Point Mutation Calling

NGS Lectures - Day 6

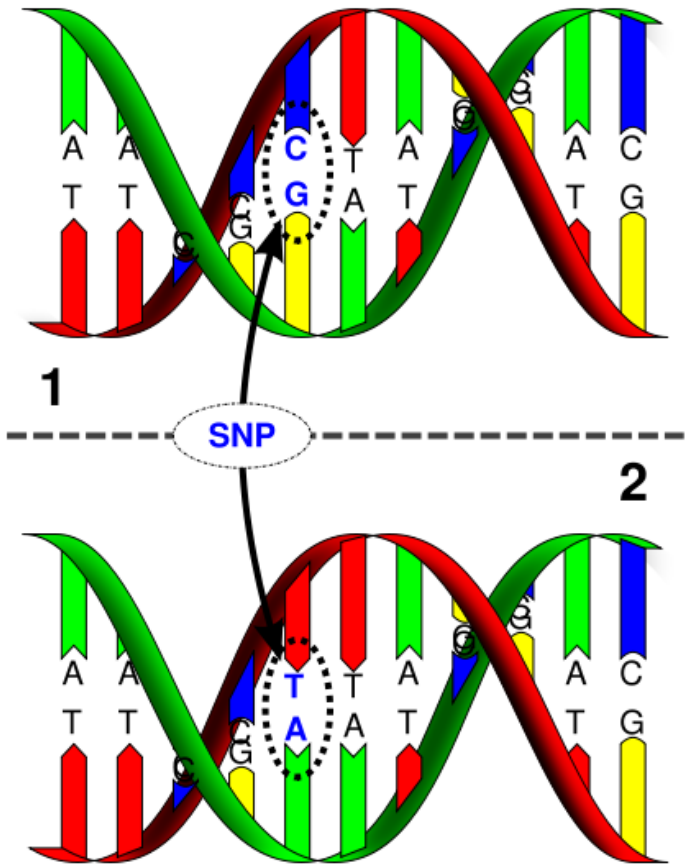
Lars Feuerbach

NGS – Point mutations

1. The biological phenomenon
2. Translation into NGS signals
3. Signal detection and evaluation

Biology of point mutations

Nucleotide differences



http://www.science.marshall.edu/murraye/341/Images/416px-Dna-SNP_svg.png

Individual humans differ approximately at every 1000th bp from the reference genome

The majority of variations are single nucleotide differences

Polymorphisms = inherited differences

Somatic variation = acquired differences

Single nucleotide variant (SNV)

A nucleotide change that is acquired during life time

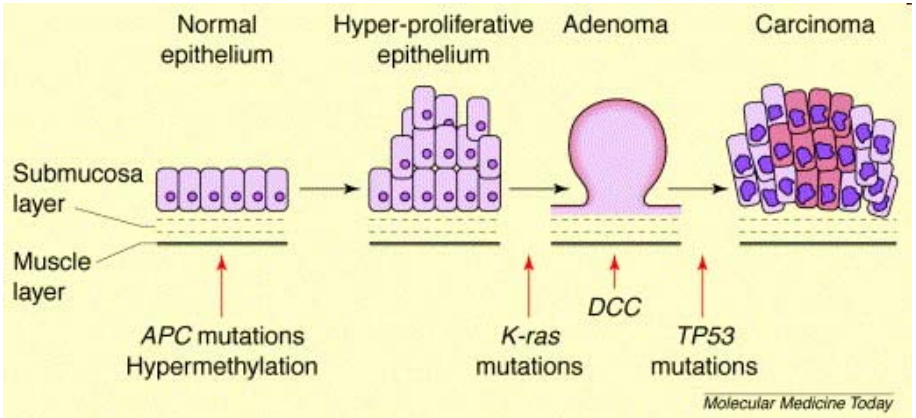
Example: TP53

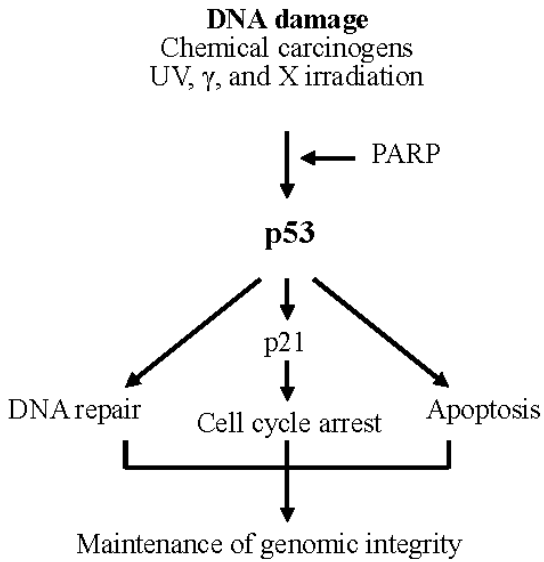
Gene is mutated in half of all human tumors

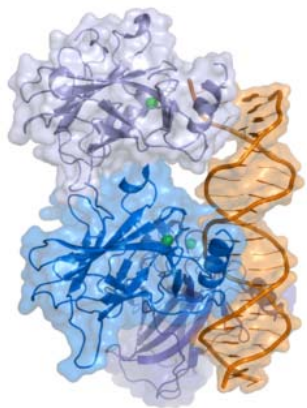
SNVs disrupt tumor suppressor function

27,580 SNVs known

http://www.p53.iarc.fr/Statistics.html







http://ars.sciencedirect.com/content/image/1-s2.0-S1357431099015981-gr1.jpg

http://herkules.oulu.fi/isbn9514270398/html/graphic22.png

http://wikipedia/p53

Single nucleotide polymorphism (SNP)

- A changed nucleotide that is distributed in the population
- An individual acquires a SNP by inheritance
- SNP frequency is often subject to natural selection

HBB Sequence in Normal Adult Hemoglobin (Hb A):

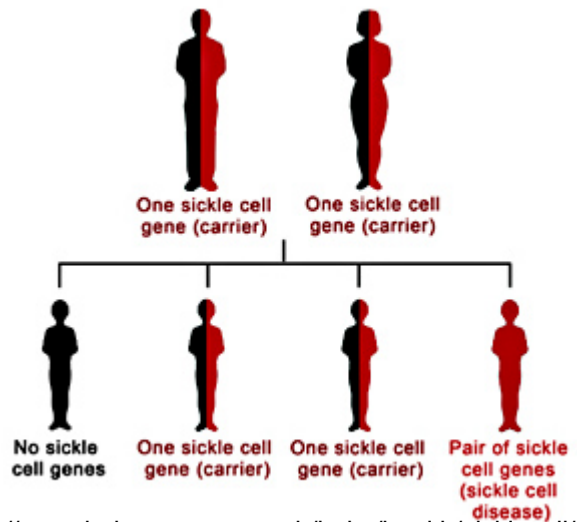
Nucleotide	CTG	ACT	CCT	GAG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Glu	Glu	Lys	Ser
	3			6			9

HBB Sequence in Mutant Adult Hemoglobin (Hb S):

Nucleotide	CTG	ACT	CCT	GTG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Val	Glu	Lys	Ser
	3			6			9

http://www.ornl.gov/sci/techresources/Human_Genome/posters/chromosome/hbb.shtml

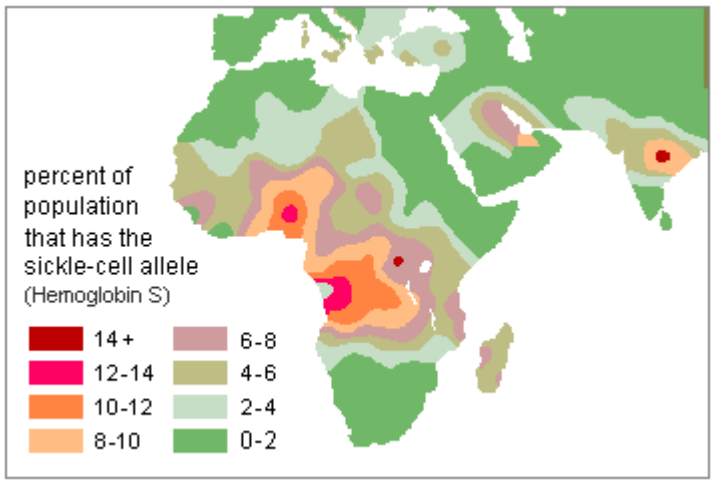
Example: Sickle cell anemia

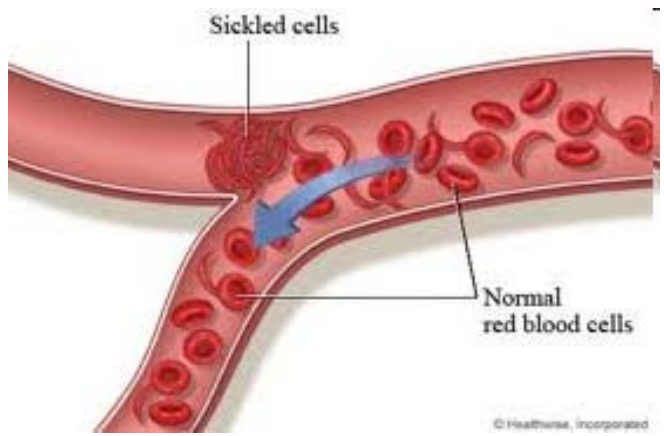


http://www.babycenter.com.ph/baby/health/sicklecell/

Single nucleotide polymorphism (SNP)

Hetrozygous carrier protected from malaria





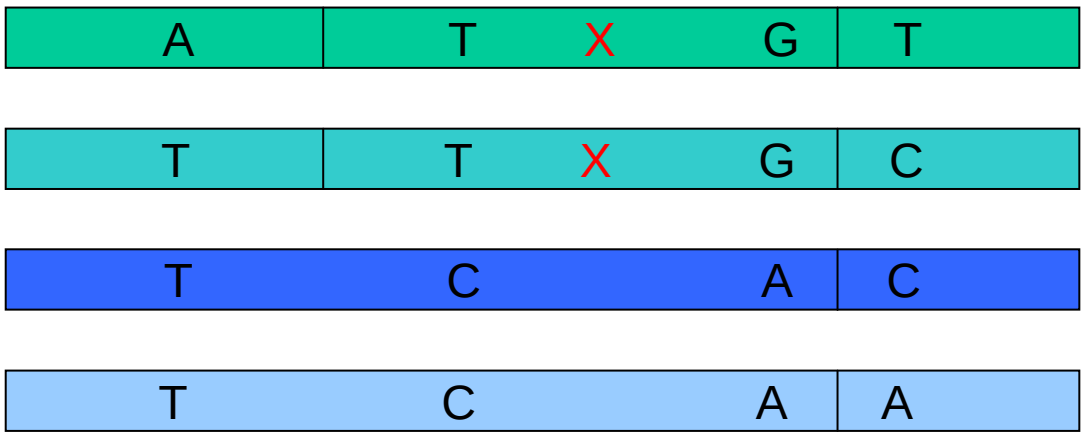
Homozygous carrier subject to sickele cell crisis

http://anthro.palomar.edu/synthetic/synth_4.htm

http://www.meghmiller.com/the-adventure-continues-an-unexpected-chapter/

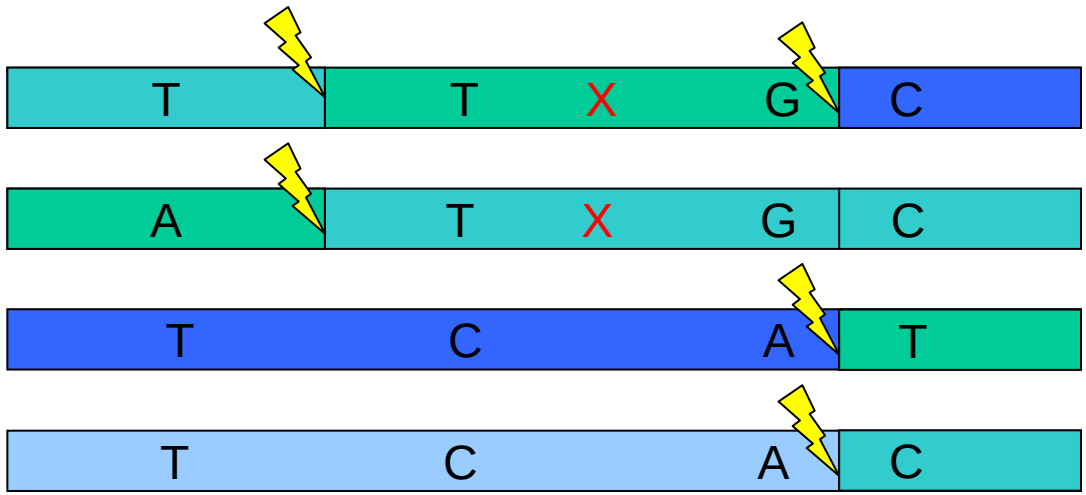
SNPs and Linkage

- Linkage = X correlates highly with close-by SNPs



SNPs and Linkage

- A haplotype is a fixed combination of alleles e.g. SNPs
- Cross-overs induce haplotypes
- Haplotypes induce linkage



Comparison

	SNP	SNV
Appearance	population	Private / tissue specific
Number	millions per individual	few to thousands per tissue
Linkage artifacts	yes	no
Hereditable	yes	no

SNV/SNP allele frequencies

Genotype	Name	Context	AF
AA / A0 / A	Reference	Various	0.0
BB	Homozygous	Diploid genome	1.0
AB	Heterozygous	Diploid genome	0.5

Genotype defines expected allele frequency

Genotype	Name	Context	AF
0	Null allele	Deletion	0.0
B0	LOH	Unbalanced	1.0
B	Hemizygous	Gonosomes (male)	1.0
BB	Homozygous	Diploid genome	1.0
AB	Heterozygous	Diploid genome	0.5
AAB	CNG	Trisomy	0.67/0.33
AABB	CNG	Genome duplications	0.5
AAAB	CNG	Tetraploidy	0.75/0.25

Cataloguing SNPs and SNVs

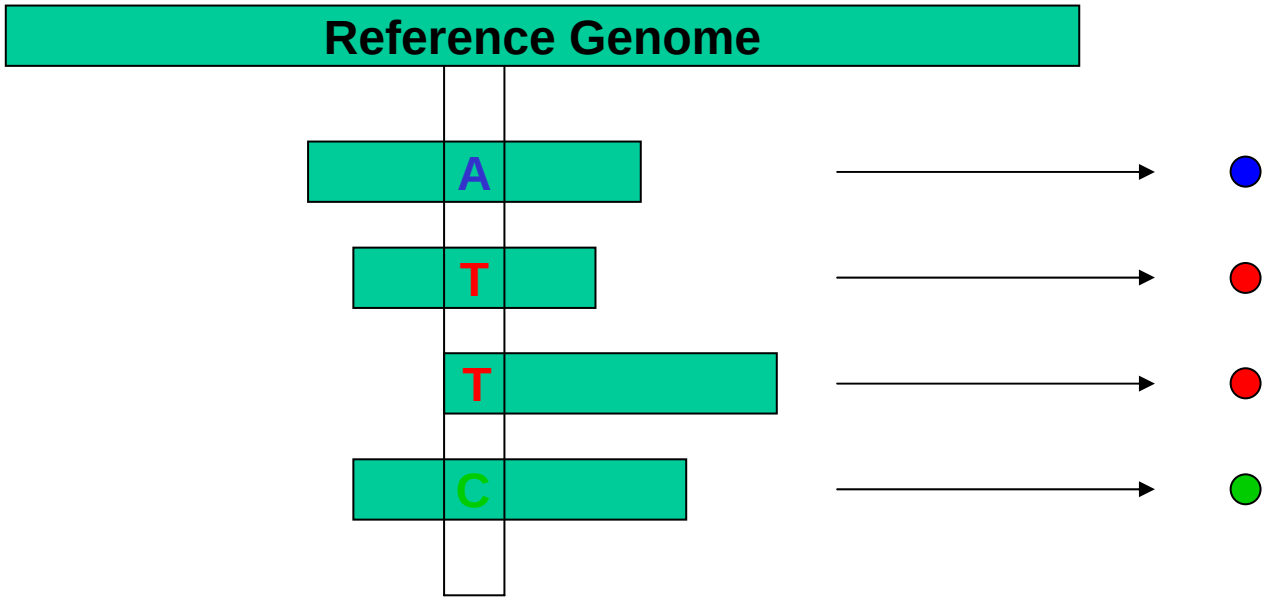
1000 Genomes
A Deep Catalog of Human Genetic Variation

International Cancer Genome Consortium

Large scale projects attempt to catalogue the most frequent SNPs and SNVs

Translation into NGS signals

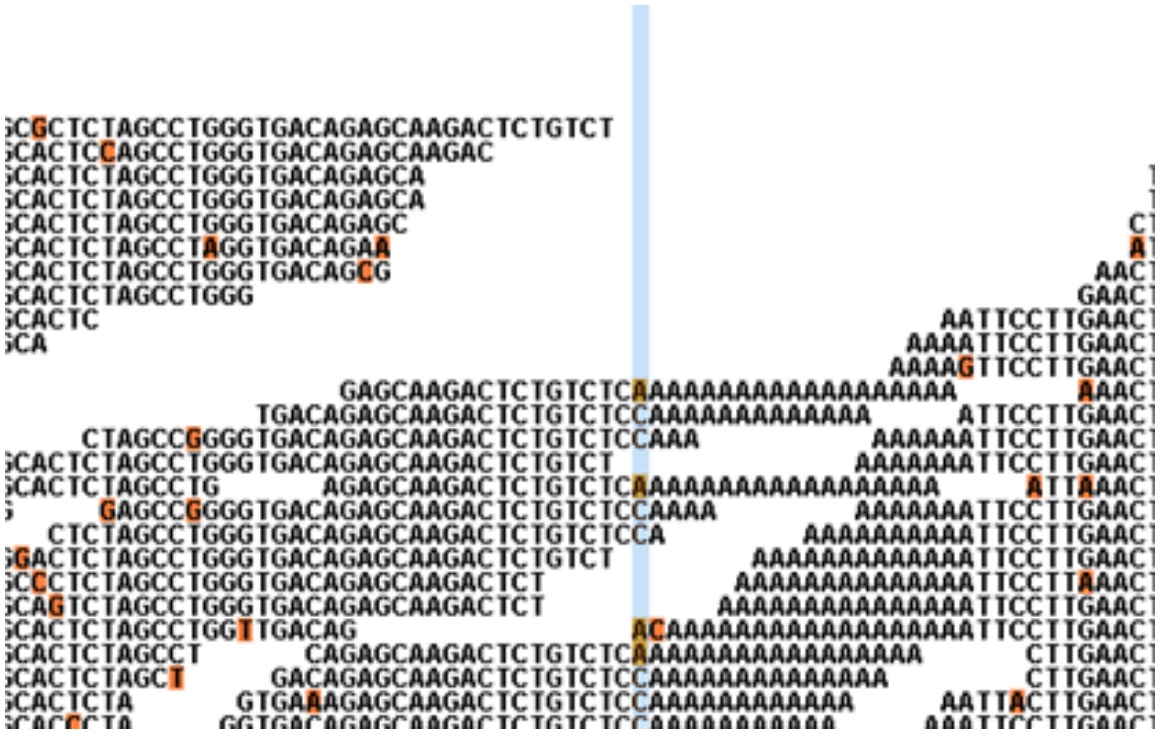
Pileup schematic



All reads that overlap with a specific position

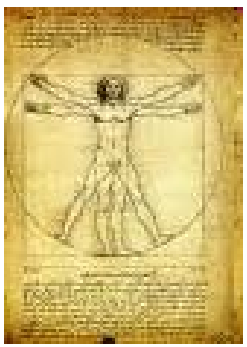
Pileup example

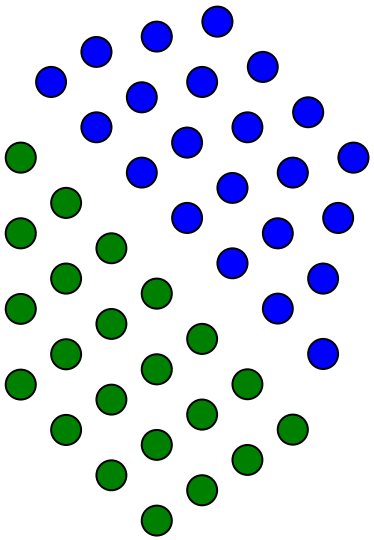
1. Sequence the sample
2. Perform the alignment
3. Check for each position how many nucleotides differ from reference

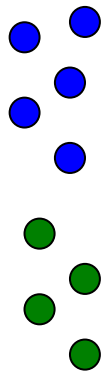


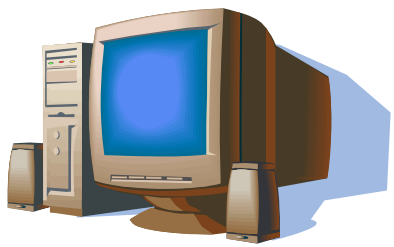
http://www.bioinfor.com/images/stories/zoom/zoom-ngs-45.png

Basic workflow







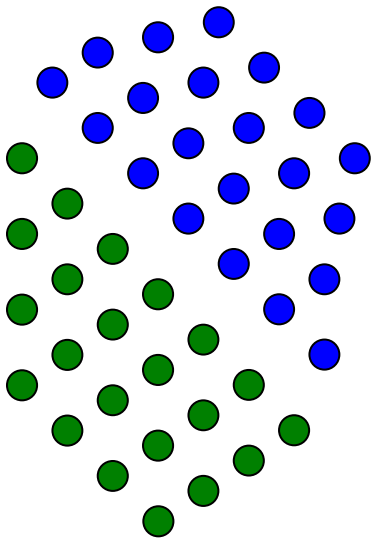


http://www.bioscientia.de/de/diagnostik/humangenetik/next-generation-sequencing/

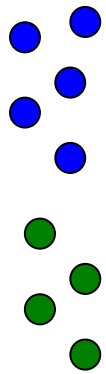
Theoretical model

Number of molecules >> Number of reads

DNA



$$P(obs) = \binom{d}{k} p^k (1-p)^{d-k}$$



Reads

Model definition

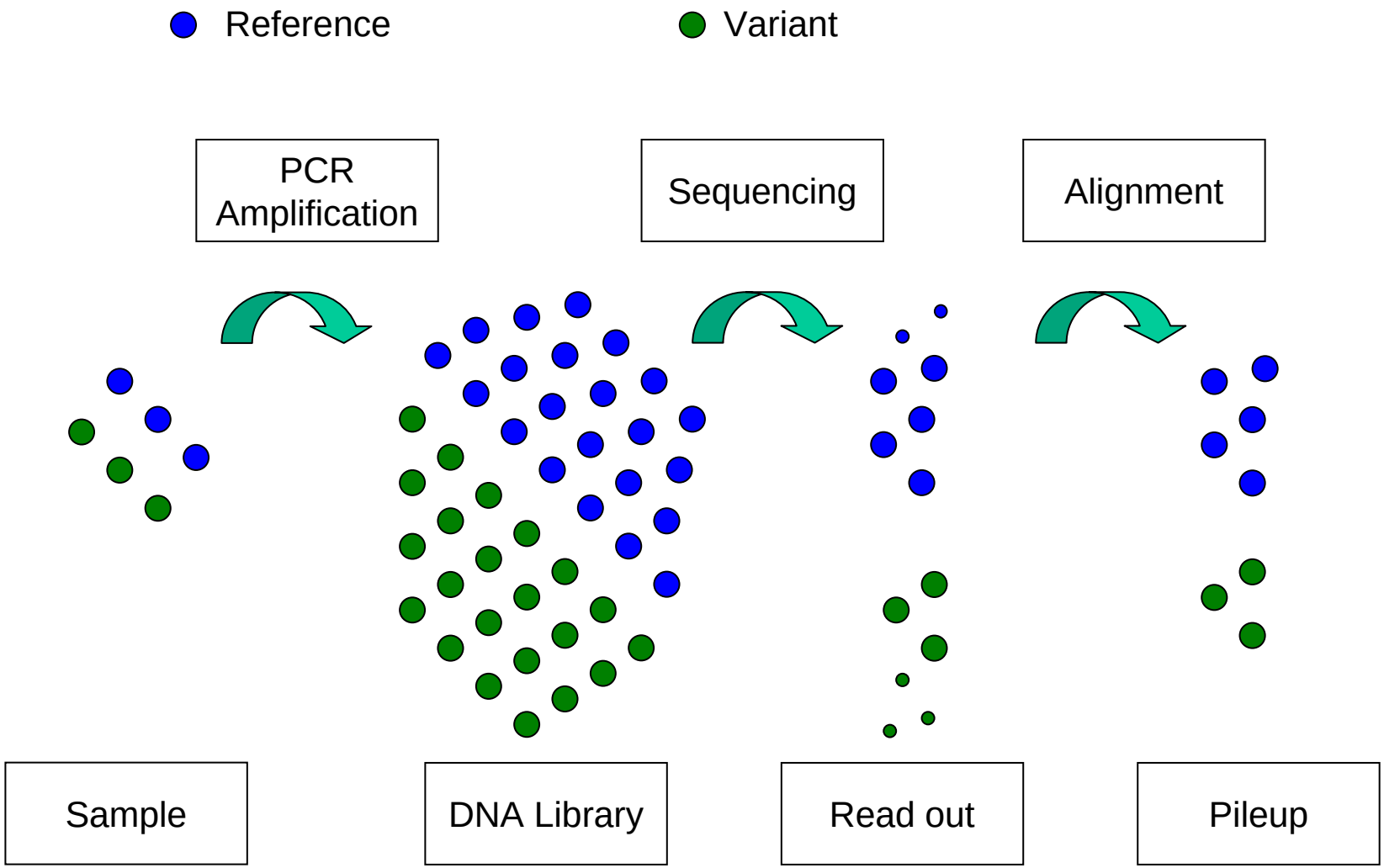
d := sequencing coverage depth

p := *allele frequency of reference*

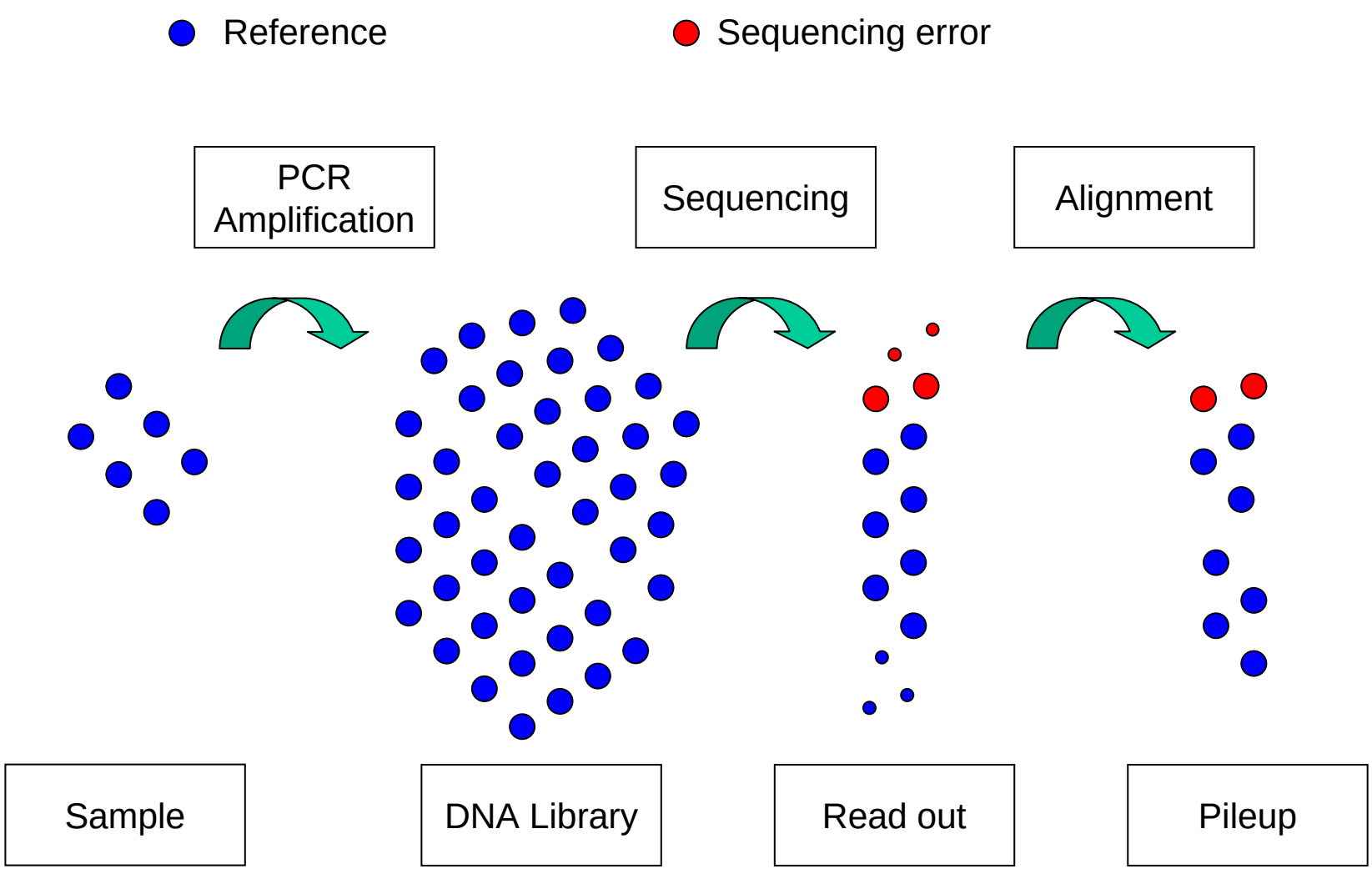
k := *reads that support reference*

http://www.bioscientia.de/de/diagnostik/humangenetik/next-generation-sequencing/

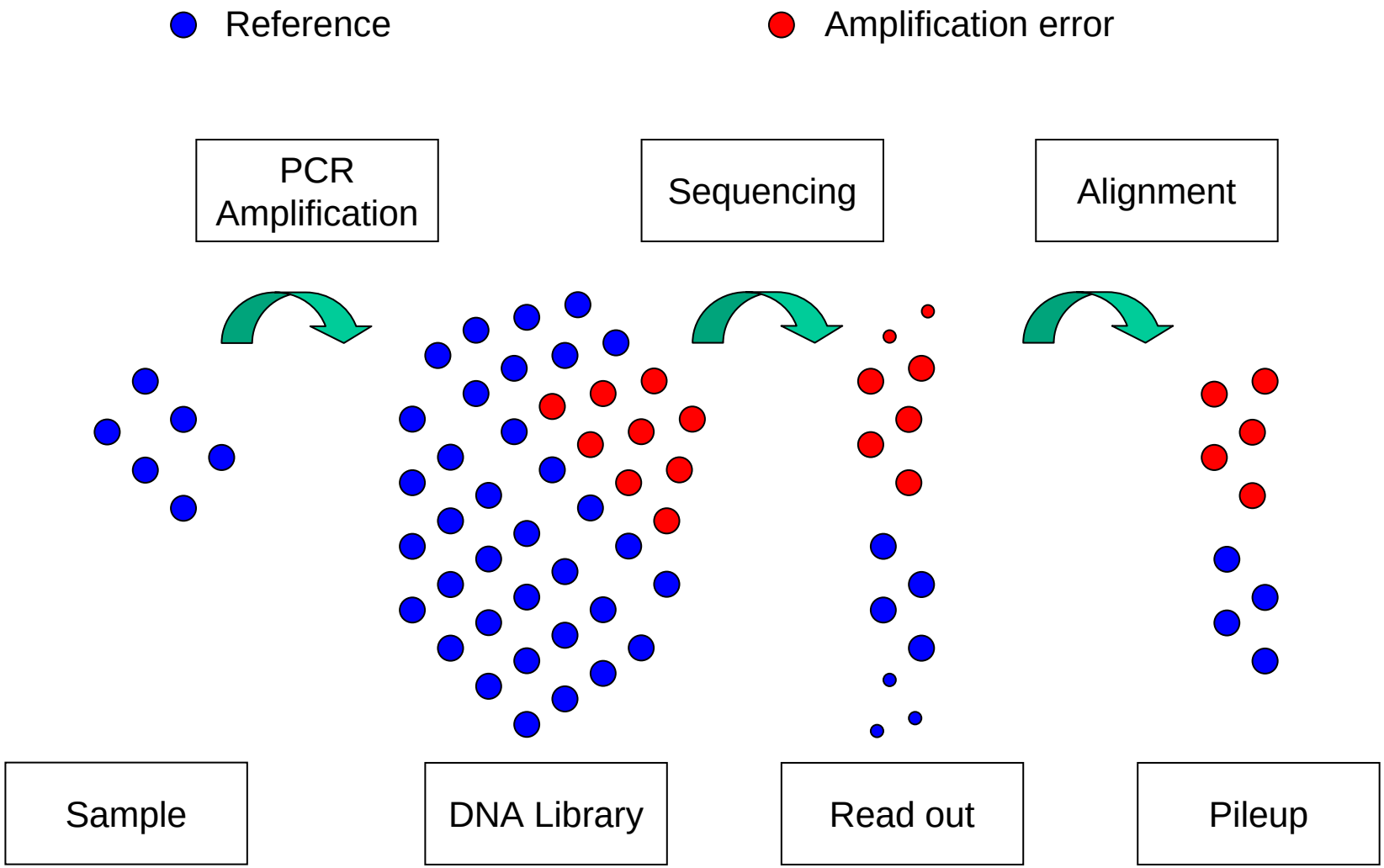
True Workflow



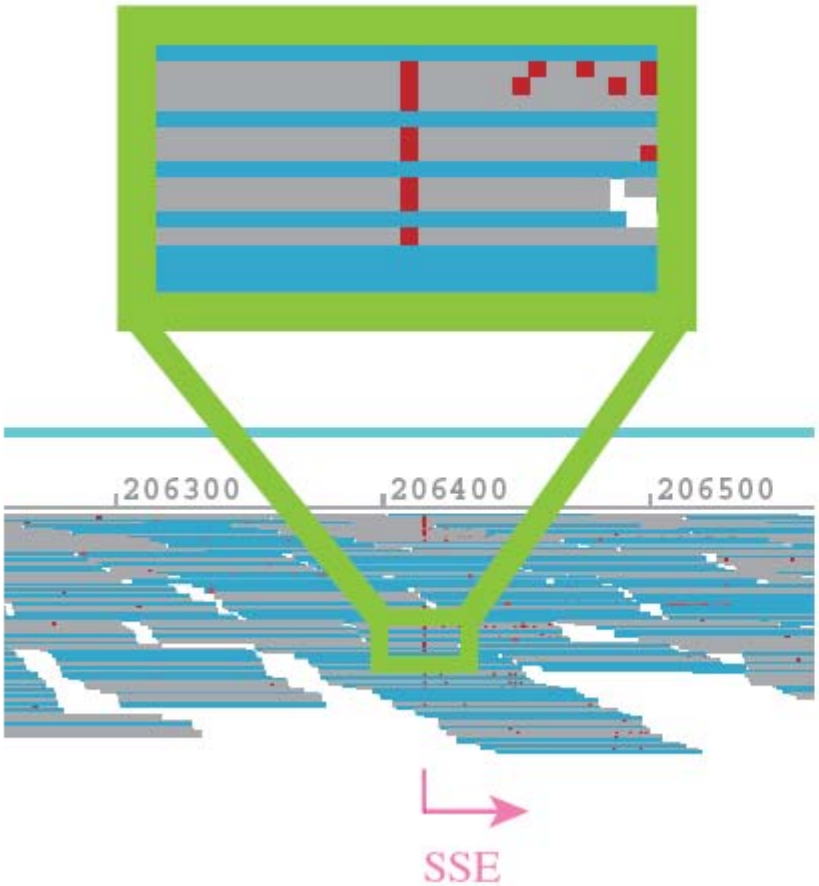
Sequencing errors and base quality

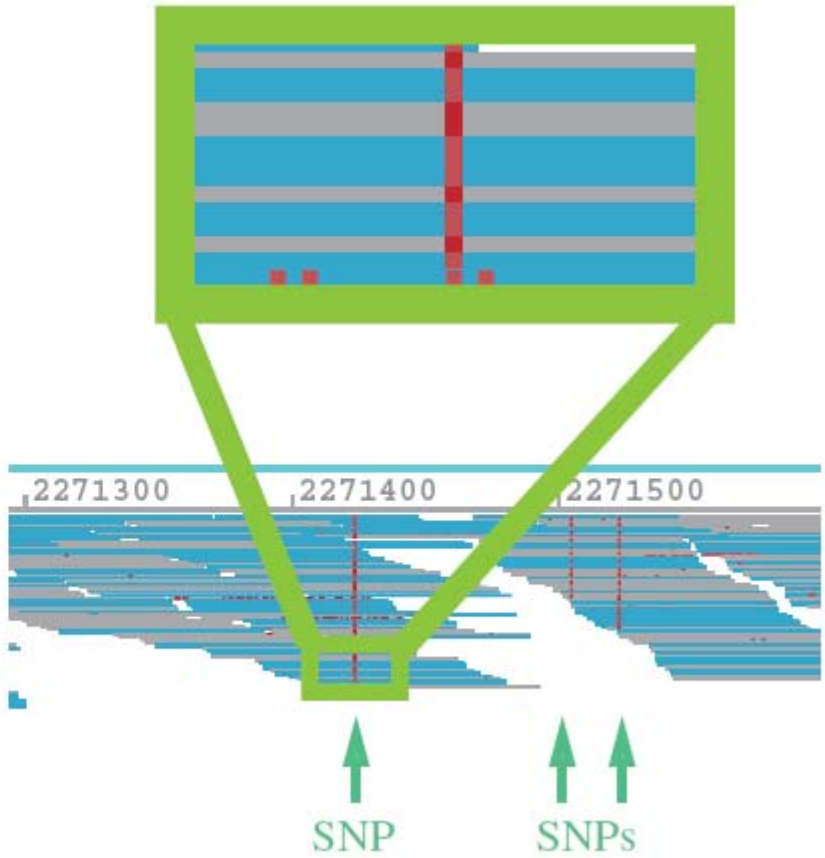


Amplification errors



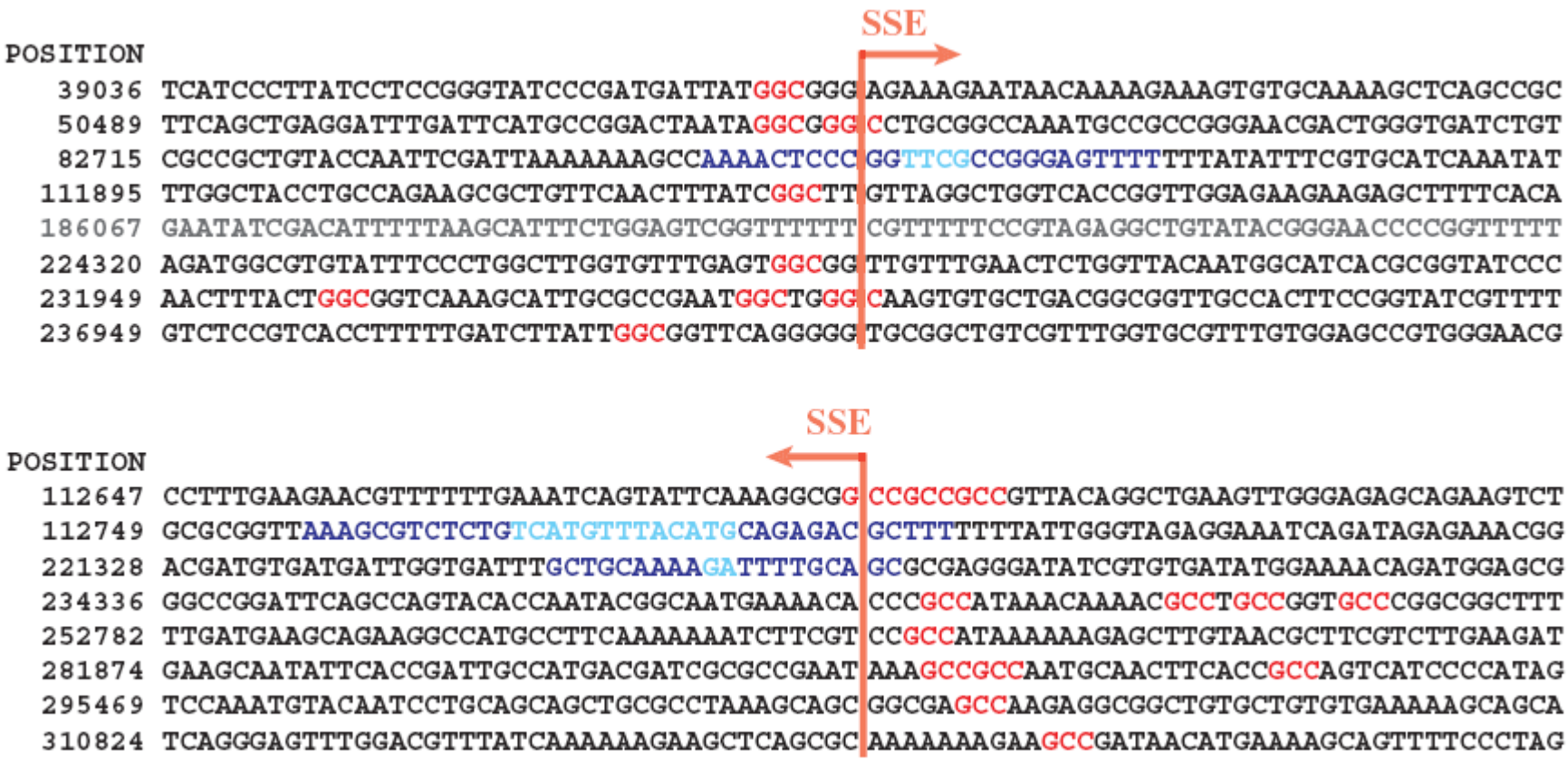
Sequence specific errors (SSE)





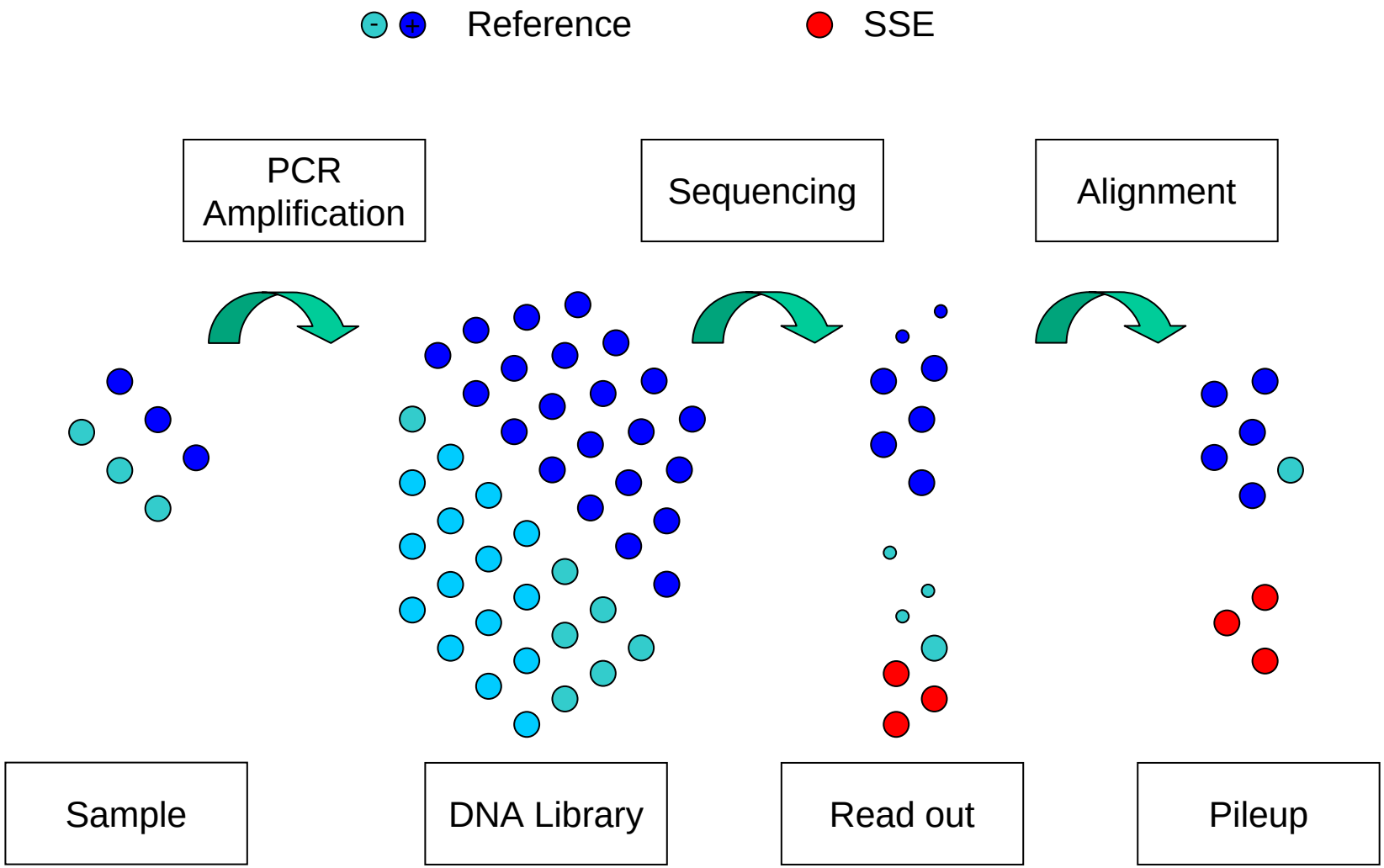
Nakamura et al., *Nucleic Acid Res* (2011)

GGC induces SSE



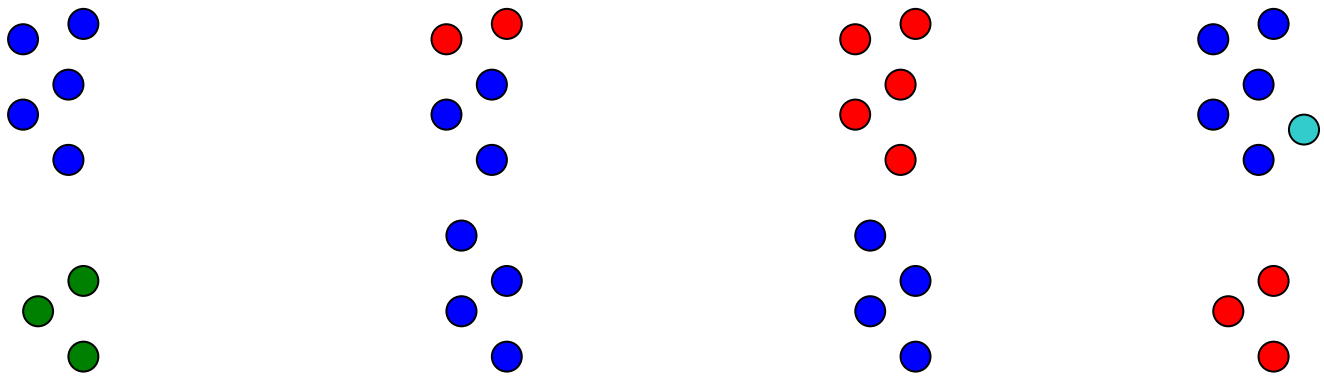
Nakamura et al., *Nucleic Acid Res* (2011)

SSE complicate variant calling



Variant calling

Search 3 million SNPs and X SNVs among 3 billion positions



3/5 2/7 5/4 3/6

Signal detection and evaluation

Tools for SNP calling

- Manual inspection is not an option
- Different statistical approaches can be used to model the sequencing process
- A number of software tools implement these strategies:
  - samtools/bcftools
  - gatk
  - varscan
  - snv-mix

Example: Samtools/bcftools

- Tool box for NGS data processing
- Samtools computes pileups from BAM files
- Bcftools calls variants

Computing genotype likelihoods

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \left[(m-g)\epsilon_j + g(1-\epsilon_j) \right] \prod_{j=l+1}^{k} \left[(m-g)(1-\epsilon_j) + g\epsilon_j \right]$$

m : ploidy

g : genotype / number of reference alleles

ϵ_j : sequencing error / 1 - base call accuracy

k : number of reads

l : reads supporting reference

k-l : reads supporting variant

$\mathcal{L}(g)$: Likelihood of g

Li, H. *Bioinformatics* (2011)

Genotype and Ploidy

Genotype	m	g	AF
AA	2	2	0.0
B0	1	0	1.0
B	1	0	1.0
BB	2	0	1.0
AB	2	1	0.5
AAB	3	2	0.67/0.33
AABB	4	2	0.5
AAAB	4	3	0.75/0.25

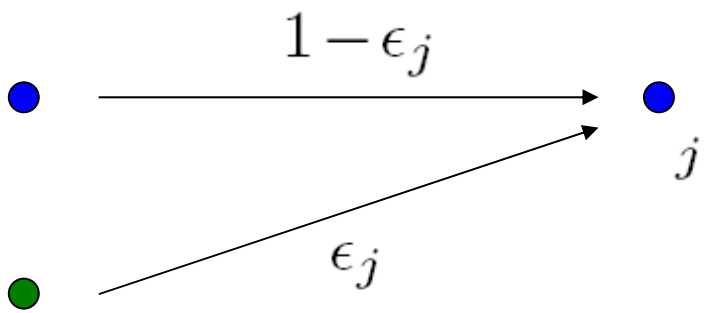
Computing genotype likelihoods

Reference Variant

$$\left[(m-g)\epsilon_j + g(1-\epsilon_j)\right]$$

m : ploidy
g : genotype

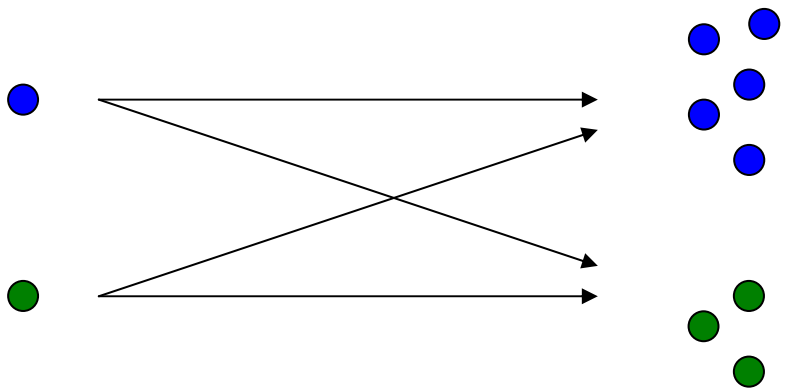
Heterozygous SNP



Li, H. *Bioinformatics* (2011)

Computing genotype likelihoods

$$\prod_{j=1}^{l} \left[(m-g)\epsilon_j + g(1-\epsilon_j) \right]$$

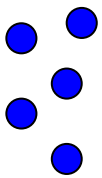


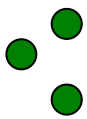
$$\prod_{j=l+1}^{k} \left[(m-g)(1-\epsilon_j) + g\epsilon_j \right]$$

Li, H. *Bioinformatics* (2011)

Example

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \left[(m-g)\epsilon_j + g(1-\epsilon_j) \right] \prod_{j=l+1}^{k} \left[(m-g)(1-\epsilon_j) + g\epsilon_j \right]$$



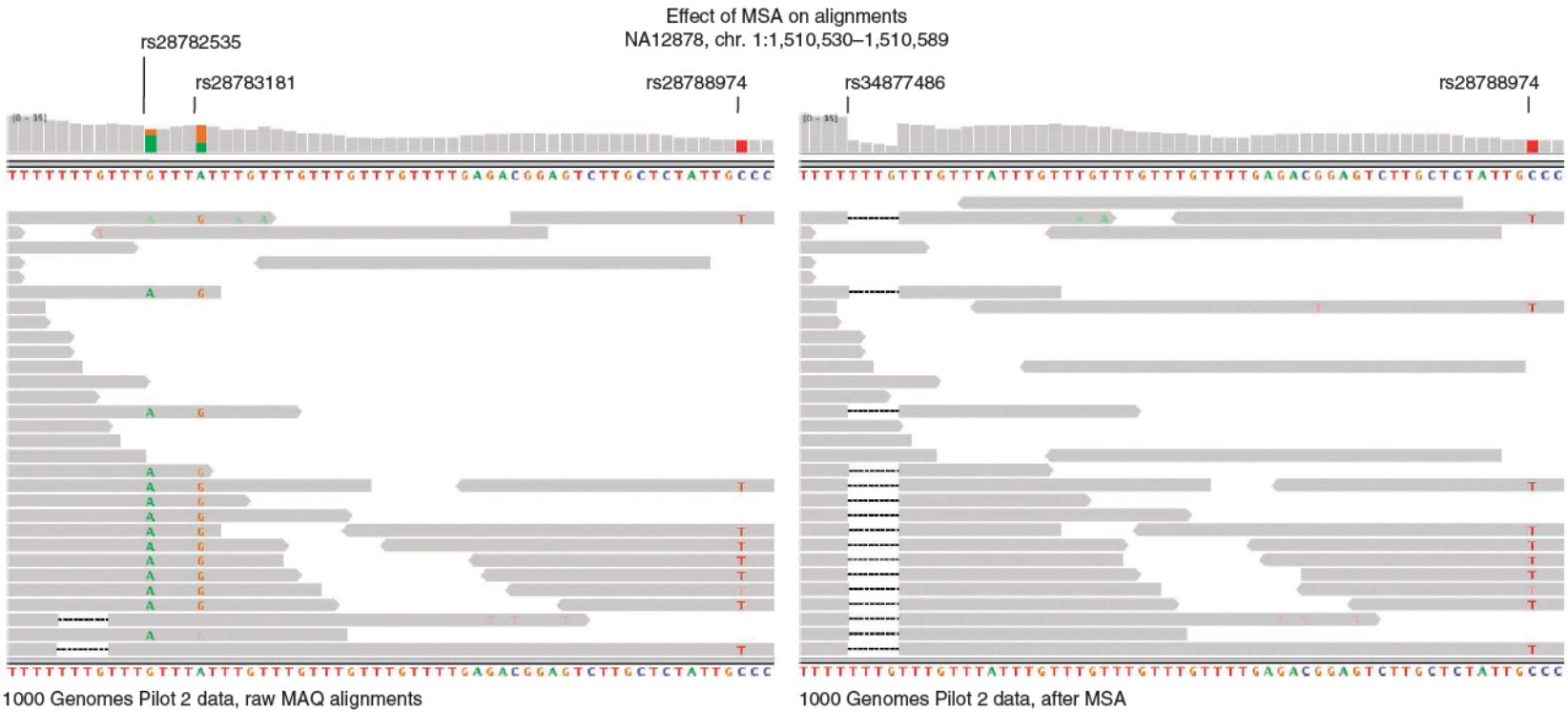


3/5

$\mathcal{L}(g)$	ϵ_j =10%	ϵ_j =1%	ϵ_j=0.1%
g=AA	5.9 e10^{-4}	9.5 e10^{-7}	1.0 e10^{-10}
g=AB	0.39 %	0.39 %	0.39 %
g=BB	7.3 e10^{-6}	9.7 e10^{-11}	1.0 e10^{-16}

Realignment

- Reads that span indels are harder to align
- Alignment heuristics that generate initial BAM files trade accuracy for speed
- Locally realigning reads can improve alignment quality

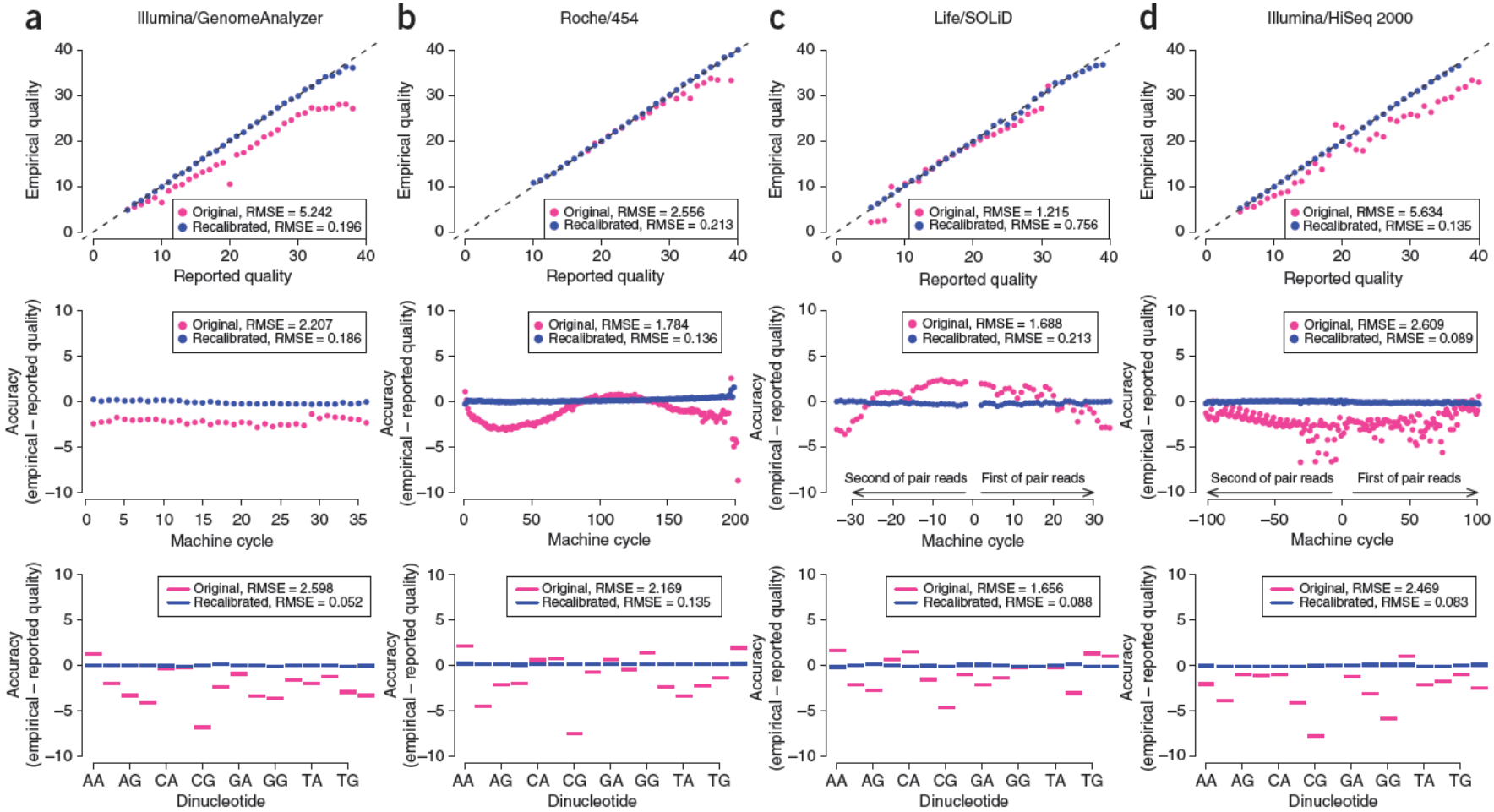


DePristo et al. Nature Genetics (2011)

Recalibration

- Native base quality score is inaccurate
- It co-varies with:
  - Sequencing technology
  - Machine cycle
  - Sequence context
- By recalibration this bias can be corrected

Recalibration

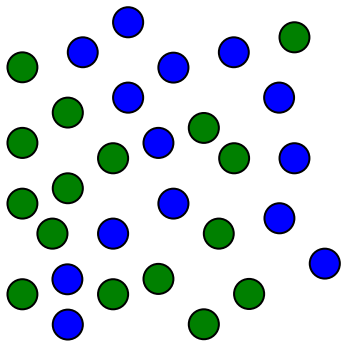


DePristo et al. Nature Genetics (2011)

SNV calling and the false negative rate

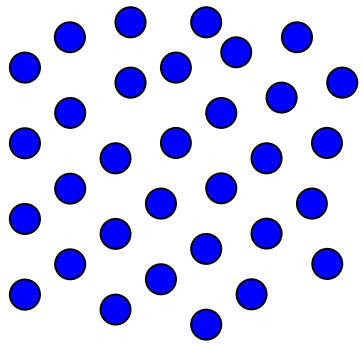
- Detection of somatic mutations is more complicated then SNP calling
- Two sample approach
  - Tissue sample
  - Germline sample
- Tissue samples often contain germline cells

Tumor



Heterozygous SNV

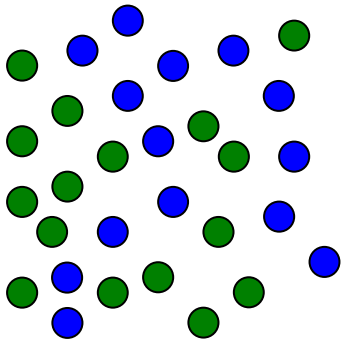
Germline



Reference

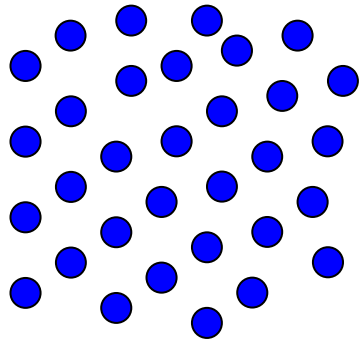
Cell mixtures

Tumor

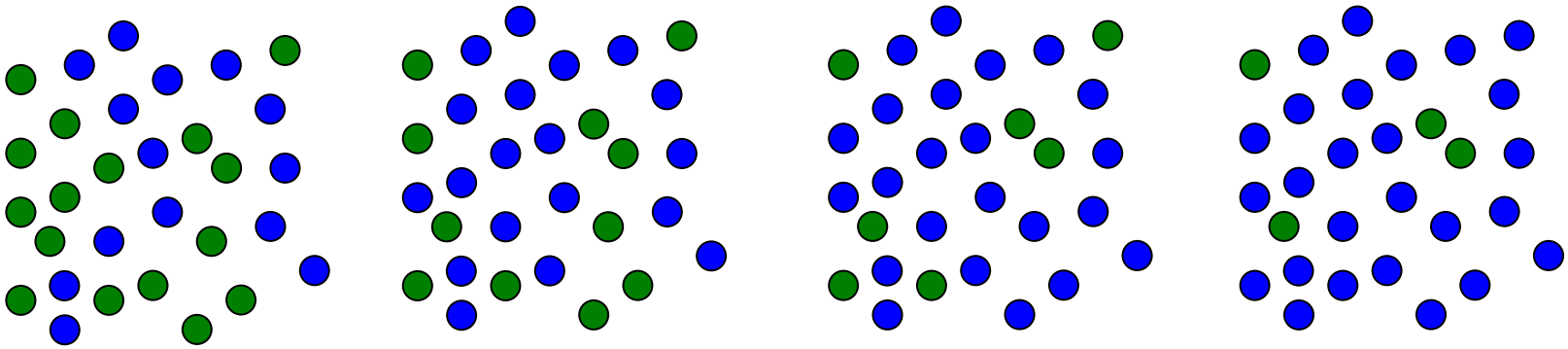


Heterozygous SNV

Germline

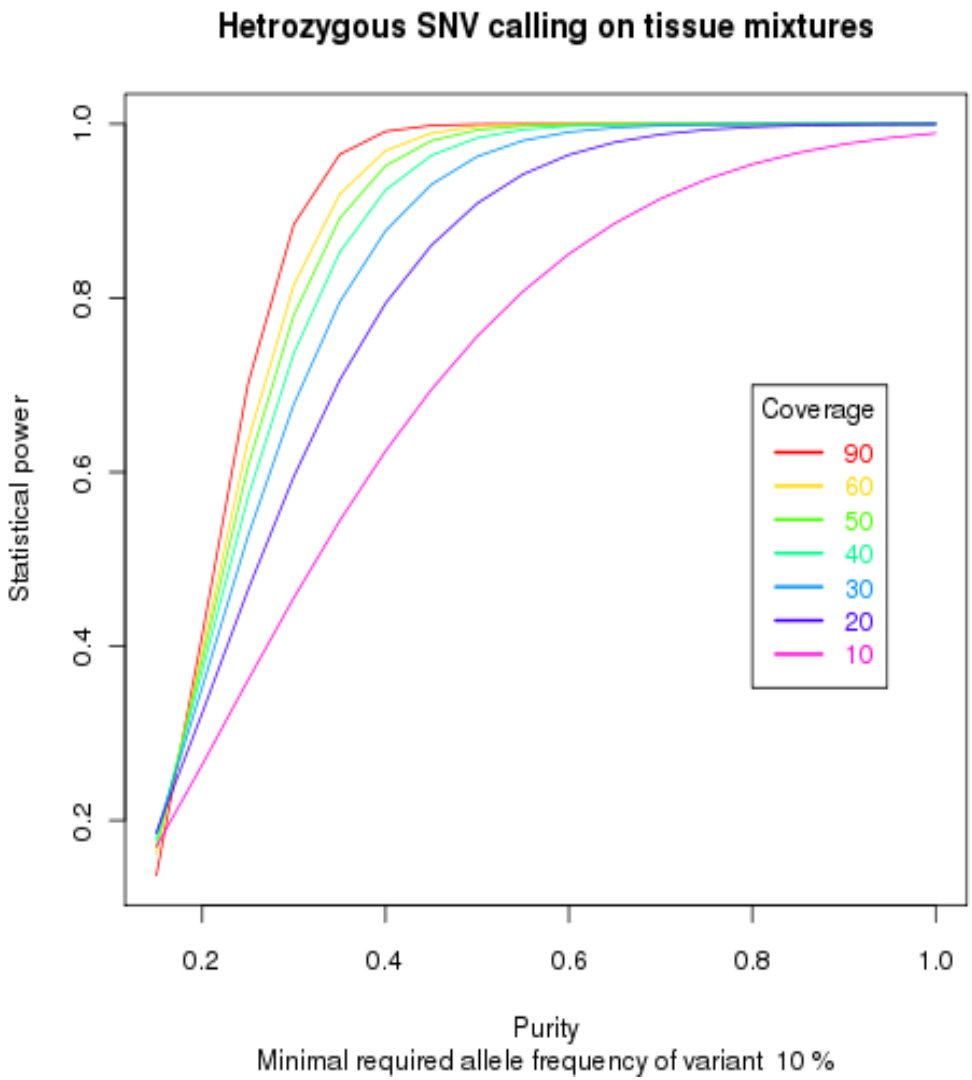


Reference



Purity: 100 % 75 % 50% 25%

Upper bound to statistical power



Model definition

c := tumor cellularity / purity

d := sequencing coverage depth

a := minimal allele frequency of variant

p := *c* · 0.5 (fraction hetrozygous variant alleles)

k := *a* · *d* (minimal required observations for call)

$$P(call) = 1 - \sum_{i=0}^{k-1} \binom{d}{i} p^i (1-p)^{d-i}$$

Summary

SNVs define differences between healthy and diseased cells

SNPs define differenecs between individuals

Both can be directly identified from NGS data

The ~3 billion repetions of the variation calling can amplify small errors to big effects

SNV calling in diseased cells is complicated by cell mixtures

SNV/SNP calling still relies on human inspection of data

References

Heng, Li

"**A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data**"

Bioinformatics (Advanced access) Sep. 2011

Nakamura, K. *et al.*

"**Sequence-specific error profile of Illumina sequences**"

Nucleic Acids Res 39 (13), e90 (2011)

De Pristo, M.A.; Banks, E.; Popolin, R.; Garimella K.V. *et al.*

"**A framework for variation discovery and genotyping using next-generation DNA sequencing data**"

Nature Genetics (2011) 43(5)