporating
   - any known indels at the site,
   - indels in reads spanning the site or
   - Smith-Waterman alignment of all reads that do not perfectly match the reference sequence.
3. For each haplotype H_i, each read R_j is aligned without gaps to H_i and assigned the **likelihood** $L(R_j|H_i)$.
4. Realign the reads to i if the log likelihood ratio satisfies

$$\log \frac{\prod_j \max[L(R_j|H_i), L(R_j|H_0)]}{L(R_j|H_0)} > 5$$

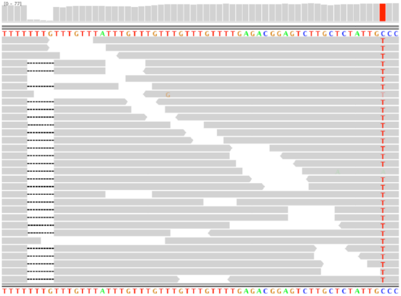
Genome Analysis Toolkit (GATK): The likelihood $L(R_j|H_i)$ in the local realignment of reads spanning an indel.

$$\begin{aligned} L(R_j|H_i) &= \prod_k L(R_{j,k}|H_{i_k}) \\ L(R_j|H_i) &= \begin{cases} 1-\epsilon_{j,k} & \text{if } R_{j,k} = H_{j,k}, \\ \epsilon_{j,k} & \text{otherwise.} \end{cases} \end{aligned} \qquad (2)$$

Read realignment in GATK points at common misalignment errors.


[0 - 77]
TTTTTTTGTTTGTTTATTTGTTTGTTTGTTTGTTTTGAGACGGAGTCTTGCTCTATTGCCC

HiSeq data, raw BWA alignments


[0 - 77]
TTTTTTTGTTTGTTTATTTGTTTGTTTGTTTGTTTTGAGACGGAGTCTTGCTCTATTGCCC

HiSeq data, after MSA

Bibliography

- R. Li et al., *SNP detection for massively parallel whole-genome resequencing*. Genome Res. 2009.
- H. Li and N. Homer, *A survey of sequence alignment algorithms for next-generation sequencing*. Briefings in Bioinformatics 2010.
- M. A. DePristo et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data*. Nature Genetics 2011.