SNP calling and genotyping

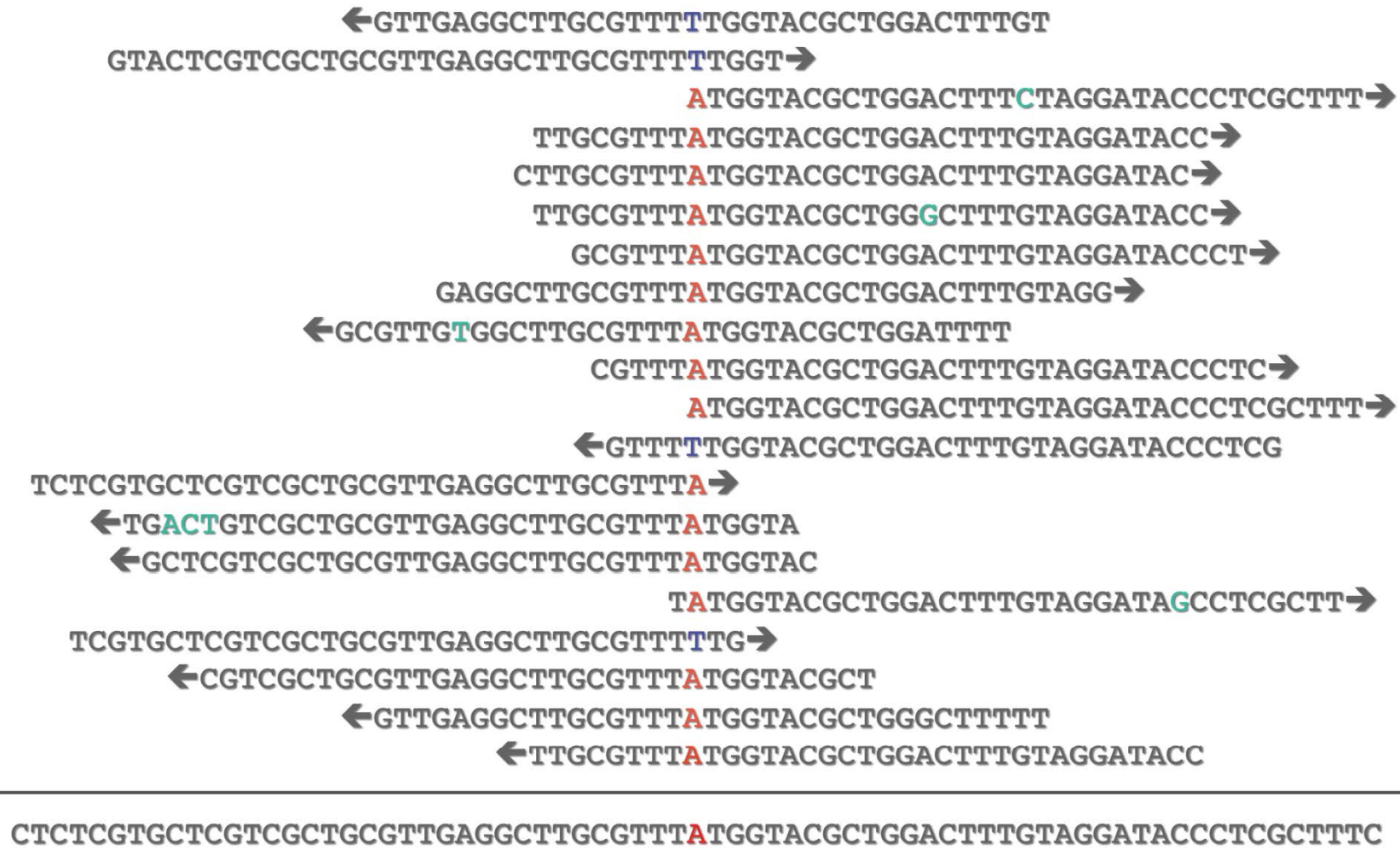
Statistical Methods for Next Generation Sequencing

ENAR 2012

Zhijin Wu zhijin_wu@brown.edu

Reads after initial mapping

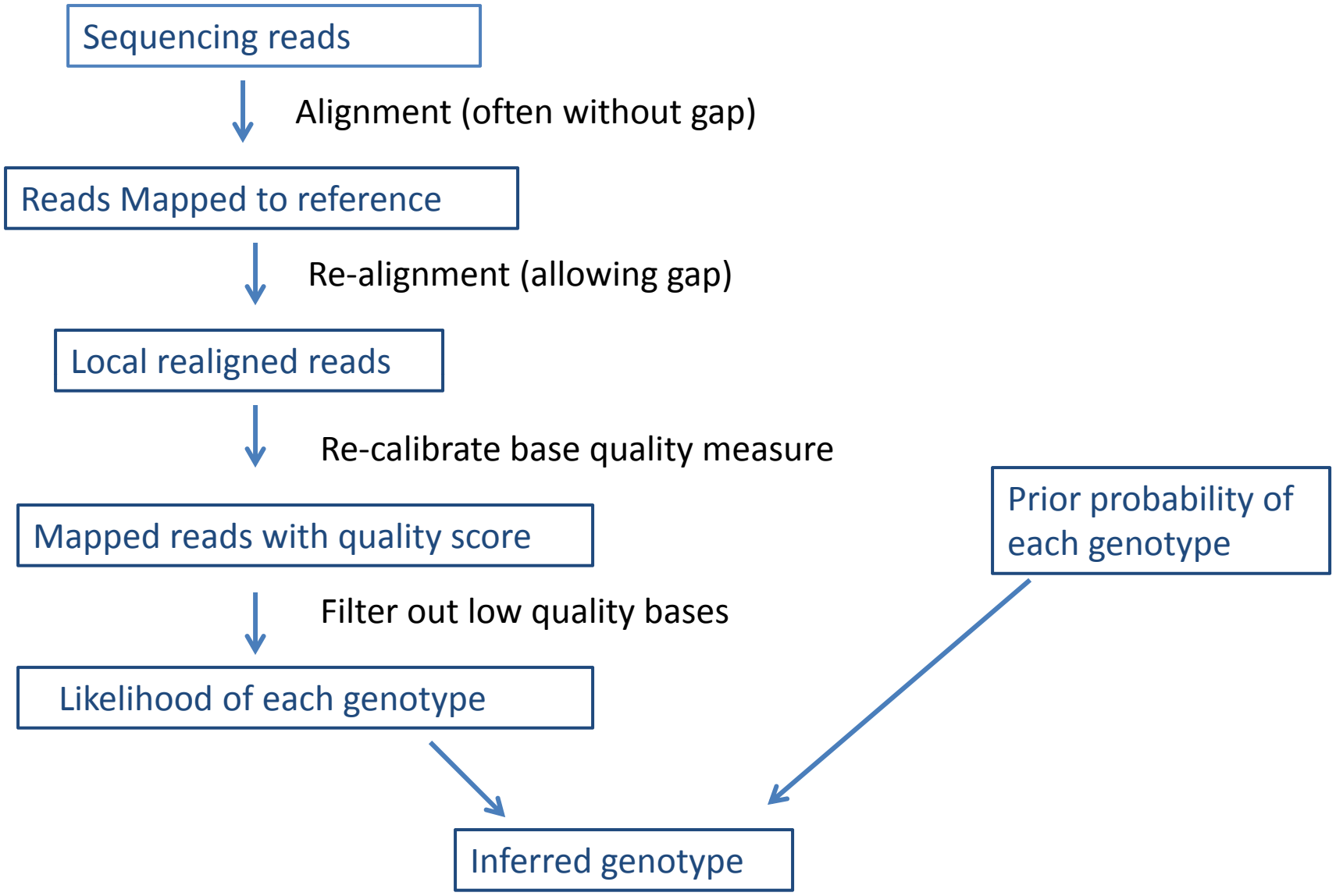
Mismatches are ***potential*** variants



Ref: CTCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC

Possible reasons for a mismatch

- True SNP
- Error generated in library preparation
- Base calling error
  - May be reduced by better base calling methods, but cannot be eliminated
- Misalignment (mapping error):
  - Local re-alignment to improve mapping
- Error in reference genome sequence


Sequencing reads
Alignment (often without gap)
Reads Mapped to reference
Re-alignment (allowing gap)
Local realigned reads
Re-calibrate base quality measure
Mapped reads with quality score
Filter out low quality bases
Likelihood of each genotype
Prior probability of each genotype
Inferred genotype

Li et al (2009)

Basic model: Bayes Theorem

$P(\text{genotype}|\text{data}) \propto P(\text{data}|\text{genotype})P(\text{genotype})$

P(genotype) : prior probability for variant

P(data|genotype): likelihood for observed(called) allele type

Error due to mapping

- Multiple alignment:
  - longer reads have higher probability of unique alignment
- Mis-alignment:
  - longer reads have lower probability of mis-alignment
- Solution
  - Filter out alleles with very low frequency
  - Filter out bases with low base call quality
  - Filter out reads with low mapping quality
  - Limit number/proportion of mismatches in the neighborhood
- For bases that pass filtering we generally treat them as correctly aligned, thus the likelihood is determined by base calling alone

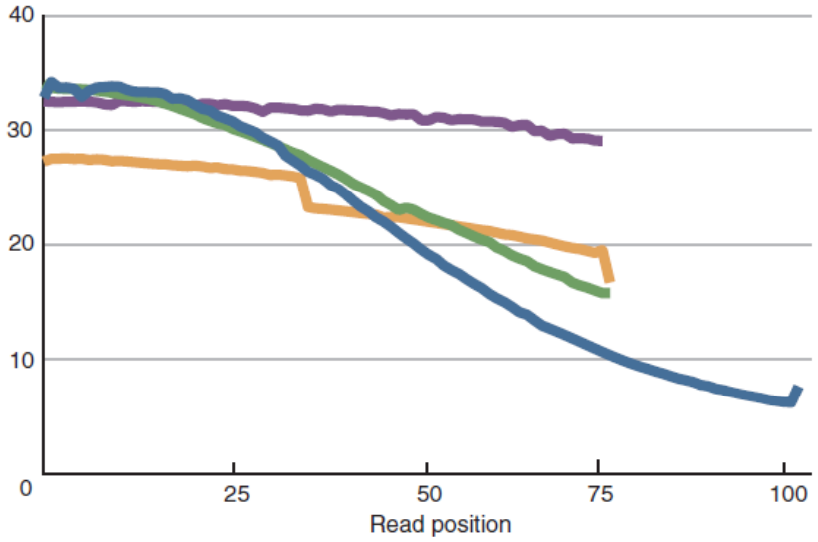
R Li et al (2009) Genome Research 19:1124-132; DePristo (2011) Nature Genetics 43(5) 491

Likelihood P(data|genotype)

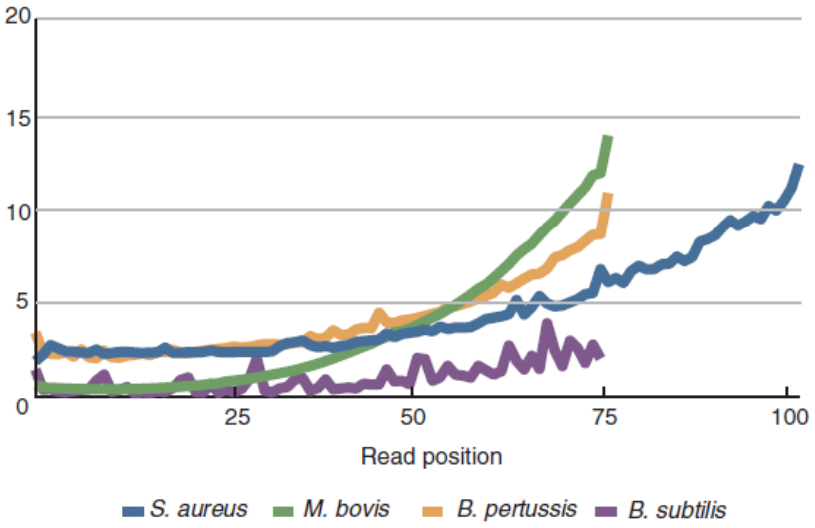
What's known to affect base calling

- Error rate increases as cycle numbers increase
- Error rate depends on substitution type
- Error rate depends on local sequence environment

Base call quality

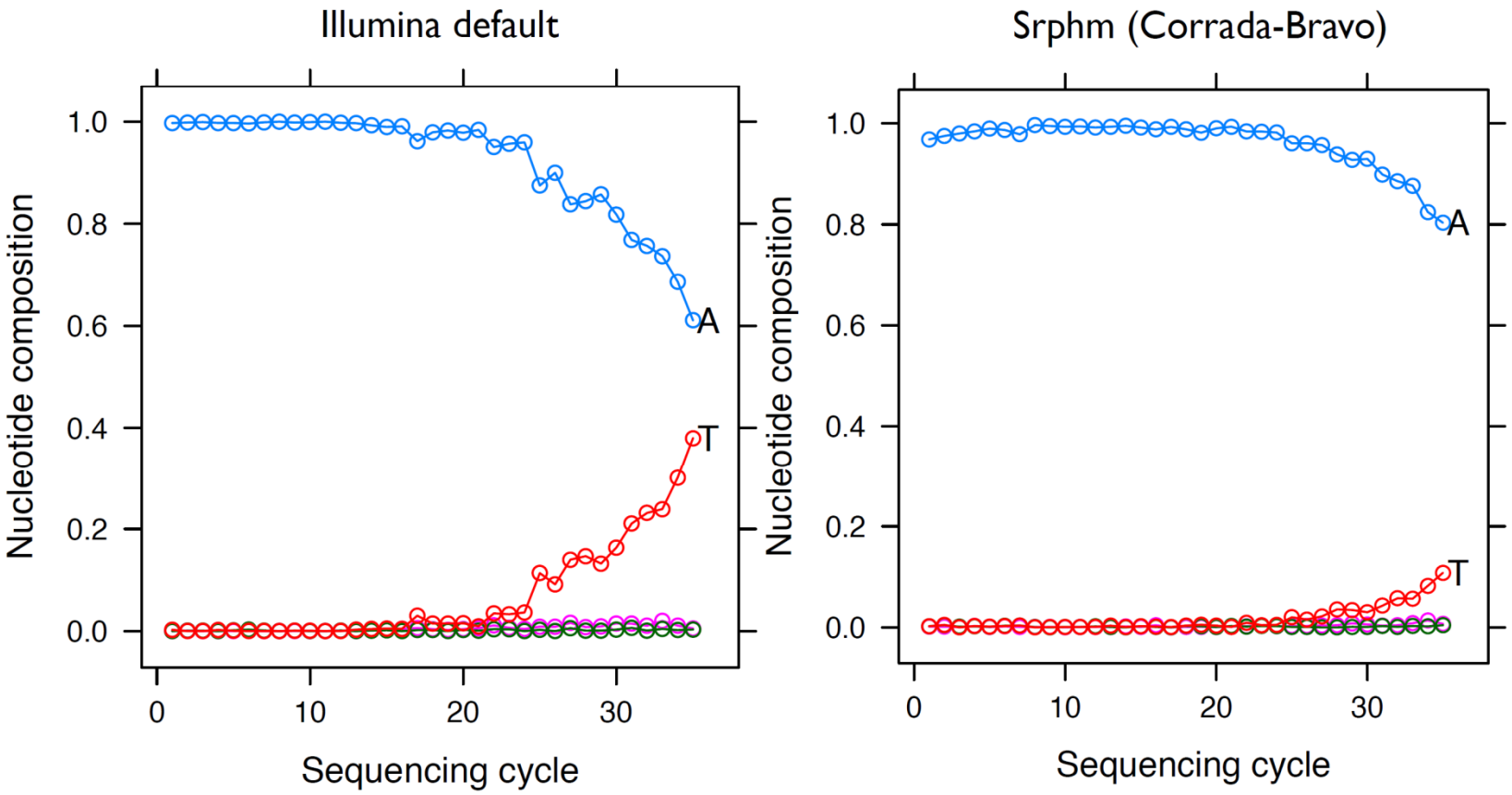


Mismatch rate



Nakamura et al (2011) NAR

Base calling can be improved but errors cannot be eliminated



Corrada-Bravo and Irizarry (2010) Biometrics. 66: 665-674

Quality score

Quality score $-10\log_{10}$ (error rate)

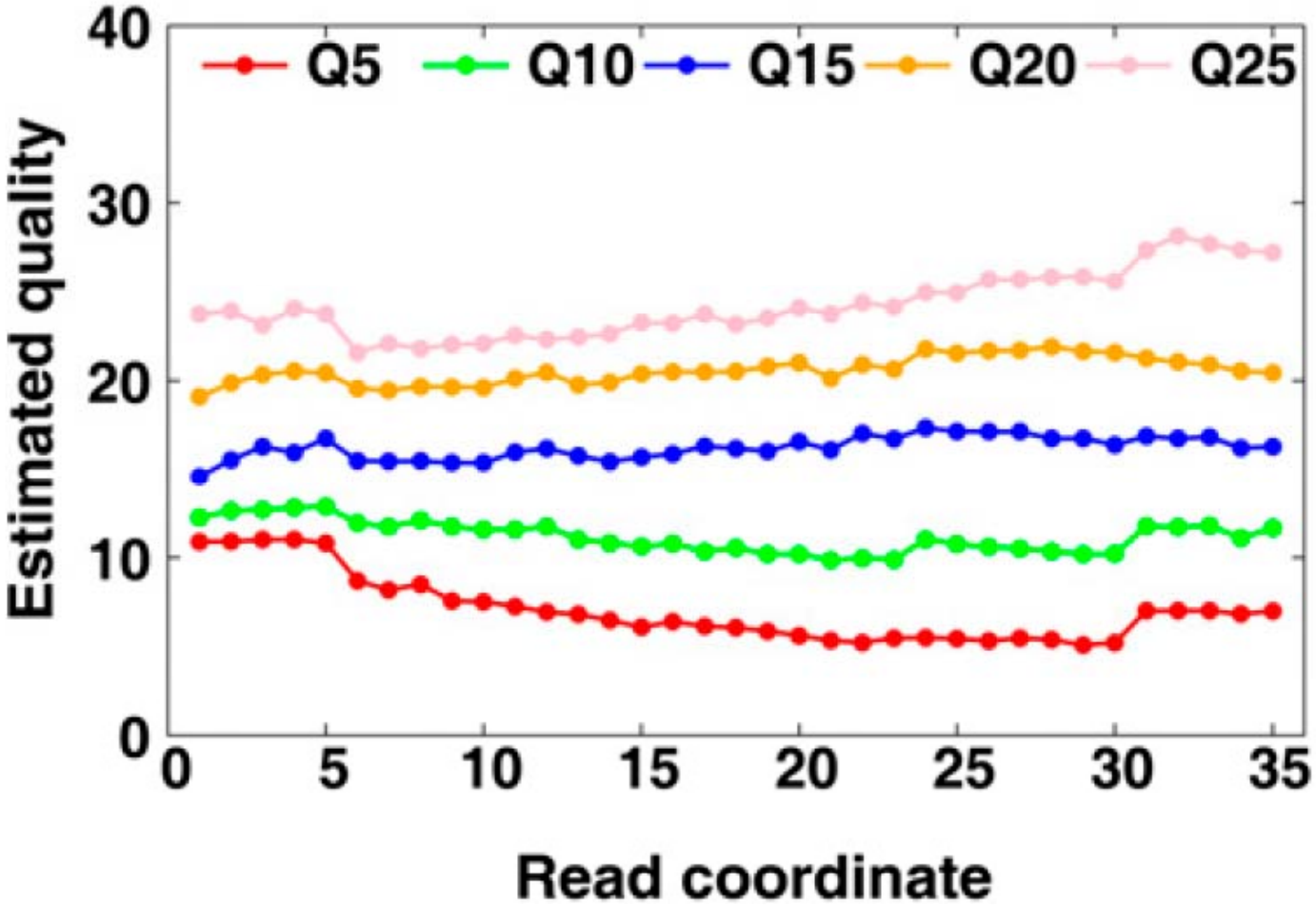
Q20: 1 in 100

Q30: 1 in 1000

Q40: 1 in 10,000

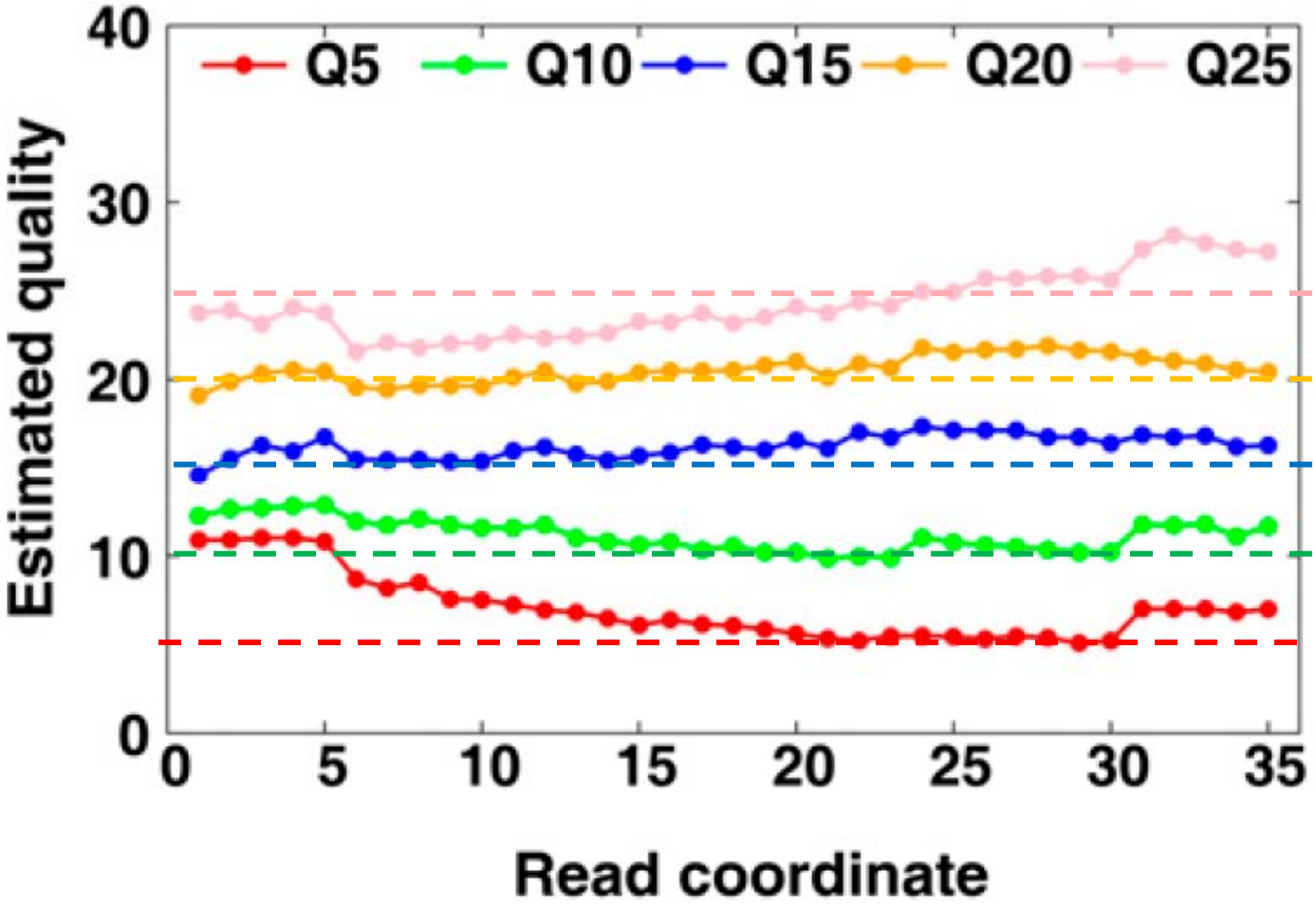
Illumina reports most reads with quality above Q30 and offers a recalibration to remove cycle effects.

Remaining cycle effect



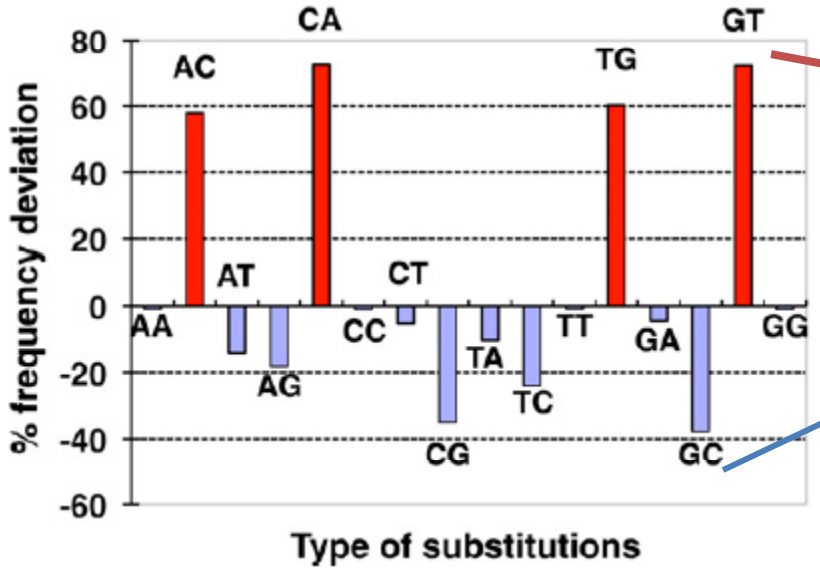
R Li et al (2009) Genome Research 19:1124-132

Remaining cycle effect



R Li et al (2009) Genome Research 19:1124-132

Substitution bias



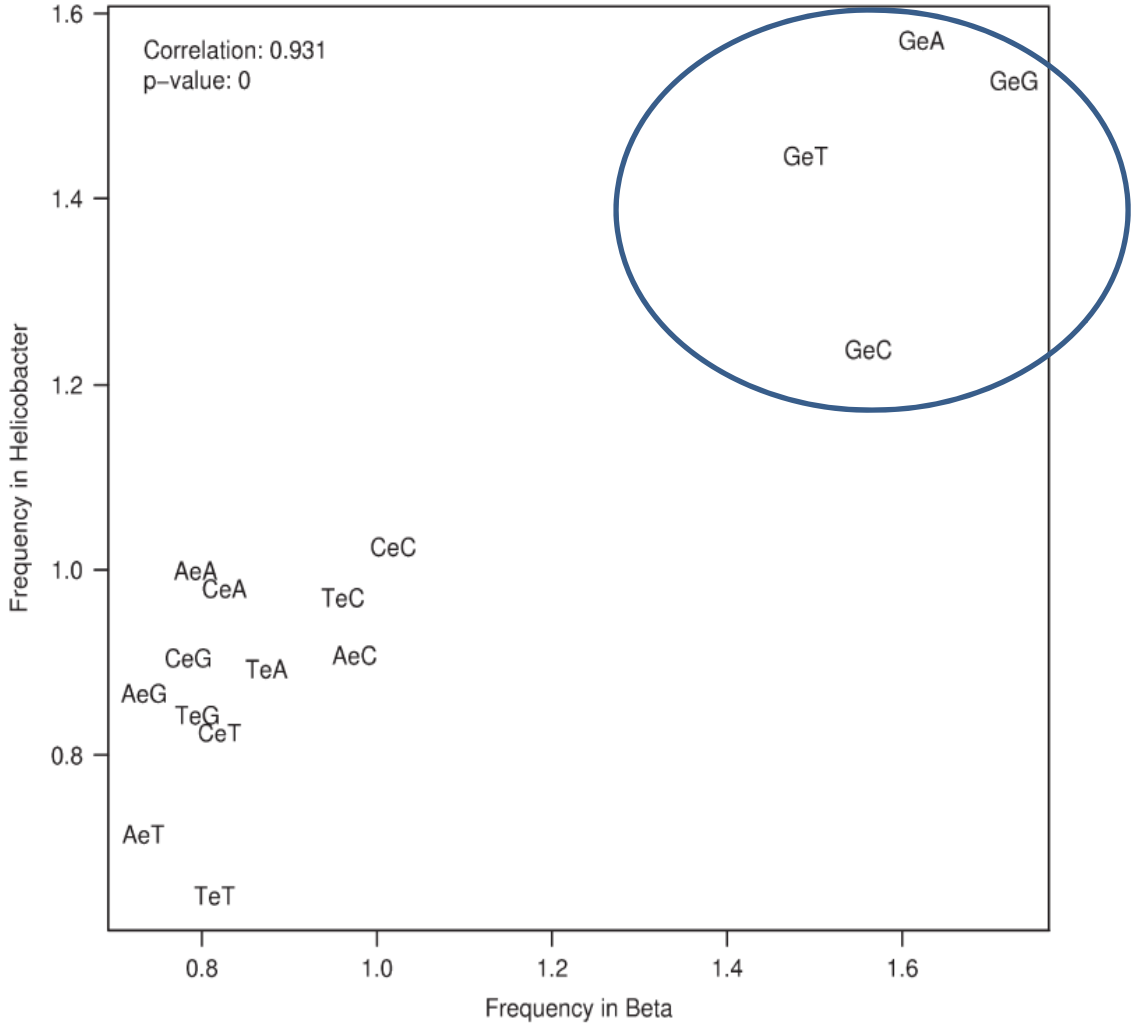
T→G mistake is probably under reported

C→G error over reported

(O-R)/R= (mismatch rate- reported error rate)/reported error rate

R Li et al (2009) Genome Research 19:1124-132

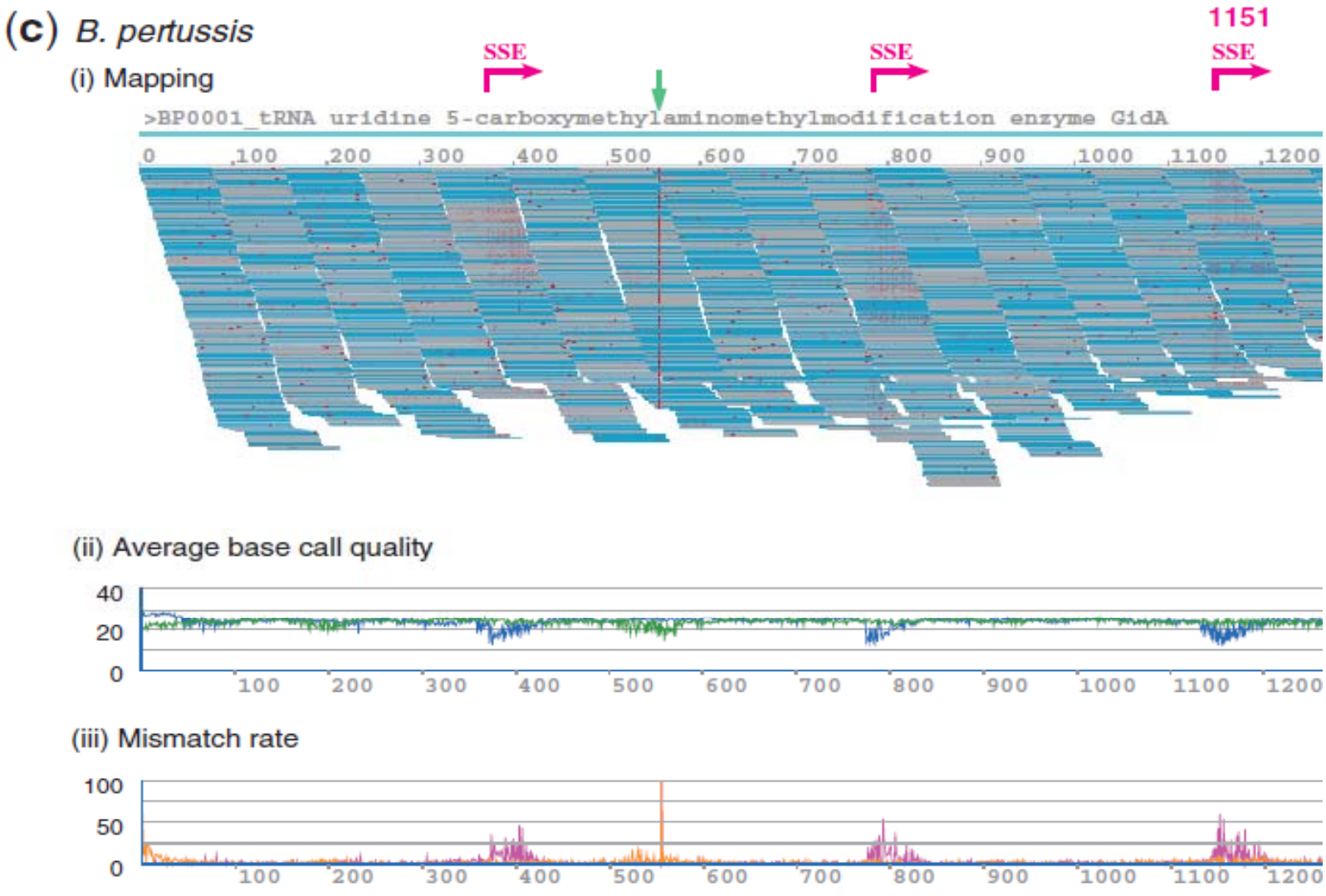
Sequence context



G is a likely base before an error

Dohm et al (2008) NAR. 36(16):e105

Sequence specific error (SSE)



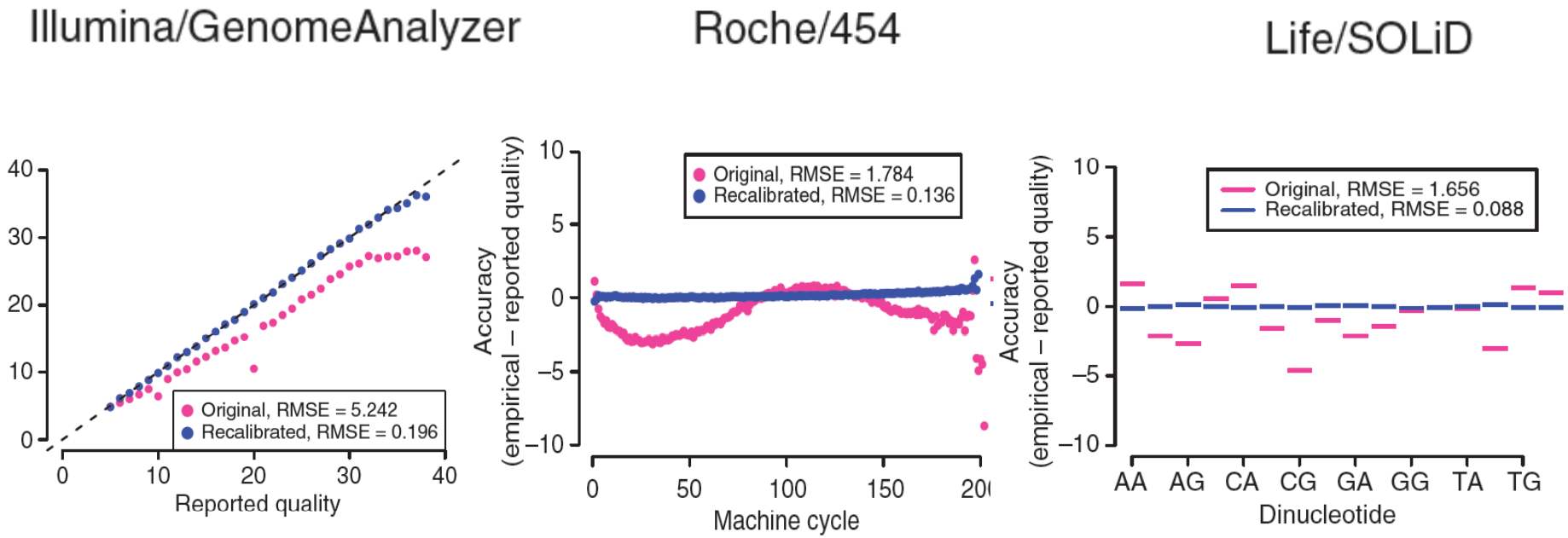
Nakamura et al (2011) NAR

Recalibrate base quality score

- Stratify bases by
  - reported quality score (q)
  - Machine cycle (C)
  - Dinucleotide context
  - Down-weighting or remove duplicate clones
- For each strata compare empirical error rate (mismatch rate) to reported error rate, compute the difference as bias in error rate
- Remove the estimated bias

DePristo (2011) Nature Genetics 43(5) 491; R Li et al (2009) Genome Research 19:1124-132

Error recalibration for various technologies



DePristo (2011) Nature Genetics 43(5) 491

Example of Re-Calibrated miscalling matrix

Illumina (GA&HiSeq)				
	A	C	G	T
A	N/A	57.7%	17.1%	25.2%
C	34.9%	N/A	11.3%	53.9%
G	31.9%	5.1%	N/A	63.0%
T	45.8%	22.1%	32.0%	N/A

Supp table 7, DePristo (2011) Nature Genetics 43(5) 491

Prior probability of genotypes

- Genome wide SNP rate
- SNP substitution type not equally likely
- Allele frequency

DePristo (2011) Nature Genetics 43(5) 491

Ti/Tv ratio

- Transition (Ti) :
  - purine<->purine (A <-> G)
  - pyrimidine<-> pyrimidine (C <-> T)
- Transversion (Tv): purine <-> pyrimidine

  A <-> C, A <-> T, G <-> C , G <-> T
- Transition is more frequent than transversion
  - Ti/Tv ~ 2.0 -2.1 for genome wide
  - Ti/Tv ~ 3.0-3.3 for exonic variations
  - Ti/Tv=2/4=0.5 for random, uniform sequencing error

DePristo (2011) Nature Genetics 43(5) 491

Prior probability of genotypes

- Example: Assuming
  - heterozygous SNP rate 0.001, homozygous SNP rate 0.0005
  - Reference allele: G
  - Transition/transversion ratio 2

	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}

R Li et al (2009) Genome Research 19:1124-132

Prior probability of genotypes

Other information that can be used in setting priors:

- Use dbSNP prior probability
- Use different polymorphism rate for different genomic regions
- Consider different Ti/Tv rate for exonic regions

An example of prior probability for a dbSNP G/T site used in Li et al (2009)

	A	C	G	T
A	$4.55*10^{-7}$	$9.11*10^{-8}$	$9.1*10^{-5}$	$9.1*10^{-5}$
C		$4.55*10^{-7}$	$9.1*10^{-5}$	$9.1*10^{-5}$
G			.454	.0909
T				.454

R Li et al (2009) Genome Research 19:1124-132

dbSNP

- A public database hosted by NCBI for SNPs (and some other variations)
- May include SNP types and allele frequency
- Quality may vary

Reference SNP(refSNP) Cluster Report: rs1579113

RefSNP

Organism: human (Homo sapiens)
Molecule Type: Genomic
Created/Updated in build: 88/135
Map to Genome Build: 37.3
Validation Status:

Allele

Variation Class: SNV: single nucleotide variation
RefSNP Alleles: C/T
Allele Origin:
Ancestral Allele: C
Clinical Source: unknown
Clinical Significance: NA
MAF/MinorAlleleCount: T=0.258/564
MAF Source: 1000 Genomes

Bayes formula

For an individual *i*, D={d_1,d_2,…,d_n}

$$P(G_i \mid D) = \frac{P(G_i)P(D \mid G_i)}{\sum_{j=1}^{J} P(G_j)P(D \mid G_j)}$$

$$P(D \mid G_j) = \prod_{k=1}^{n} P(d_k \mid G_j)$$

Haploid genotypes: G1 $\in$ {A,T,G,C}, J=4

Diploid genotypes: G2 $\in$ {AA,CC,GG,TT,AC,AG,AT,CG,CT,GT}, J=10

$$P(d_k = "A" \mid G2 = "GA") = \frac{P(d_k = "A" \mid G1 = "G") + P(d_k = "A" \mid G1 = "A")}{2}$$

R Li et al (2009) Genome Research 19:1124-132

Multiple sample SNP calling

1000 genome project

- Low coverage (~4x)
  - 60 of European ancestry from Utah (CEU)
  - 59 from a Nigeria population (YRI)
  - 30 of Han Chinese ancestry (CHB)
  - 30 of Japanese ancestry (JPT)
- High coverage (42X) trio
  - two parent-offspring trios

Multiple sample SNP calling

- Phase I: Likelihood for each individual *i*

$$P(D_i|G_i) = \prod P(D_{i,j}|G_i)$$

$$P(D_{i,j} = d|G_i = B) = \begin{cases} 1 - \epsilon_{ij} & d = B \\ \epsilon_{ij} P(B \rightarrow d|miscalled) & \text{otherwise} \end{cases}$$

$$P(D_{i,j}|G_i = AB) = \frac{P(D_{i,j}|A) + P(D_{i,j}|B)}{2}$$

DePristo (2011) Nature Genetics 43(5) 491

Multiple sample SNP calling

- Phase II: combine all samples

For $q_i \in \{0, 1, 2\}, q = \sum_{i=1}^{N} q_i, X \in \{0, 1, \ldots, 2N\}$

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

a population genetic prior for allele frequency

$$p(q = X) = \begin{cases} \theta/X & X > 0 \\ 1 - \theta \sum_{i=1}^{2N} 1/i & \text{otherwise} \end{cases}$$

DePristo (2011) Nature Genetics 43(5) 491

Infinite sites Wright-Fisher model

- A classical model in population genetics for genetic drift (the stochastic fluctuations in allele frequency due to random sampling in a finite population)

- Under the infinite site neutral variation model, the allele frequency spectrum (AFS) of segregating sites is

$$\phi(x) = \begin{cases} \theta/x & x > 0 \\ 1 - \theta \sum_{x=1}^{2N} 1/x & \text{otherwise} \end{cases}$$

where θ is the expected heterozygosity

Multiple sample SNP calling

- Phase II: combine all samples

For $q_i \in \{0, 1, 2\}, q = \sum_{i=1}^{N} q_i, X \in \{0, 1, \ldots, 2N\}$

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

$$p(q = X) = \begin{cases} \theta/X & X > 0 \\ 1 - \theta \sum_{i=1}^{2N} 1/i & \text{otherwise} \end{cases}$$

$$P(D|q = X) = \sum_{\mathbf{G} \in \Gamma_X} P(D|\mathbf{G})P(\mathbf{G}|q = X) = \sum_{\mathbf{G} \in \Gamma_X} \prod_i^N P(D_i|G_i)P(\mathbf{G}|q = X)$$

$$\Gamma_X = \{(G_1, G_2, \ldots, G_N) \text{ such that } \sum_i q_i = X\}$$

DePristo (2011) Nature Genetics 43(5) 491

- The probability $P(D|q = X)$ is often approximated to avoid evaluating all combinations in the set

$$\Gamma_X = \{(G_1, G_2, \ldots, G_N) \text{ such that } \sum_i q_i = X\}$$

- DePristo (2011) uses an EM like algorithm with Hardy-Weinberg Equilibrium assumption that emits the both P(q|D) as well as **G**, the genotype assignments
- The probability of having a SNP is represented in a quality score

$$QUAL = -10 \log_{10}[P(q = 0|D)]$$

DePristo (2011) Nature Genetics 43(5) 491

Hardy-Weinberg equilibrium

- For a large population under random mating, if the allele frequencies are

$$P(A)=p,\ P(a)=1-p=q$$

The genotype frequency is

$$P(AA)=p^2 \quad P(Aa)=2pq \quad P(aa)=q^2$$

References

[1] H.C. Bravo and R.A. Irizarry. Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics*, 66(3):665–674, 2010.

[2] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. Del Angel, M.A. Rivas, M. Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491–498, 2011.

[3] J.C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic acids research*, 36(16):e105–e105, 2008.

[4] H. Li, J. Ruan, and R. Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–1858, 2008.

[5] R. Li, Y. Li, X. Fang, H. Yang, J. Wang, K. Kristiansen, and J. Wang. Snp detection for massively parallel whole-genome resequencing. *Genome research*, 19(6):1124–1132, 2009.

[6] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[7] K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M.C. Linak, A. Hirai, H. Takahashi, et al. Sequence-specific error profile of illumina sequencers. *Nucleic acids research*, 39(13):e90–e90, 2011.

[8] R. Nielsen, J.S. Paul, A. Albrechtsen, and Y.S. Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.

References

[1] H.C. Bravo and R.A. Irizarry. Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics*, 66(3):665–674, 2010.

[2] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. Del Angel, M.A. Rivas, M. Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491–498, 2011.

[3] J.C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic acids research*, 36(16):e105–e105, 2008.

[4] H. Li, J. Ruan, and R. Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–1858, 2008.

[5] R. Li, Y. Li, X. Fang, II. Yang, J. Wang, K. Kristiansen, and J. Wang. Snp detection for massively parallel whole-genome resequencing. *Genome research*, 19(6):1124–1132, 2009.

[6] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[7] K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M.C. Linak, A. Hirai, H. Takahashi, et al. Sequence-specific error profile of illumina sequencers. *Nucleic acids research*, 39(13):e90–e90, 2011.

[8] R. Nielsen, J.S. Paul, A. Albrechtsen, and Y.S. Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.