ECCB‘12 Tutorial 4

Inferring genetic diversity from next-generation sequencing data: Computational methods and biomedical applications

Niko Beerenwinkel
Karin Metzner
Volker Roth

Basel, September 9, 2012

About us

Niko Beerenwinkel
ETH Zurich

Computational Biology

Karin Metzner
University Hospital Zurich

Virology

Volker Roth
University of Basel

Computer Science

Goals

- learn about important **biomedical applications** of genetic diversity estimation
  - cancer: intra-tumor diversity
  - infectious diseases: intra-patient viral diversity
- understand the tasks of **SNV calling** and **haplotype reconstruction**
- appreciate the **opportunities and limitations of different NGS technologies** for diversity estimation
- survey existing approaches and **software**
- understand the **basic computational and statistical principles** underlying haplotype inference

Schedule

9:00	**Introduction, Motivation, Case Studies, NGS Technology**
10:30	*Coffee break*
11:00	**Local Diversity Estimation**
12:30	*Lunch*
13:30	**Global Diversity Estimation**
15:00	*Coffee break*
15:30	**Comparative Assessment of Methods, Demonstration of Case Studies**
17:00	*End of workshop*






Main references

- Beerenwinkel N and Zagordi O (2011). Ultra-deep sequencing for the analysis of viral populations. *Current Opinion in Virology* 1:413–418.
  doi: 10.1016/j.coviro.2011.07.008

- Beerenwinkel N, Günthard HF, Roth V and Metzner KJ (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology* 3:329.
  doi: 10.3389/fmicb.2012.00329

ECCB 2012 Tutorial 4

Introduction: Motivation, Case Studies, NGS Technology

Karin J. Metzner

Division of Infectious Diseases and Hospital Epidemiology

University of Zurich[UZH]

UniversityHospital Zurich

Topics

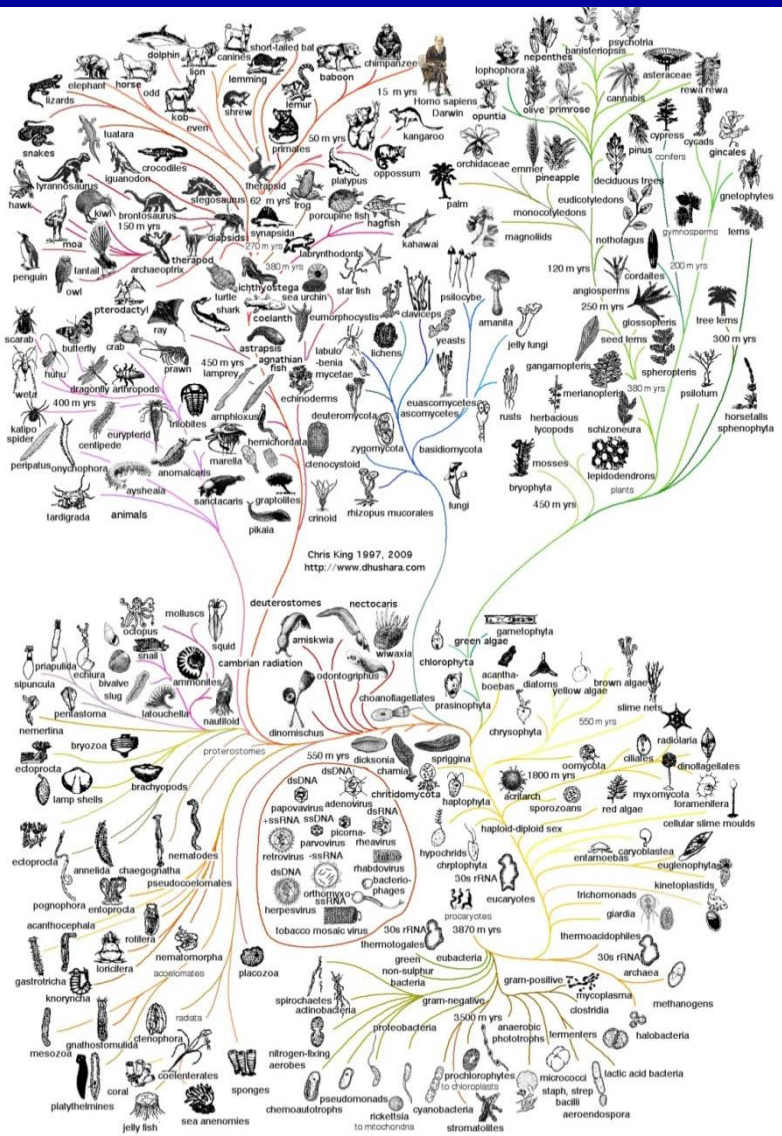
- genetic diversity
  - definitions and motivation
- case studies
  - cancer
  - viruses
- NGS technologies
  - techniques (mainly 454 and Illumina)
  - error pattern and quality scores

Topics

- genetic diversity
  - definitions and motivation
- case studies
  - cancer
  - viruses
- NGS technologies
  - techniques (mainly 454 and Illumina)
  - error pattern and quality scores

The global view on diversity

The tree of life

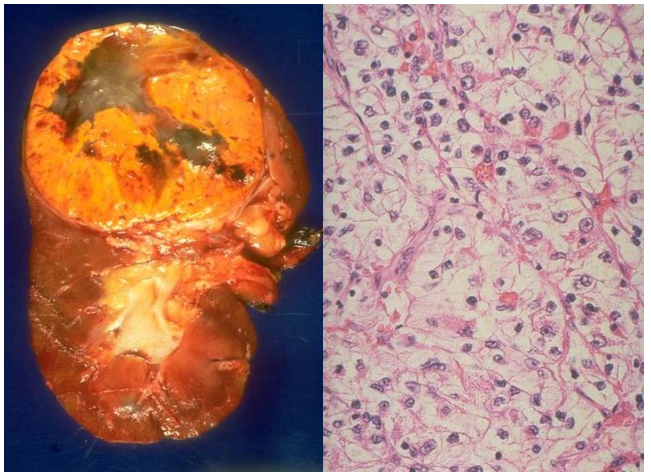


Taxonomy

domain	eukaryote	unassigned
kingdom	animal	unassigned
phylum	chordate	unassigned
class	mammalia	unassigned
order	primate	unassigned
family	hominidae	retroviridae
genus	homo	lentivirus
species	homo sapiens	HIV-1 (example 2: in one host)
individuum	example 1	

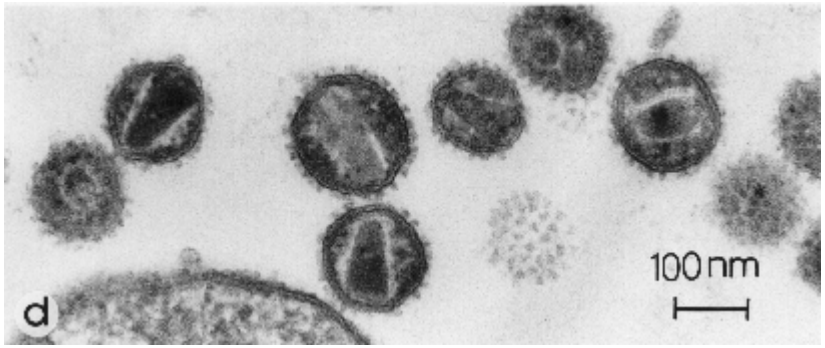
Case studies/examples

- Genetic diversity in an individuum:
  intra-tumour diversity



Copyright 2004-2012 University of Washington, UW Medicine Pathology

- Genetic diversity of a species in a defined environment:
  intra-patient diversity of HIV-1



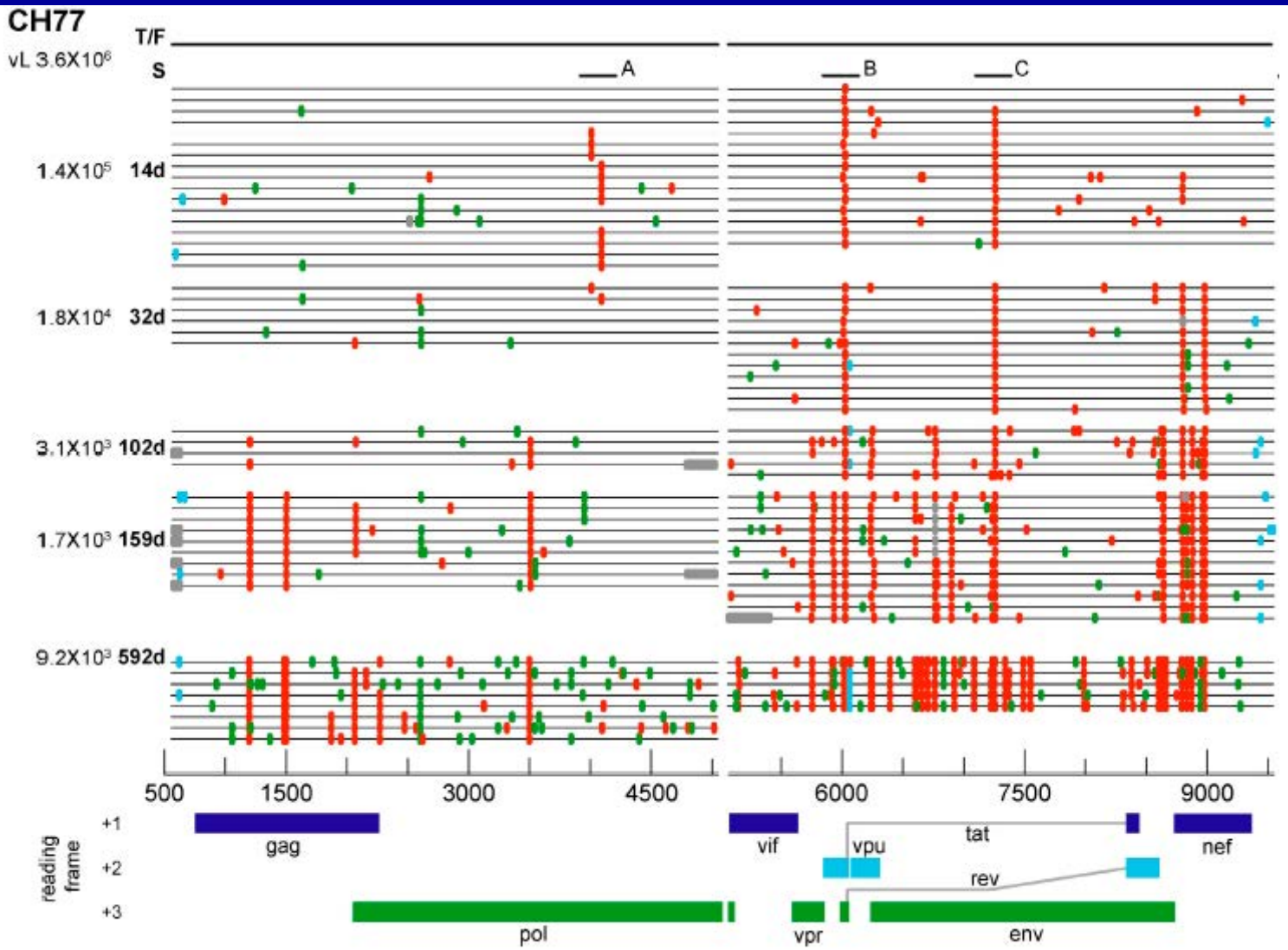
HR Gelderblom *et al*., Virology 1987

Genetic diversity _ divergence

- Genetic diversity:
  genetic characteristics at a certain time point

- Genetic divergence:
  changes in genetic diversity over time

HIV-1 population dynamics within a host

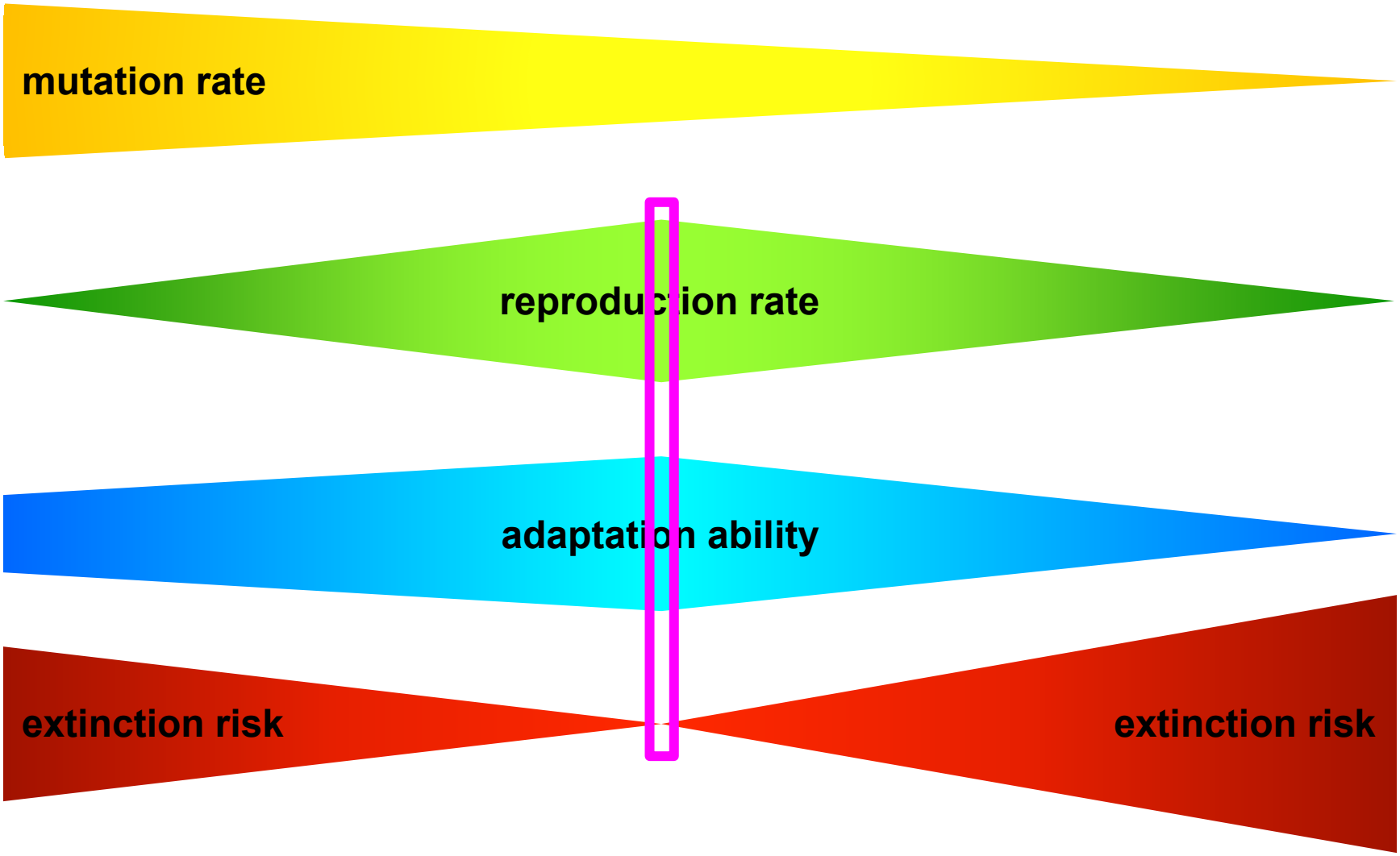


JF Salazar-Gonzalez *et al*., JEM 2009

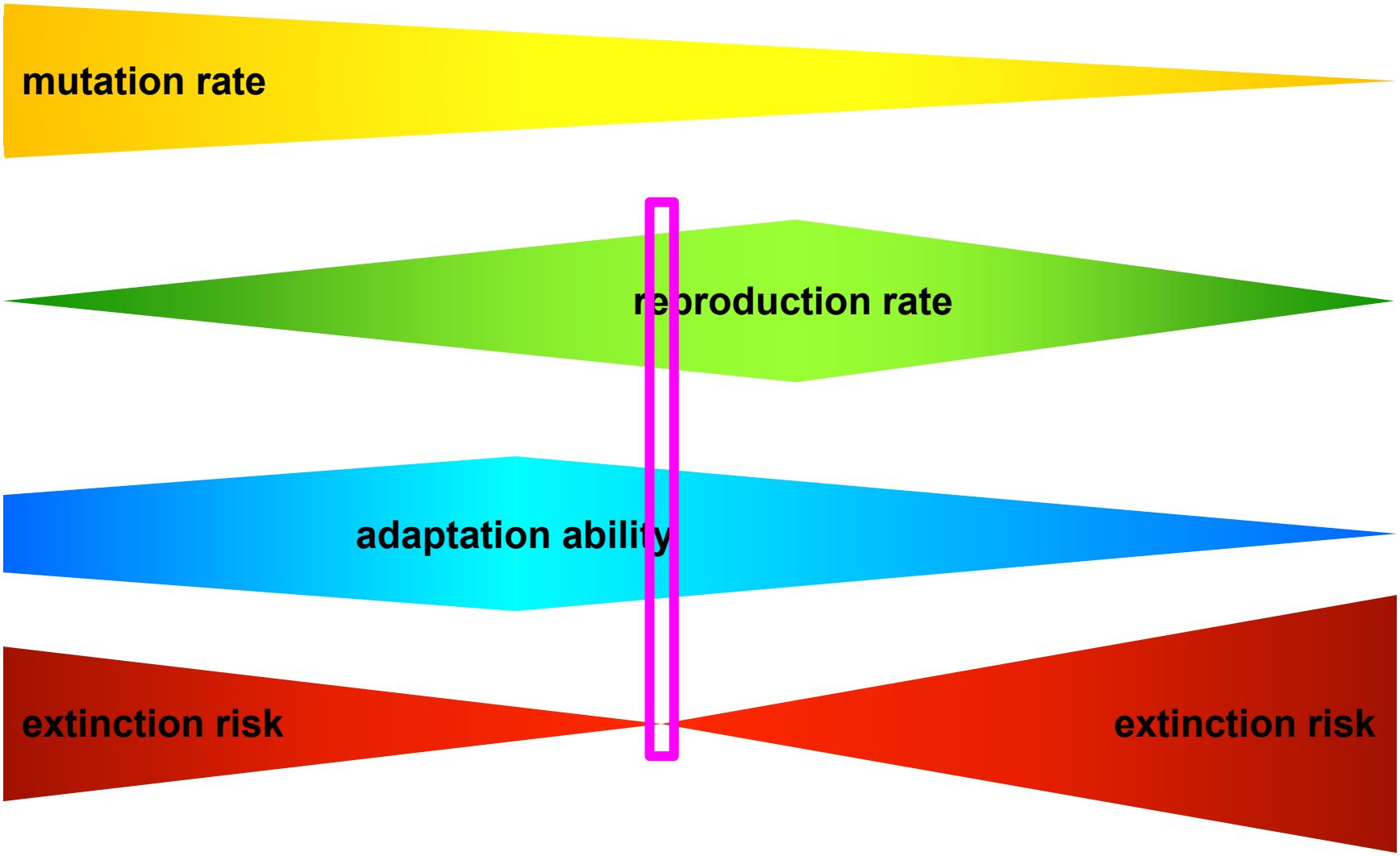
The needs of the many

M Kohn, Nature 2008

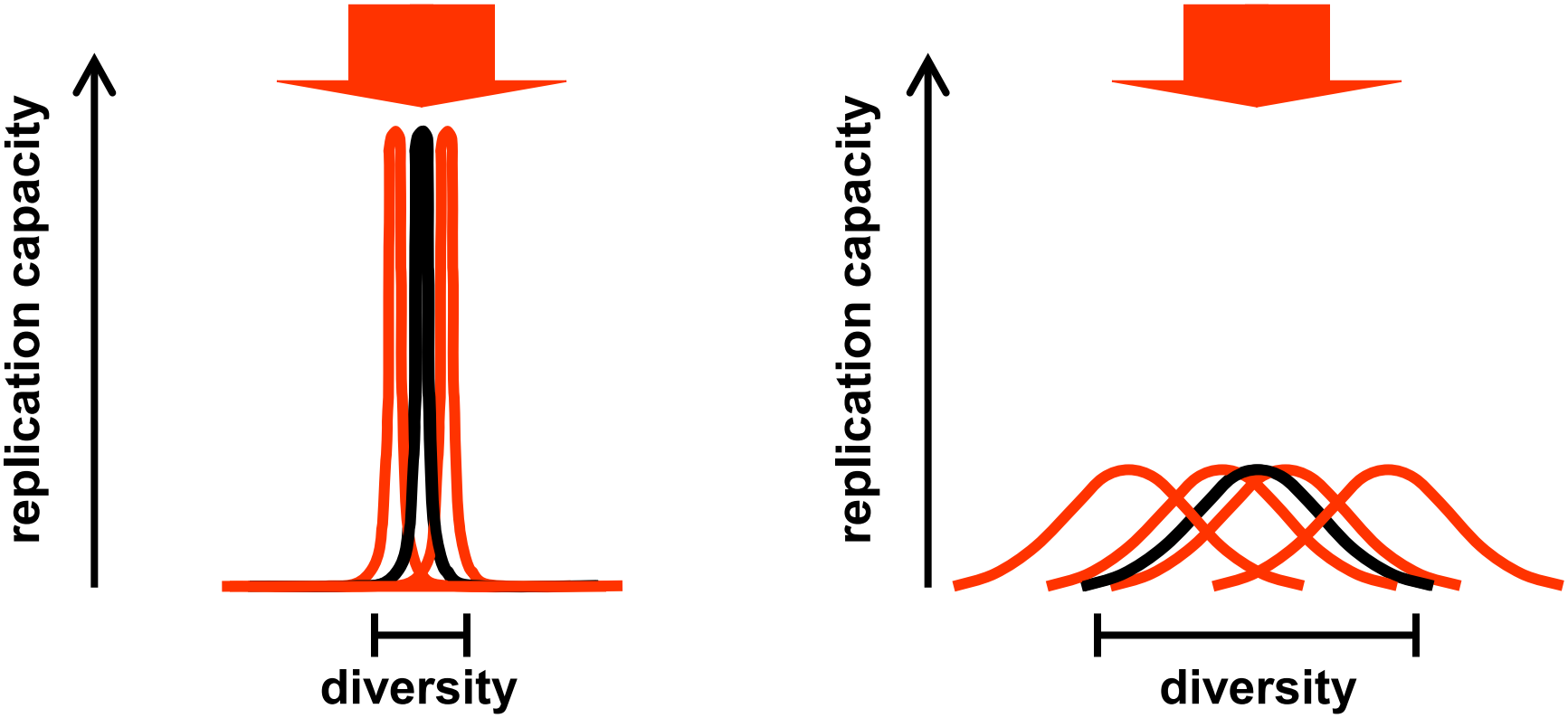
Chances and risks of evolution



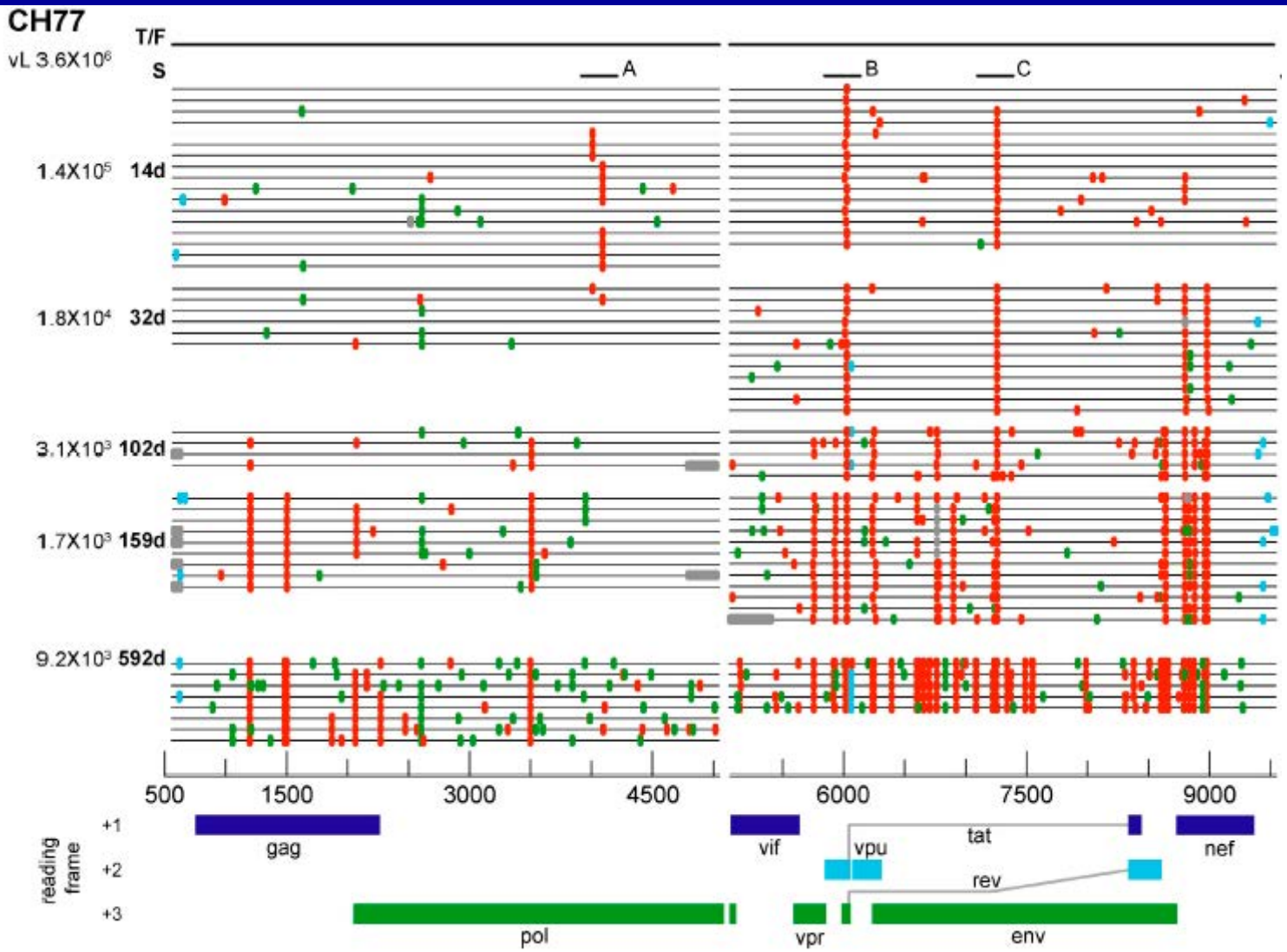
Chances and risks of evolution



Survival of populations



HIV-1 population dynamics within a host mainly driven by immune responses

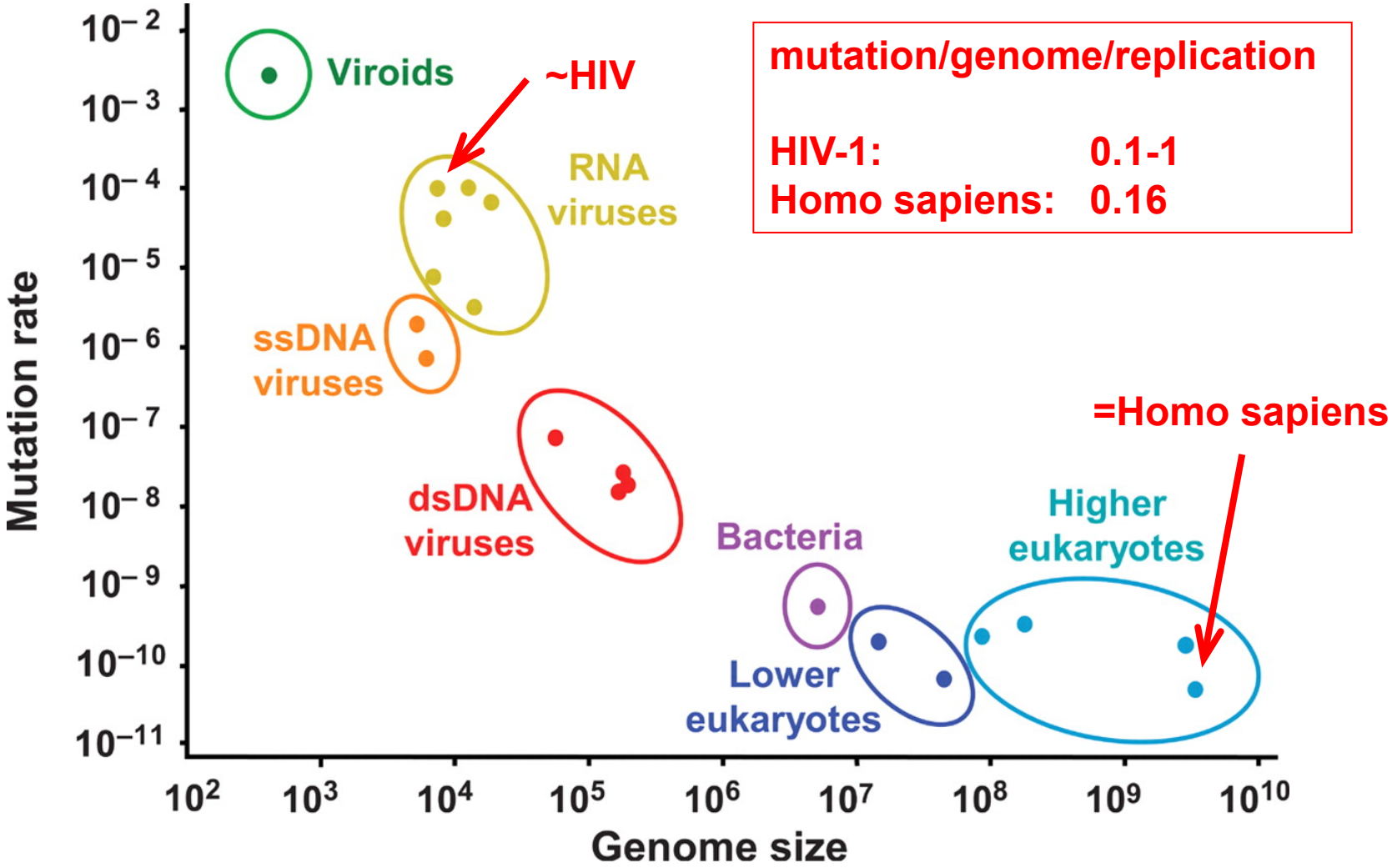


JF Salazar-Gonzalez *et al.*, JEM 2009

Viral quasispecies

- quasispecies model of molecular evolution
  (M Eigen and P Schuster, 1979)
- selection pressure on the whole population rather than on single individuums
- viral quasispecies = viral population = mutant cloud = swarm
  → all virus variants within one host interconnected by mutations
- virus variant = viral haplotype

Mutation rate correlates to genome size



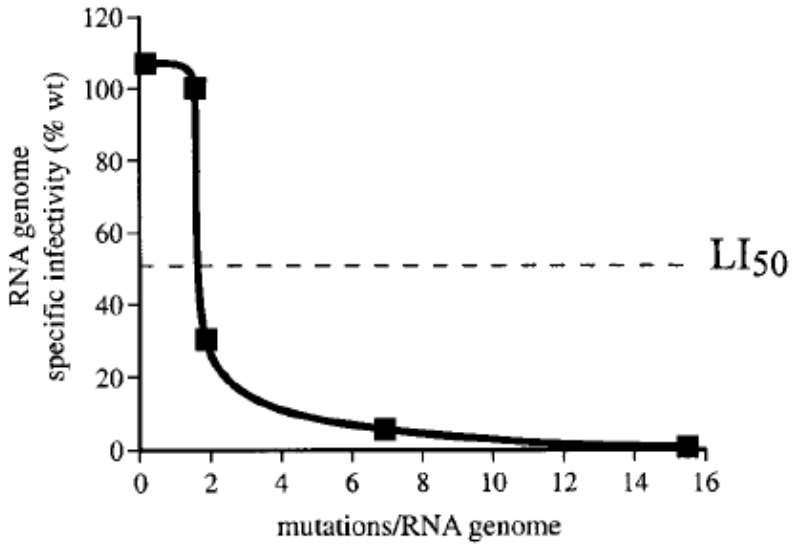
mutation/genome/replication

HIV-1:	**0.1-1**
Homo sapiens:	**0.16**

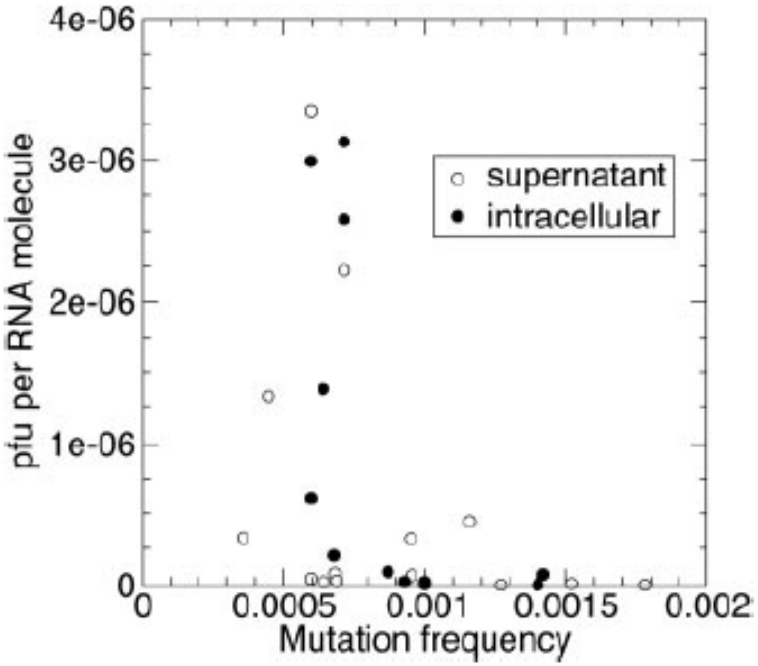
adapted from S Gago *et al.*, Science 2009

Error threshold

poliovirus



cytomegalovirus



S Crotty *et al*., PNAS 2001; A Grande-Pérez *et al*., PNAS 2005

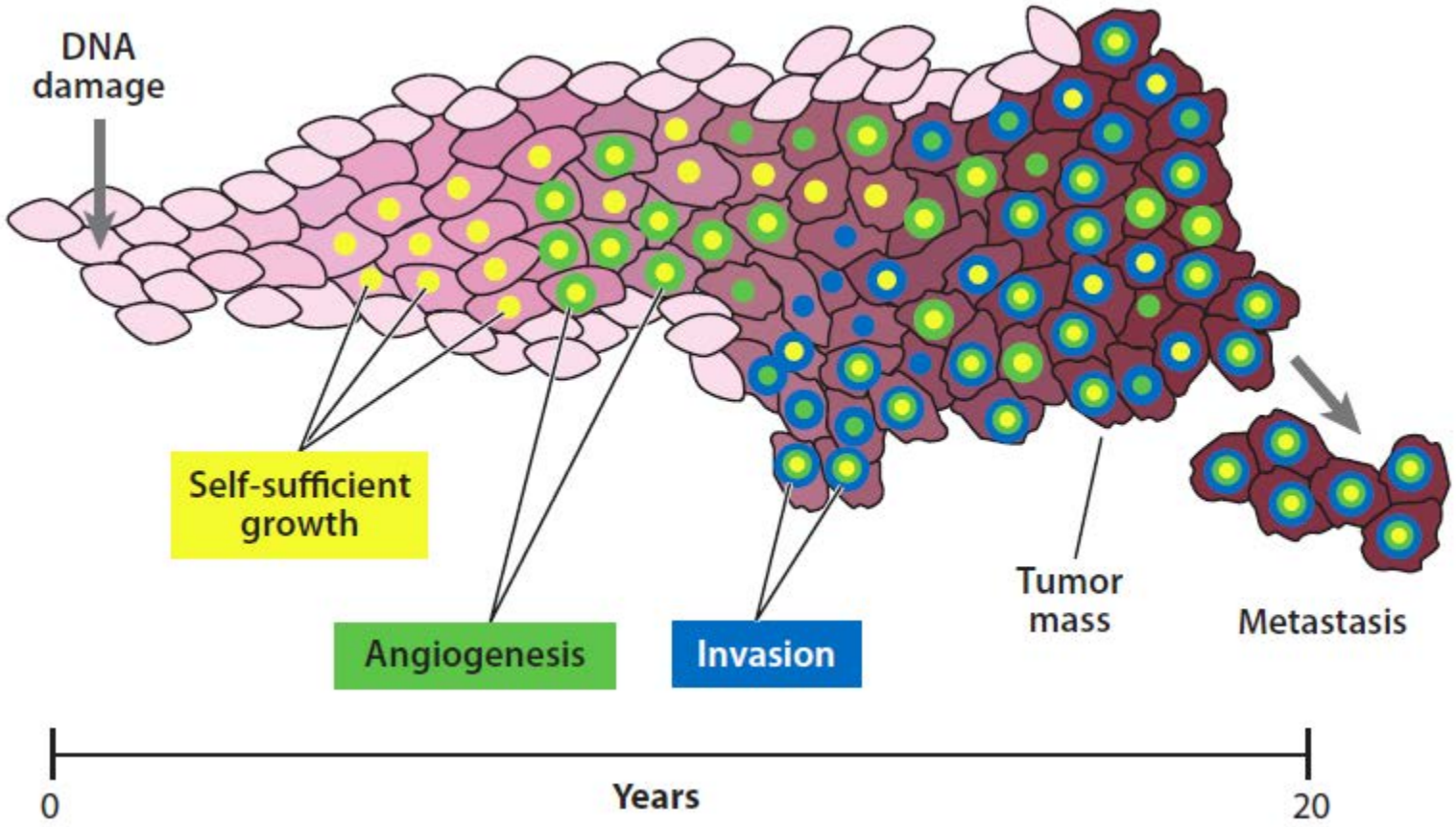
Topics

- genetic diversity
  - definitions and motivation
- case studies
  - cancer
  - viruses
- NGS technologies
  - techniques (mainly 454 and Illumina)
  - error pattern and quality scores

Case studies/examples

- Genetic diversity in tumors: Detecting low-frequency single-nucleotide variants (SNVs)
- Genetic diversity in virus populations: Local and global haplotype reconstruction

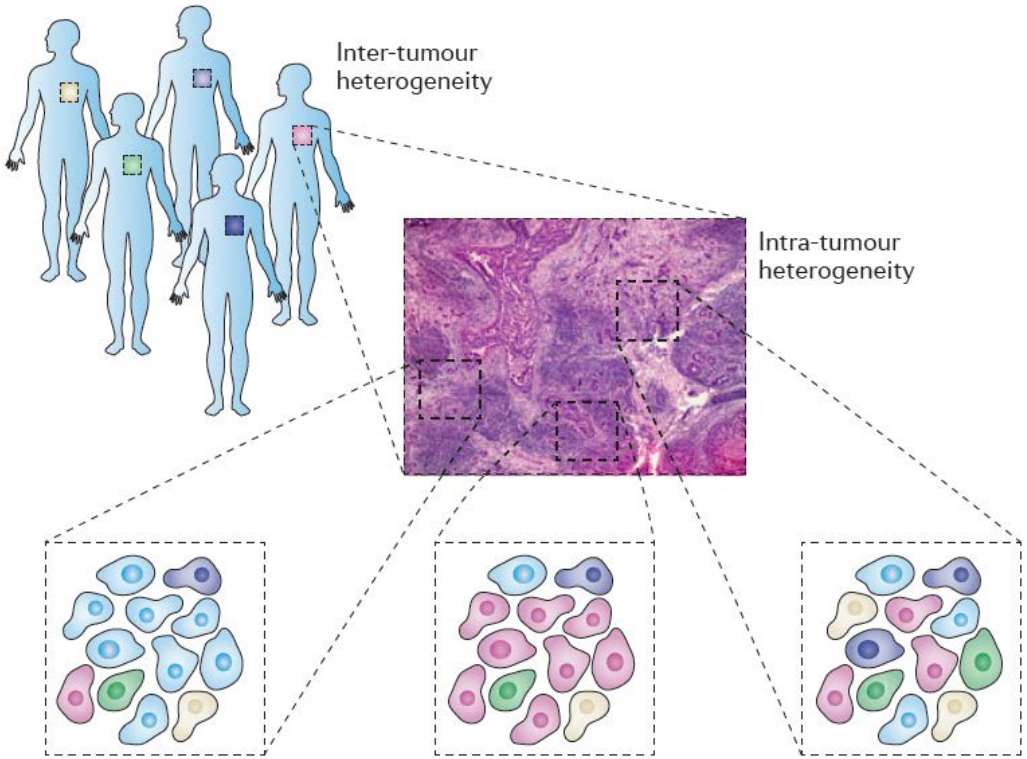
Cancer is a somatic evolutionary process



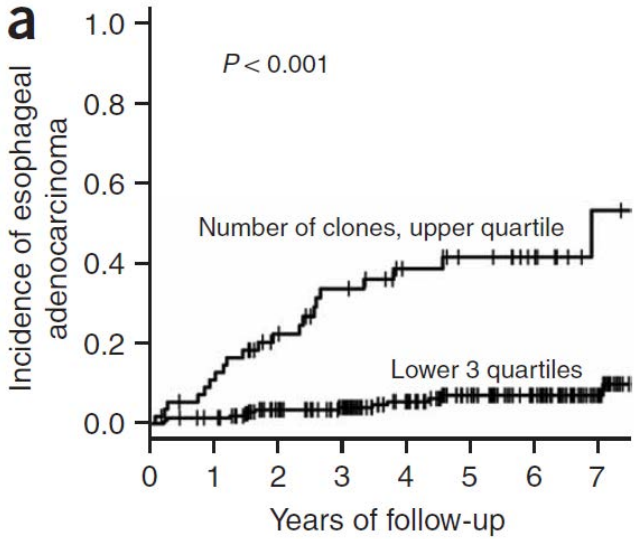
JJ Salk *et al*., Annu Rev Pathol 2010

Intra-tumor diversity

Diagnostics



Prognostics



→ Personalized medicine

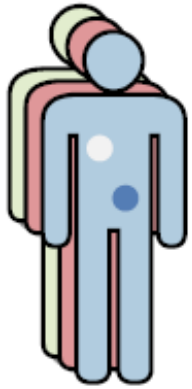
A Marusyk *et al.*, Nat Rev Cancer 2012, CC Maley *et al.*, Nat Genet 2006

Intra-patient genetic diversity of tumors

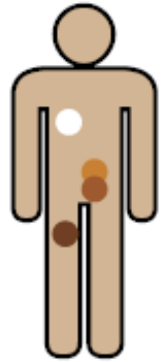
- Evolutionary dynamics
  - mutation rate can be elevated (genetic instability)
  - high turn-over
  - large population size
- Disease progression
- Drug resistance

Cancer: Case study

- renal cell cancer
  - Three matched tumour-normal samples
  - one case with biopsies from multiple lesions
- Illumina genome sequencing
- detection of low-frequency single-nucleotide variants (SNVs)



3x Tumour-normal:
Tumour 1
Tumour 2
Tumour 3



1x Multiple lesions
Primary 1
Primary 2
Metastasis

M Gerstung *et al*., Nat Comm 2012

Topics

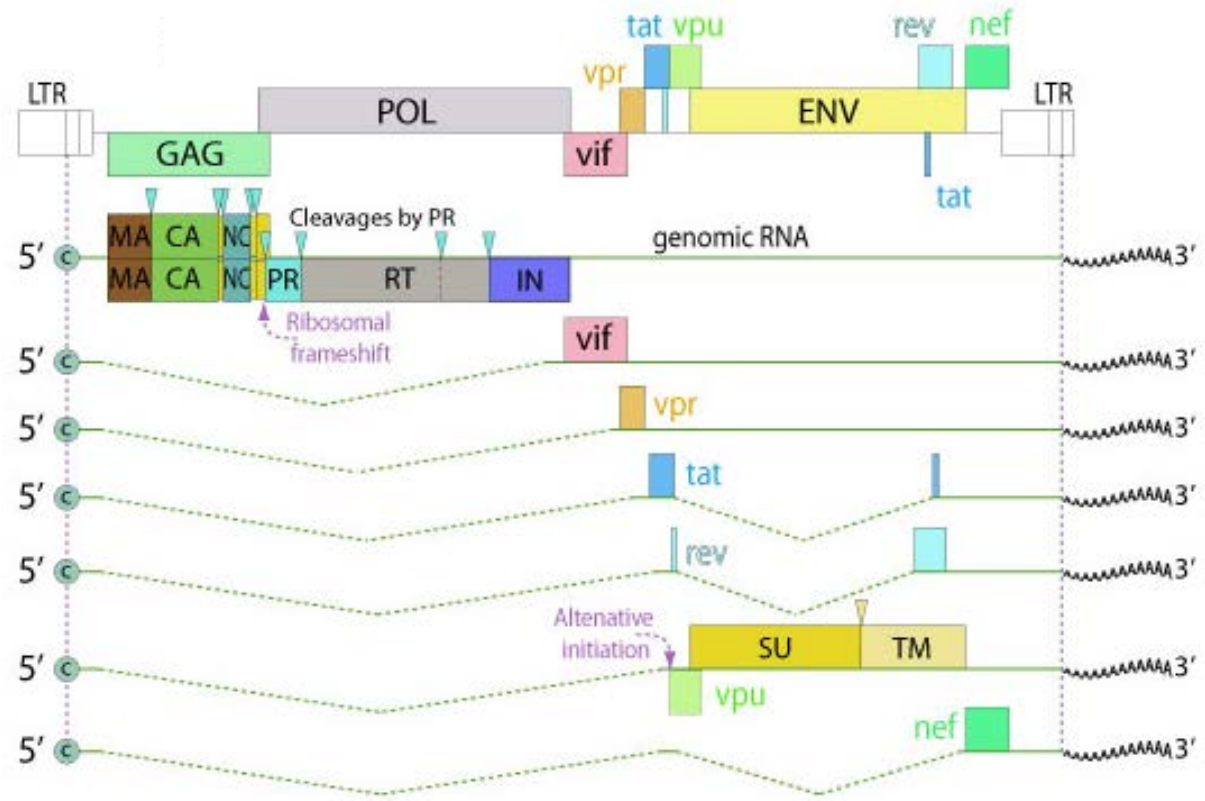
- genetic diversity
  - definitions and motivation
- case studies
  - cancer
  - viruses
- NGS technologies
  - techniques (mainly 454 and Illumina)
  - error pattern and quality scores

Advantages of viruses to study evolution

- low genome size
- rapid replication cycles
- broad knowledge available in viral proteins, their functions, pathogenesis, virus-host interactions, etc.
- established cell culture and animal models available
- huge amount of sequence data

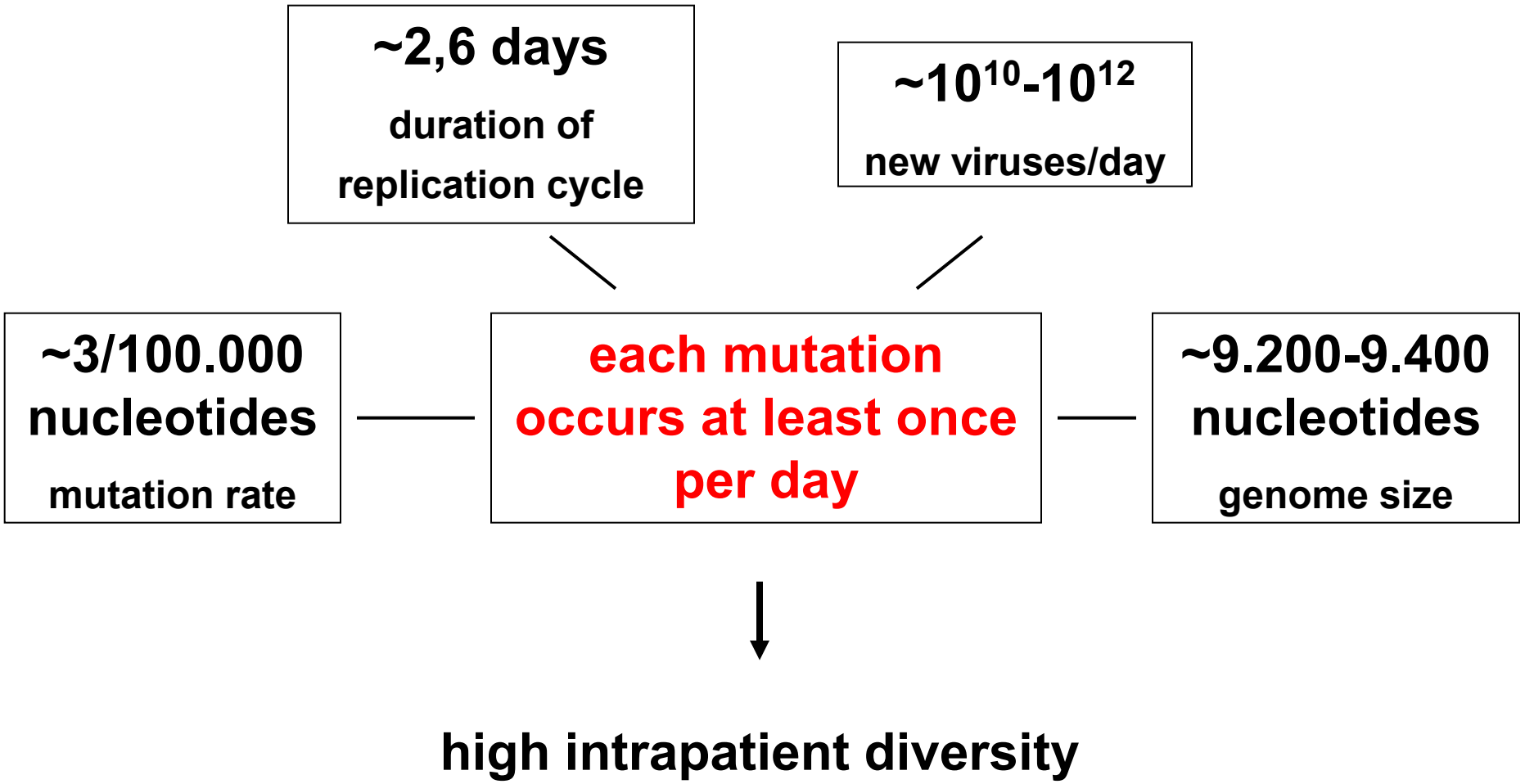
HIV-1 genome

- single-stranded, plus-sense RNA genome
- organized in a highly sophisticated way



ViralZone www.expasy.org/viralzone, Swiss Institute of Bioinformatics

Evolution and diversity of HIV

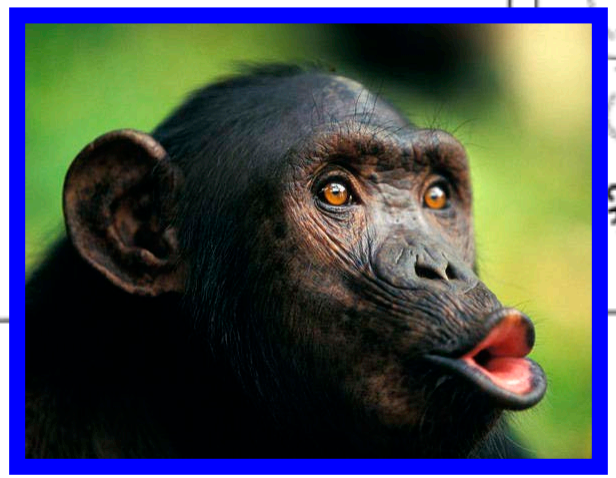


The natural variability of HIV-1 is extensive

Influenza Sequences
Hemagglutinin (H3), 1997
n = 96

HIV-1 Single
Subtype B, n
6 yrs post ser

Multiple Individuals
ngo, 1997
3

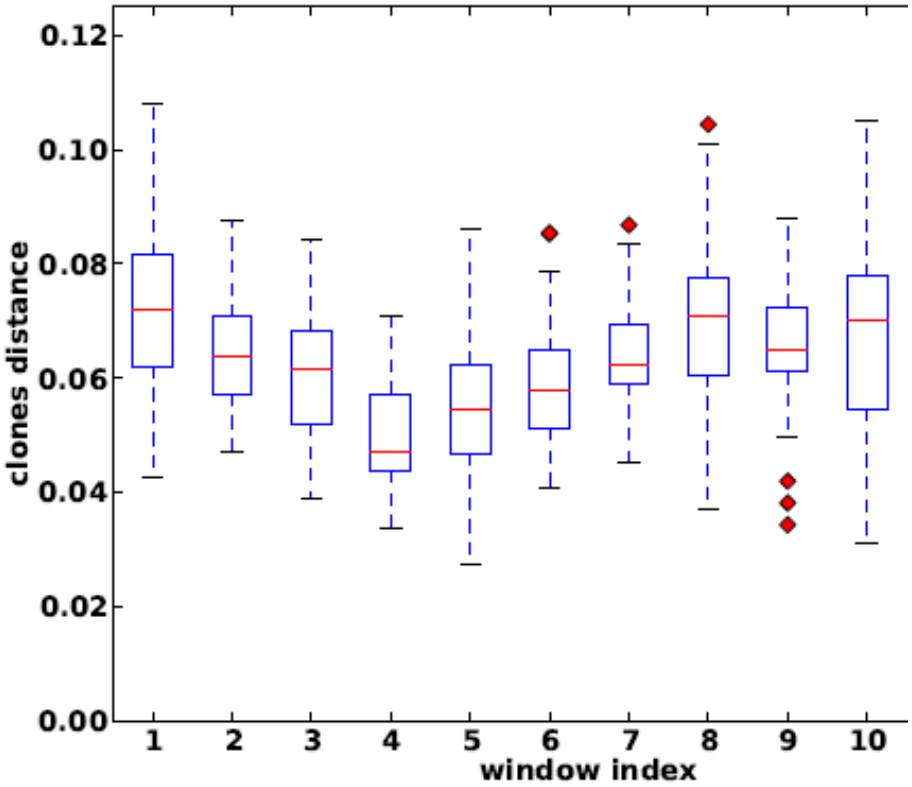


B Korber *et al*., BMB 2001

HIV: Case study

- mixture of 10 well characterized patients' HIV-1 clones ('haplotypes')
  - O Zagordi et al., Nucl Acids Res 2010
- the data set is publicly available
  - https://wiki-bsse.ethz.ch/display/ShoRAH/Data
- 454 sequencing of HIV-1 pol (~1'500 bp)
- Local and global haplotype reconstruction

Local diversity of the original haplotypes



average distance of 6.8% among the 45 pairs

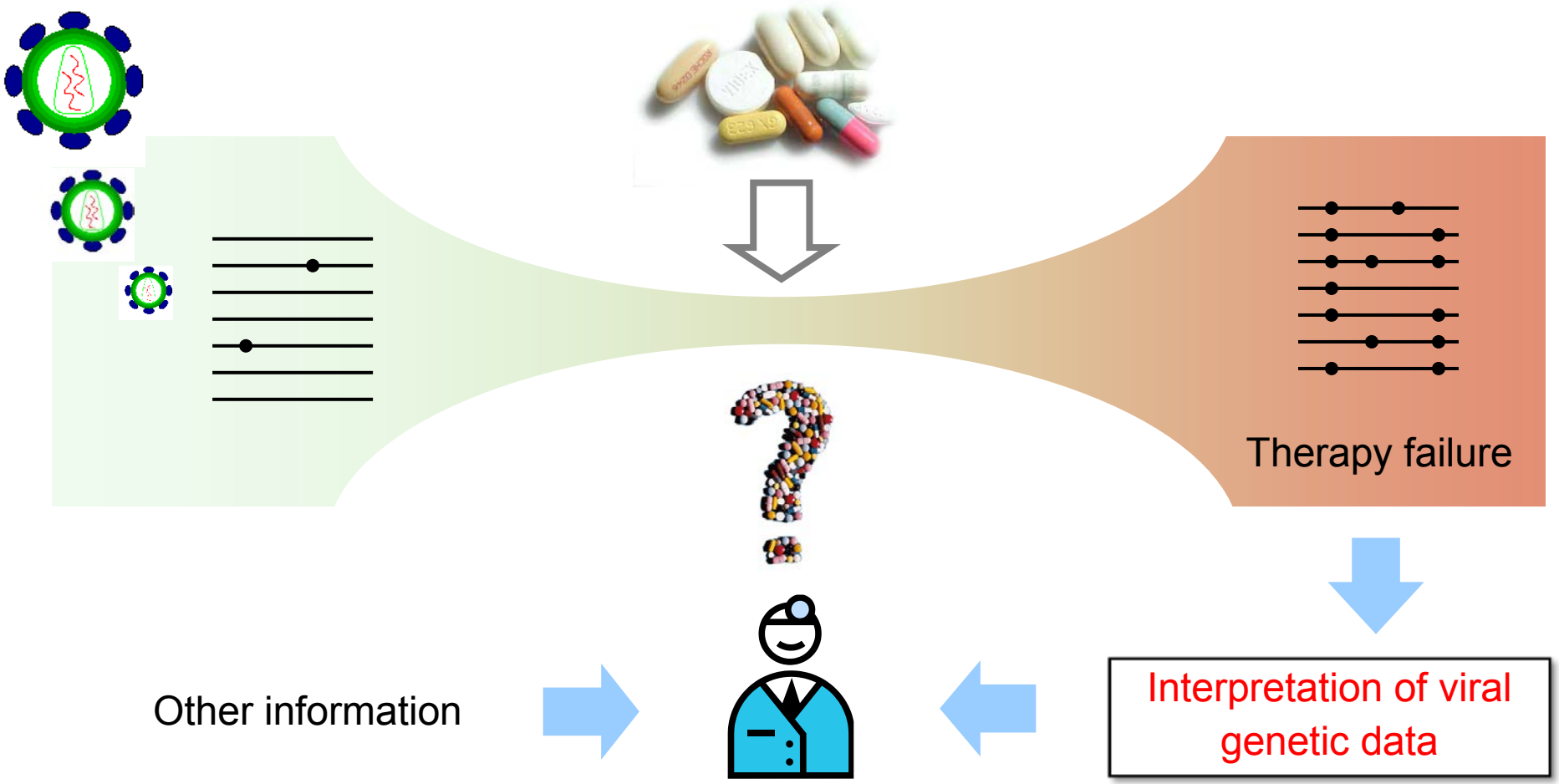
O Zagordi *et al*., Nucl Acids Res 2010

Frequencies of the original haplotypes

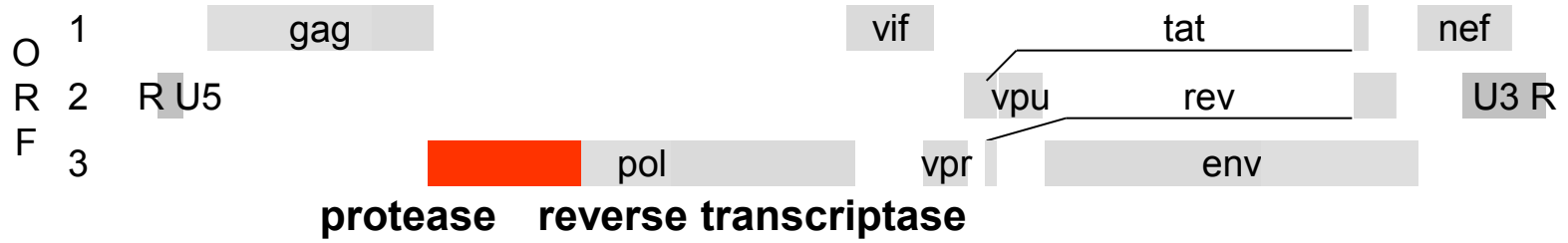
- mixture of 10 haplotypes (%)
  - 50
  - 25
  - 12.5
  - 6.25
  - 3.125
  - 1.563
  - 0.781
  - 0.391
  - 0.195
  - 0.098

O Zagordi *et al*., Nucl Acids Res 2010

HIV drug resistance, individualized treatment

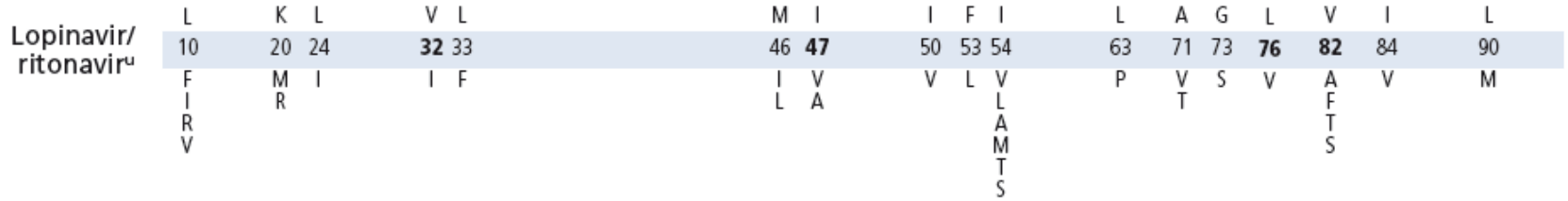


HIV-1 drug resistance



RT

PR

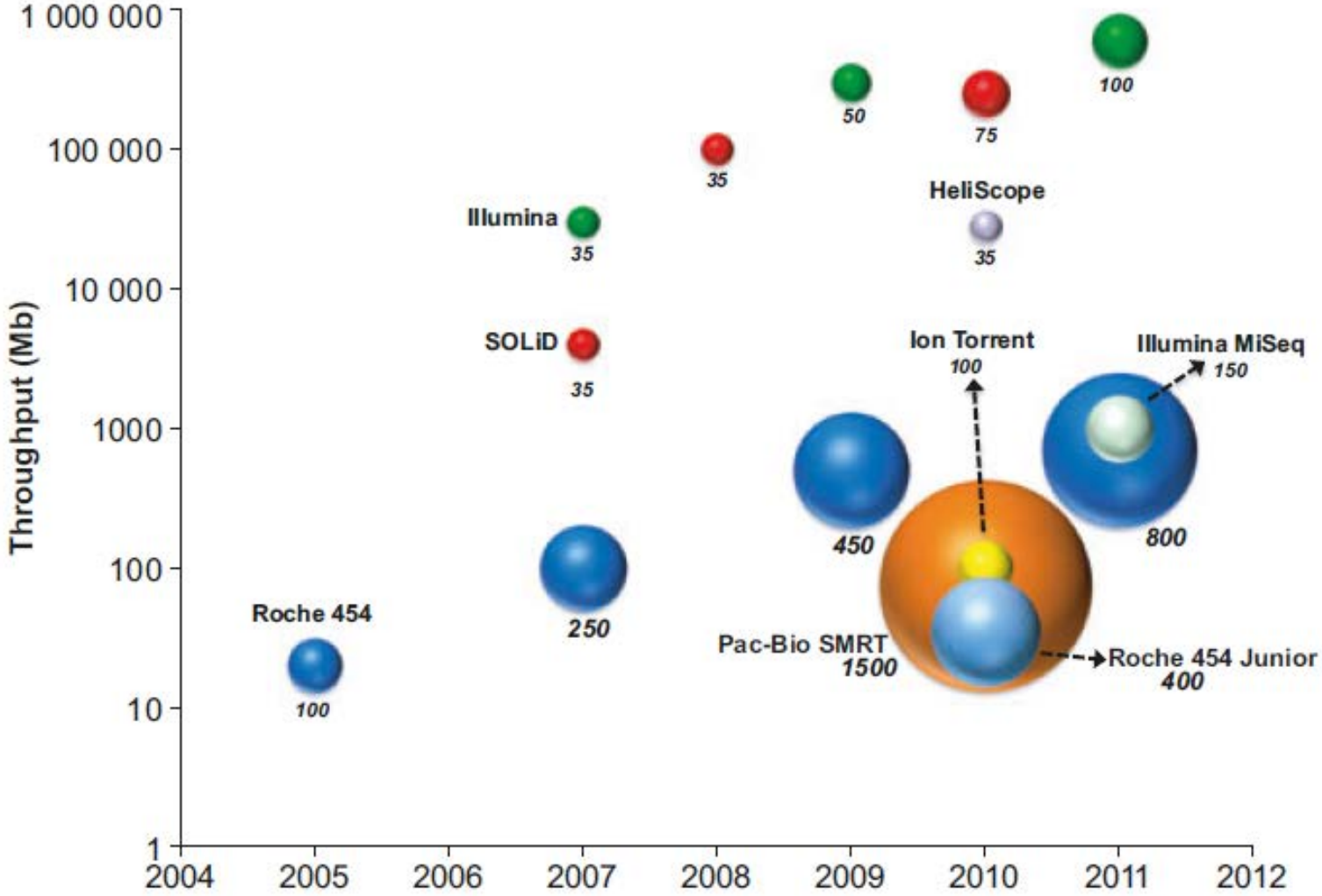


VA Johnson *et al.*, Top HIV Med 2009

Topics

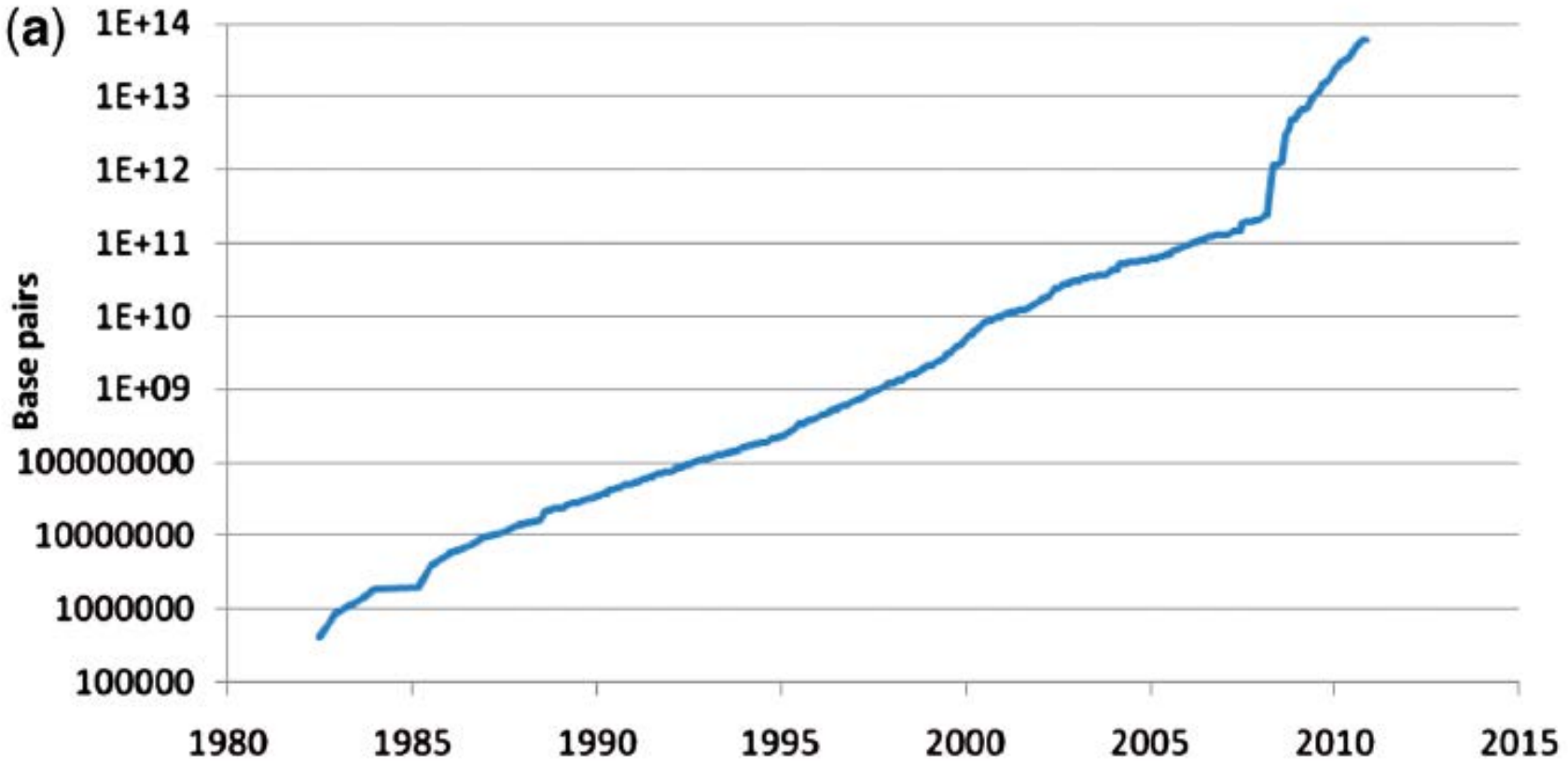
- genetic diversity
  - definitions and motivation
- case studies
  - cancer
  - viruses
- NGS technologies
  - techniques (mainly 454 and Illumina)
  - error pattern and quality scores

Historical development of next-generation sequencing technologies



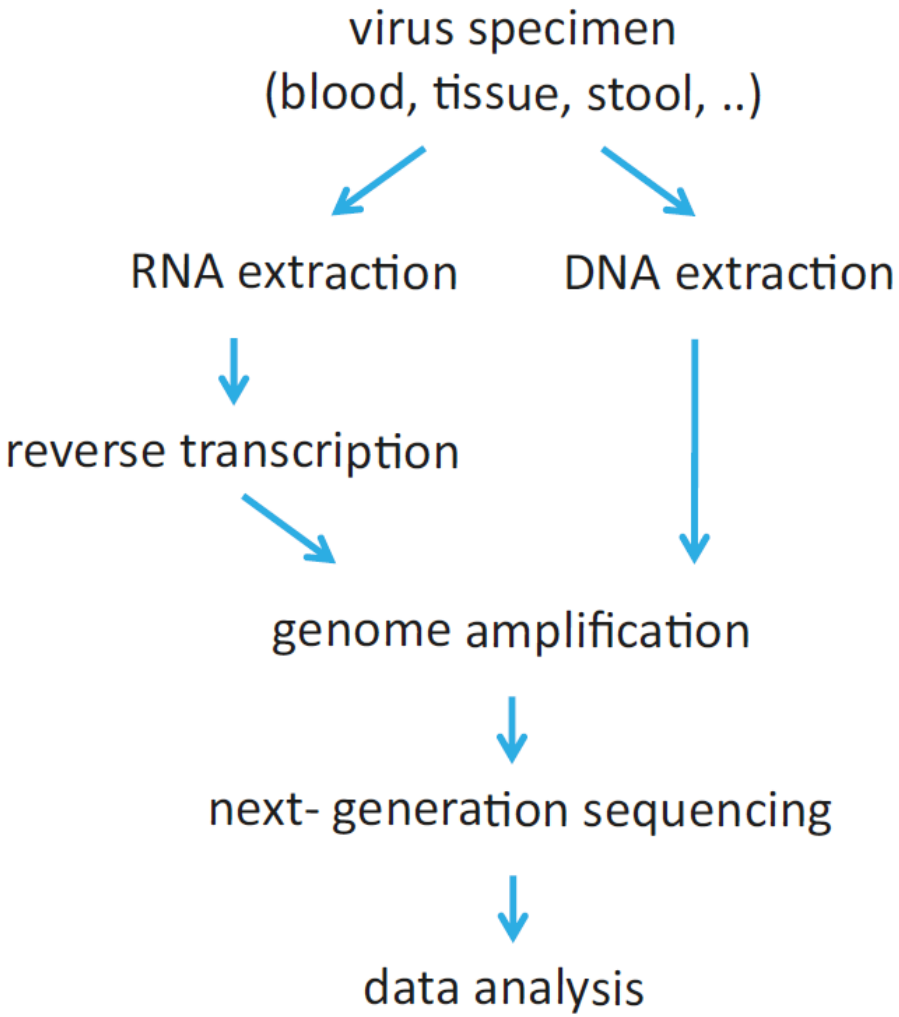
S Shokralla *et al*., Mol Ecol 2012

Cumulative data volume in base pairs over time



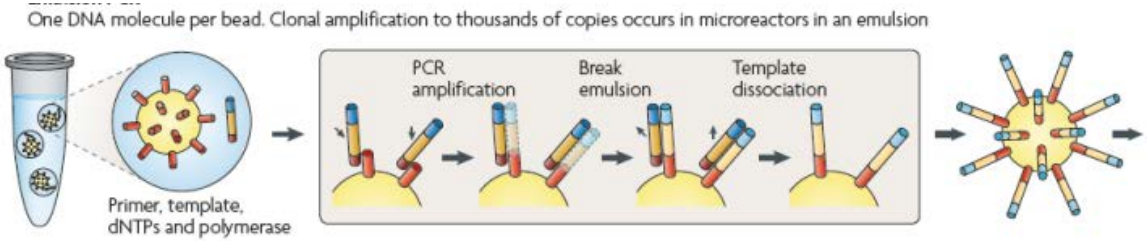
The International Nucleotide Sequence Database Collaboration, NAR 2010

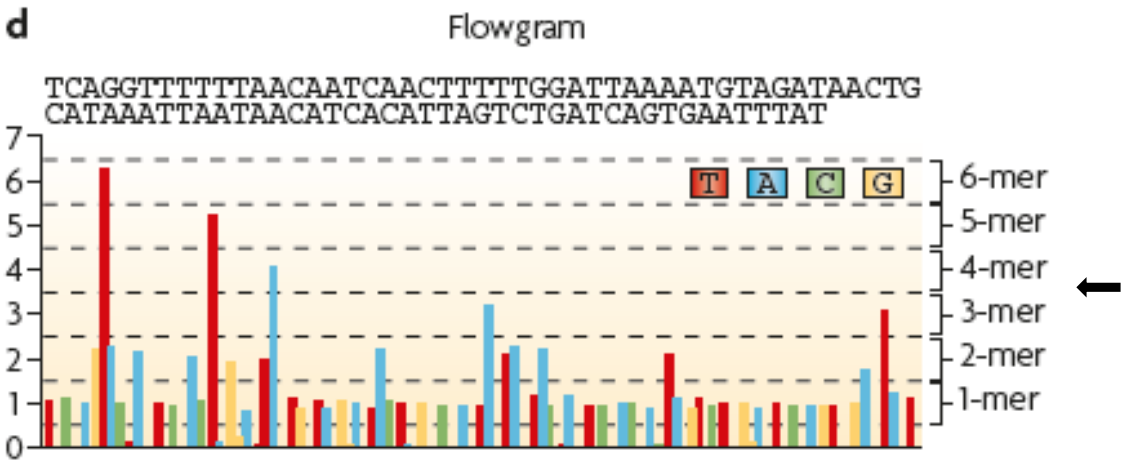
Biological sample to genotype

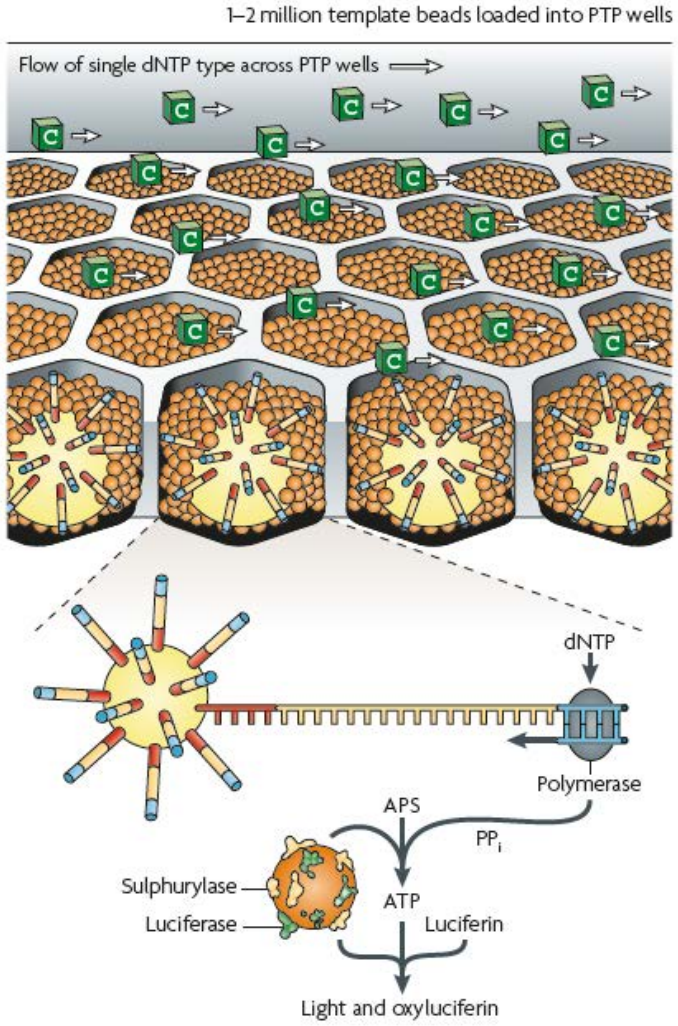


N Beerenwinkel *et al*., Frontiers Microbiol 2012

454 Life Sciences/Roche

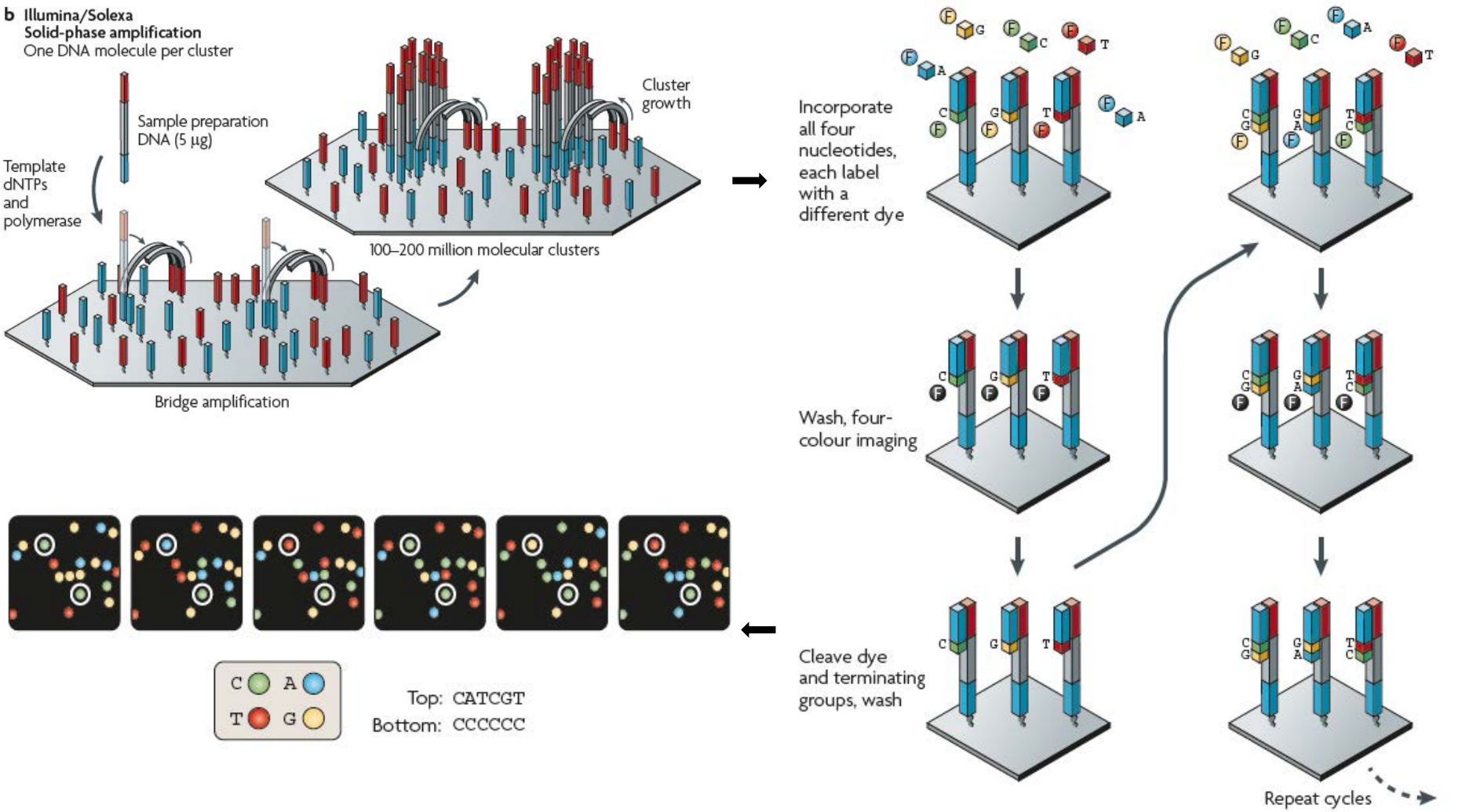






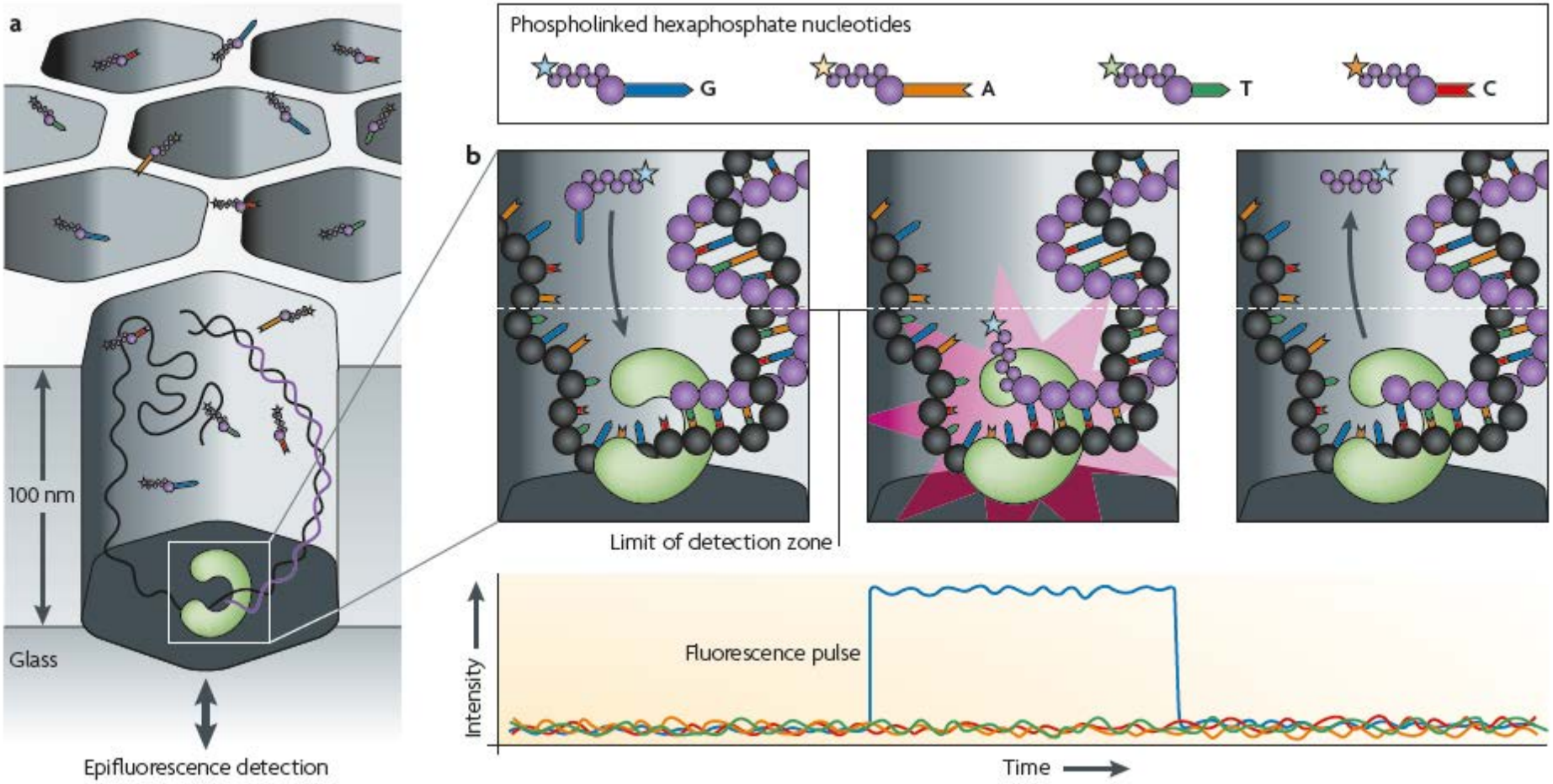
ML Metzker, Nature Review Genetics 2010

Illumina/Solexa



ML Metzker, Nature Review Genetics 2010

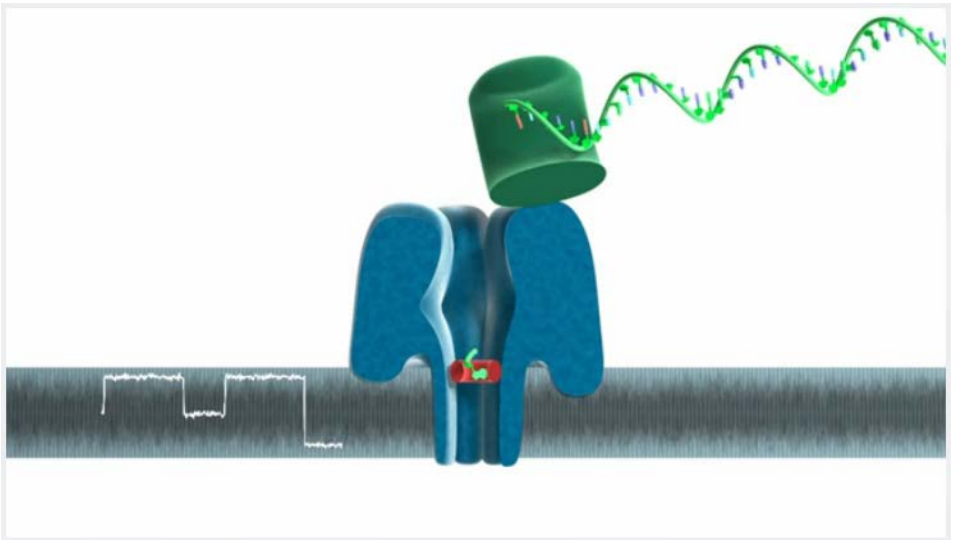
Pacific Biosciences/PacBio RS



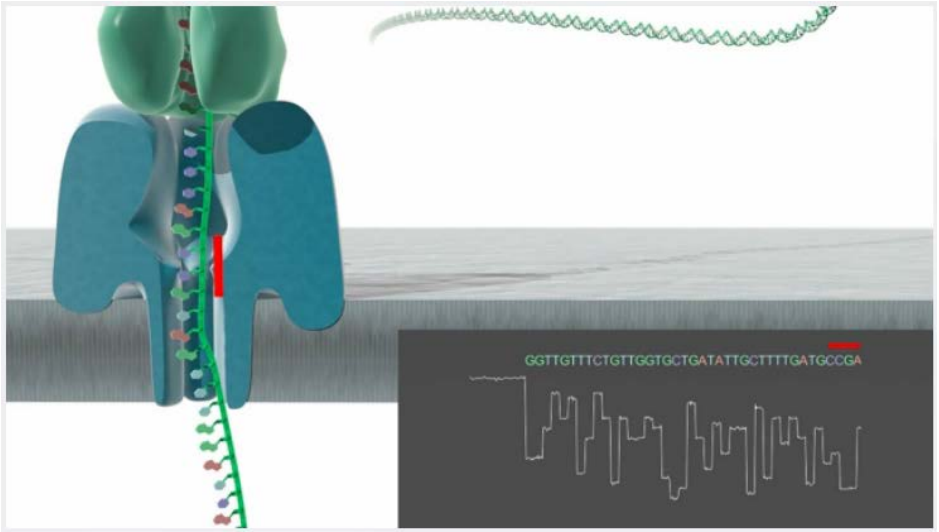
ML Metzker, Nature Review Genetics 2010

Oxford Nanopore Technologies

exonuclease sequencing



strand sequencing



Comparison of next-generation sequencing platforms

454/Roche GS-FLX:
up to 1’000’000 sequences/run
length: 400-700 bp/read

Illumina HiSeq 2000:
up to 1’500’000’000 seq./run
length: 2x100 bp/read

ABI 5500 SOLiD :
up to 900’000’000 seq./run
length: 50-75 bp/read

Pacific Biosciences RS:
up to 800’000 seq./run
length: ~1’500 bp/read

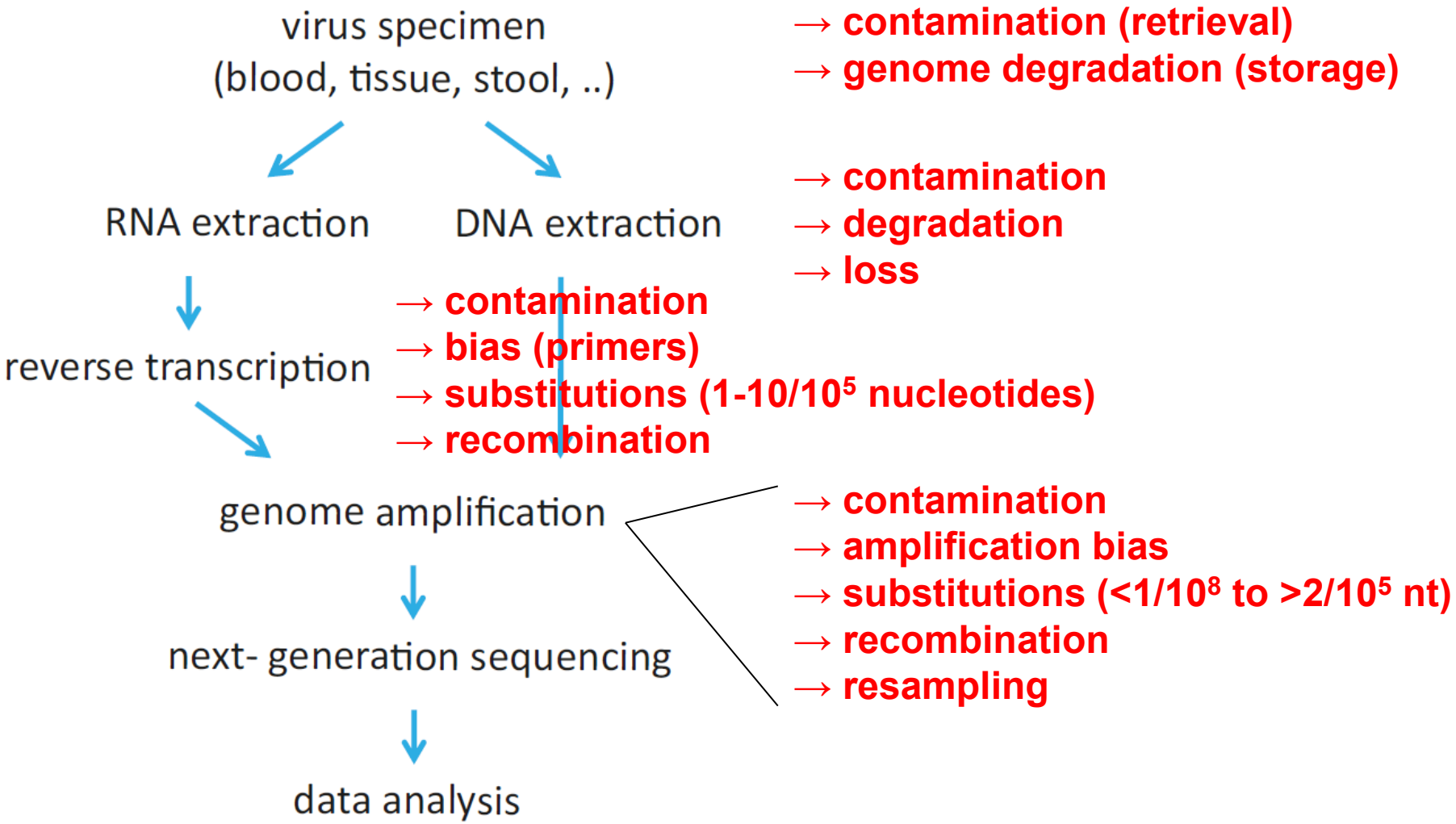
Ion Torrent PGM:
up to 5’000’000 seq./run
length: 35-200 bp/read

Helicos HeliScope:
up to 800’000’000 seq./run
length: 25-55 bp/read

Topics

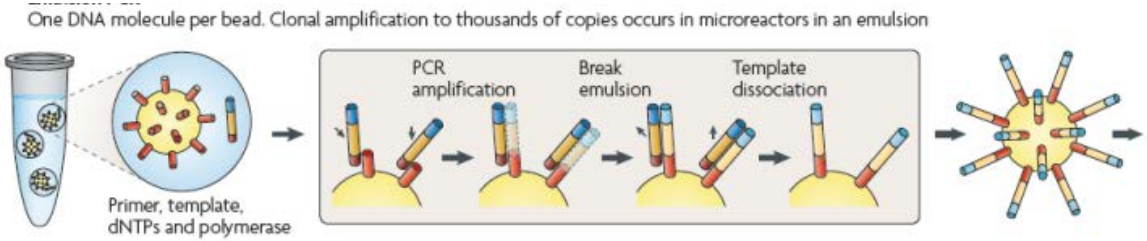
- genetic diversity
  - definitions and motivation
- case studies
  - cancer
  - viruses
- NGS technologies
  - techniques (mainly 454 and Illumina)
  - error pattern and quality scores

Error sources on the way from a biological sample to a pre-NGS sample

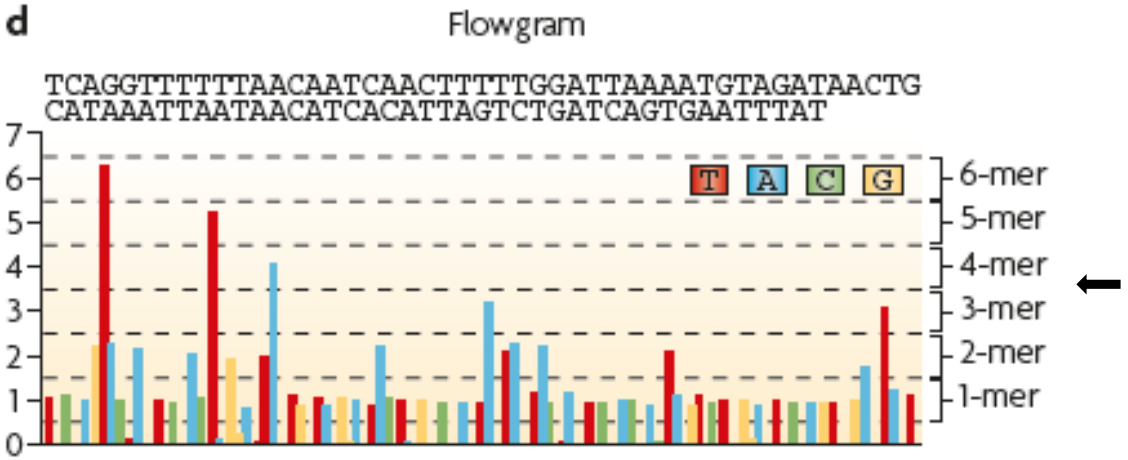


N Beerenwinkel *et al*., Frontiers Microbiol 2012

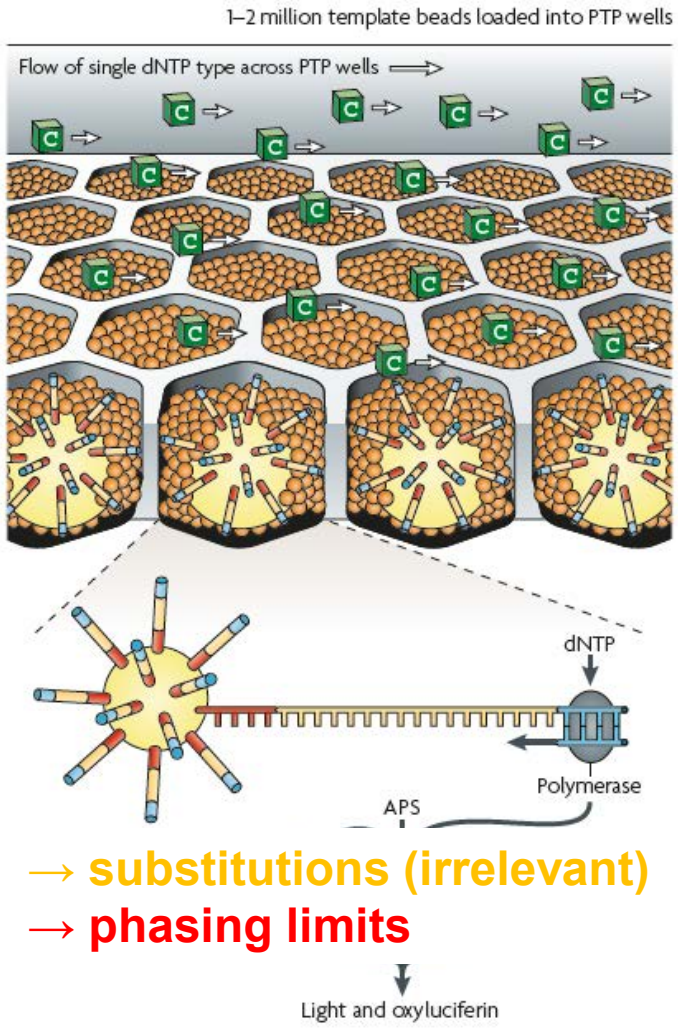
NGS platform dependent errors: 454 Life Sciences/Roche



→ substitutions (irrelevant)
→ recombination (irrelevant)
→ mixed beads (irrelevant)



→ insertions and deletions (indels) in homopolymeric regions!!
→ neighbor interference

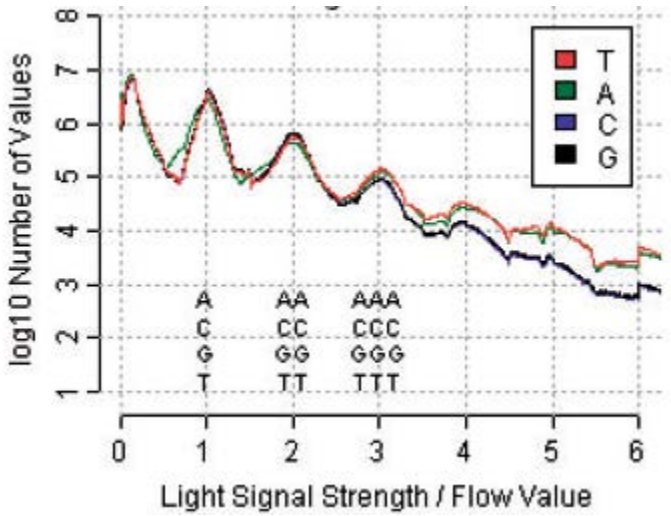


→ substitutions (irrelevant)
→ phasing limits

ML Metzker, Nature Review Genetics 2010

Pyrosequencing errors

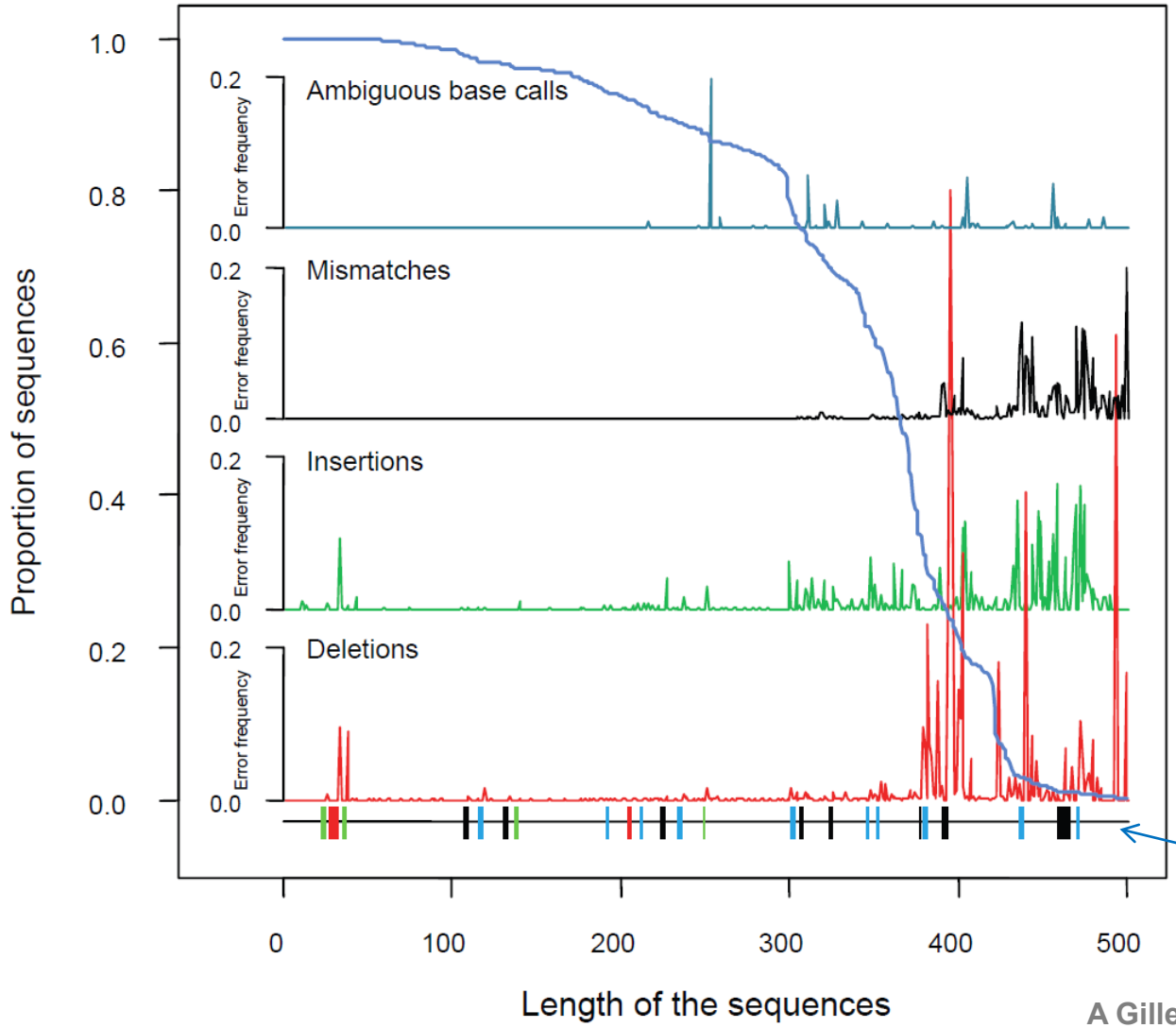
Individual read insertion error rate	0.44%
Individual read deletion error rate	0.15%
Individual read substitution error rate	0.004%
All errors	0.60%



- In long homopolymeric regions, linearity between signal intensity and number of nucleotides incorporated fails.

M Margulies *et al.*, Nature 2005; S Balzer *et al.*, Bioinf 2010

Error rates increase with read length



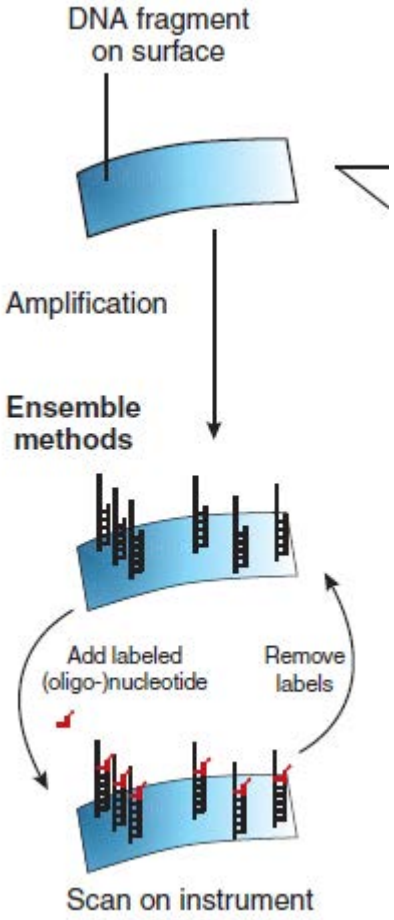
→ **phasing limits**

homopolymeric regions (A, T, G, C)

A Gilles *et al.*, BMC Genomics 2011

Phasing limits

phasing = “maintaining synchronous synthesis among all the identical templates of the ensemble”



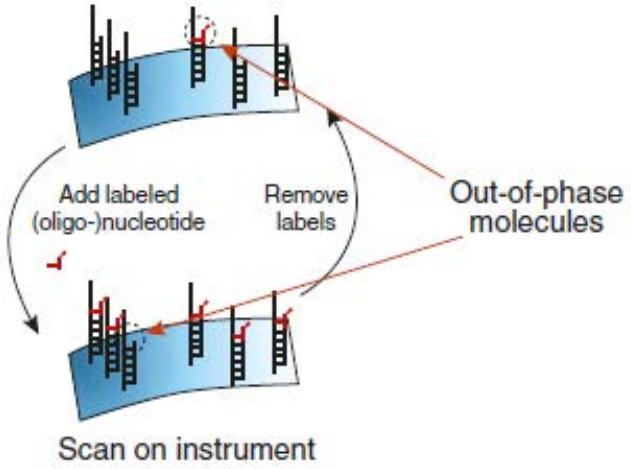


Figure 5 Read-length and phasing. For ensemble sequencing methods, thousands of identical copies of a DNA sequence are synthesized together. Whenever synthesis lags behind or steps ahead, signal is lost and background increases. Ultimately, this limits the length over which synthesis can be read.

CW Fuller *et al*., Nature Biot 2009

NGS platform dependent errors: Illumina/Solexa

→ substitutions (irrelevant)

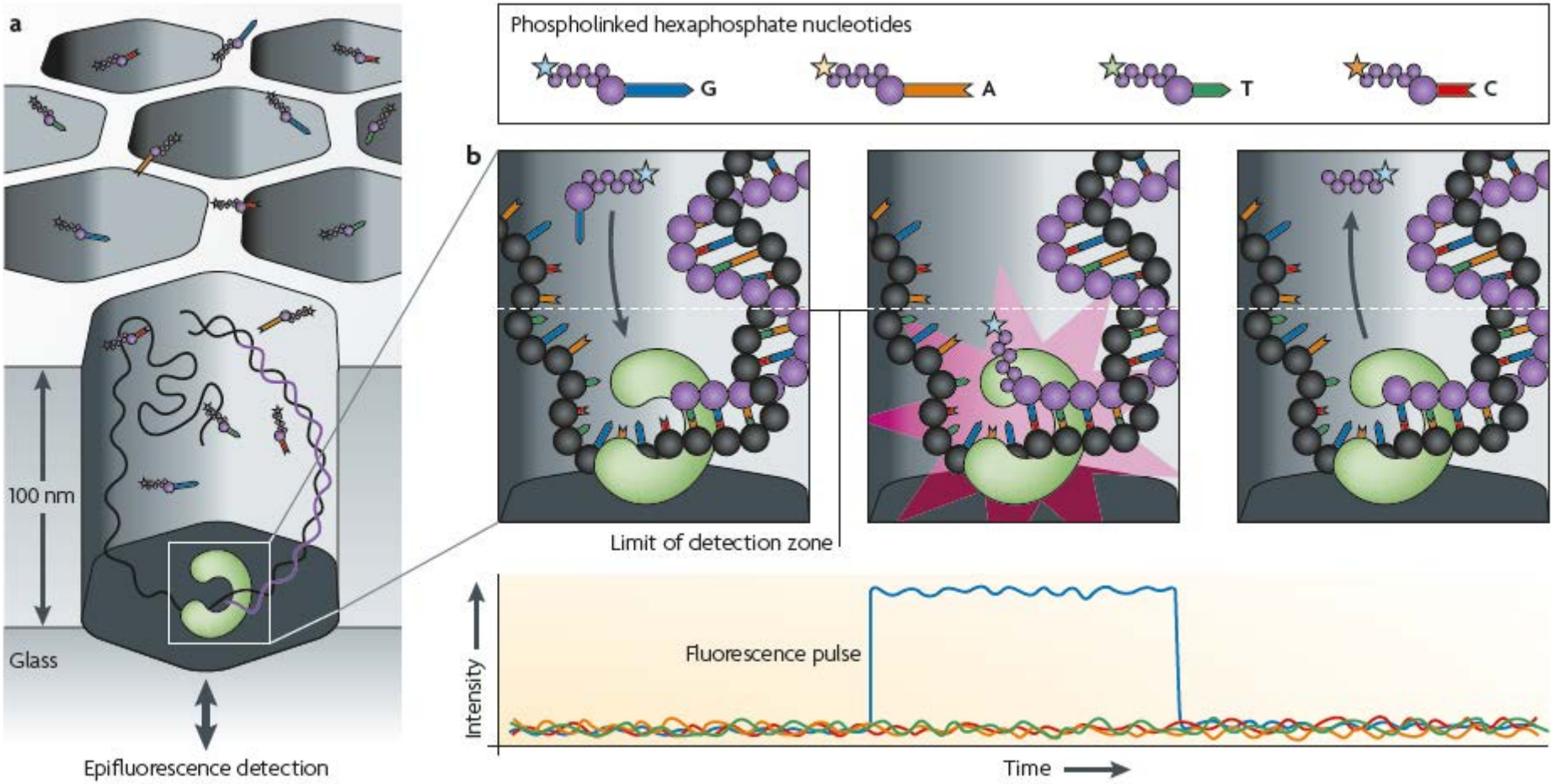
→ recombination (irrelevant)

→ mixed clusters/neighbor interference

→ phasing limits

ML Metzker, Nature Review Genetics 2010

NGS platform dependent errors: Pacific Biosciences/PacBio RS



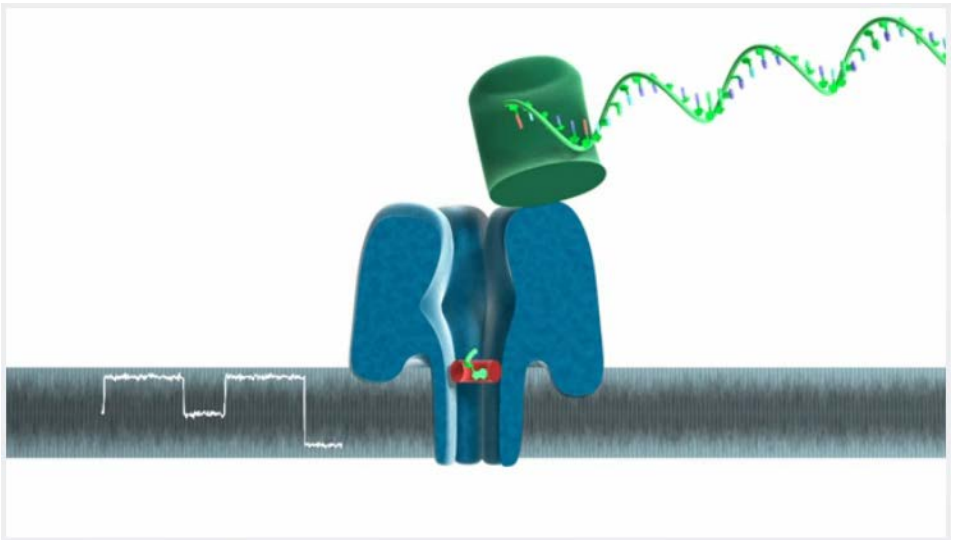
→ **substitutions !! (repeated sequencing of the same molecule necessary)**

→ **others?**

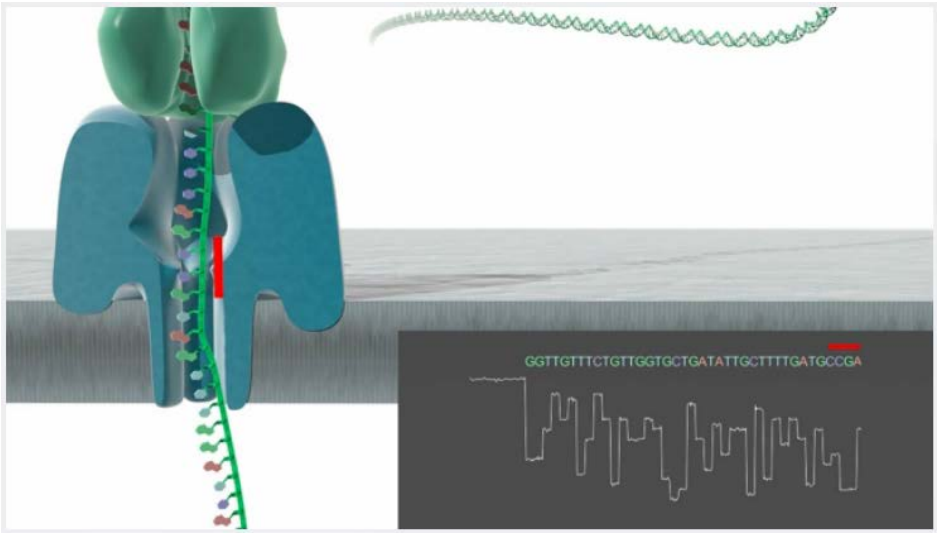
ML Metzker, Nature Review Genetics 2010

NGS platform dependent errors: Oxford Nanopore Technologies

exonuclease sequencing



strand sequencing



→ 0.1 - 4% raw read error rates are reported in press releases

→ error sources?

Oxford Nanopore Technologies

A lot of reads are incorrect

- Let us assume (454 data)
  - a sequencing error of 0.1% per base pair, and
  - an average read length of 500 bp
- Then the fraction of reads with at least one error is

$$1-\left(1-0.001\right)^{500}=0.394$$

- Thus, ~40% of the reads are incorrect!

Thanks for your attention

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

SIB Swiss Institute of Bioinformatics

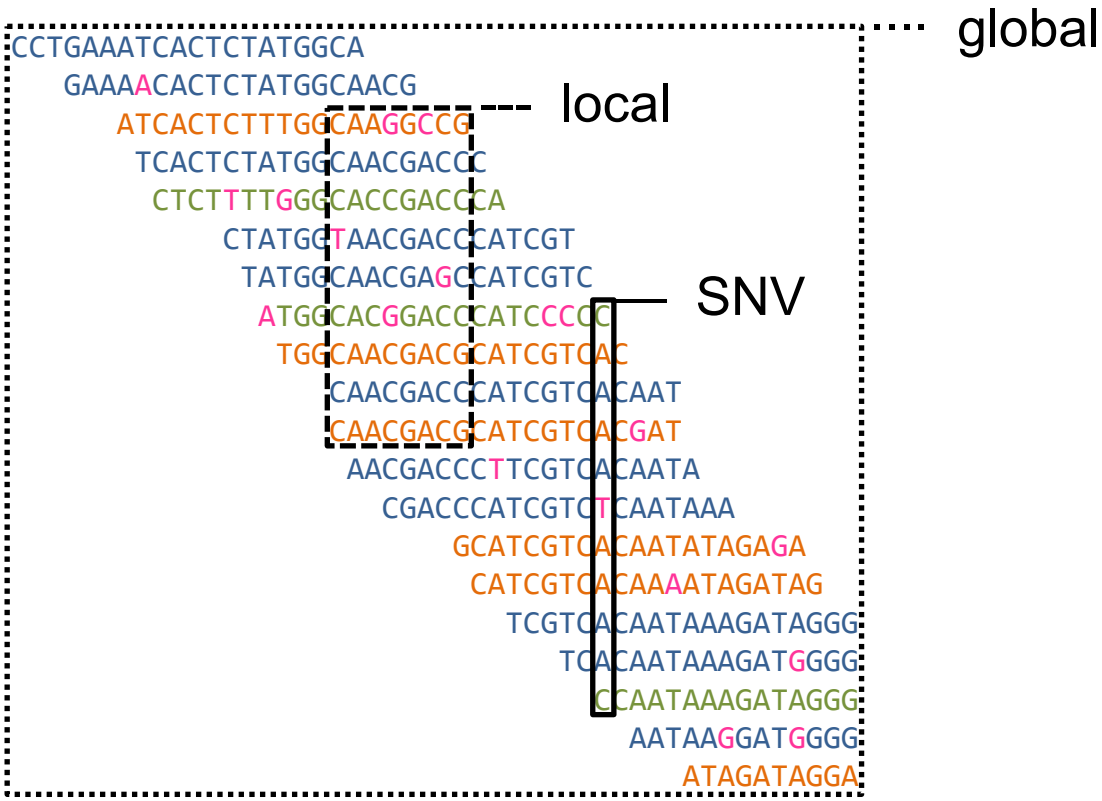
D-BSSE
Department of Biosystems Science and Engineering

Local Diversity Estimation

Niko Beerenwinkel, ETH Zurich

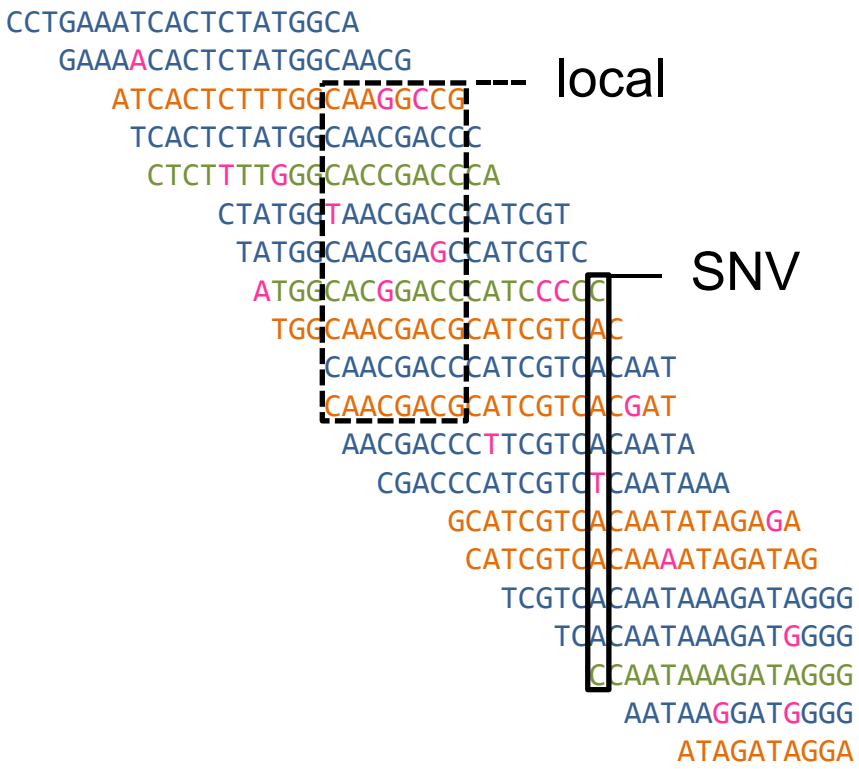
niko.beerenwinkel@bsse.ethz.ch

Spatial scales of diversity estimation



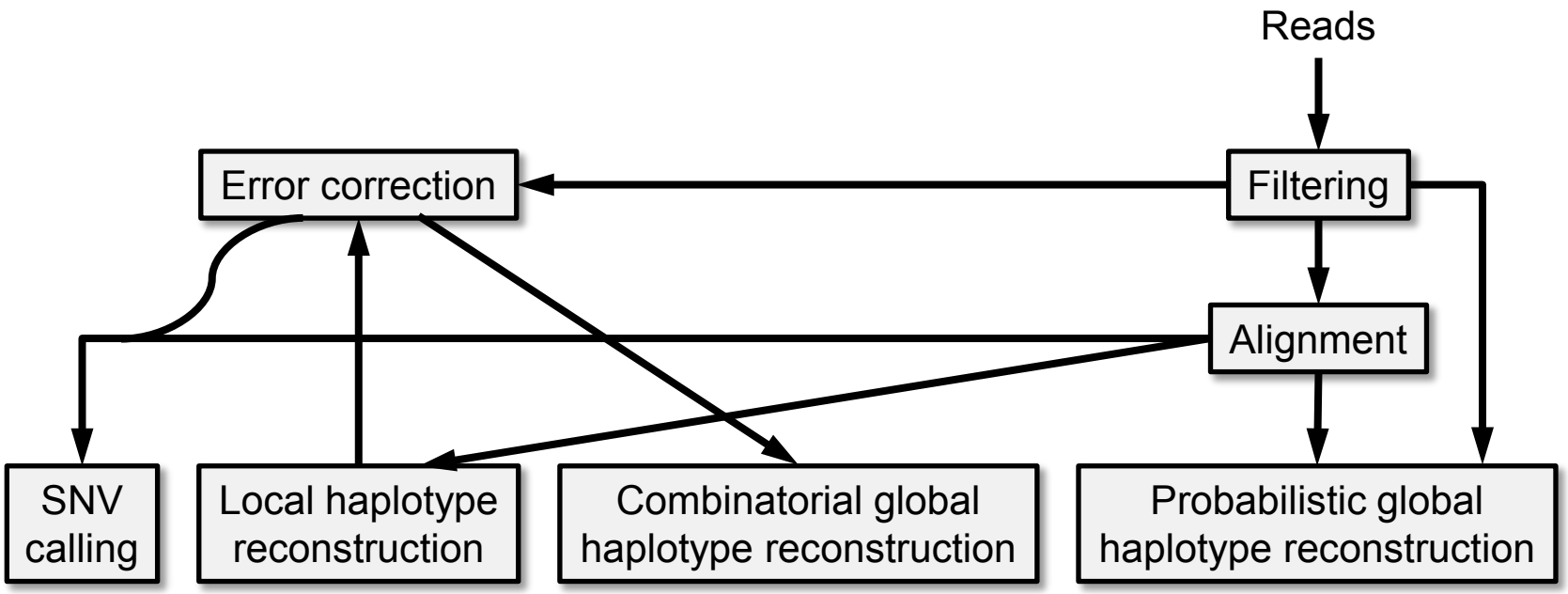
Beerenwinkel et al (Front Microbio 2012)

This session: SNVs and local diversity

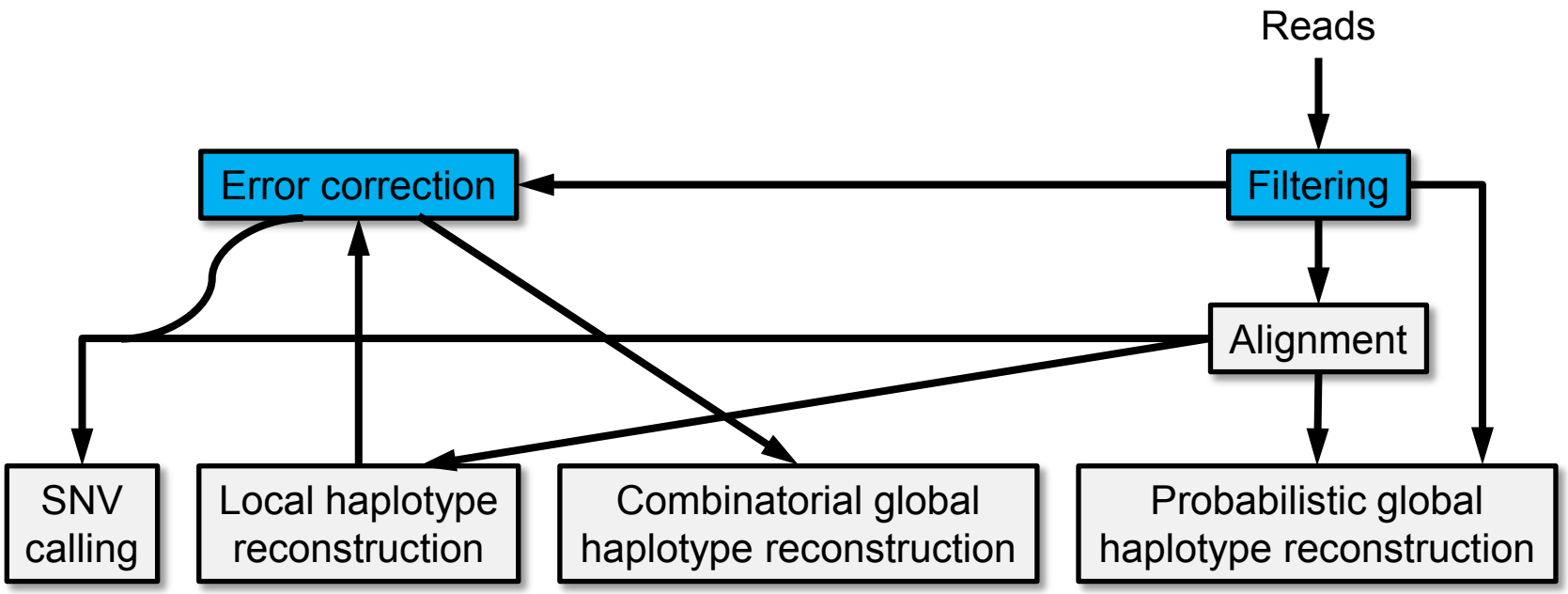


Beerenwinkel et al (Front Microbio 2012)

Overview

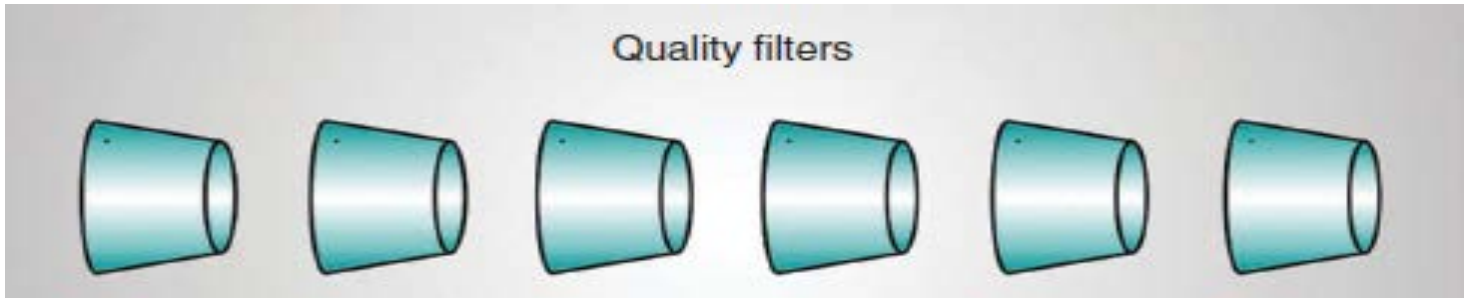

Reads
Error correction
Filtering
Alignment
SNV calling
Local haplotype reconstruction
Combinatorial global haplotype reconstruction
Probabilistic global haplotype reconstruction

Overview


Reads
Error correction
Filtering
Alignment
SNV calling
Local haplotype reconstruction
Combinatorial global haplotype reconstruction
Probabilistic global haplotype reconstruction

Filtering

before alignment	after alignment



Low phred score

Many N's

Short or long reads (relative to amplicon size)

Dissimilar to external reference sequence

Dissimilar to internal reference sequences

Many indels

Reumers et al (Nat Biotech 2011), Skums et al (BMC Bioinformatics 2012)

Alignment-free, k-mer-based read error correction

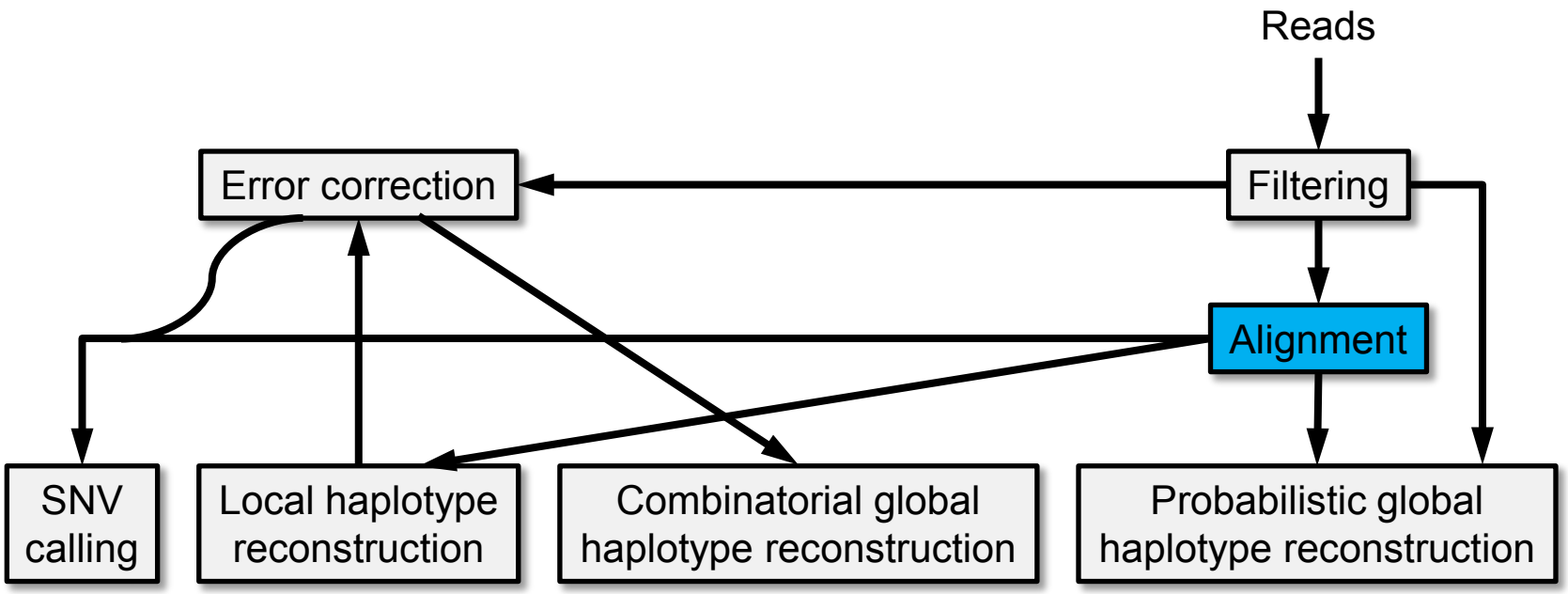
- k-mer = substring of length k
- Idea: Rare ("weak") k-mers are likely to contain errors
- KEC algorithm:
  1. Calculate k-mers and their frequencies, called k-counts.
  2. Determine the threshold k-count which distinguishes solid k-mers from weak k-mers.
  3. Find error regions.
  4. Correct the errors in error regions.

- Example:

Sample:	M1	M2	M3	M4	M5
No. of reads before error correction:	4220	4222	4418	4344	4426
No. of unique/unique maximal reads after correction:	306/8	502/18	385/8	483/9	179/2

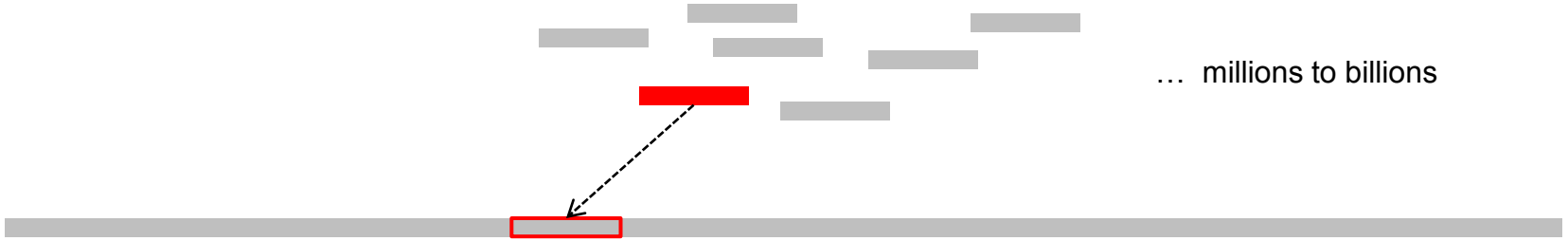
Skums et al (BMC Bioinformatics 2012)

Overview


Reads
Error correction
Filtering
Alignment
SNV calling
Local haplotype reconstruction
Combinatorial global haplotype reconstruction
Probabilistic global haplotype reconstruction

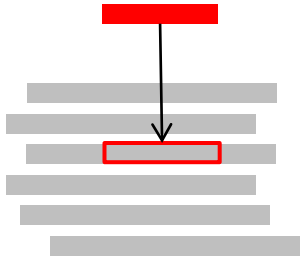
Read alignment (mapping)

- Task: Find the location of each read in a given reference genome in the presence of errors



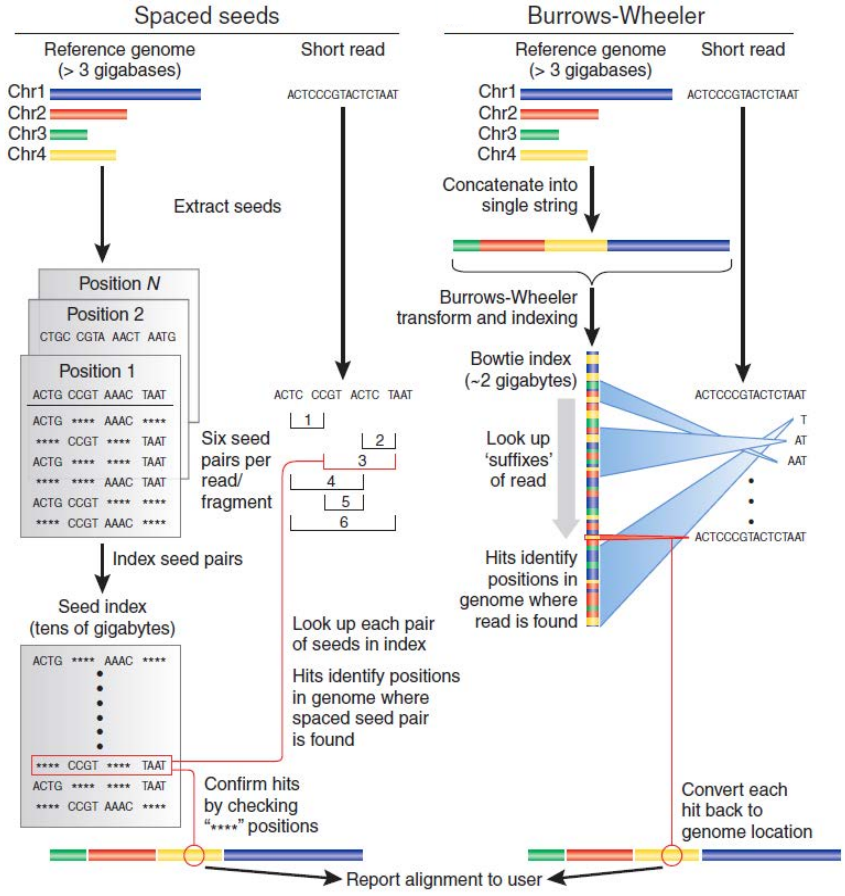
  - dominated by sequencing errors, genetic diversity of the sequenced species, and uncertainty in reference genome assembly

- By comparison, “traditional alignment”
  - finds matches in (remote) homologous sequences in large databases (Smith-Waterman, BLAST, FASTA)
  - uses evolutionary models (DNA/protein substitution models, phylogenetic trees)



Read mapping

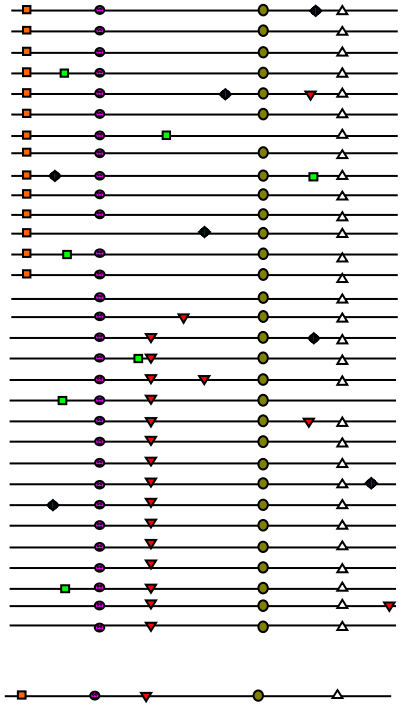
- Challenges:
  - many short reads
  - long genomes
  - errors
  - repetitive DNA
- Read mappers are based on indexing techniques to locate reads, followed by high-quality local alignments.
- Around 50 mapping programs currently available
  - see [wikipedia list]



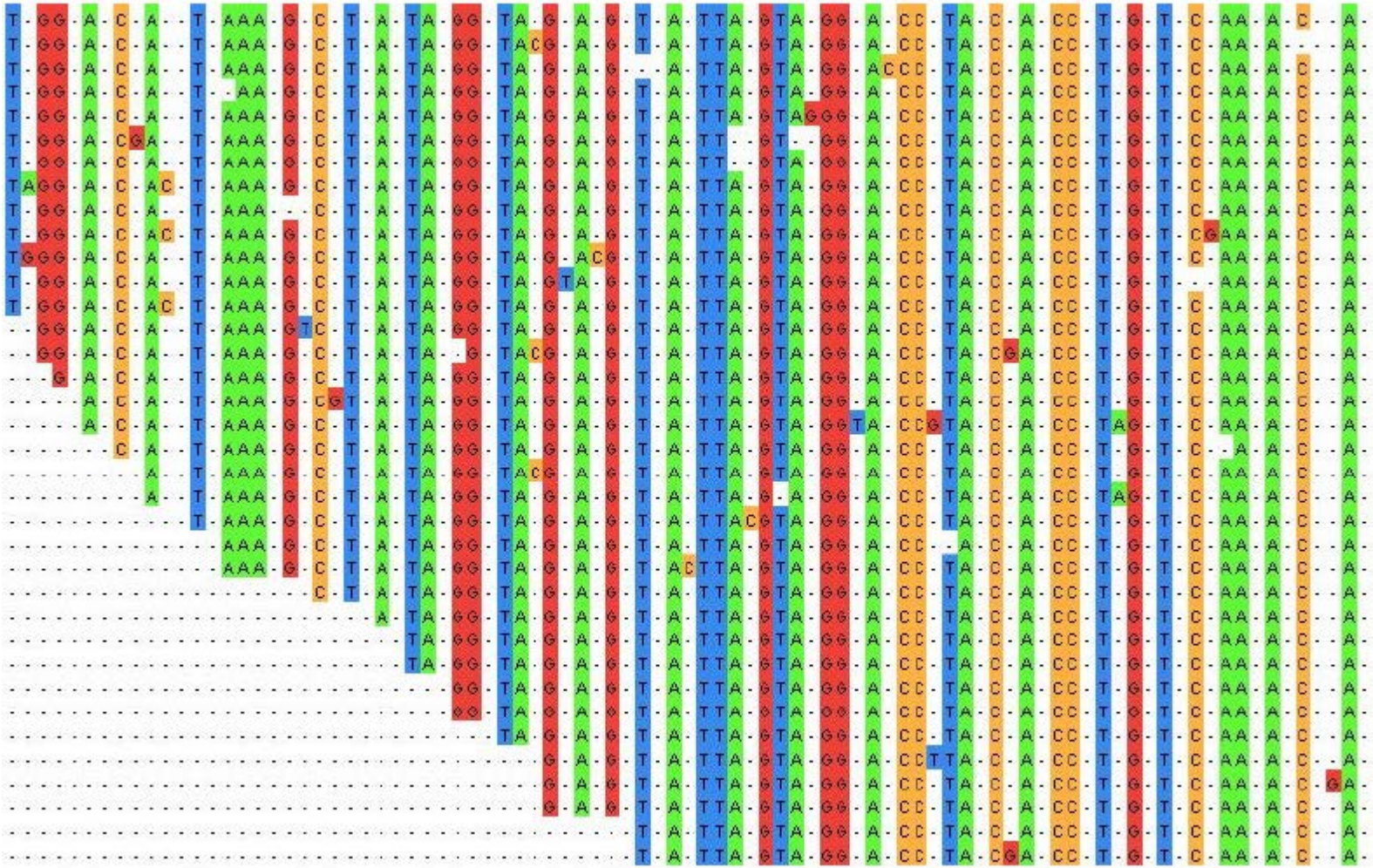
Trapnell & Salzberg (Nat Biotech 2009)

Read mapping of diverse populations at high coverage

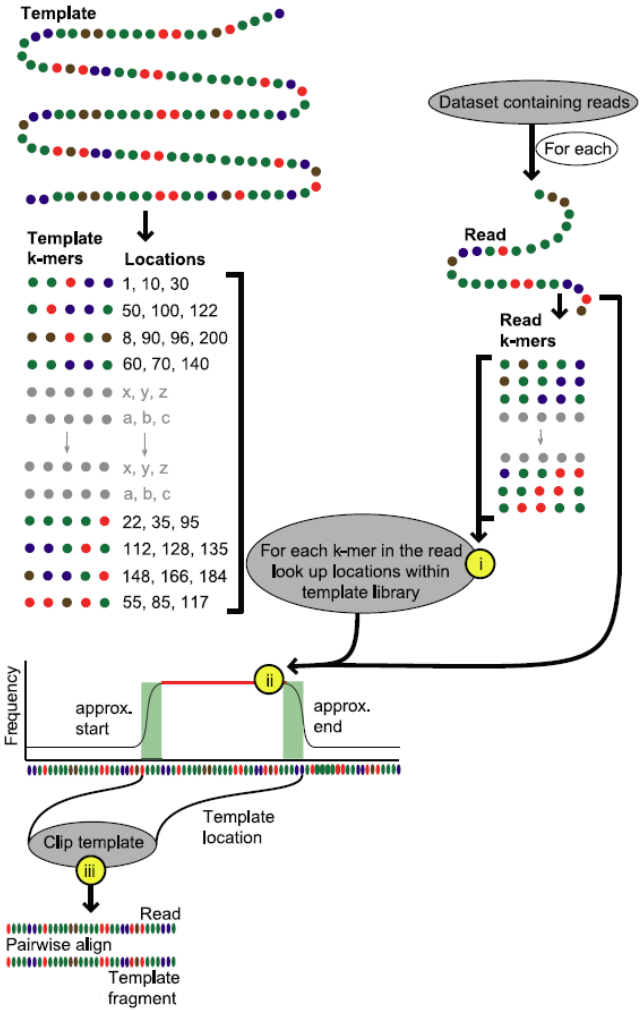
- Goals:
  1. Multiple alignment of all reads
  2. Error correction
- Strategies:
  - Mapping to reference genome
  - Mapping to consensus sequence
  - Pairwise local alignments (Smith-Waterman) to reference or consensus
  - Multiple sequence alignment (MSA)
  - De novo assembly
  - Account for quality scores, error patterns
- In practice, strategies are often combined.



MSA of reads – snapshot



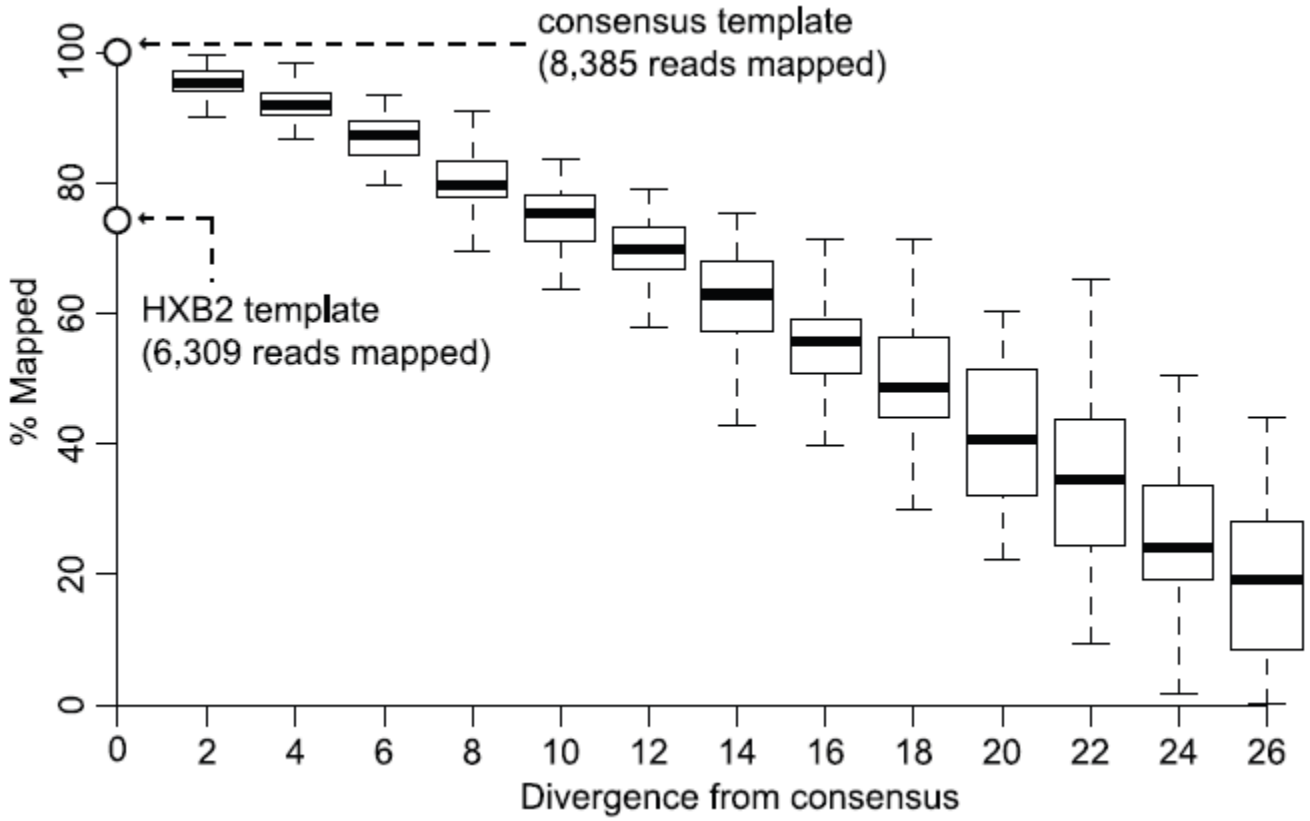
Example: HIV *env* gene



1. Locate reads by k-mer matching on reference (template) sequence (here: HIV-1 HXB2)
2. Build MSA in windows of size 70nt with overlap 20nt
3. Generate in-frame consensus sequences, concatenate
4. Align reads locally to consensus sequence (Smith-Waterman)
5. Remove indels causing frameshifts

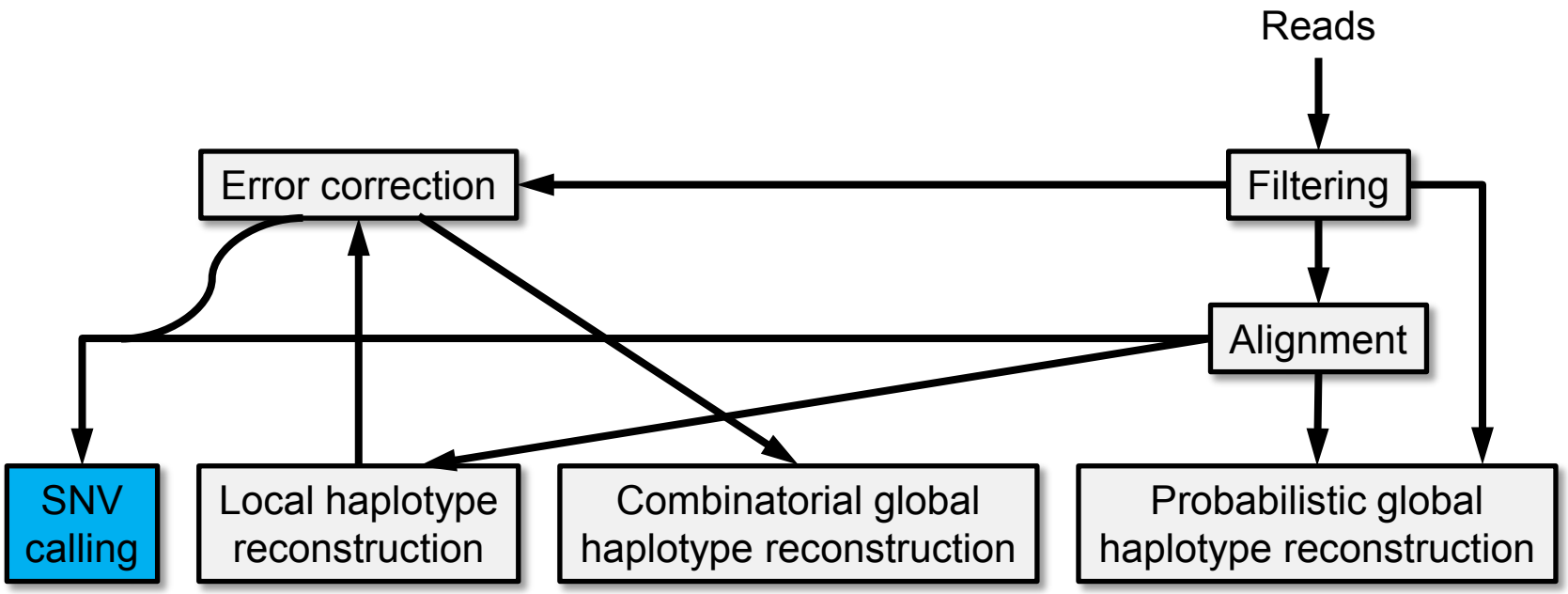
Archer et al (PLoS Comput Biol 2010)

Read mapping improves with consensus over reference template sequence

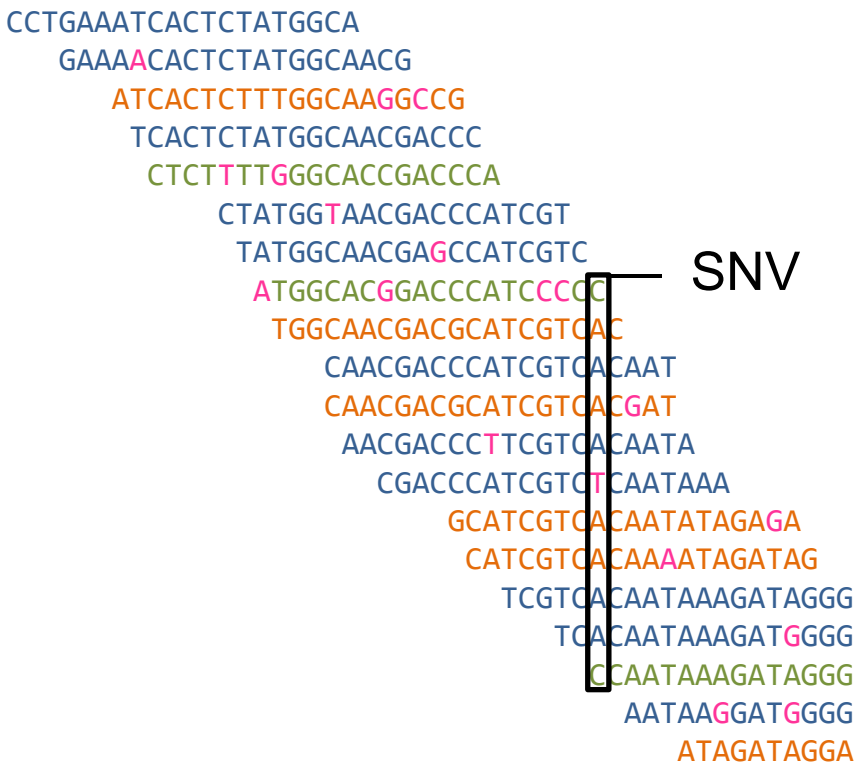


Archer et al (PLoS Comput Biol 2010)

Overview


Reads
Error correction
Filtering
Alignment
SNV calling
Local haplotype reconstruction
Combinatorial global haplotype reconstruction
Probabilistic global haplotype reconstruction

SNV calling



Beerenwinkel et al (Front Microbio 2012)

SNV detection in a mixed sample

- Let q be the per-site error rate.
- The number of errors at position i is

$$X_i \sim \text{Binom}(n_i, q)$$

where n_i is the coverage.

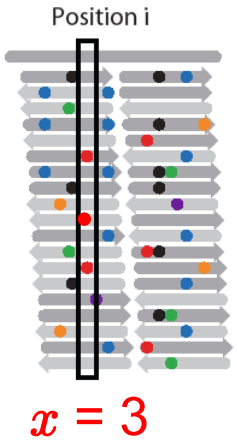
- With $\lambda_i := n_i q = \text{E}[X_i]$, approximately,

$$X_i \sim \text{Pois}(\lambda_i)$$

- For calling an allele observed x times, consider

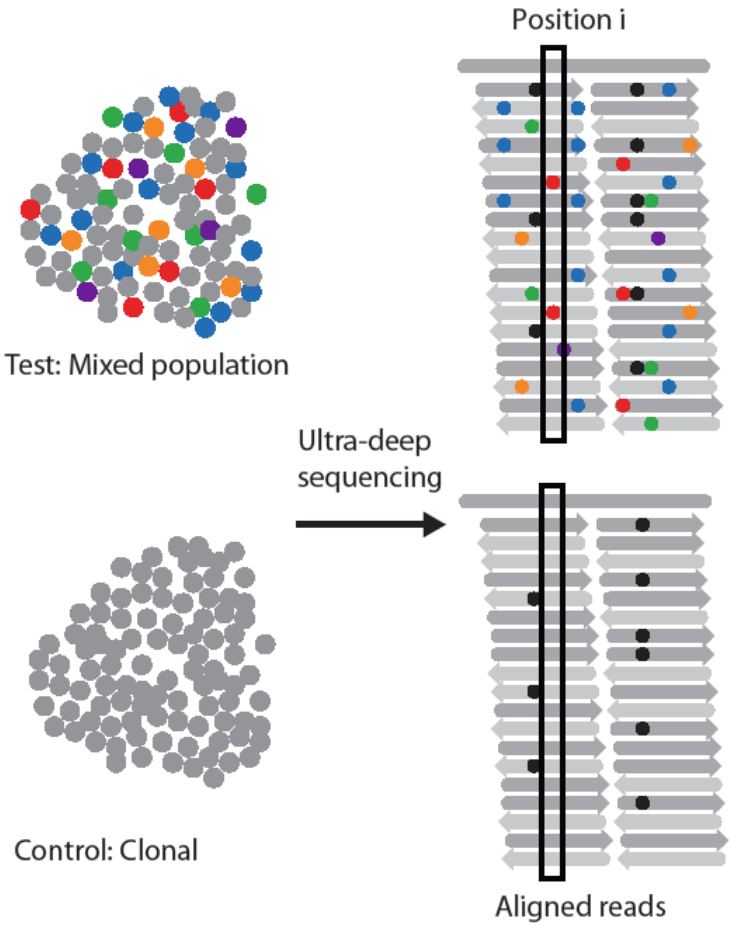
$$P(X_i \geq x) = 1 - \sum_{k=0}^{x-1} \frac{\lambda_i^k \, e^{-\lambda_i}}{k!}$$

n_i = 19



x = 3

SNV detection via comparative sequencing



- Task: For each allele, decide whether its frequency in the tumor is higher than in normal control tissue.
- If so, the allele is called (i.e., likely to be a true biological variant), otherwise it is more likely to be an experimental error (i.e., noise).
- Requires a statistical framework for comparing allele counts

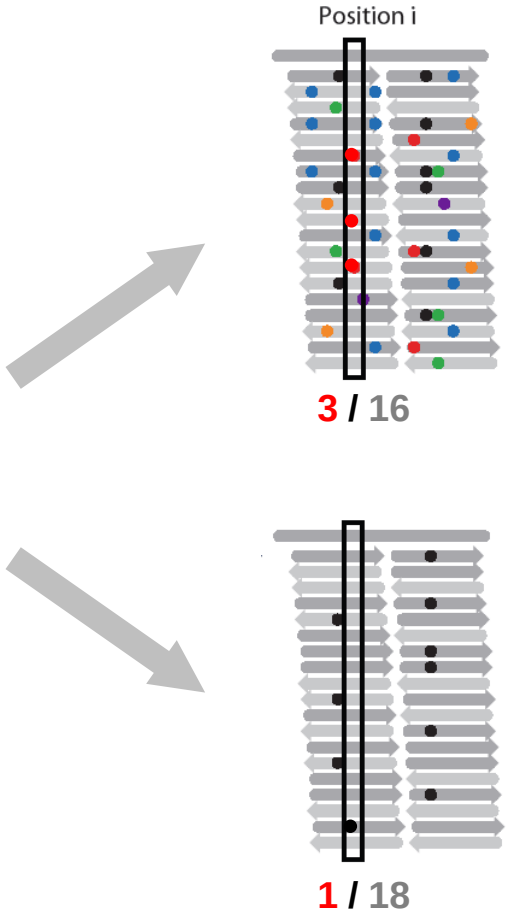
Gerstung et al (Nat Commun 2012)

Simple approach (Varscan 2)

- For each allele, Fisher's exact test on

	No. of variant alleles	No. of all other alleles
Tumor	3	16
Normal	1	18

$P = 0.34$



- Correct for multiple testing

Koboldt et al (Genome Res 2012)

Independent Poisson distributions (vipR)

- Allele count in tumor:

$$X_i \sim \text{Pois}(\mu_i = n_i q)$$

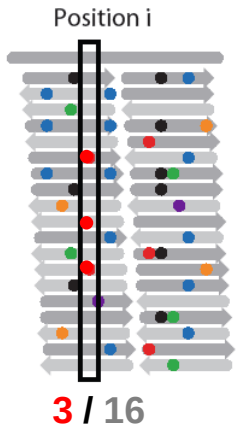
- Allele count in normal control:

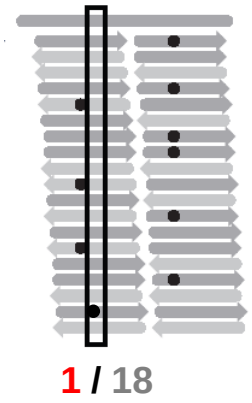
$$Y_i \sim \text{Pois}(\lambda_i = m_i q)$$

- For calling an allele observed x times in the tumor and y times in the control, consider

$$P(X_i - Y_i \geq x - y) = 1 - \sum_{k=-\infty}^{x-y-1} \text{Skel}(k \mid \mu_i, \lambda_i)$$

where Skel is the Skellam distribution.





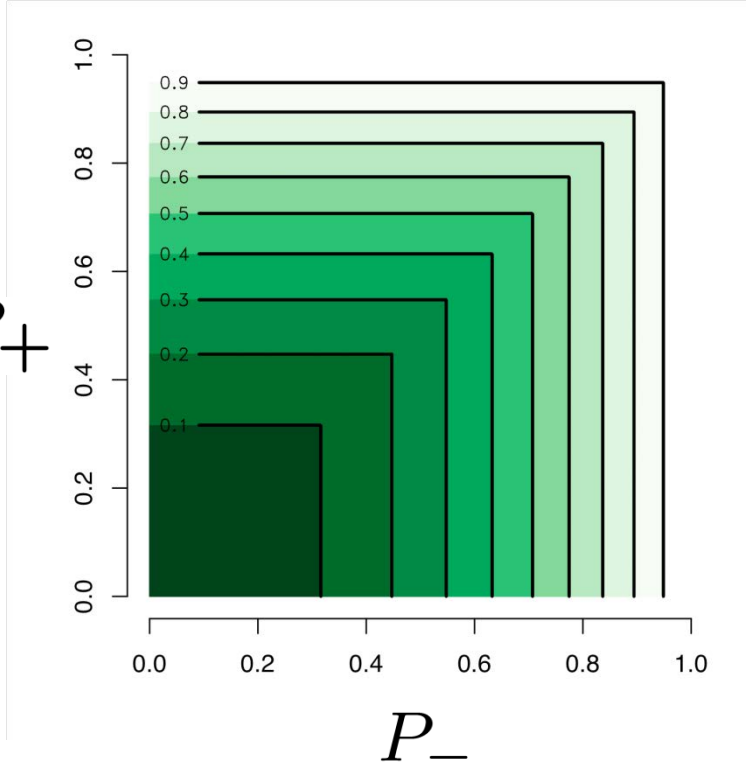
Altmann et al (Bioinformatics 2011)

Strand specificity

- Sequencing errors often occur predominantly on one strand,

```
      →
      GTAAAAA…
5′    GTAAAAGCGTATG  3′
3′    CATTTTCGCATAC  5′
            …CGCATAC
                   ←
```

whereas true variants do not.



$$P_{\text{combined}} = \max(P_+, P_-)^2$$

Altmann et al (Bioinformatics 2011), Gerstung et al (Nat Commun 2012)

Independent, but non-identical, error rates

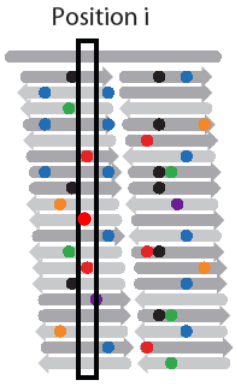
- Let q_{ik} be the error rate at position i on read k.
- If E_{ik} indicates an error at position i on read k, $P(E_{ik} = 1) = q_{ik}$, then

$$X_i = \sum_{k=1}^{n_i} E_{ik}$$

is, in general, not binomial, but its distribution can be computed recursively using the discrete convolution formula.

- In the special case that $q_{ik} = q_i$ for all reads k,

$$X_i \sim \text{Binom}(n_i, q_i)$$



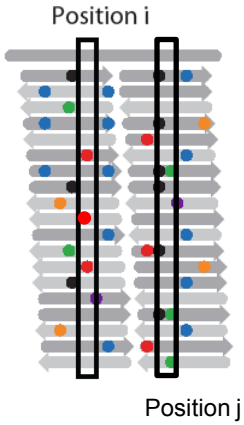
Macalalad et al (PLoS Comput Biol 2012)

Non-independent and non-identical error rates

- Let E_{ijk} indicate the joint occurrence of errors at positions i and j on read k, $P(E_{ijk} = 1) = q_{ijk}$.
- Then the distribution of the number of joint errors

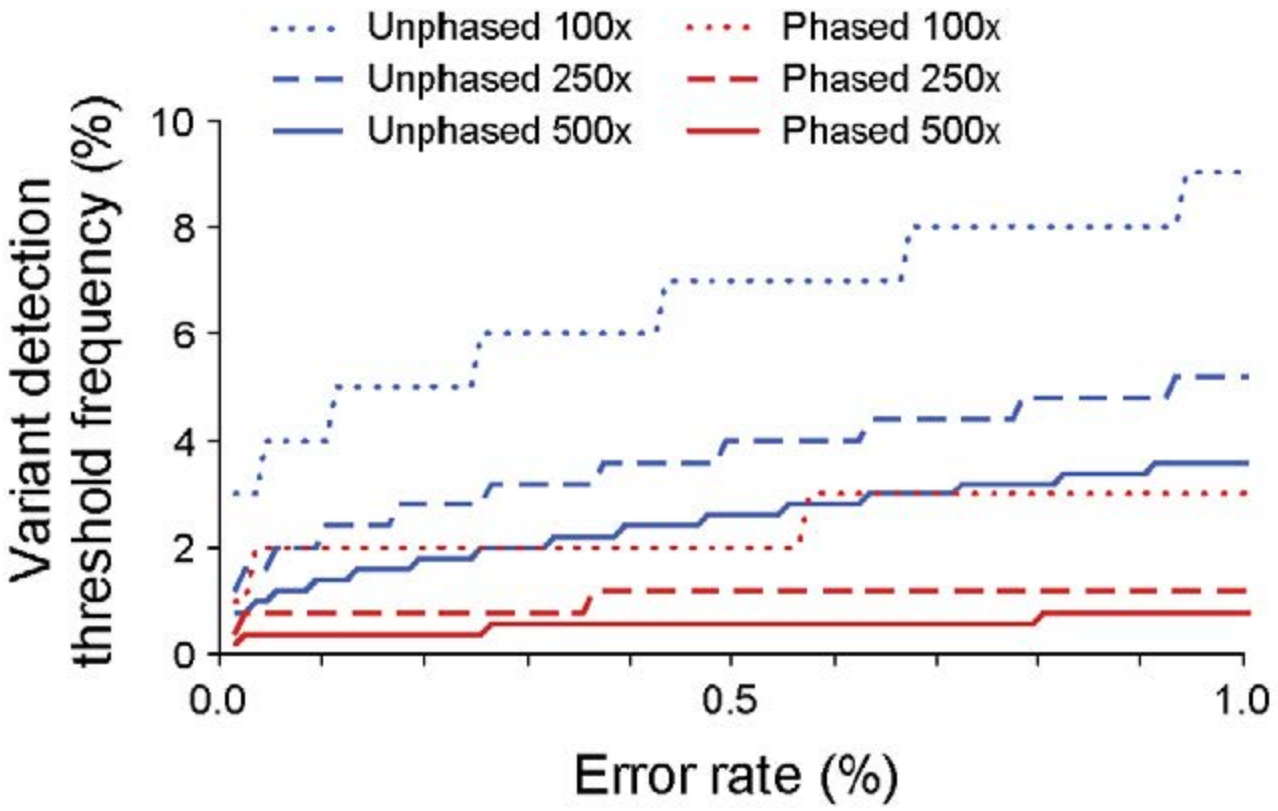
$$X_{ij} = \sum_{k=1}^{n_{ij}} E_{ijk}$$

  can still be computed recursively using the discrete convolution formula.
- Positions i and j are *phased*.
- Software: V-Phaser



Macalalad et al (PLoS Comput Biol 2012)

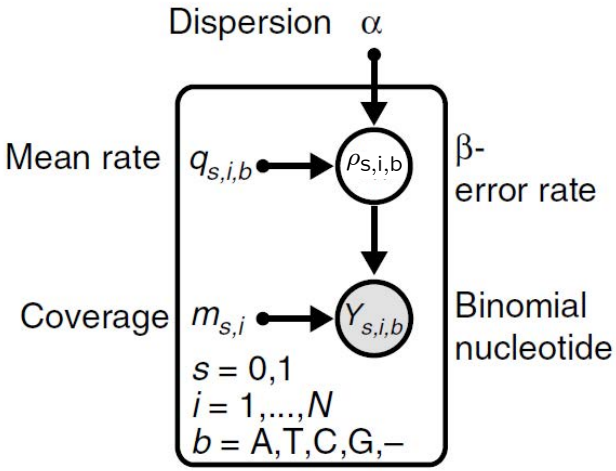
Phasing improves SNV calling



Macalalad et al (PLoS Comput Biol 2012)

Beta-binomial model (deepSNV)

- For each strand s, position i, and nucleotide b:



$$\rho_{s,i,b} \sim \text{Beta}(q_{s,i,b}, \alpha)$$

$$Y_{s,i,b} \sim \text{Binom}(m_{s,i}, \rho_{s,i,b})$$

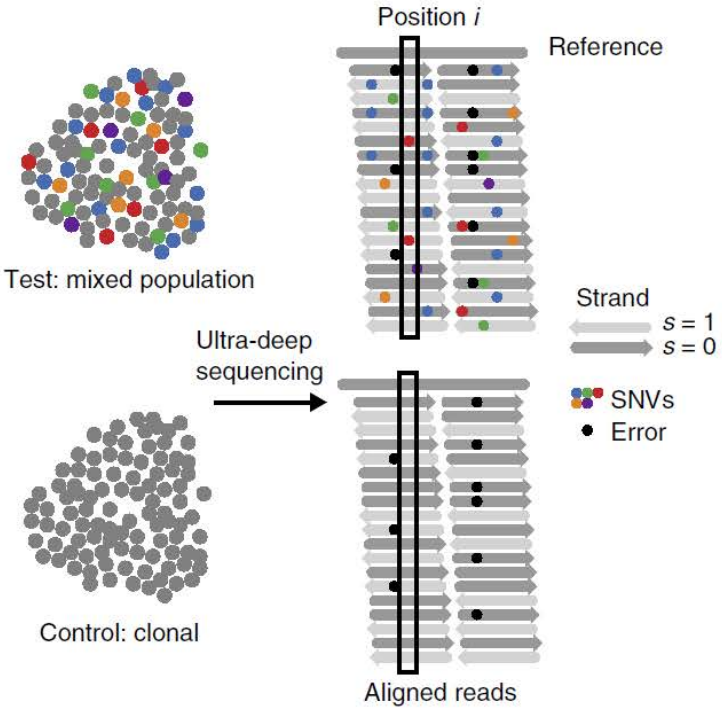
$$\Rightarrow \quad Y_{s,i,b} \sim \text{BetaBin}(m_{s,i}, q_{s,i,b}, \alpha)$$

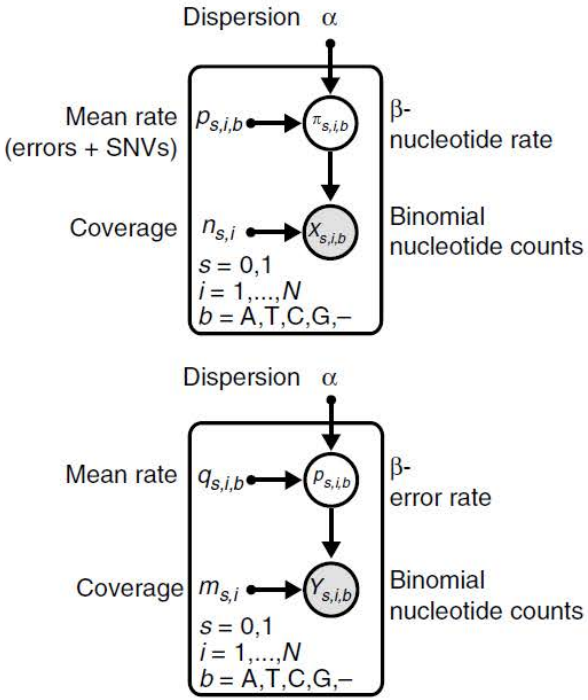
$$\text{E}[Y_{s,i,b}] = m_{s,i} q_{s,i,b}$$

$$\text{Var}[Y_{s,i,b}] \approx m_{s,i} q_{s,i,b} + (m_{s,i} q_{s,i,b})^2/\alpha, \quad \text{and CV} \approx 1/\alpha$$

Gerstung et al (Nat Commun 2012)

Minor allele frequency test





deepSNV algorithm

1. Test for each strand *s*, position *i*, and nucleotide *b*:

   Null-hypothesis: no SNV

   $p_{s,i,b} = q_{s,i,b}$

   Alternative: SNV *b*, frequency *f*

   $p_{s,i,b} = q_{s,i,b} + f > q_{s,i,b}$

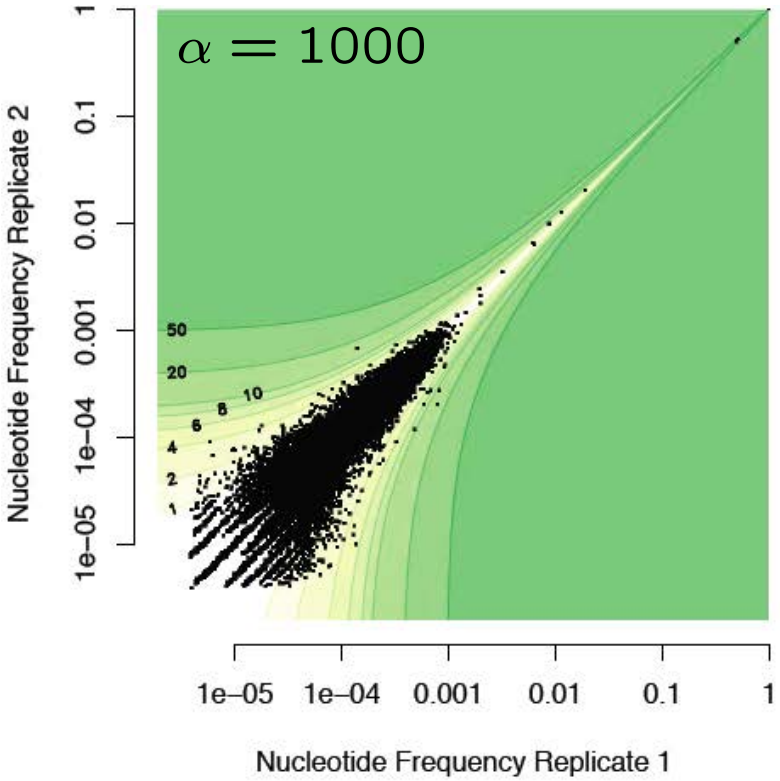
2. Combine *P*-values from each strand
3. Adjust *P*-values for multiple testing

$$-2 \log \frac{L(X_{s,i,b}, Y_{s,i,b} \mid H_0)}{L(X_{s,i,b}, Y_{s,i,b} \mid H_1)} \sim \chi^2_1$$

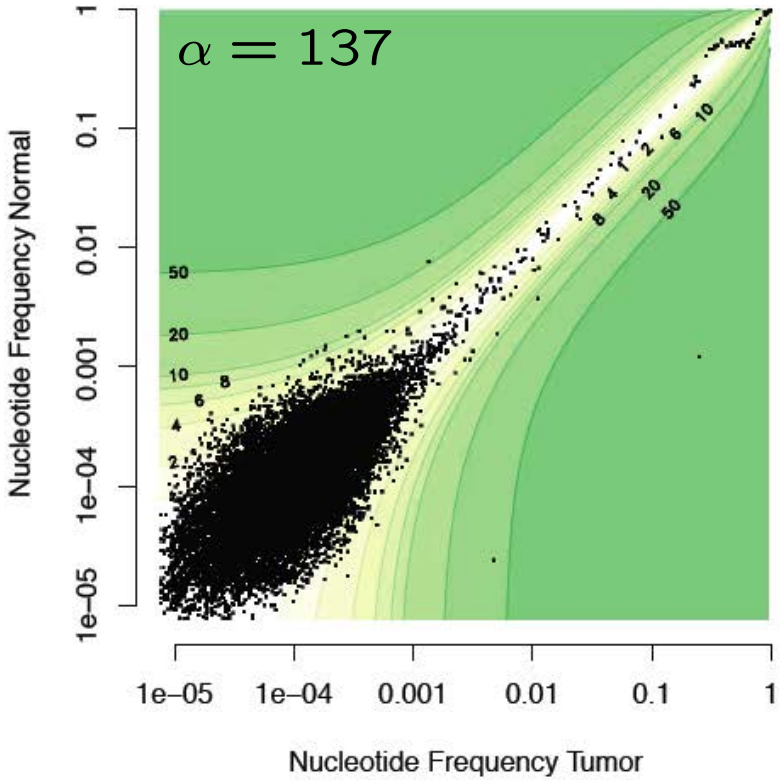
Gerstung et al (Nat Commun 2012)

Overdispersion

phiX replicates

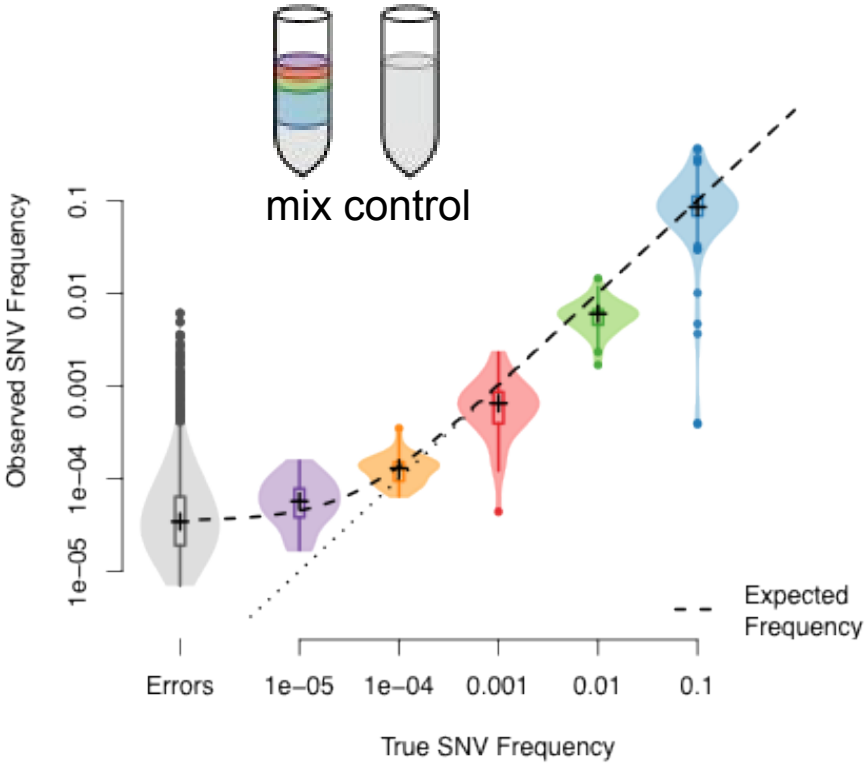


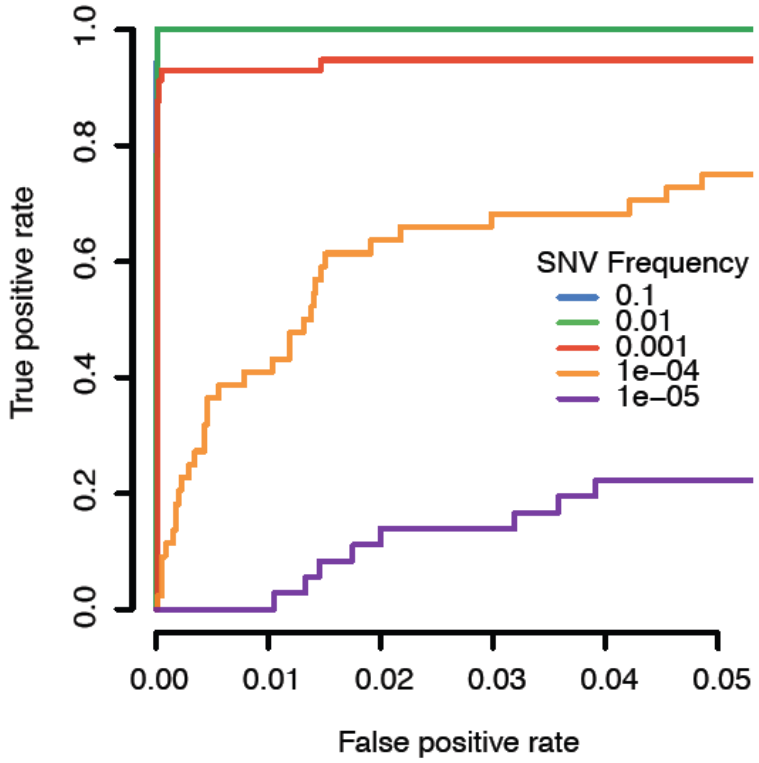
tumor vs. normal



Gerstung et al (Nat Commun 2012)

Test data: Known mix of 5 clones, coverage 10^5

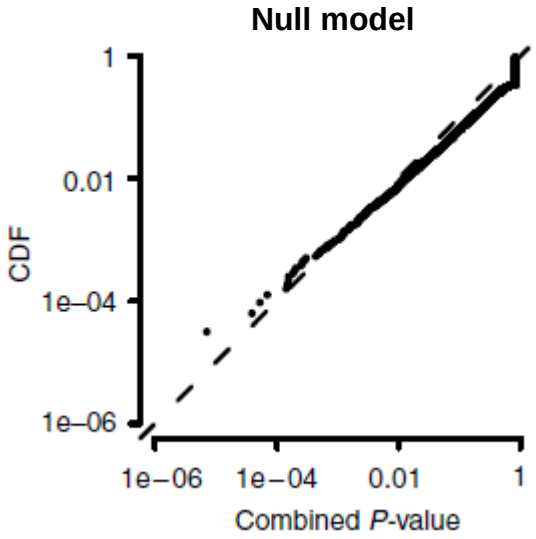




Gerstung et al (Nat Commun 2012)

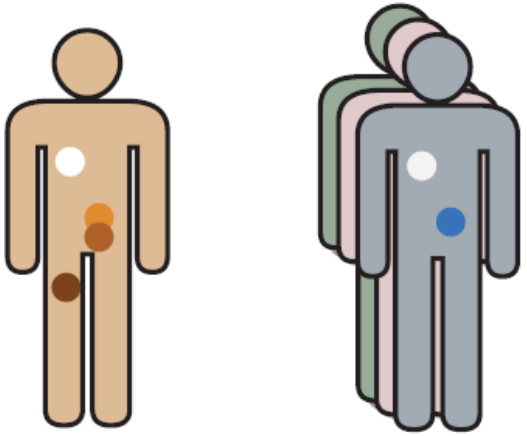
Performance comparison

	SNV frequency					Errors	CPU time
	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}		
Truth	101	46	57	44	36	5,740*	
deepSNV FDR < 0.05	101	46	53	3	0	2	141 s
deepSNV FWER < 0.05	99	46	49	0	0	0	141 s
VarScan[17] pileup2snp	96	42	26	32	8	472	361 s†
VarScan somatic	50	29	34	1	0	33	439 s†
CRISP[18]	91	43	46	0	0	16	44 h
vipR[19]	98	43	30	0	0	1	279 s†



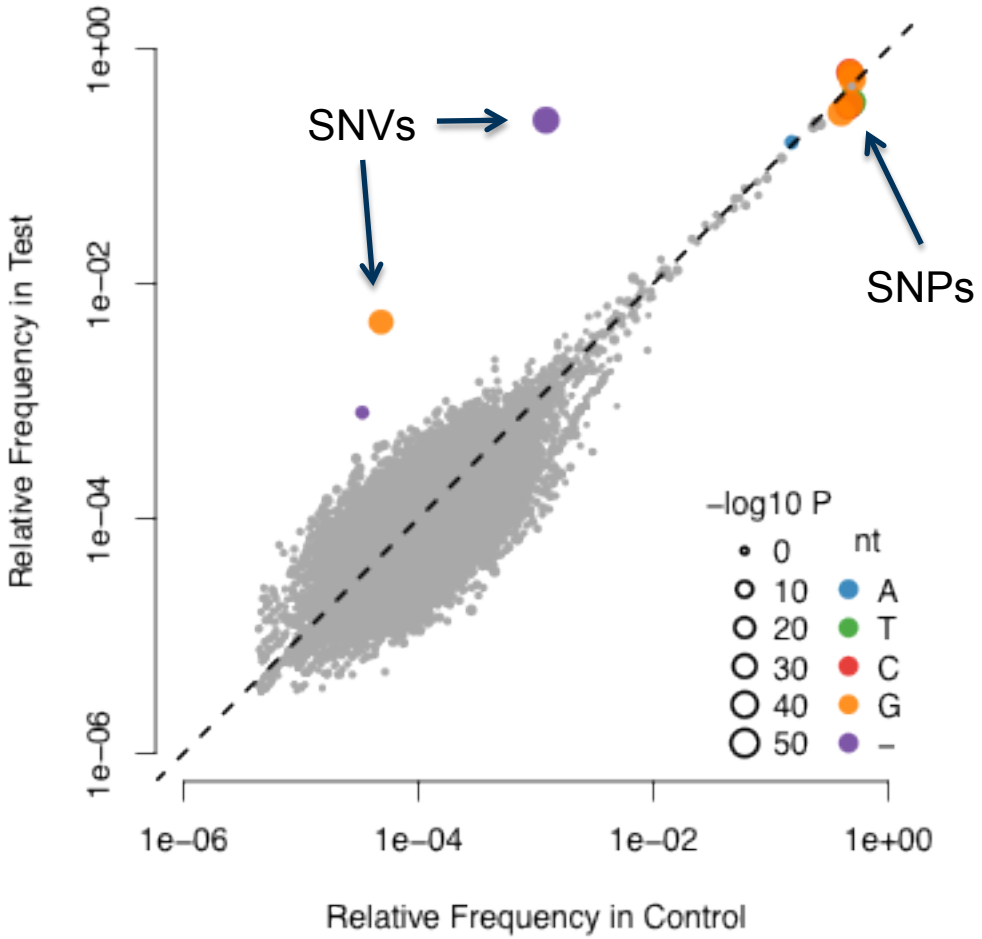
Gerstung et al (Nat Commun 2012)

Application: Renal cell carcinoma



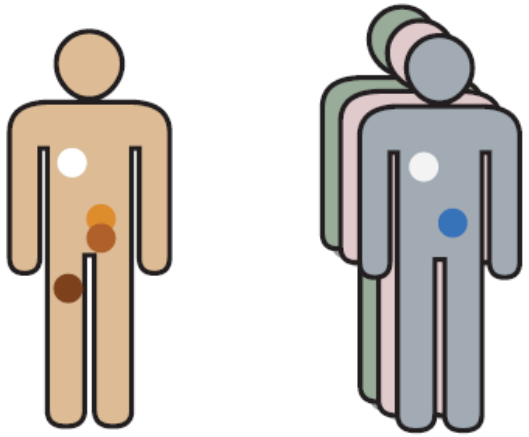
1x multiple lesions
Primary 1
Primary 2
Metastasis

3x tumor-normal:
Tumor 1
Tumor 2
Tumor 3



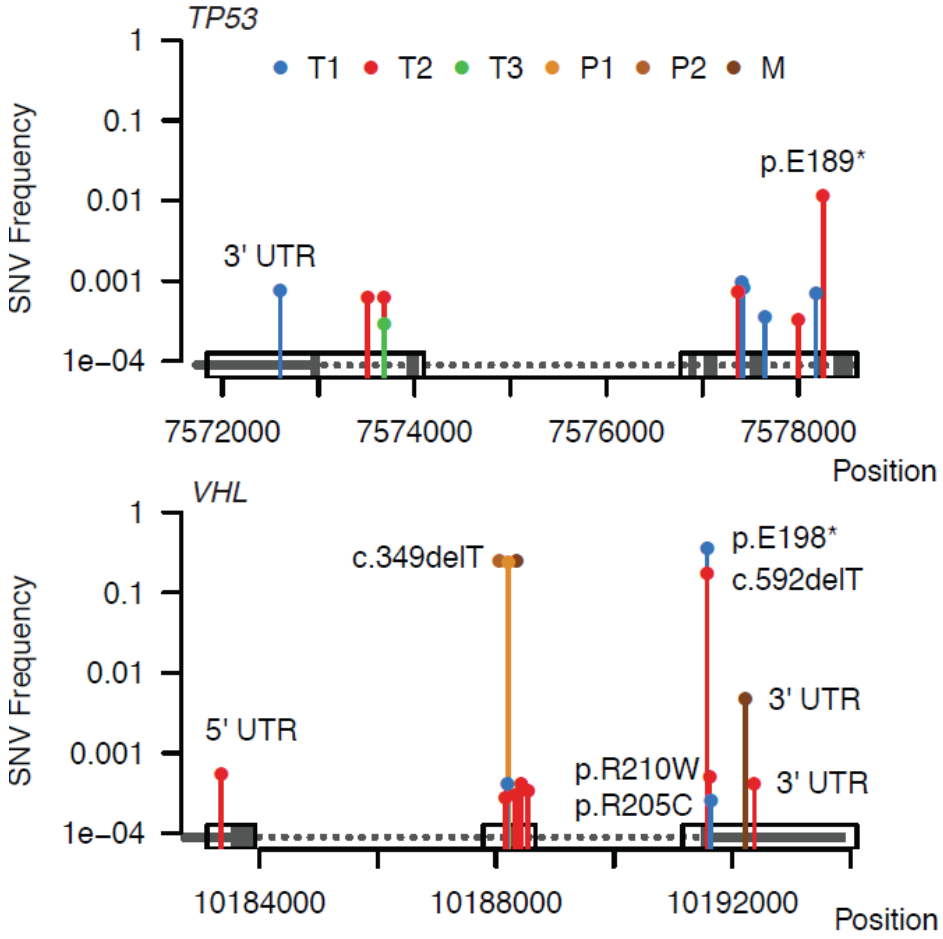
Gerstung et al (Nat Commun 2012)

Application: Renal cell carcinoma



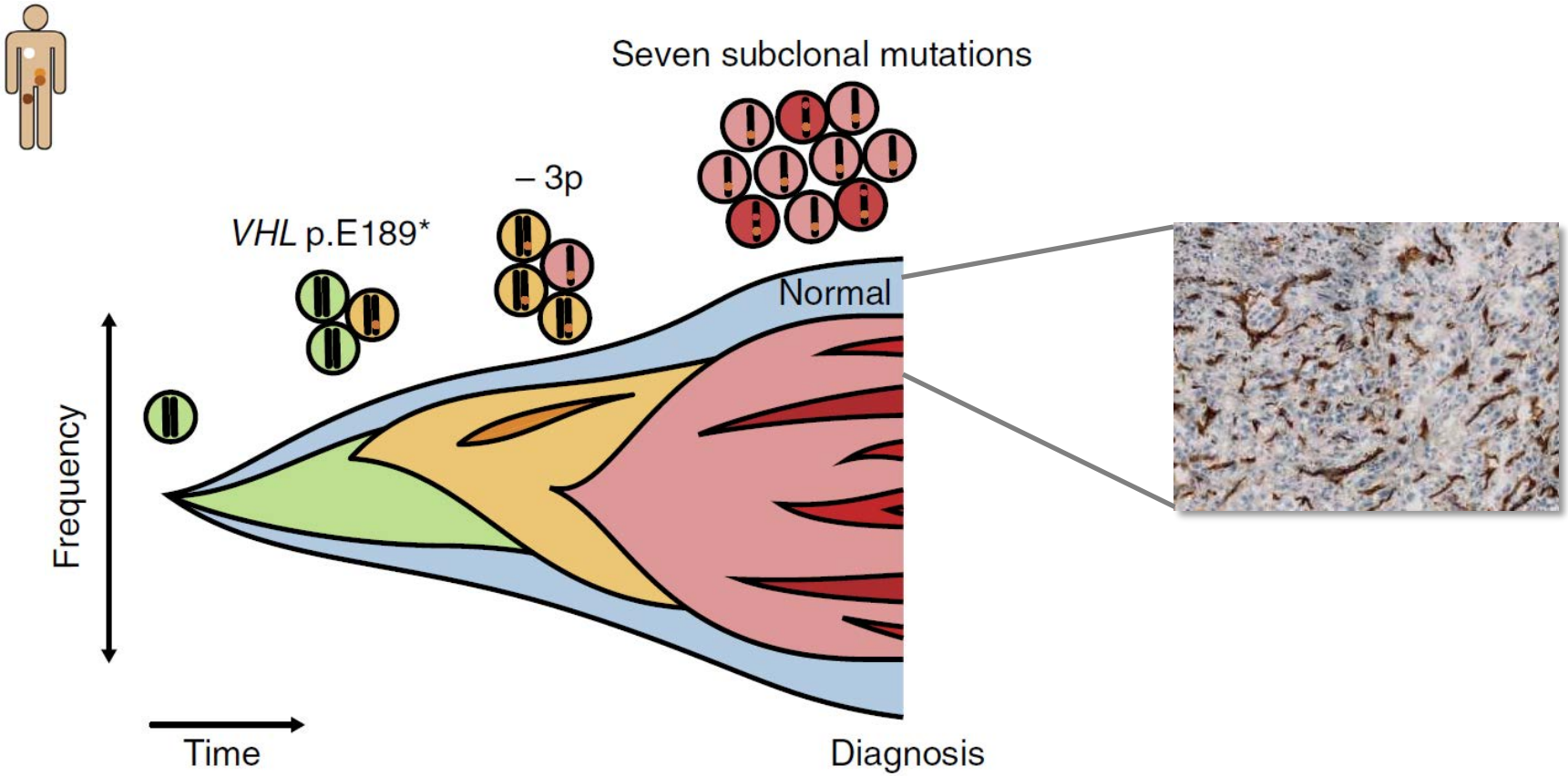
1x multiple lesions
Primary 1
Primary 2
Metastasis

3x tumor-normal:
Tumor 1
Tumor 2
Tumor 3



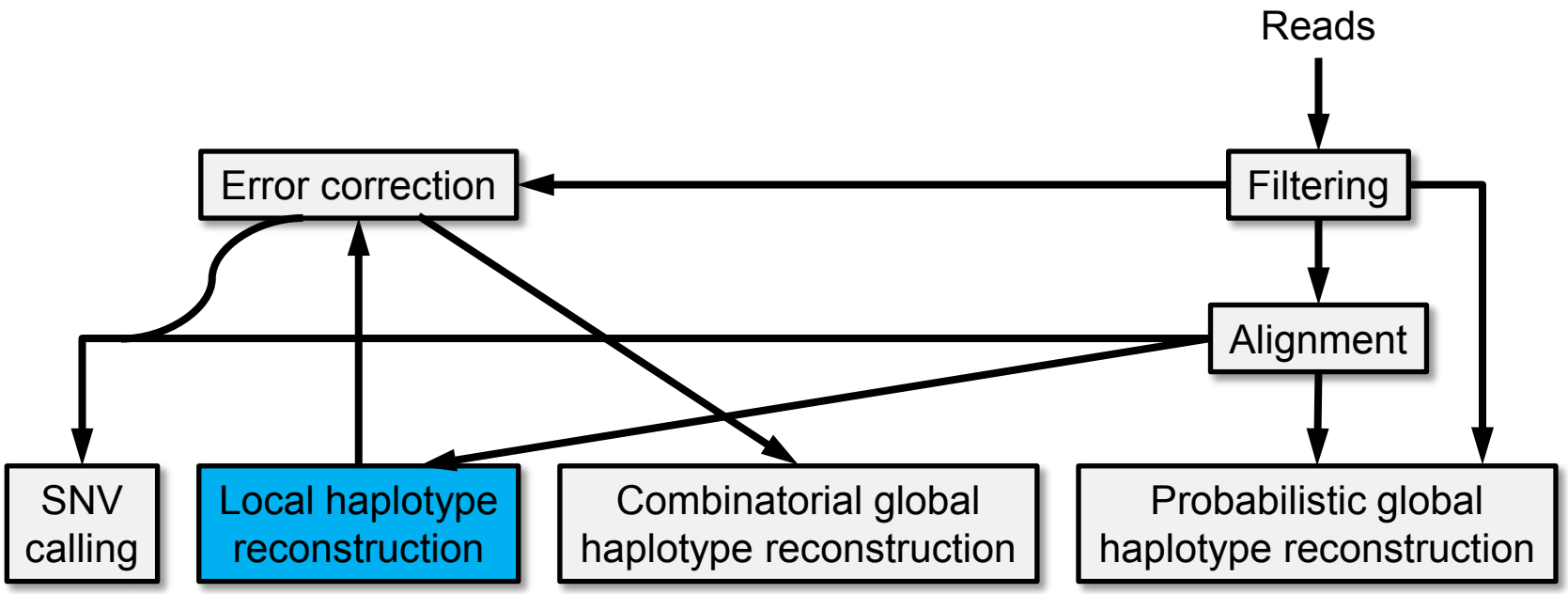
Gerstung et al (Nat Commun 2012)

Evolutionary history



Gerstung et al (Nat Commun 2012)

Overview

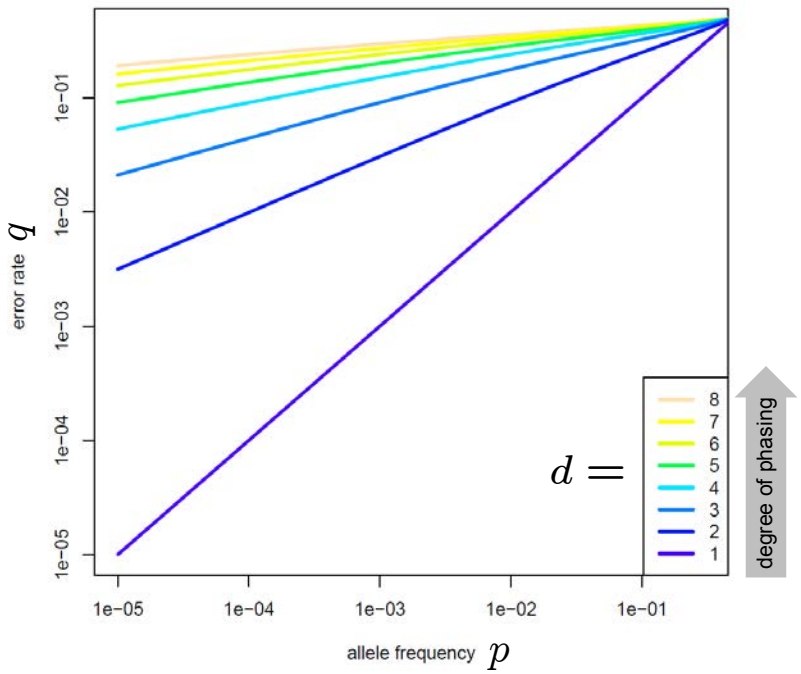

Reads
Error correction
Filtering
Alignment
SNV calling
Local haplotype reconstruction
Combinatorial global haplotype reconstruction
Probabilistic global haplotype reconstruction

Local haplotype inference

Beerenwinkel et al (Front Microbio 2012)

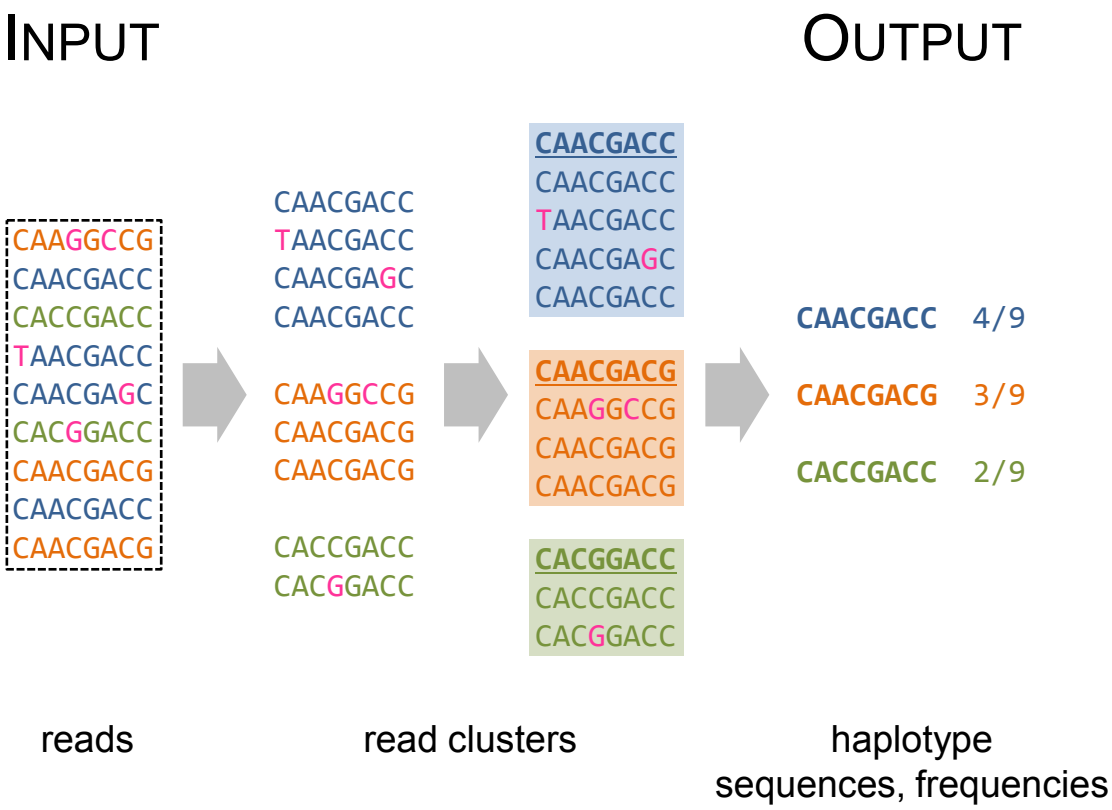
Phasing improves limit of detection

- In a simple model, d phased SNVs (i.e., a haplotype) of frequency p can be called correctly at an i.i.d. error rate q, if and only if $q^d < p$.



McElroy et al (submitted)

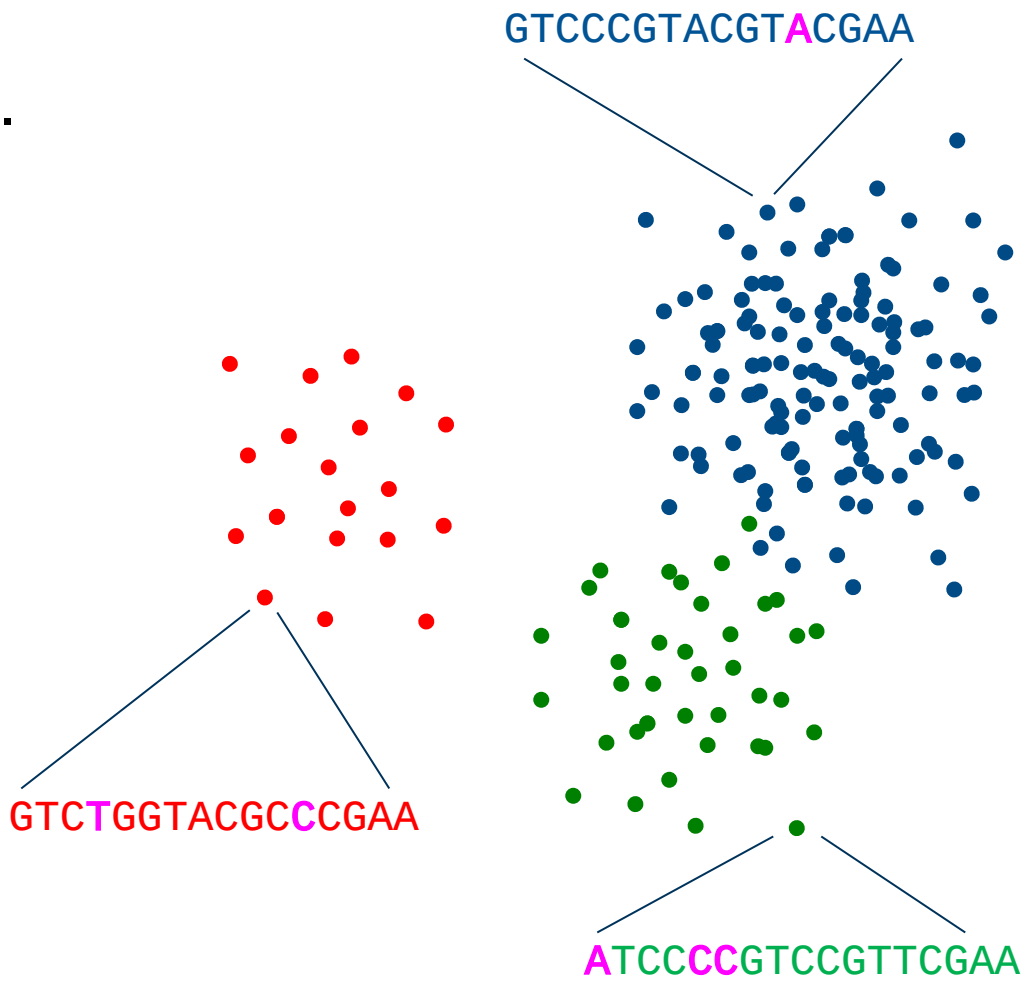
Local haplotype inference via read clustering



Beerenwinkel et al (Front Microbio 2012)

Probabilistic clustering

- We assume an i.i.d. error rate $1 - \theta$.
- Main problem: number of clusters (haplotypes) unknown
- Bayesian approach
  - Dirichlet process mixture (Chinese restaurant process)
  - Gibbs sampler



Zagordi et al (J Comput Biol 2010)

Likelihood

- The probability of observed reads $\mathbf{r}$, given haplotype assignments $\mathbf{c}$ and haplotypes $\mathbf{h}$, is

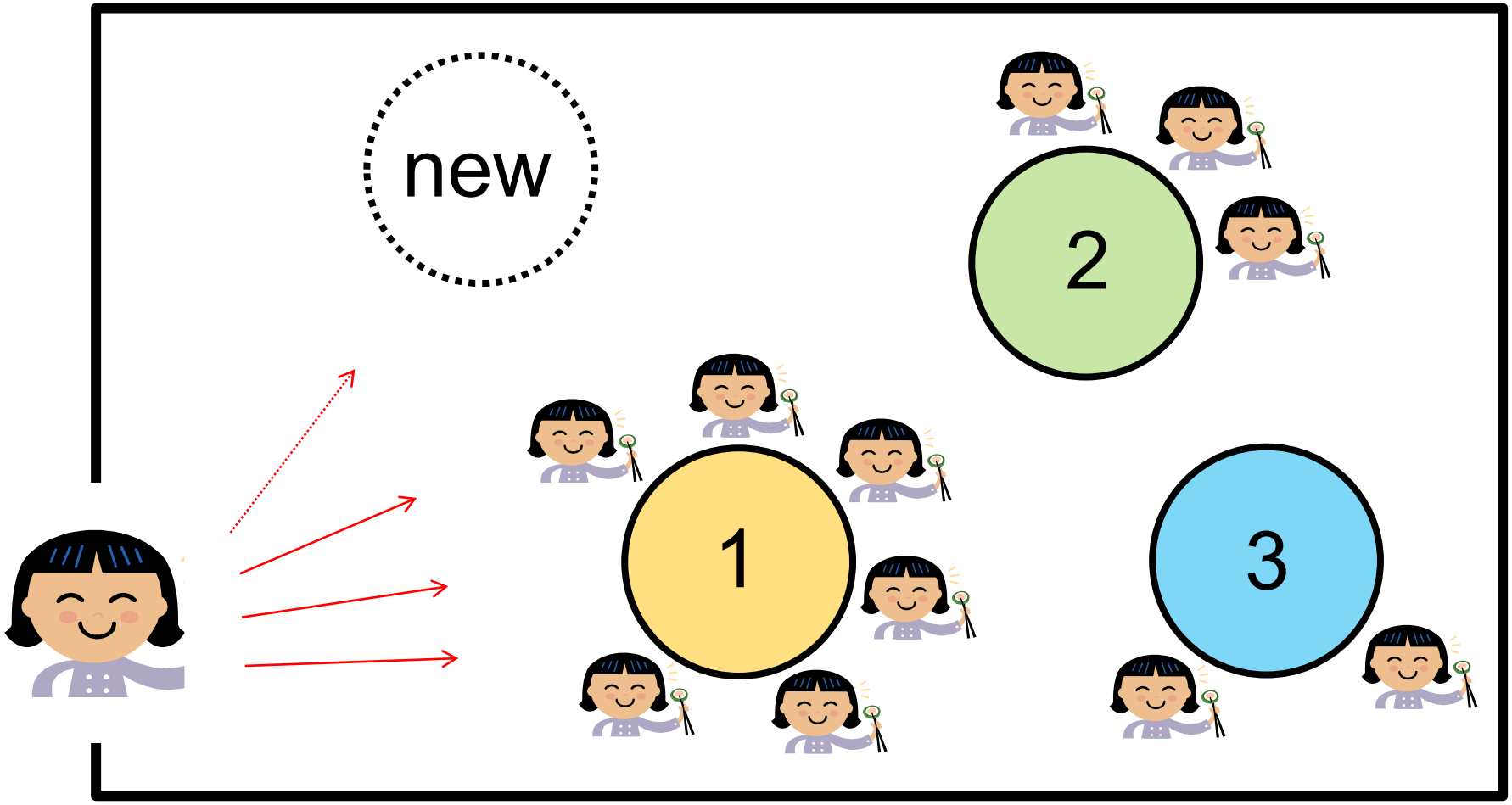
$$P(\mathbf{r} \mid \mathbf{c}, \mathbf{h}) = \prod_k \theta^{m_k} \left(\frac{1-\theta}{|B-1|} \right)^{m'_k},$$

$$m_k = \sum_{i,j} \mathbb{I}(r_{i,j} = h_{k,j}) \mathbb{I}(c_i = k) \quad \text{matches}$$

$$m'_k = \sum_{i,j} \mathbb{I}(r_{i,j} \neq h_{k,j}) \mathbb{I}(c_i = k) \quad \text{mismatches}$$

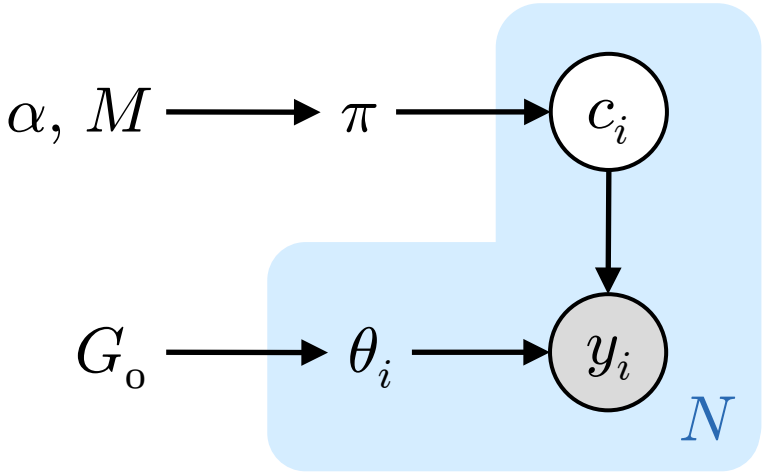
where B is the alphabet size.

Chinese restaurant process



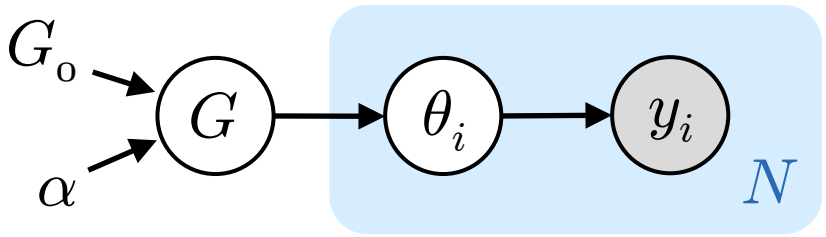
Finite mixture model, K < ∞

$$y_i \sim \sum_{j=1}^{K} \pi_j F(y_i \mid \theta_j)$$



$$\begin{aligned} y_i \mid c_i, \theta &\sim F(y \mid \theta_{c_i}) \\ c \mid \pi &\sim \text{Discrete}(\pi) \\ \theta_i &\sim G_0(\theta) \\ \pi &\sim \text{Dir}(\alpha, M) \end{aligned}$$

Dirichlet process mixture, K → ∞

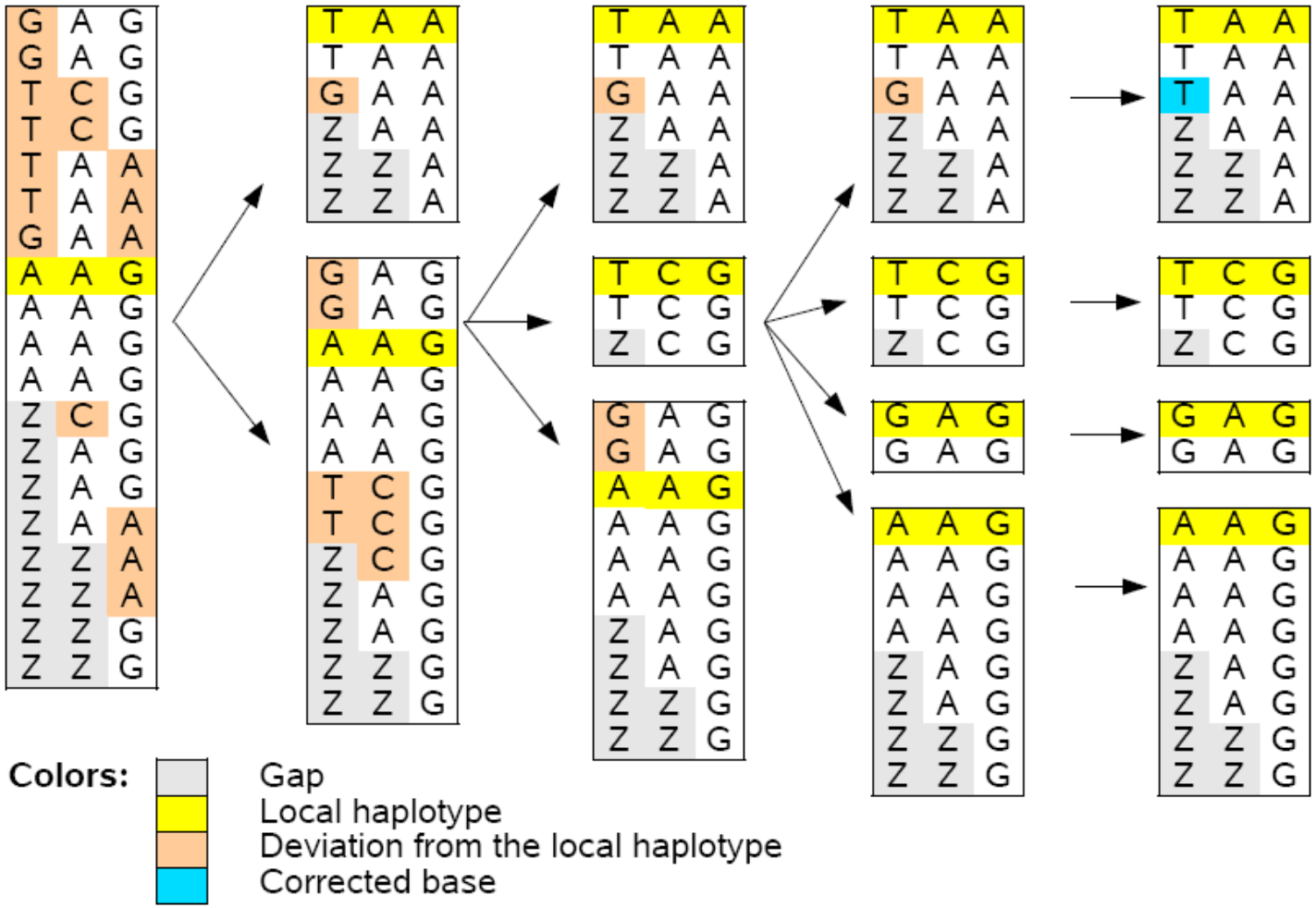


$$
\begin{aligned}
y_i \mid \theta_i &\sim F(y \mid \theta_i) \\
\theta_i \mid G &\sim G(\theta) \\
G &\sim \mathrm{DP}(\alpha, G_0(\theta))
\end{aligned}
$$

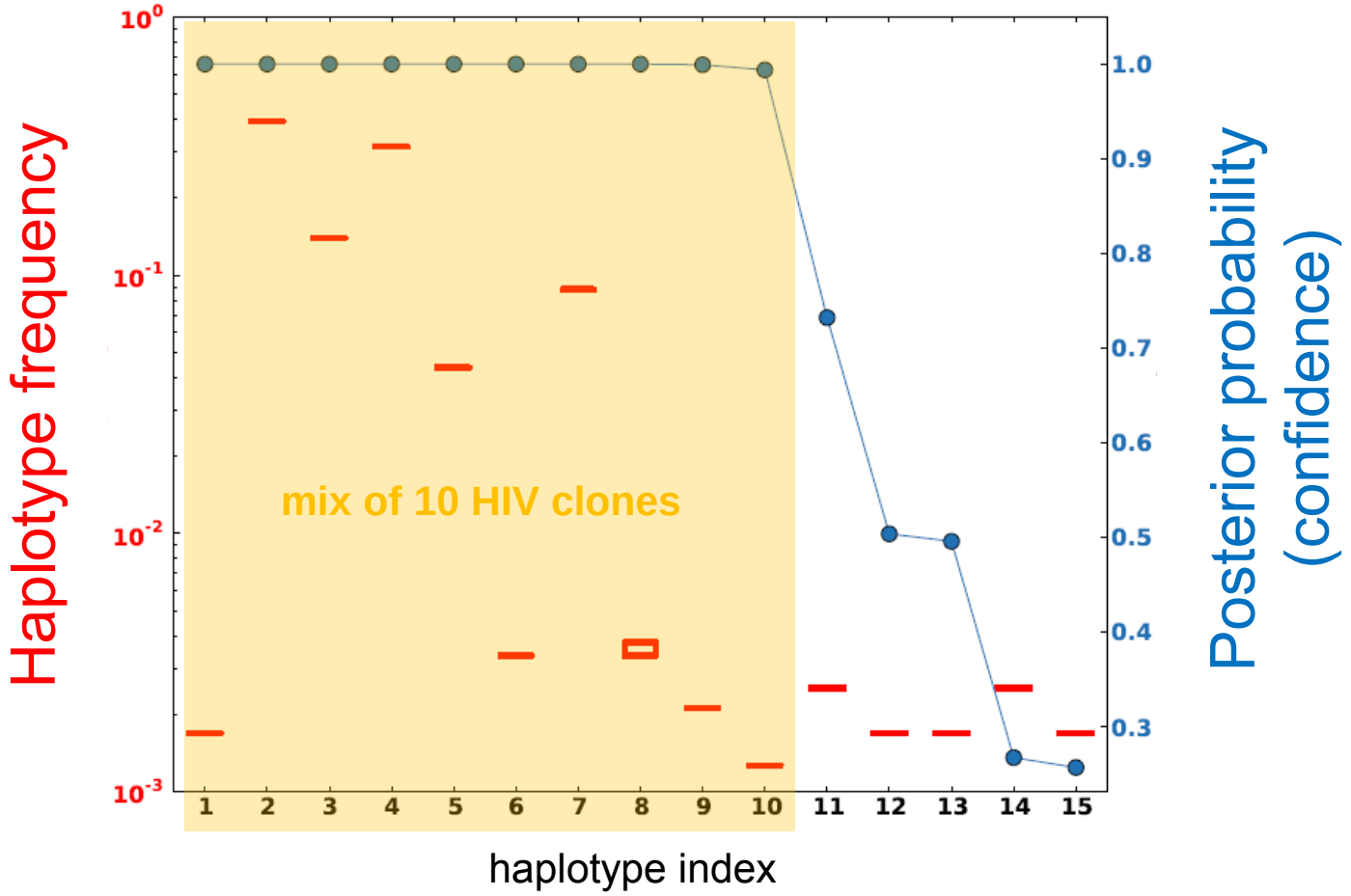
$$P(c_i = c \mid c_{1:i-1}) \rightarrow \frac{n_{i,c}}{i - 1 + \alpha}$$

$$P(c_i \neq c_j \text{ for all } j < i \mid c_{1:i-1}) \rightarrow \frac{\alpha}{i - 1 + \alpha}$$

Example: 19 reads of length 3

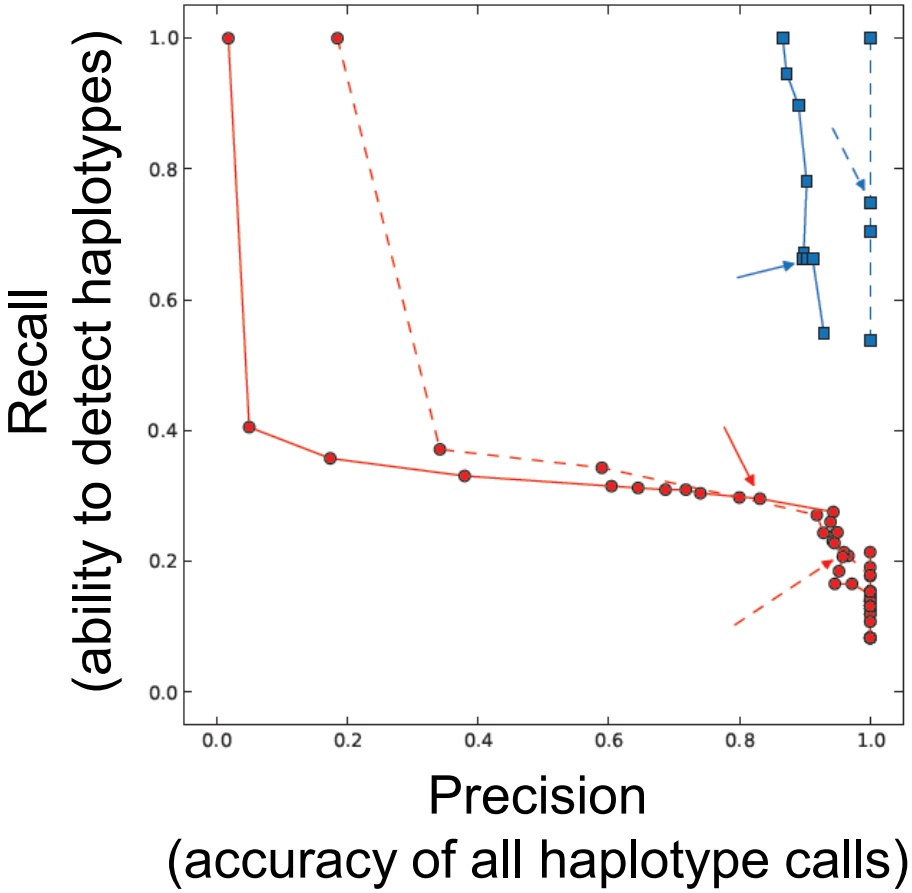


Output: haplotypes, frequencies, posterior



Zagordi et al (Nucl Acids Res 2010)

Performance of haplotype reconstruction, ROC curve (control experiment of 10 cloned viruses)



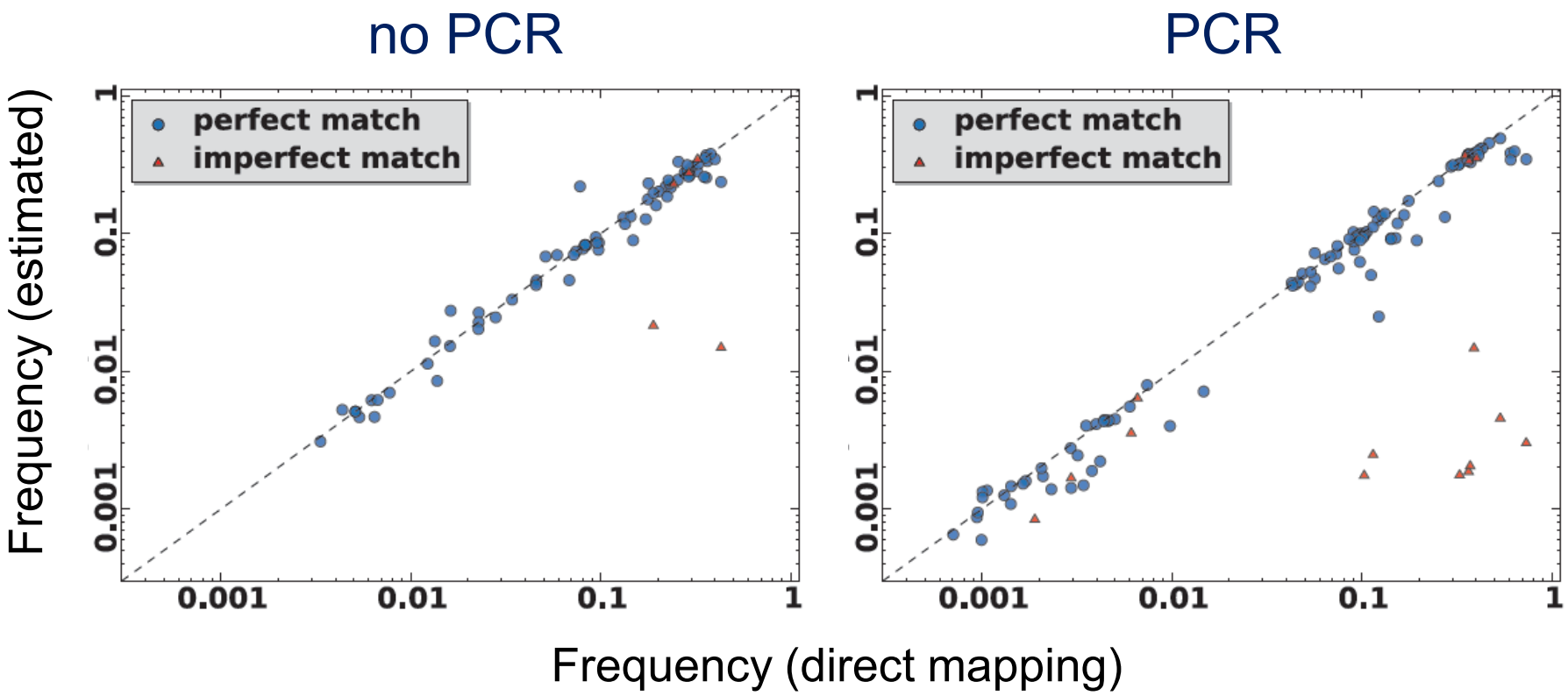
cut-off
(↗ default 50)

posterior
(↗ default 0.9)

— PCR
--- no PCR

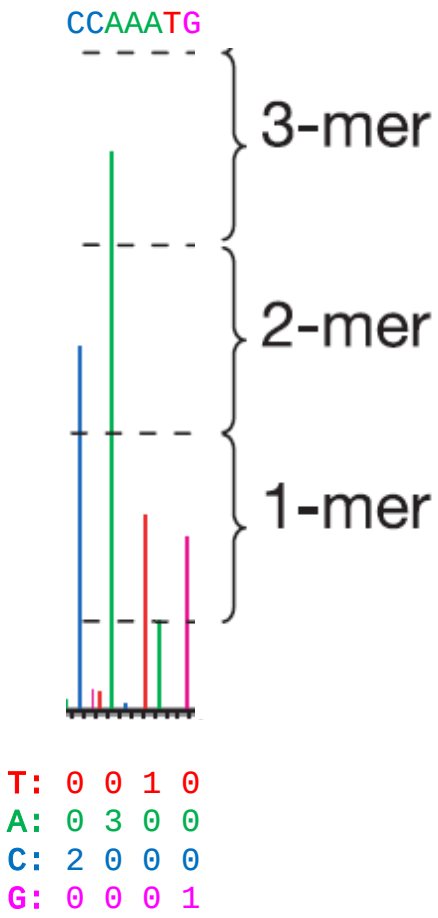
Zagordi et al (Nucl Acids Res 2010)

Haplotype frequency estimation

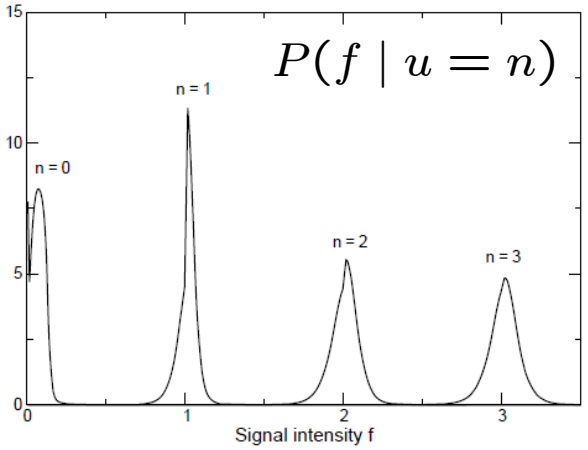


Zagordi et al (Nucl Acids Res 2010)

Clustering flograms (AmpliconNoise)



- Observed flowgrams f are obtained from ideal flowgrams u subject to noise (which is estimated from control data).



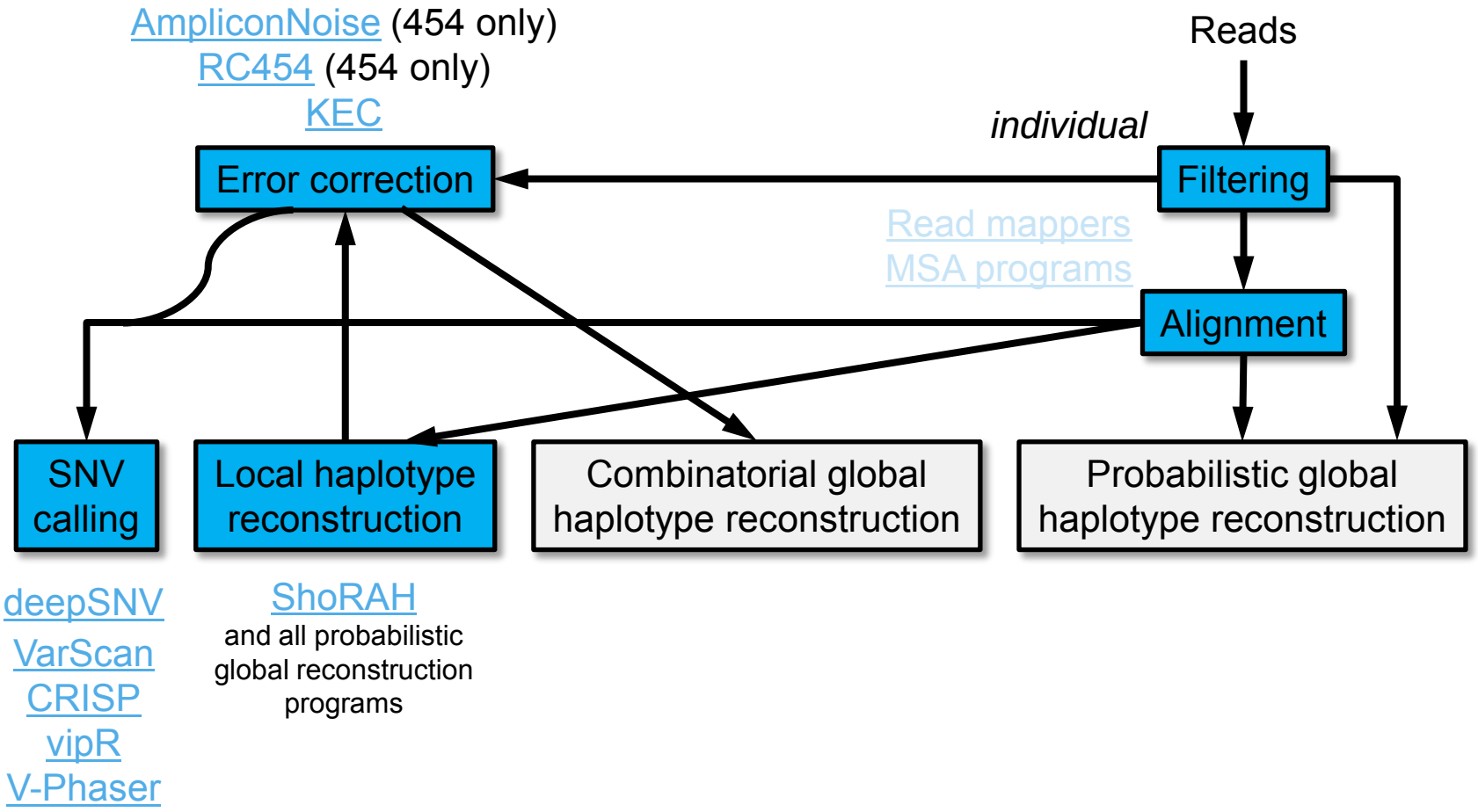
- Flowgrams are clustered (around ideal flowgram centers) with the distance measure

$$d(f,u) = -\frac{1}{M}\sum_{i=1}^{M} \log P(f_i \mid u_i)$$

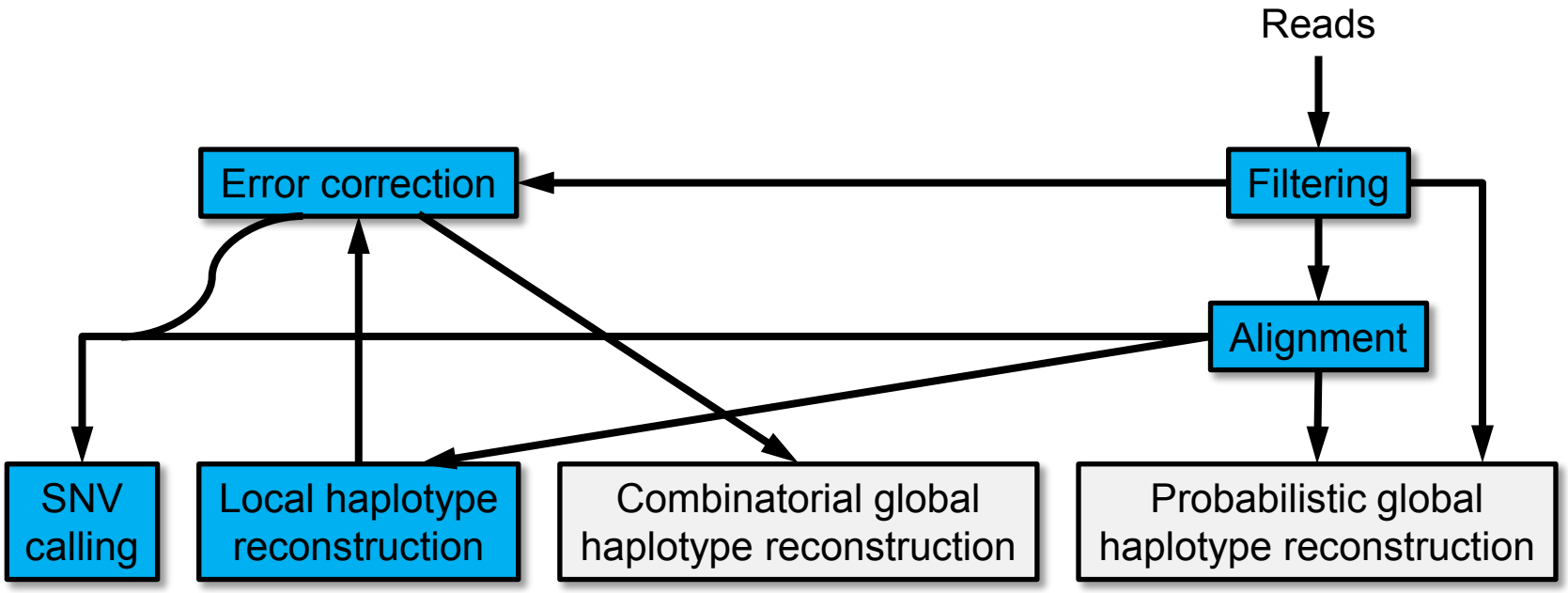
using an EM algorithm.

Quince et al (Nat Meth 2009, BMC Bioinformatics 2011)

Software


AmpliconNoise (454 only)
RC454 (454 only)
KEC
Reads
individual
Error correction
Filtering
Read mappers
MSA programs
Alignment
SNV calling
Local haplotype reconstruction
Combinatorial global haplotype reconstruction
Probabilistic global haplotype reconstruction
deepSNV
VarScan
CRISP
vipR
V-Phaser
ShoRAH
and all probabilistic global reconstruction programs

Summary and discussion



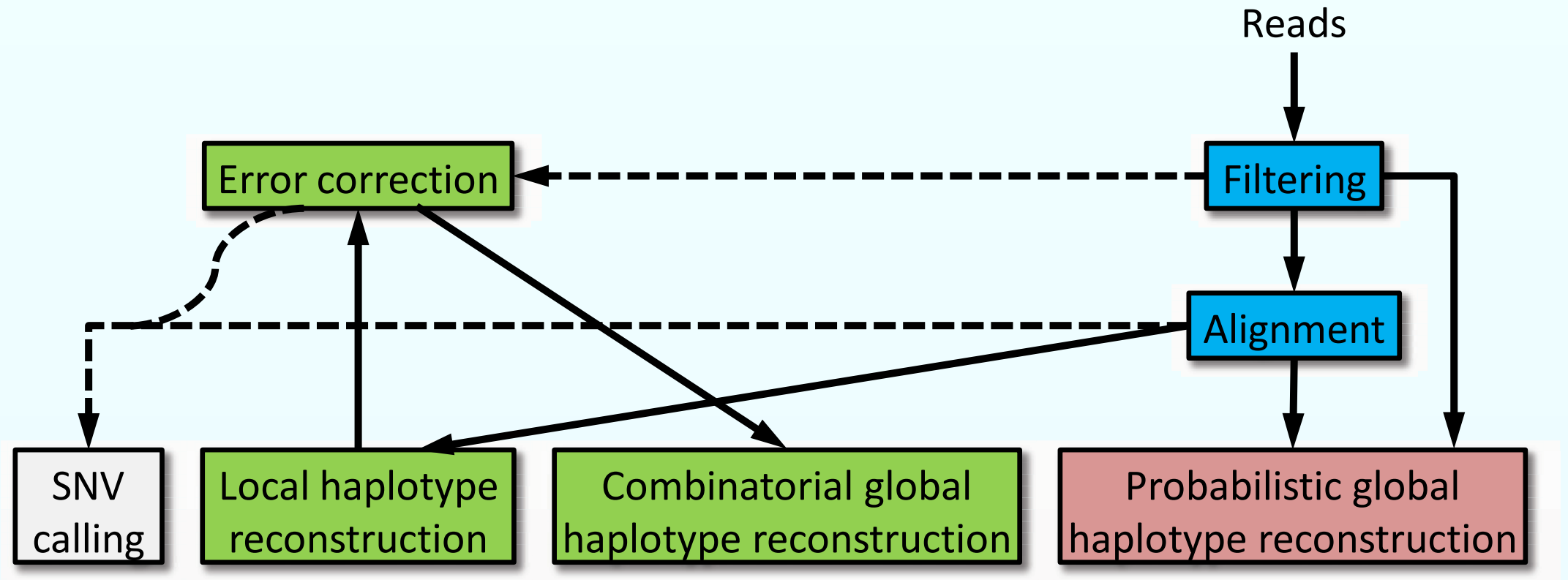
- Preprocessing, uncertainty
- Direct SNV calling vs. local-to-SNV

ECCB 2012 Tutorial 4

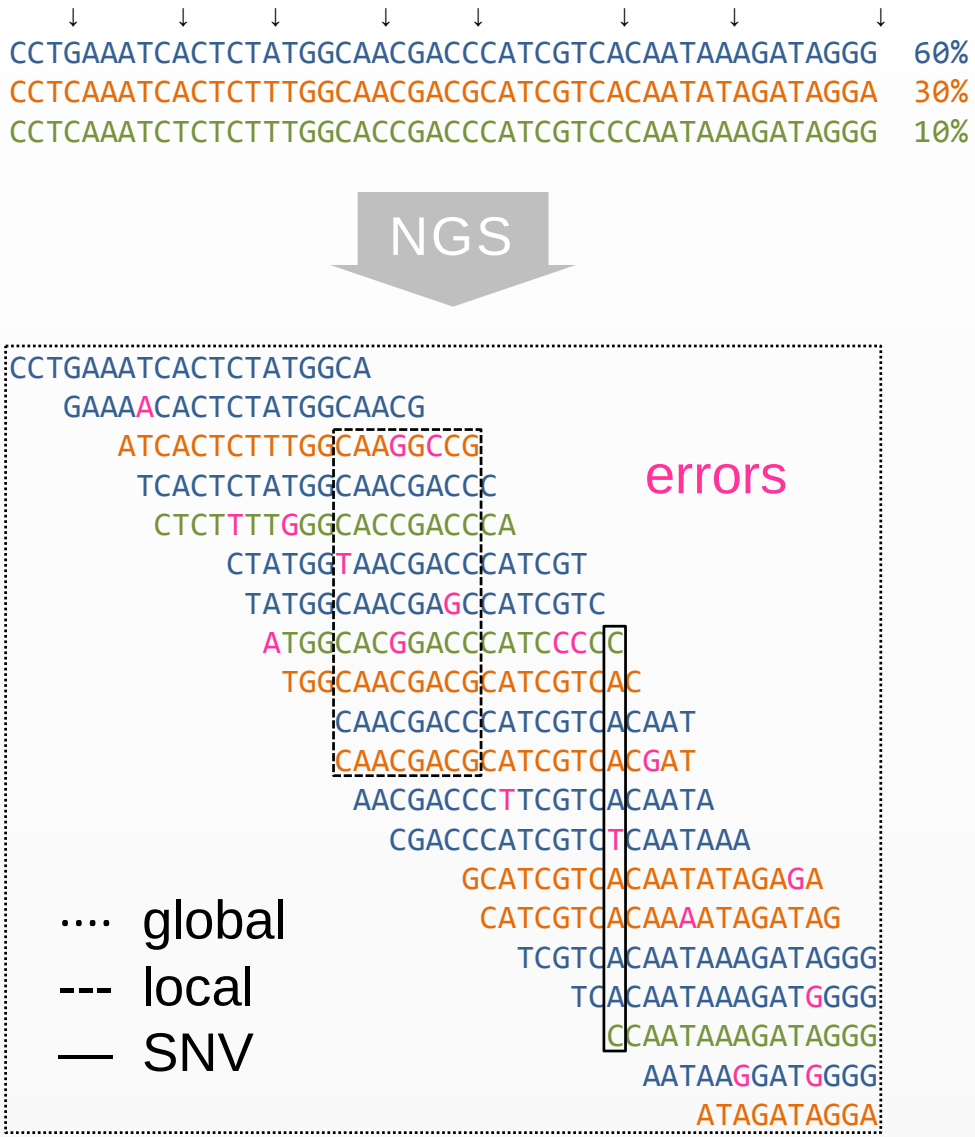
Global Haplotype Assembly

Volker Roth, Department of Mathematics and Computer Science,
University of Basel

Overview


Reads
Error correction
Filtering
Alignment
SNV calling
Local haplotype reconstruction
Combinatorial global haplotype reconstruction
Probabilistic global haplotype reconstruction

Global Haplotype Assembly



Global Haplotype Assembly

Combinatorial assembly

- Network flow
  (Westbrooks et al, 2008)
- Minimal path cover
  (Eriksson et al, 2008)
- Greedy paths sampling
  (Prosperi et al., 2011, 2012)
- Graph coloring
  (Huang et al., 2012)

☺ **well-studied graph-theoretic background.**

☹ **requires error correction prior to assembly.**

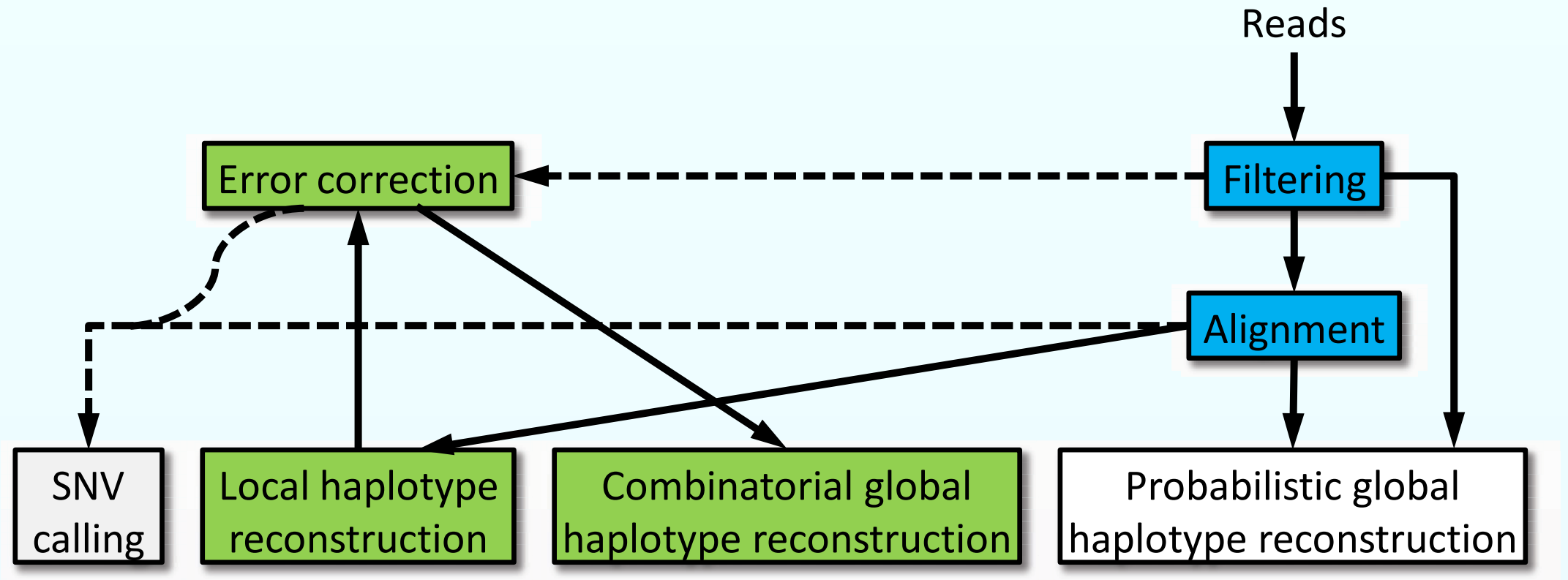
Probabilistic assembly

- Integrating alignment
  (Jojic et al., 2008)
- Local-to-global mixture model
  (Prabhakaran et al., 2010)
- Modeling recombinants
  (Beerenwinkel et al., 2012)

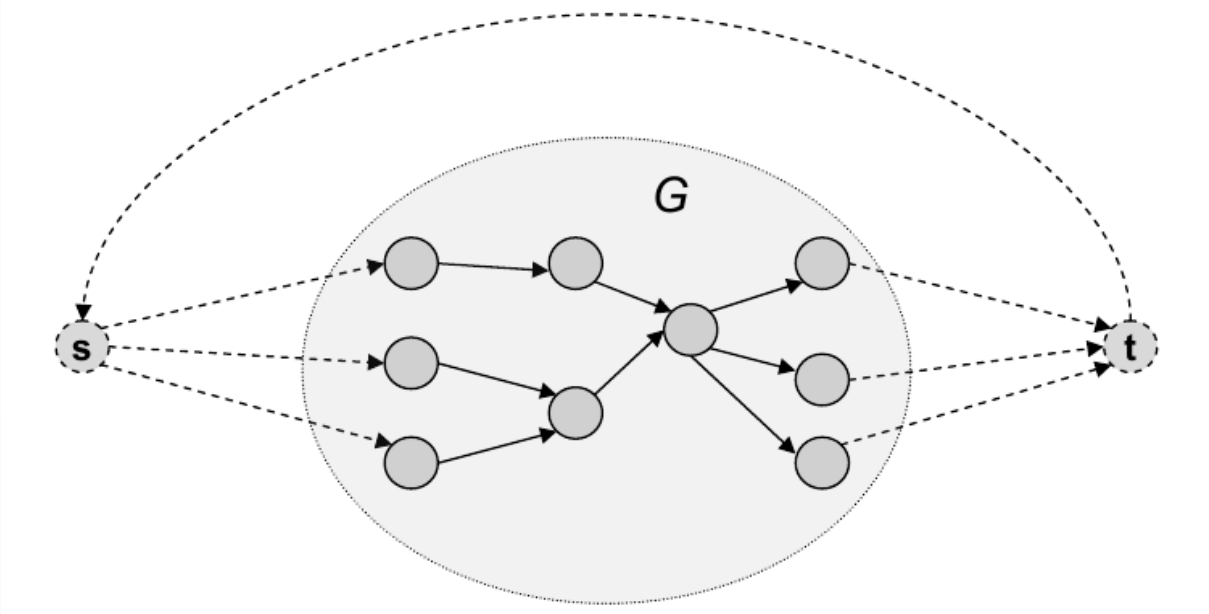
☺ **"integrated", no separate "hard" error correction.**

☹ **computational problems, approximations needed.**

Overview


Reads
Error correction
Filtering
Alignment
SNV calling
Local haplotype reconstruction
Combinatorial global haplotype reconstruction
Probabilistic global haplotype reconstruction

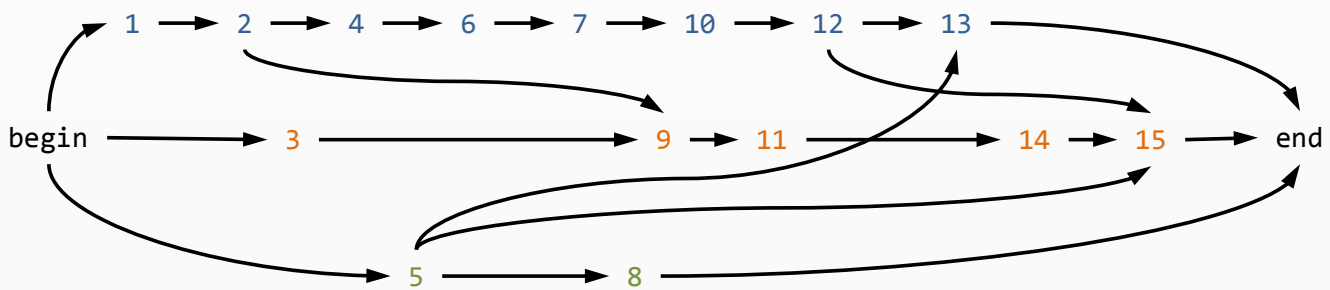
Combinatorial Assembly: Network Flow



Westbrooks et al (ISBRA 2008)

Combinatorial Assembly: Read Graph

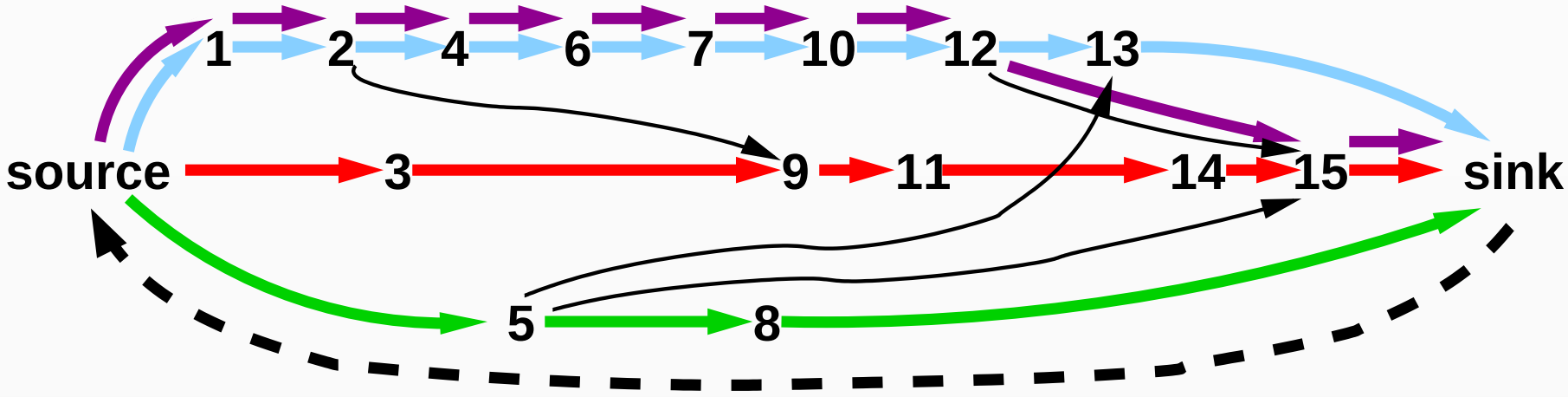
```
      ↓     ↓     ↓     ↓     ↓         ↓     ↓         ↓
A  CCTGAAATCACTCTATGGCAACGACCCATCGTCACAATAAAGATAGGG  60%
B  CCTCAAATCACTCTTTGGCAACGACGCATCGTCACAATATAGATAGGA  30%
C  CCTCAAATCTCTCTTTGGCACCGACCCATCGTCCCAATAAAGATAGGG  10%

 1 CCTGAAATCACTCTATGGCA
 2    GAAATCACTCTATGGCAACG
 3       ATCACTCTTTGGCAACGACG
 4        TCACTCTATGGCAACGACCC
 5         CTCTCTTTGGCACCGACCCA
 6             CTATGGCAACGACCCATCGT
 7              TATGGCAACGACCCATCGTC
 8               TTGGCACCGACCCATCGTCC
 9                TGGCAACGACGCATCGTCAC
10                   CAACGACCCATCGTCACAAT
11                   CAACGACGCATCGTCACAAT
12                    AACGACCCATCGTCACAATA
13                      CGACCCATCGTCACAATAAA
14                          GCATCGTCACAATATAGATA
15                           CATCGTCACAATATAGATAG
```



Each path is a potential haplotype

Network Flow

- A HT h corresponds to a **path from source to sink** in the read graph.
- Each path can be viewed as a **flow** *source* $\rightarrow$ $\{$*reads from* $h\}$ $\rightarrow$ *sink*.
- The value of the (circular) **flow** f through a read is the **number of haplotypes that contain the read**.
- **Main idea:** minimizing flow $\rightsquigarrow$ most parsimonious HT assembly.



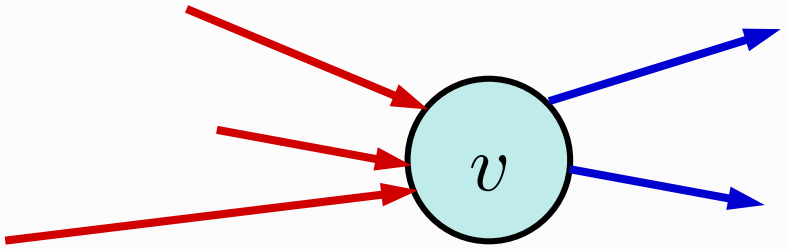
Quasispecies assembly via network flows

LP for Most Parsimonious Quasispecies Assembly:

Objective: Minimize **backflow** $f(\text{sink}, \text{source}) \rightsquigarrow$ **parsimonious:** every unit of flow from a single HT should pass through $(\text{sink}, \text{source})$ **once**.

Subject to:

- **Flow conservation:**



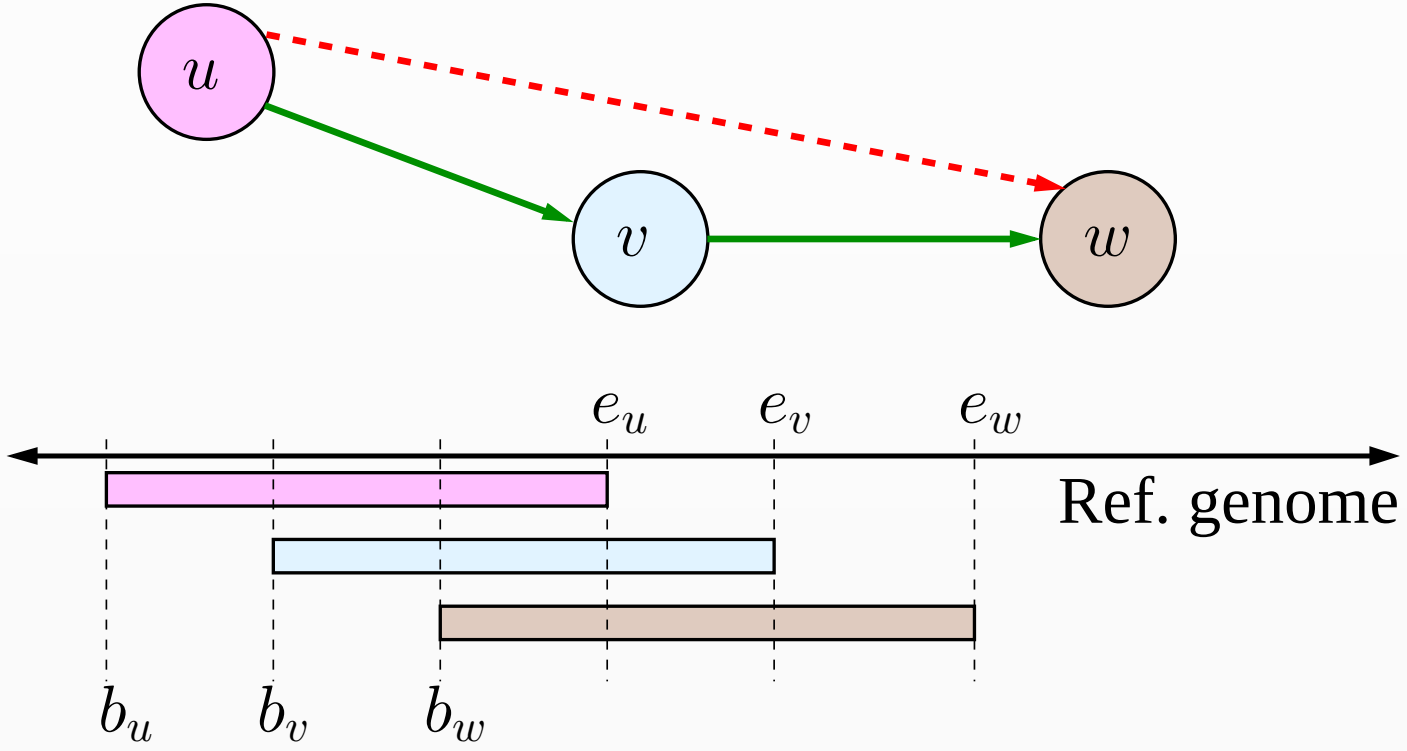
$$\sum_{e_{\text{in}}} f(e) = \sum_{e_{\text{out}}} f(e)$$

- **Each read covered by at least one haplotype**

Extension: include **cost terms** for the individual flows.

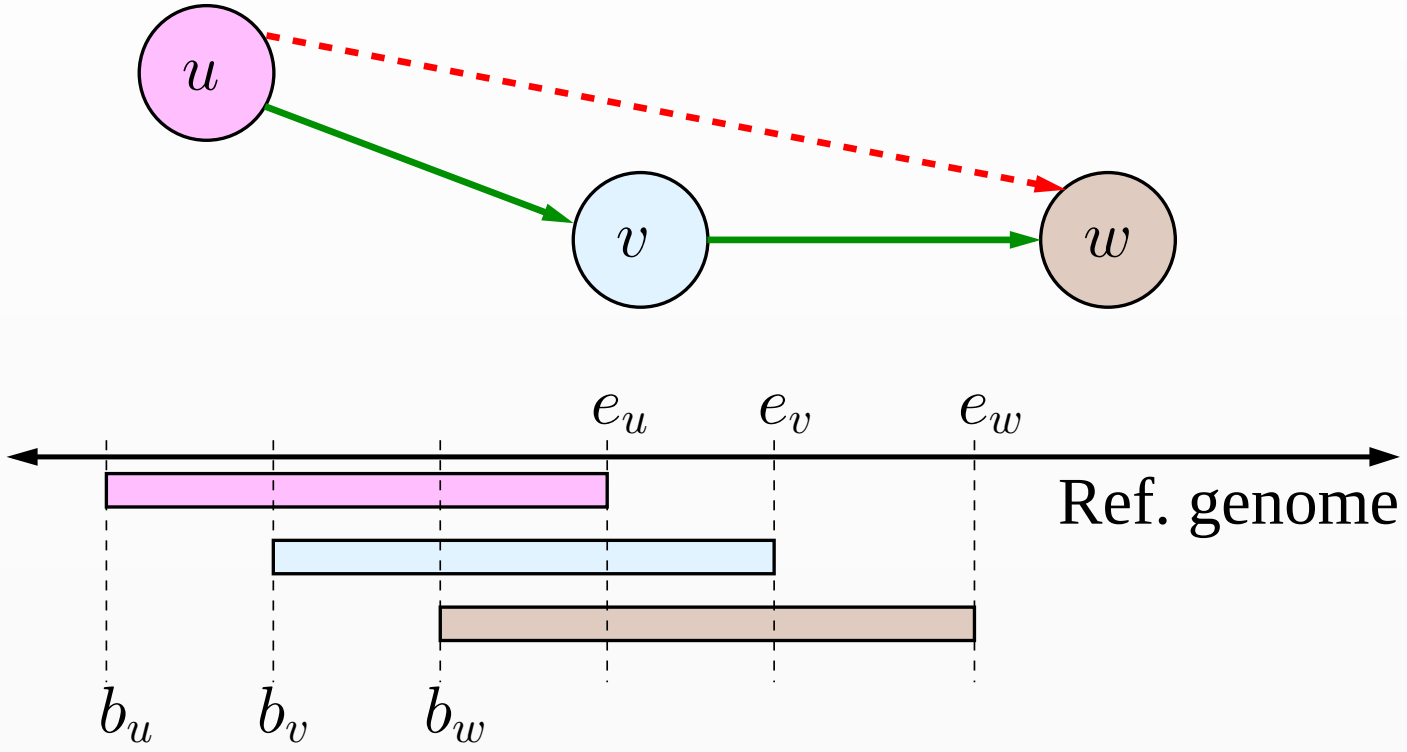
Transitive Reduction of the Read Graph

- Edge $u \to w$ **logically follows** from edges $u \to v$ and $v \to w$.
- Drop $u \to w$ from consideration – **no information**, any HT containing u and w will also have v.

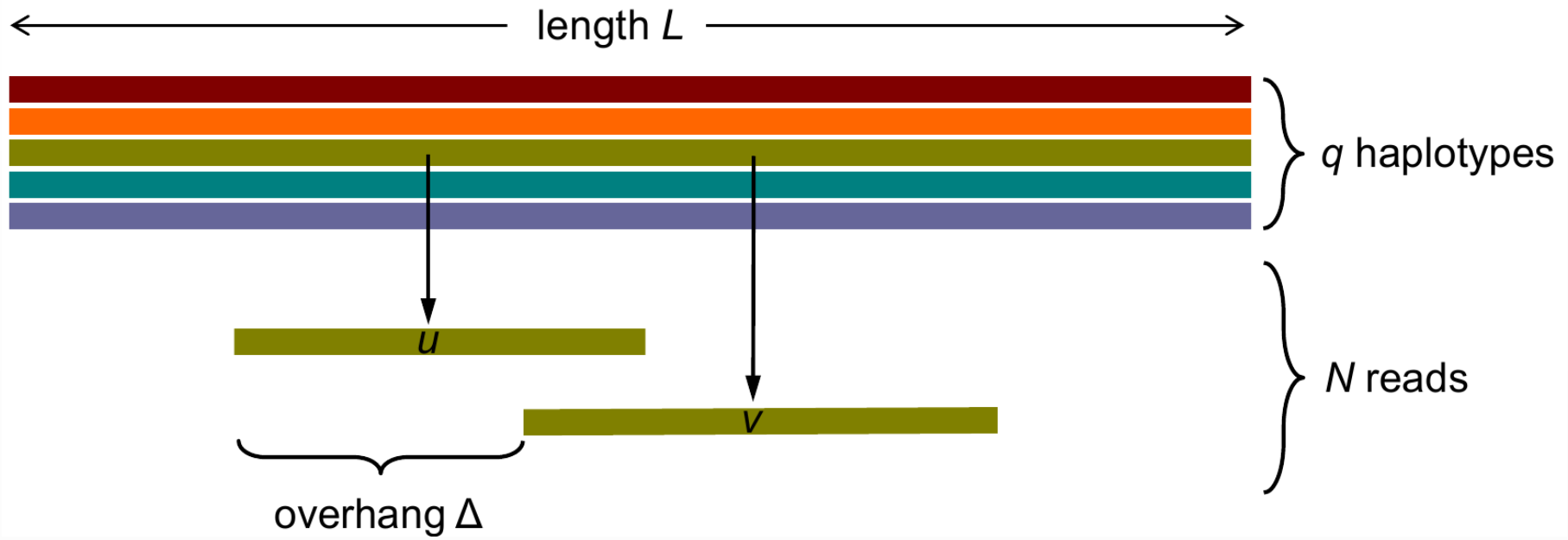


Transitive Reduction (2)

- In transitively reduced read graph:
  reads u, v from HT h are **connected by an edge** $u \longrightarrow v$ implies that there is **no other read** w **from the same HT** h **with** $b_u < b_w < b_v$.



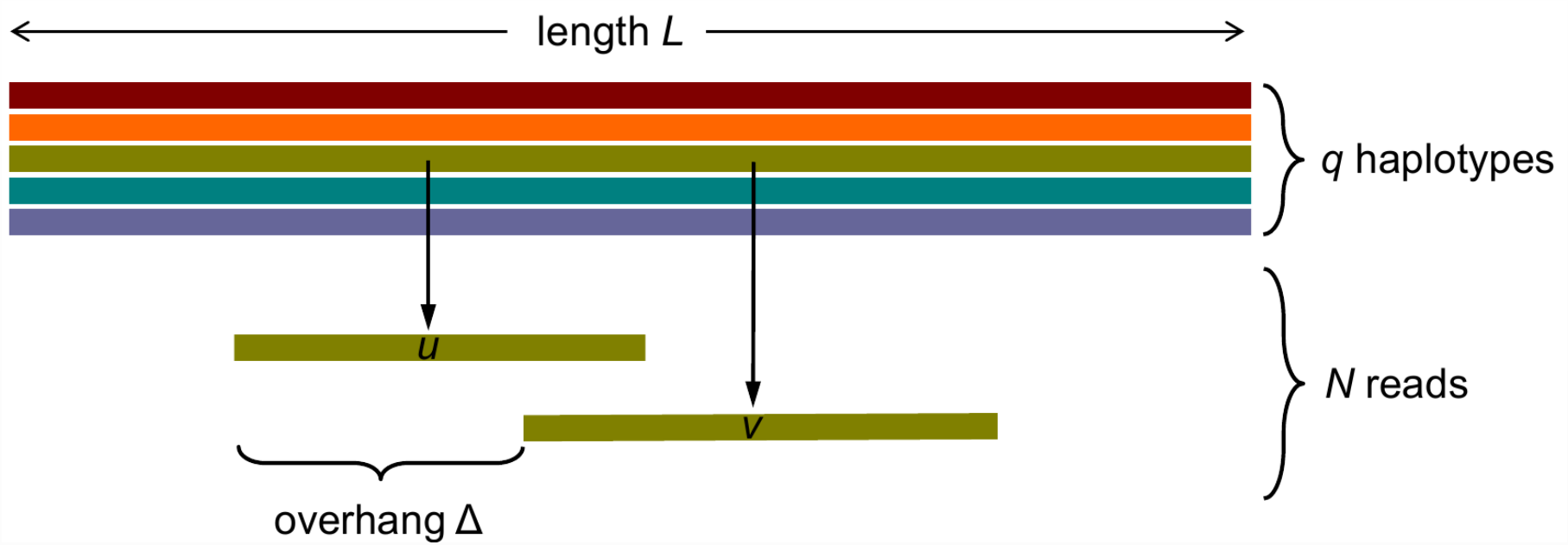
Probability of an edge in the read graph



- **Prob**(read u from HT h starts at b_u) $= N/(Lq)$.
- Event $(\Delta := b_v - b_u > k)$ = event that $b_u + 1, \ldots, b_u + k$ are **not beginnings of reads from HT** h.
- **Random starting positions:**
  Prob$(\Delta > k) =: p_k = (1 - N/(LQ))^k \approx \exp\left(-(kn)/(Lq)\right)$.

Westbrooks et al (ISBRA 2008)

Probability of an edge in the read graph (2)



- **Intuition:** if $u \longrightarrow v$ is a "true" overlap in HT h, Δ should be small.
- For **any two reads** with overhang Δ,

$$cost(u \longrightarrow v) := 1/p_{\Delta} \approx \exp\left((\Delta n)/(Lq)\right)$$

measures the **implausibility** that $u \longrightarrow v$ is a **true edge**.

Westbrooks et al (ISBRA 2008)

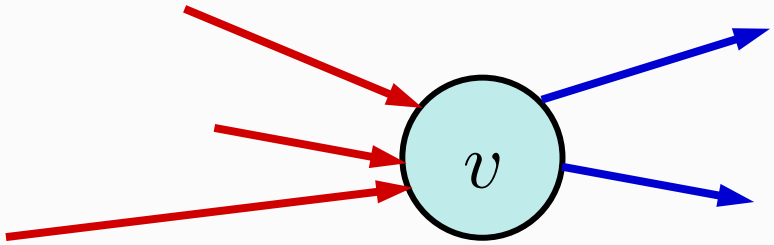
Quasispecies assembly via network flows (ViSpA)

LP for Minimum Cost Quasispecies Assembly:

Objective: Minimize the total cost $= \sum_e \text{cost}(e) f(e)$ over all edges e in the read graph.

Subject to:

- **Flow conservation:**

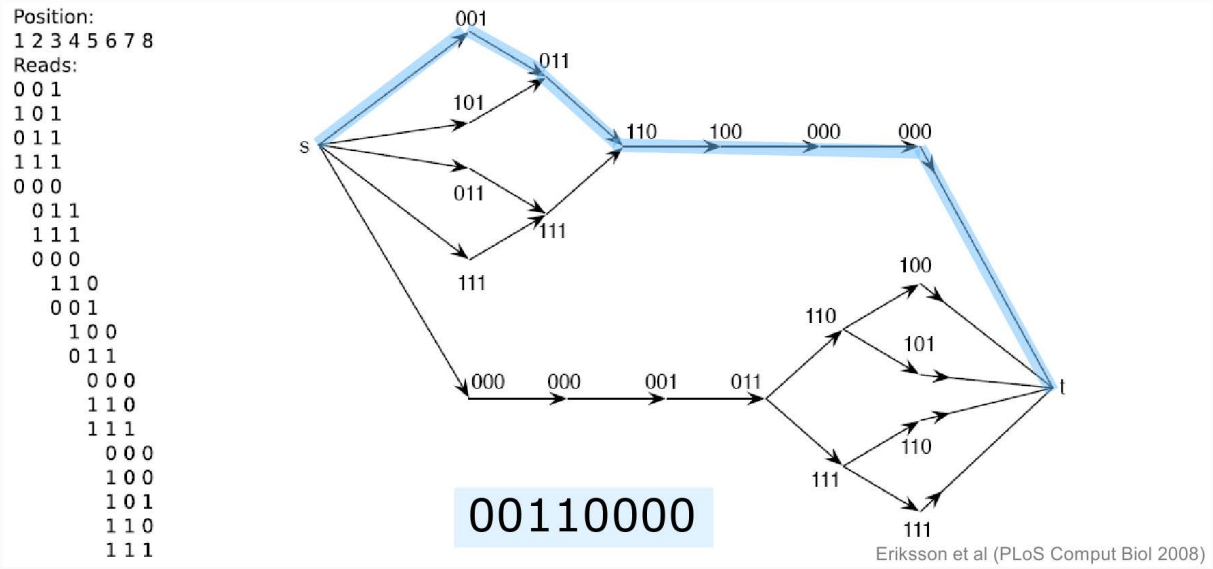


$$\sum_{e_{\text{in}}} f(e) = \sum_{e_{\text{out}}} f(e)$$

- **Each read covered by at least one haplotype**

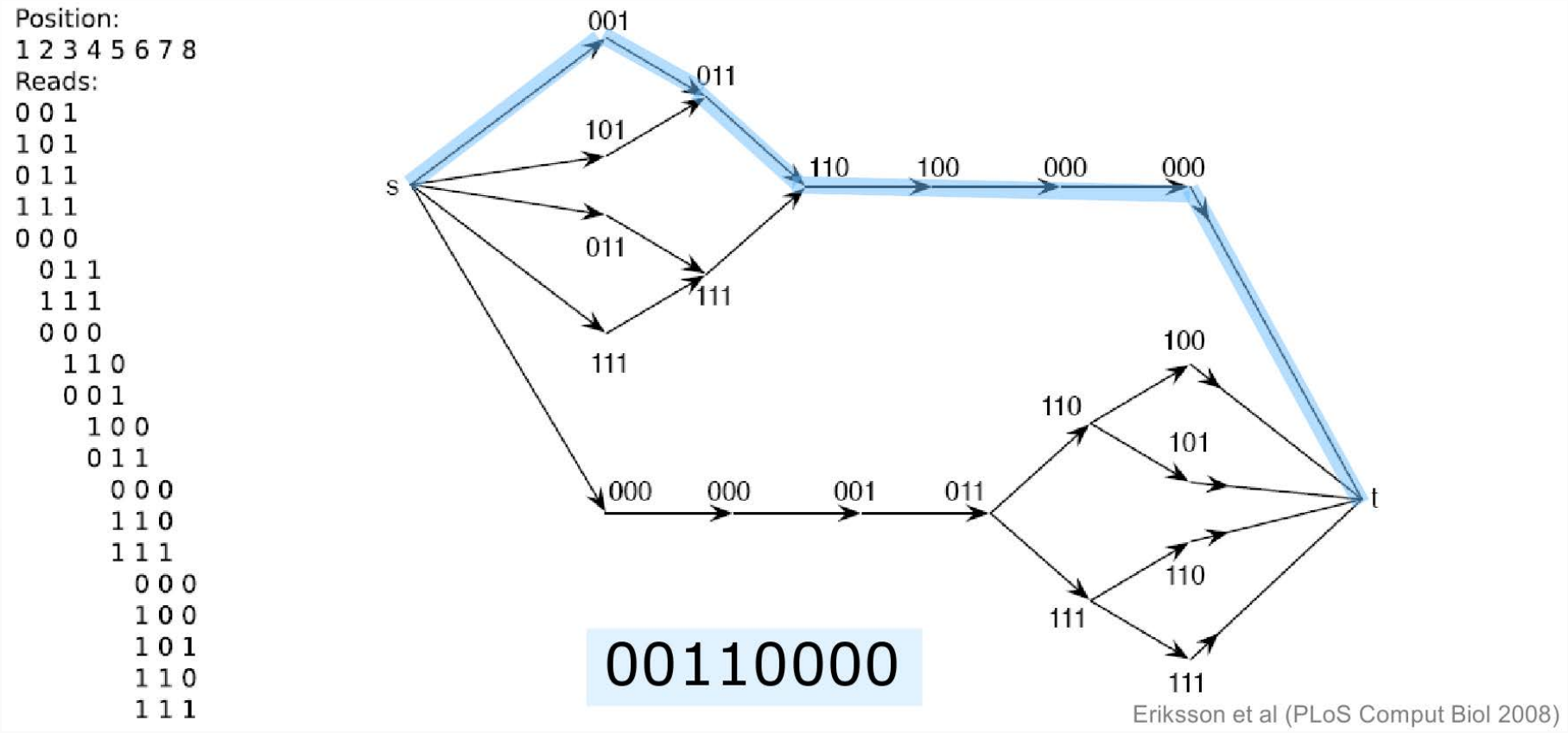
Westbrooks et al (ISBRA 2008), Astrovskaya et al (BMC Bioinf 2011)

Combinatorial Assembly: Path Cover



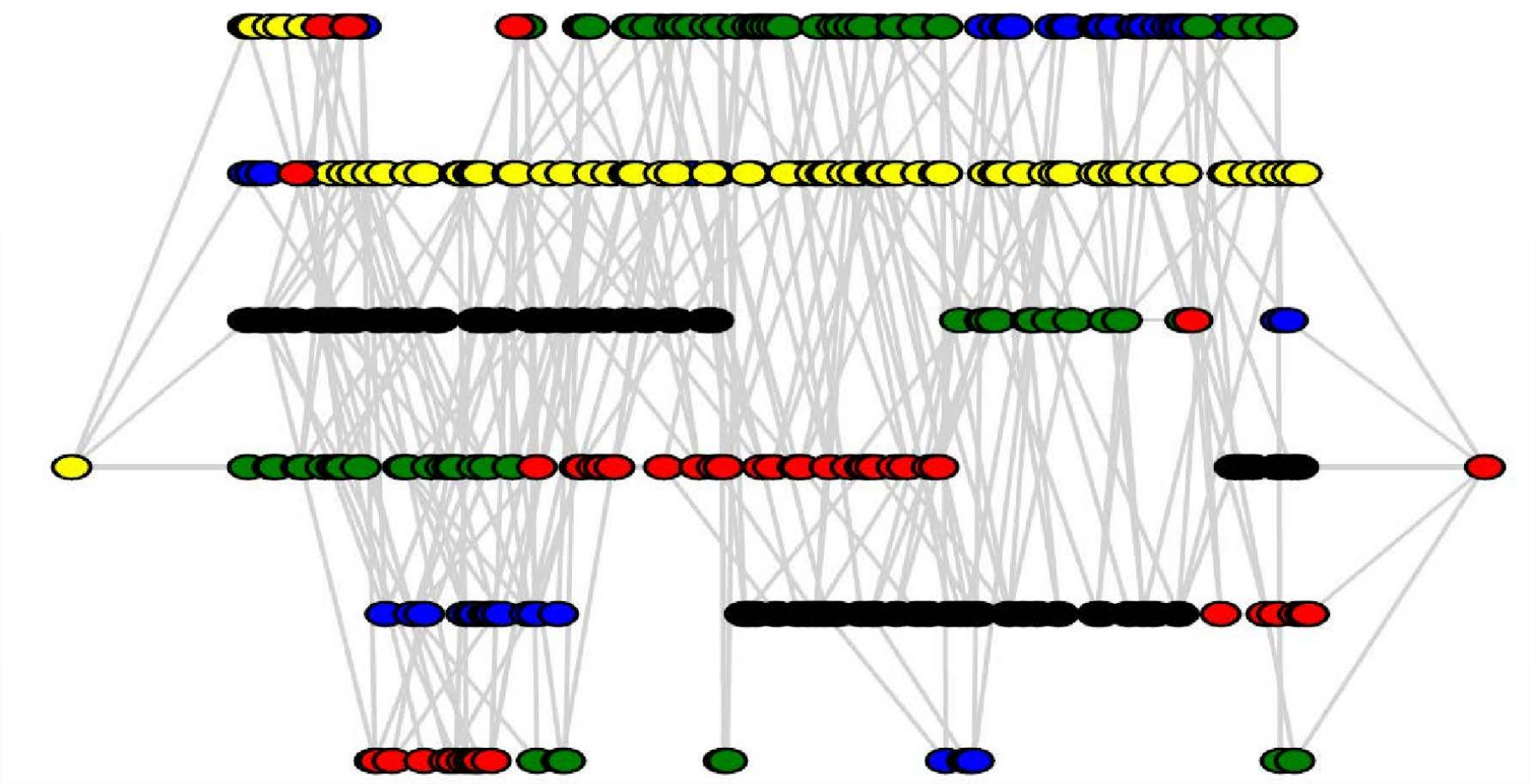
Quasispecies assembly as a minimal path cover

Theorem (Dilworth, 1950; Hopcroft and Karp, 1973)
(1) Every minimal cover of the read graph has the same cardinality
(2) A minimal path cover can be computed in time $O(N^3)$.



Chain decomposition

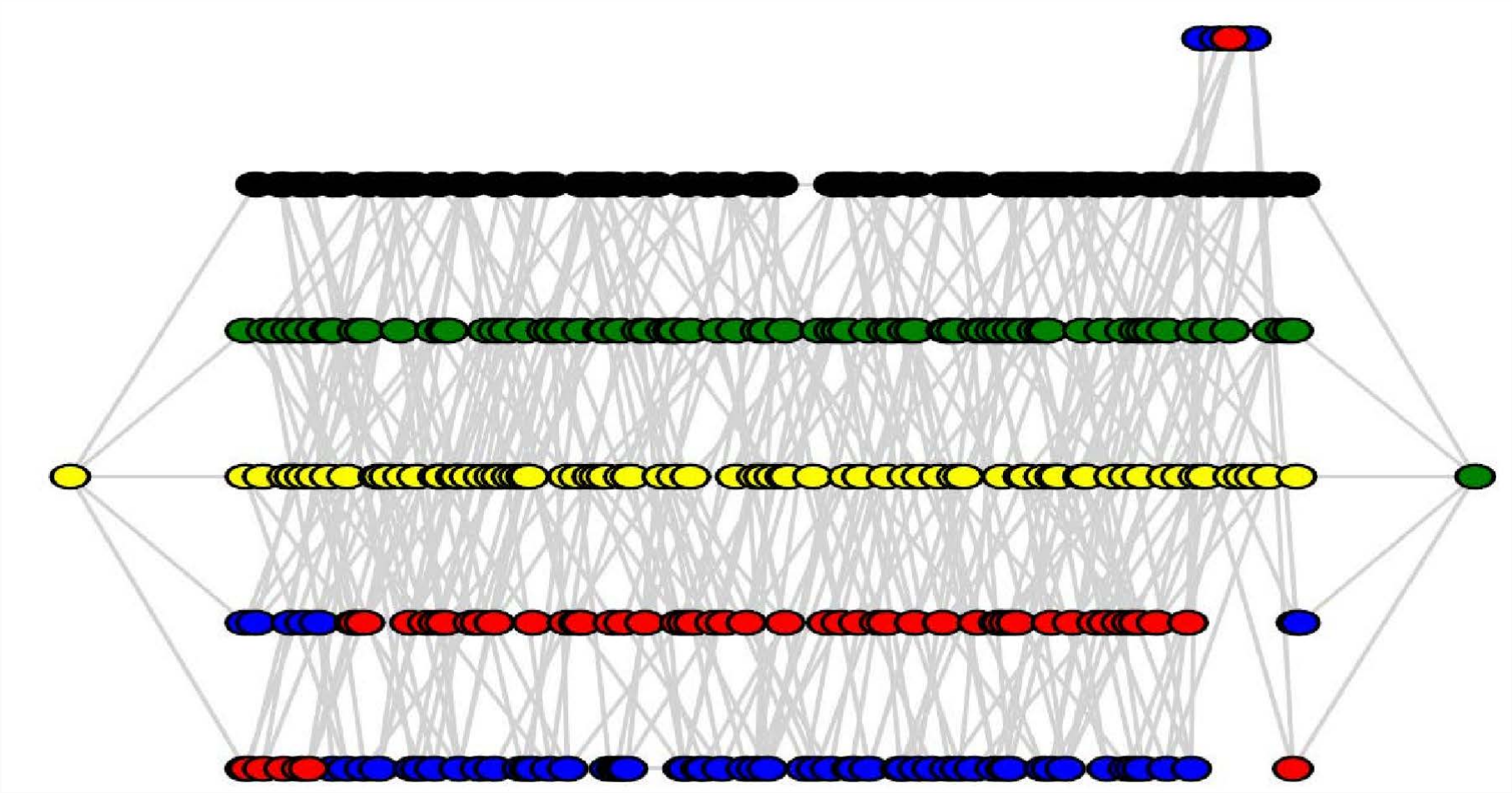
1000 reads from 5 haplotypes at 3% diversity



Eriksson et al (PLoS Comput Biol 2008)

Chain decomposition

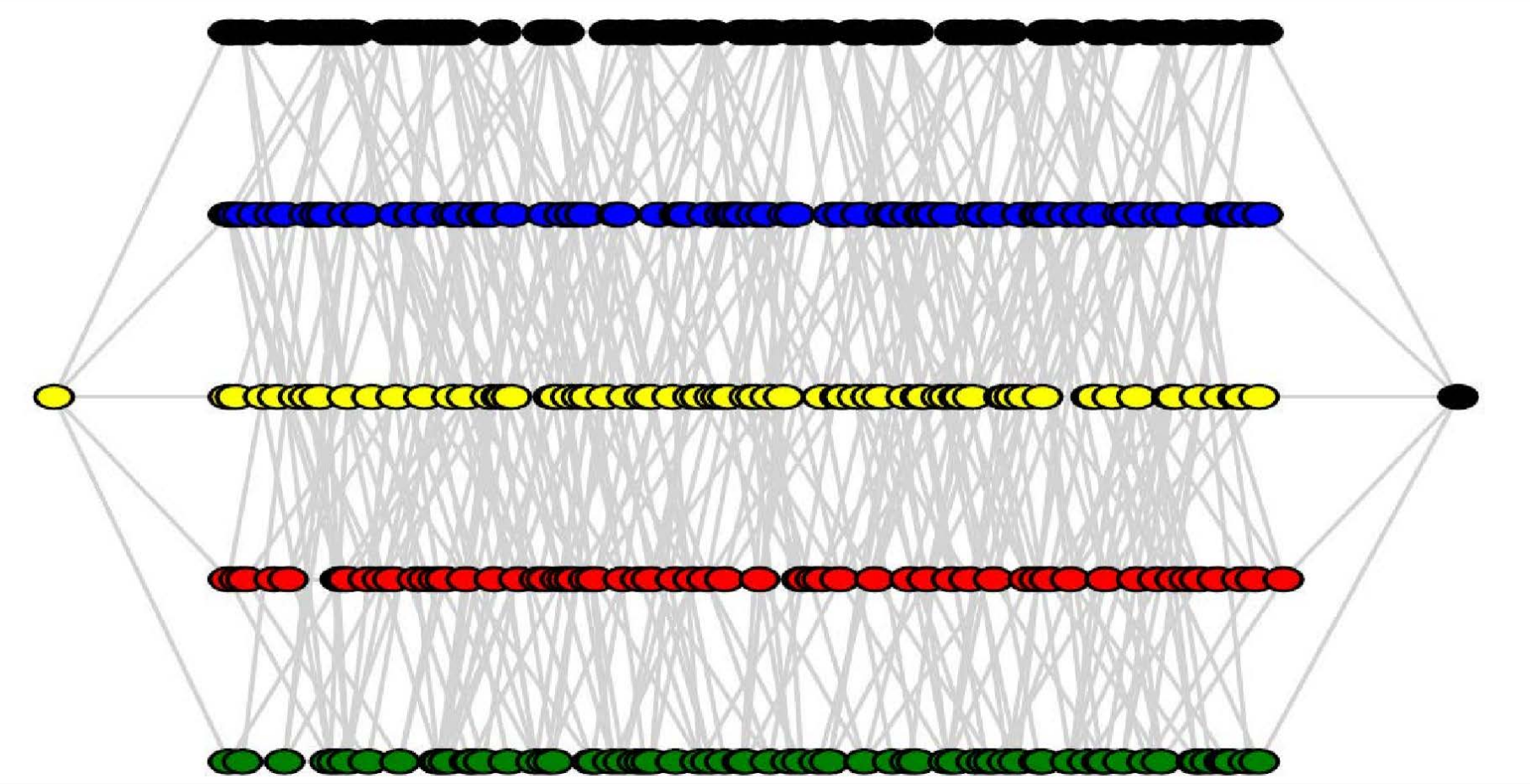
1000 reads from 5 haplotypes at 5% diversity



Eriksson et al (PLoS Comput Biol 2008)

Chain decomposition

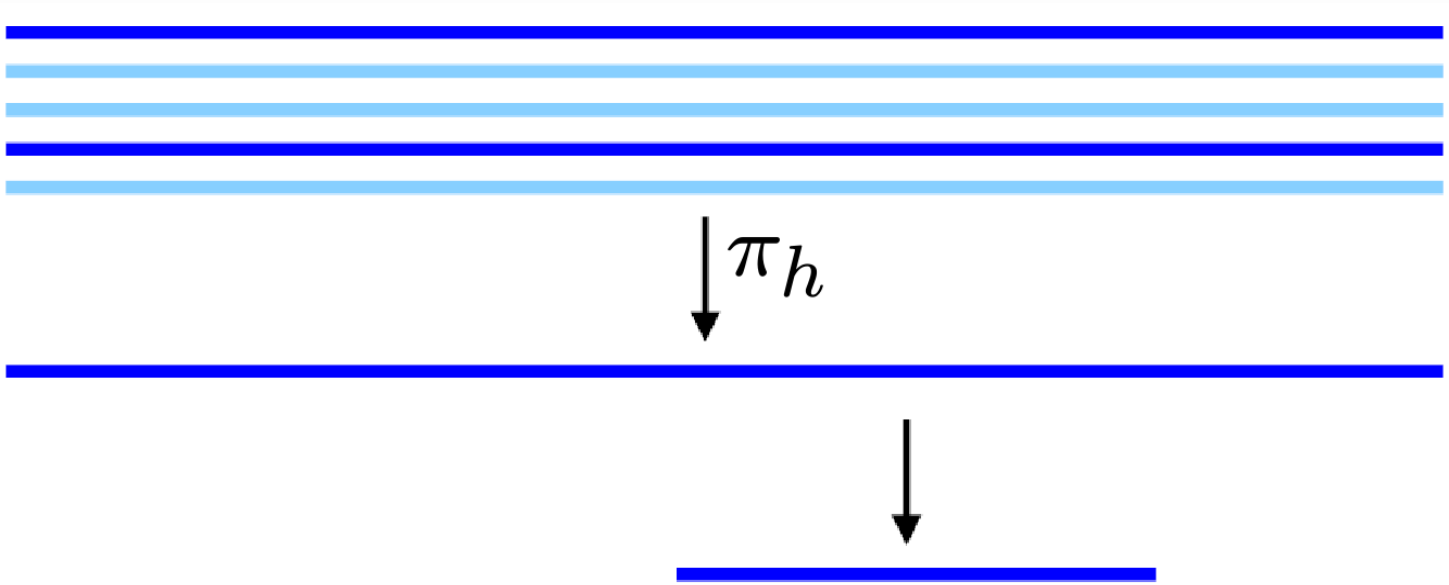
1000 reads from 5 haplotypes at 7% diversity



Eriksson et al (PLoS Comput Biol 2008)

Haplotype frequencies: Generating reads from a (small) set of candidate haplotypes

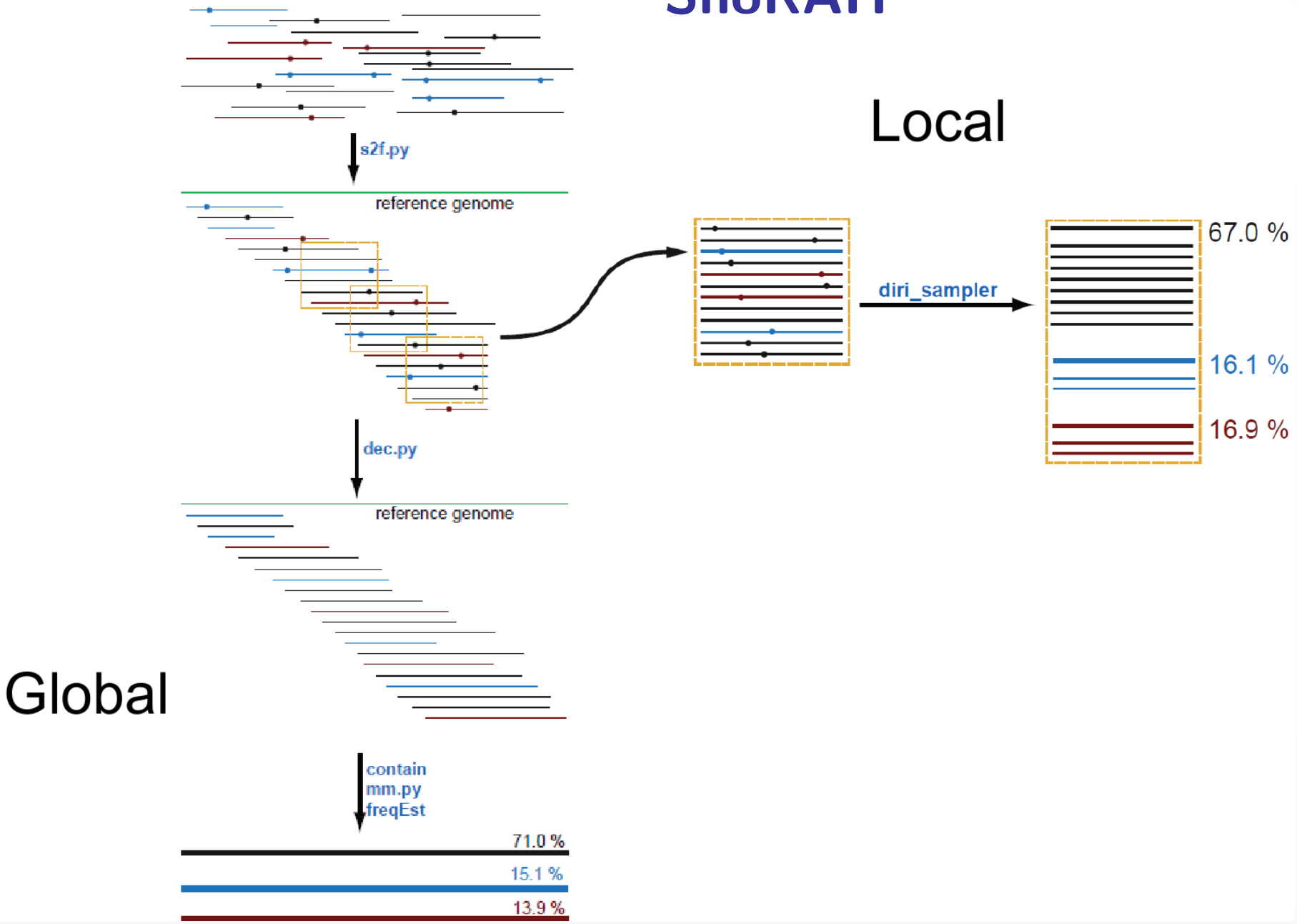
Estimate haplotype frequencies π_h using the EM algorithm



$$p(r) = \sum_{h \in \mathcal{H}} \pi_h p(r \mid h)$$

Eriksson et al (PLoS Comput Biol 2008)

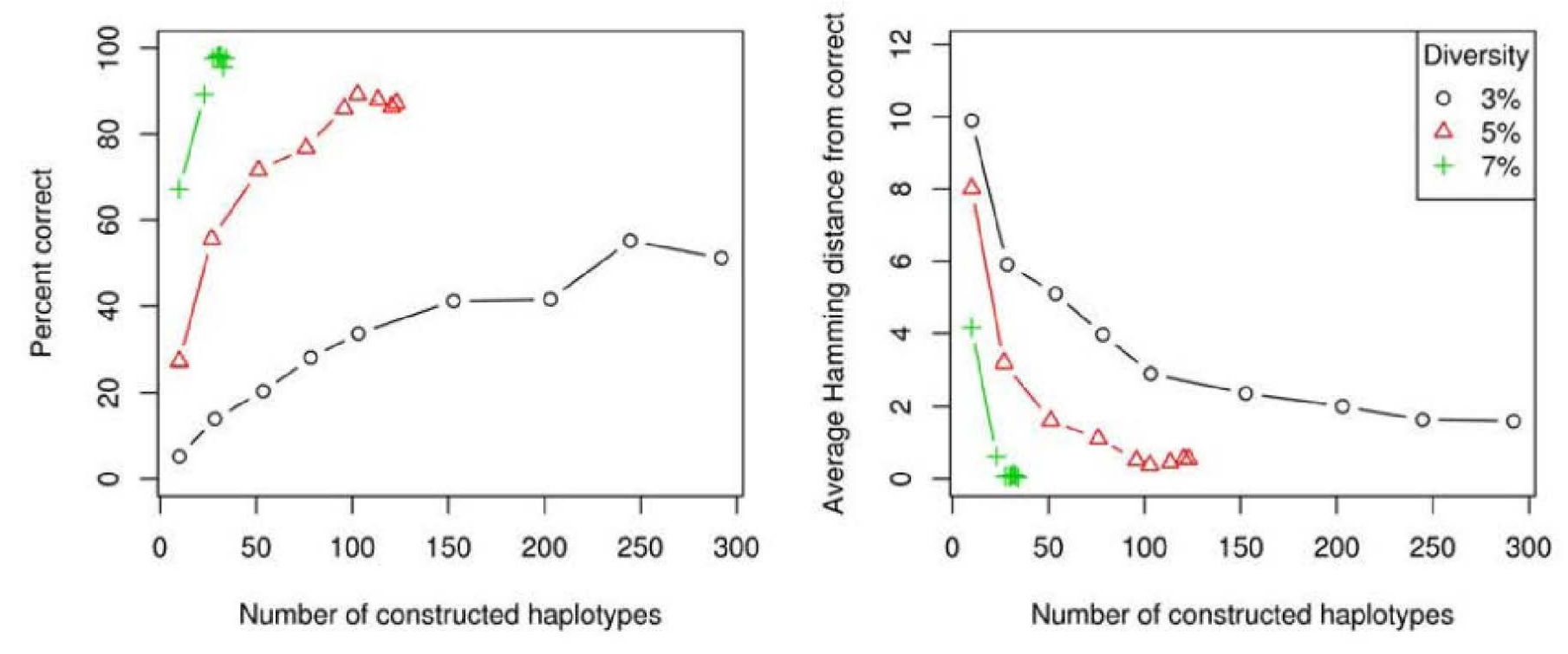
ShoRAH



Zagordi et al (BMC Bioinf 2010)

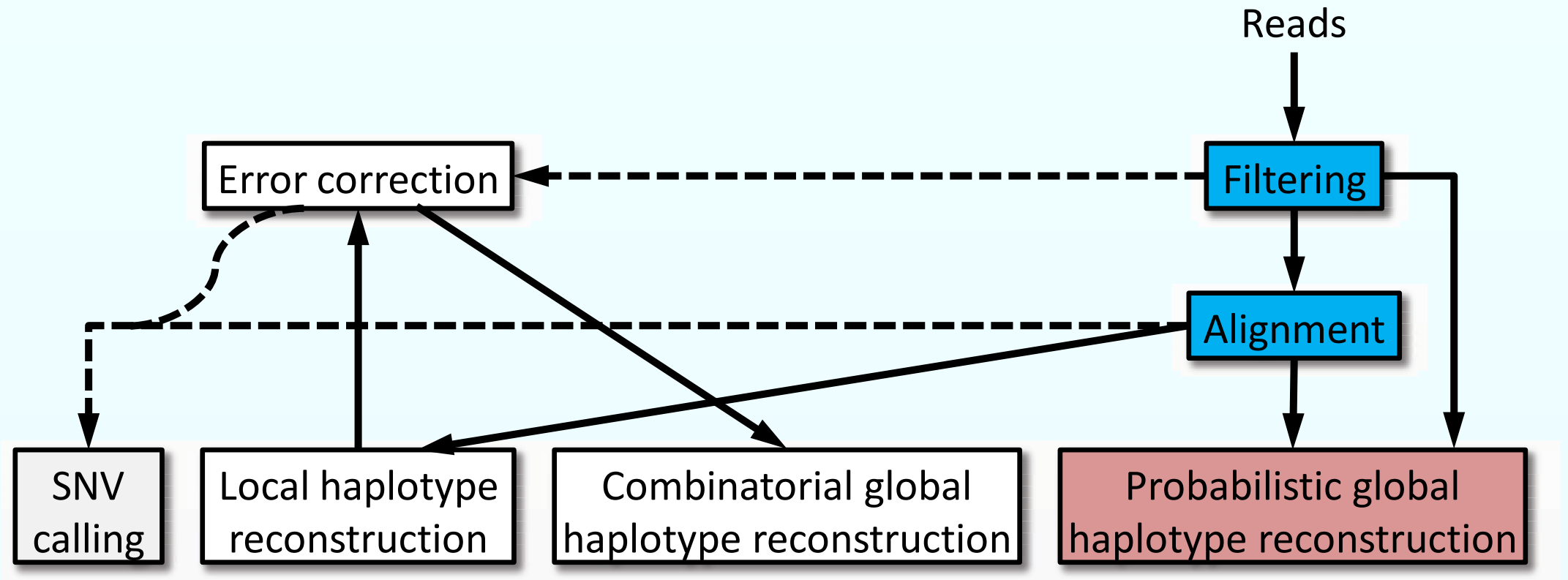
ShoRAH: Performance of haplotype reconstruction

Ten haplotypes at equal frequencies, varying distances

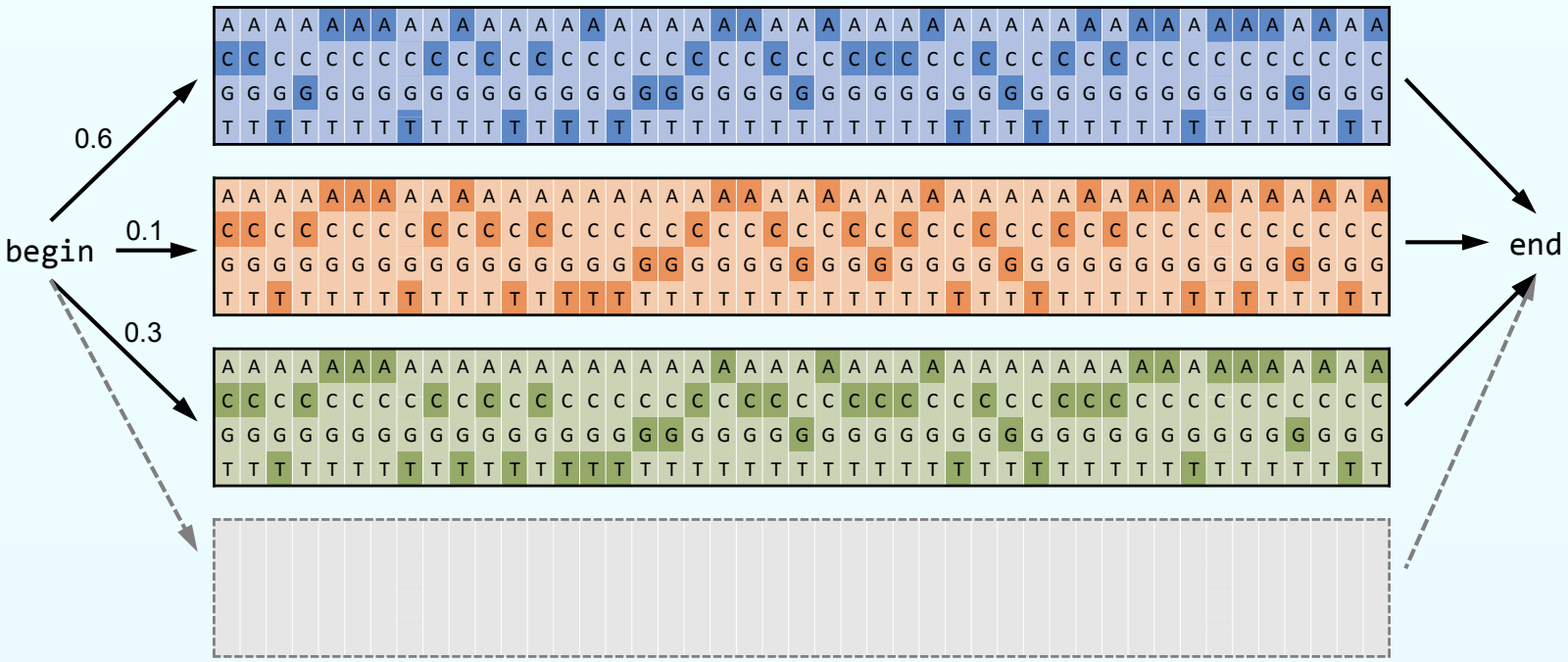


Eriksson et al (PLoS Comput Biol 2008)

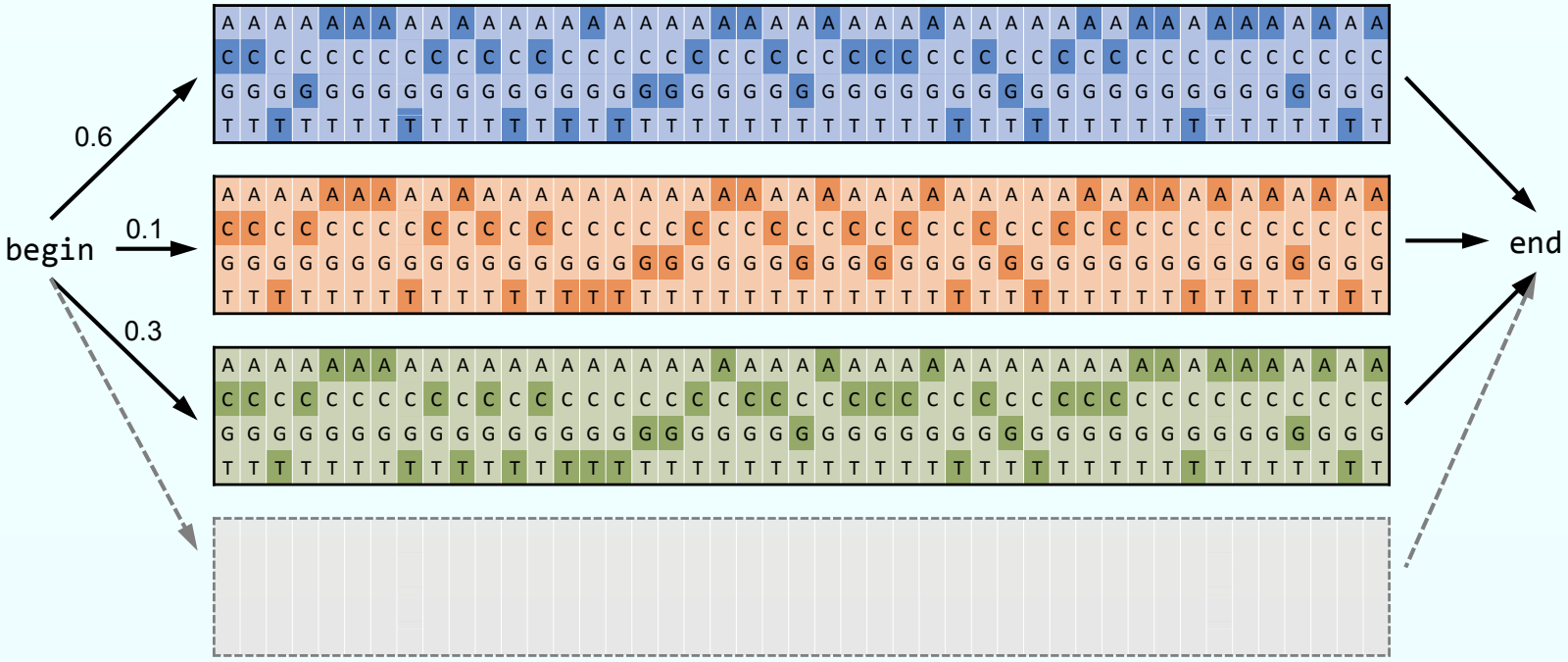
Overview


Reads
Error correction
Filtering
Alignment
SNV calling
Local haplotype reconstruction
Combinatorial global haplotype reconstruction
Probabilistic global haplotype reconstruction

Probabilistic Assembly: Mixture Model



PredictHaplo: Generative Model for Reads

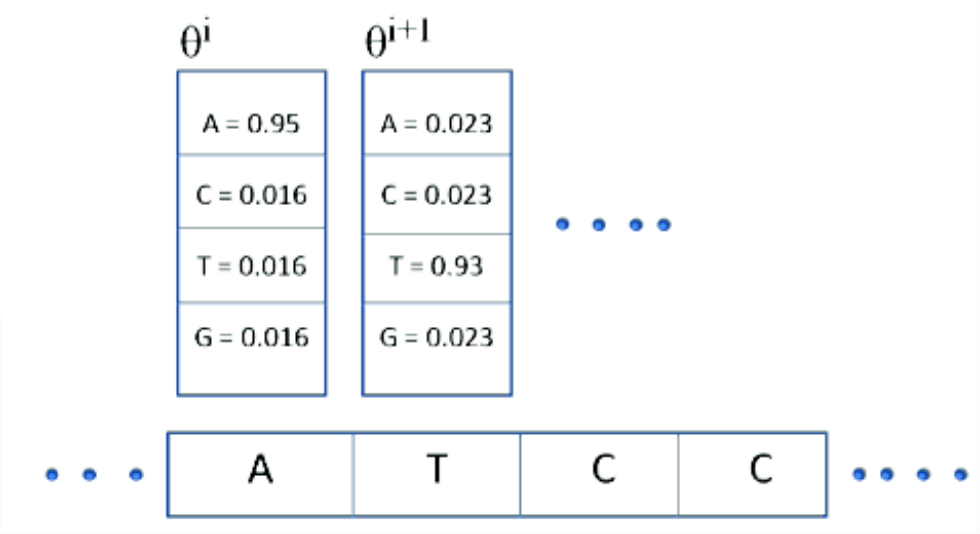


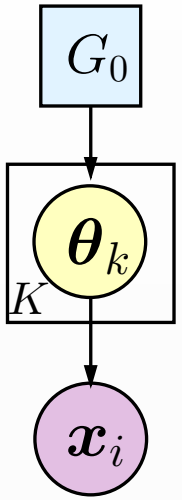
Bayesian mixture model: assign **priors** on class proportions and component distributions, integrate out latent variables.

Infinite mixture model: allow infinite number of classes...

Component Distributions

Haplotype: **position-wise multinomial probability tables θ**:





Prior parameter for $\boldsymbol{\theta}$

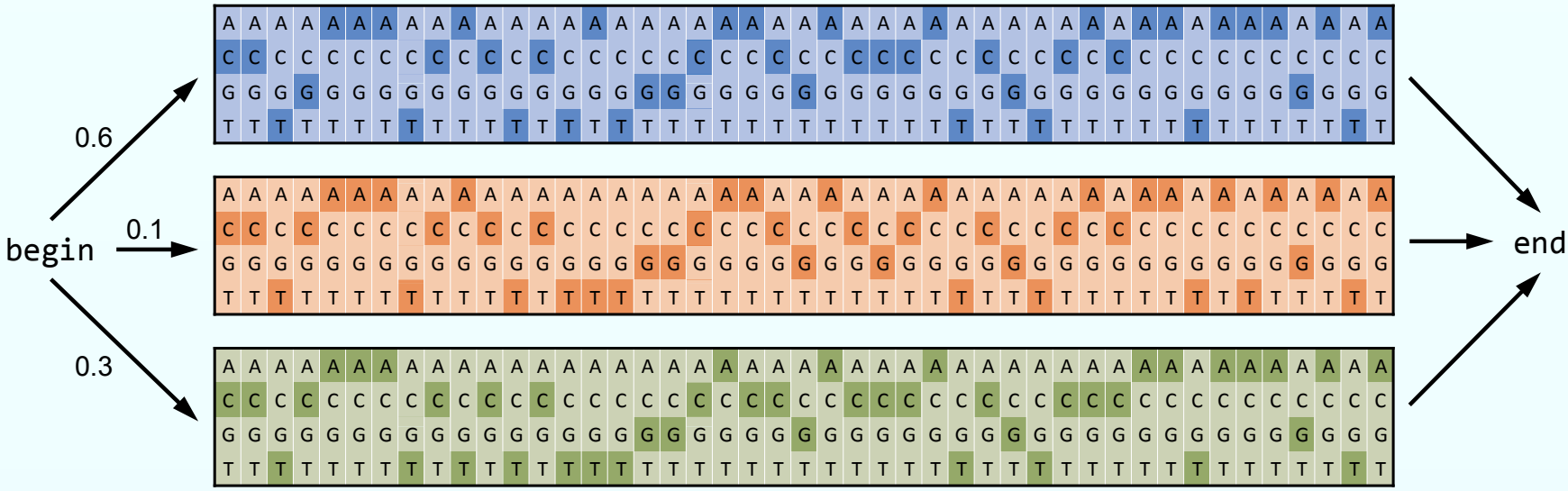
k-th haplotype $\boldsymbol{\theta}_k$

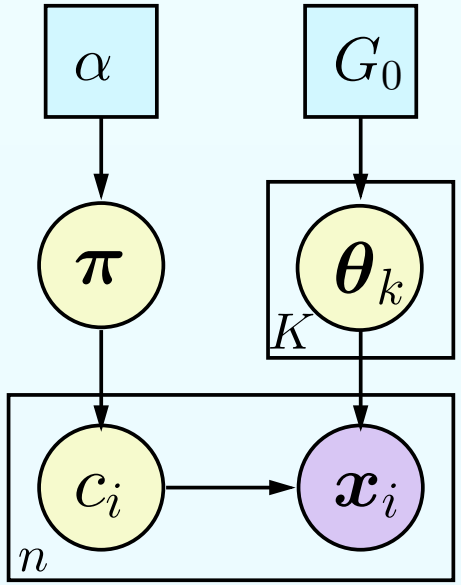
i-th read from k-th haplotype

- Parameters for k-th haplotype: $\boldsymbol{\theta}_k = (\theta_k^1, \ldots, \theta_k^L)$
- Position-wise independence assumption: i-th read $\boldsymbol{x}_i$ ranging from position a to b drawn from k-th haplotype:

$$\boldsymbol{x}_i \sim P(\boldsymbol{x}|\boldsymbol{\theta}_k) = \prod_{j=a}^{b} \text{Mult}(\theta_k^j)$$

Finite Mixture of Haplotypes





$$\boldsymbol{\pi} \sim \text{Dir}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$$

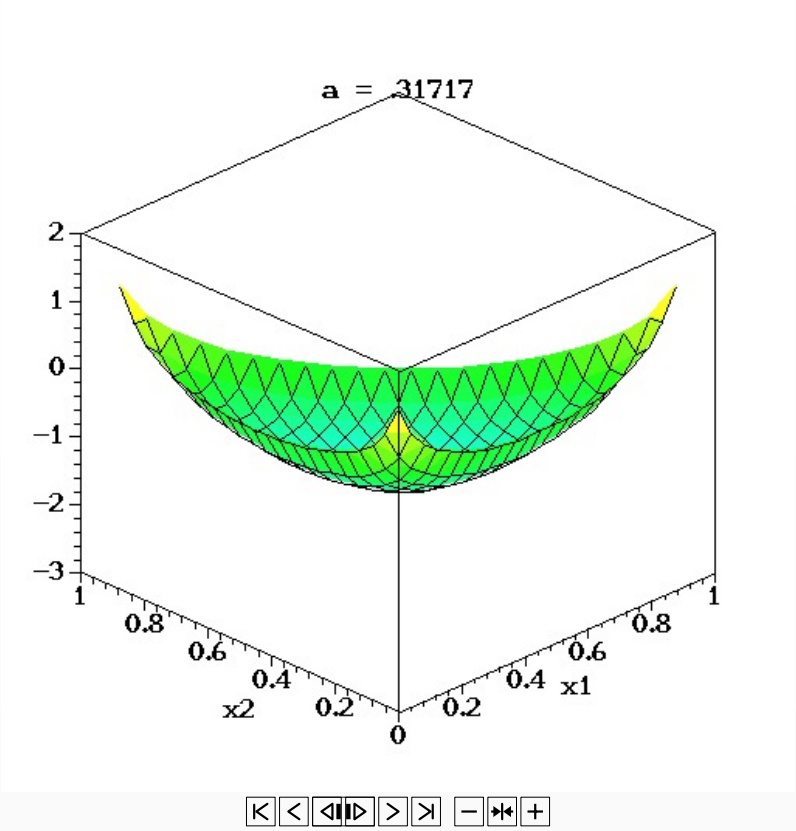
$$\boldsymbol{\theta}_k \sim G_0$$

$$c_i \sim \text{Mult}(\boldsymbol{\pi})$$

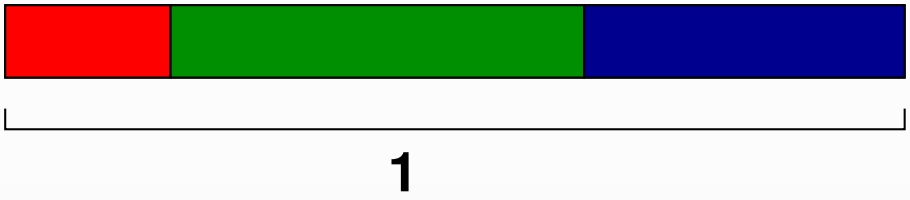
$$\boldsymbol{x}_i \sim P(\boldsymbol{x}_i | \boldsymbol{\theta}_{c_j})$$

Dirichlet Priors for Mixture Proportions

- Assignment variables $c_i \sim \text{Mult}_k(\boldsymbol{\pi})$.
- **Dirichlet prior** $\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$



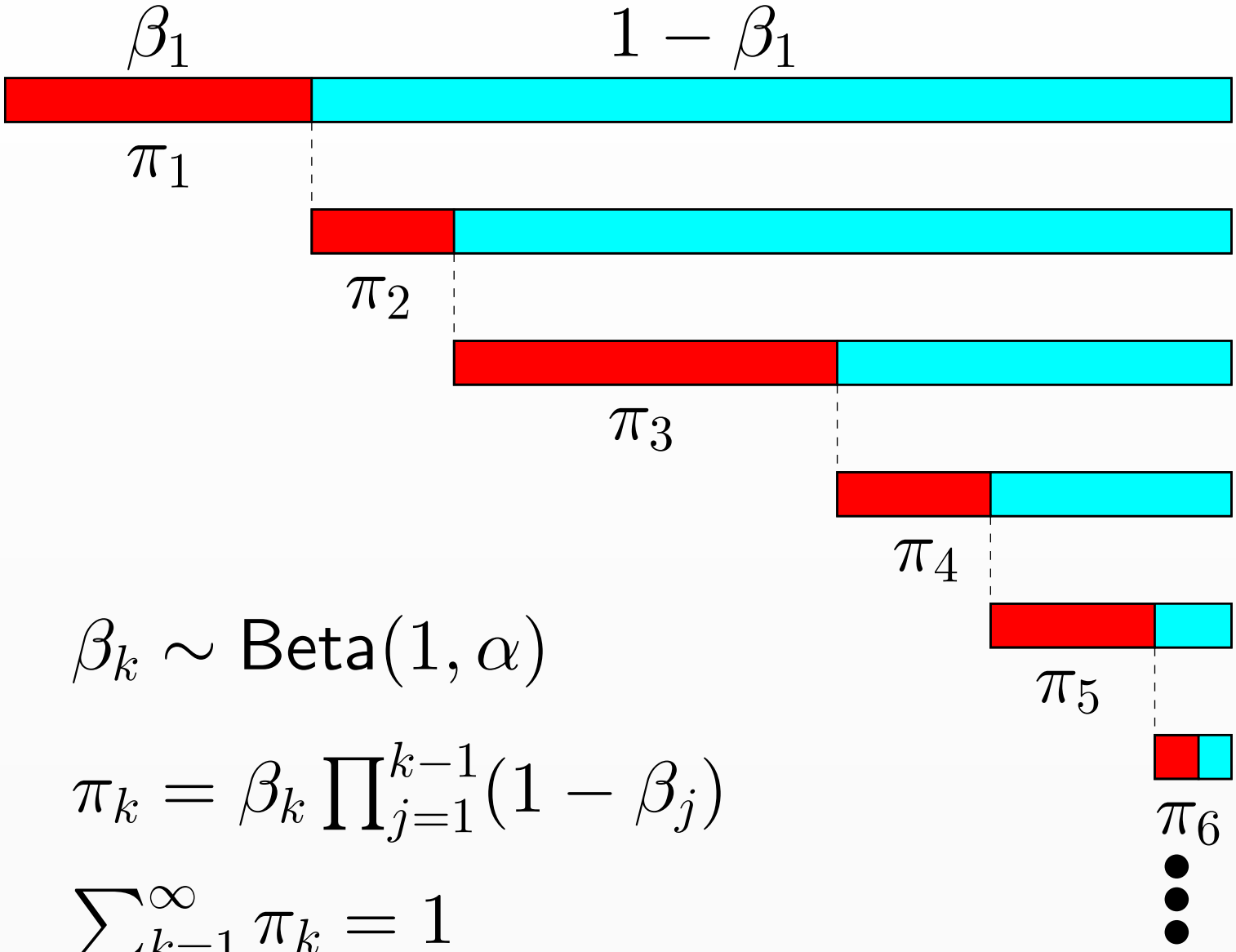
Interpretation: breaking a stick of length 1 into $K = 3$ parts



- **Problem:** don't want to specify fixed number of haplotypes... but what happens when $K \to \infty$?

http://commons.wikimedia.org/wiki/File:LogDirichletDensity-alpha_0.3_to_alpha_2.0.gif

Infinite Mixtures: Stick Breaking Construction

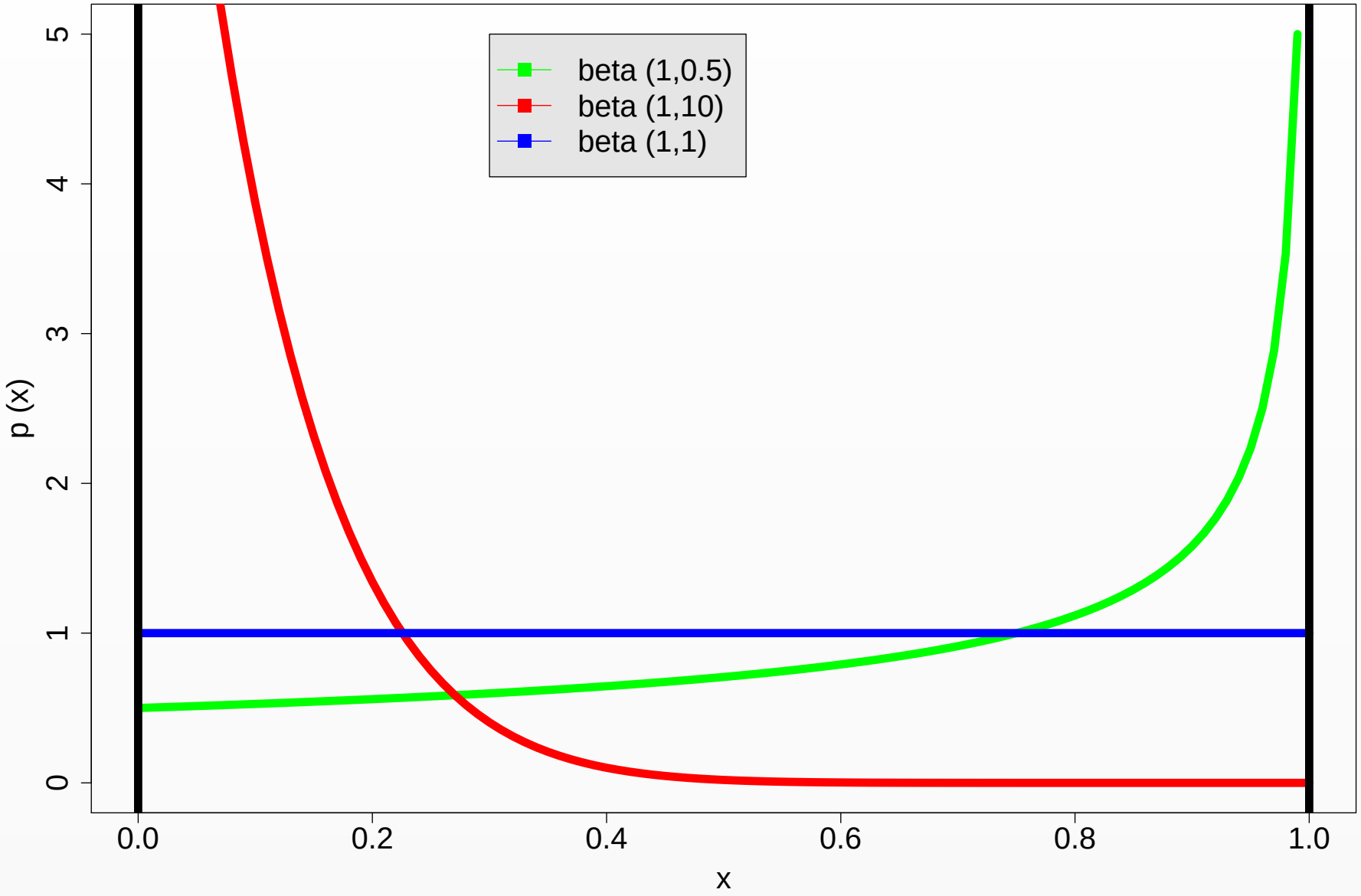


$$\beta_k \sim \mathsf{Beta}(1, \alpha)$$

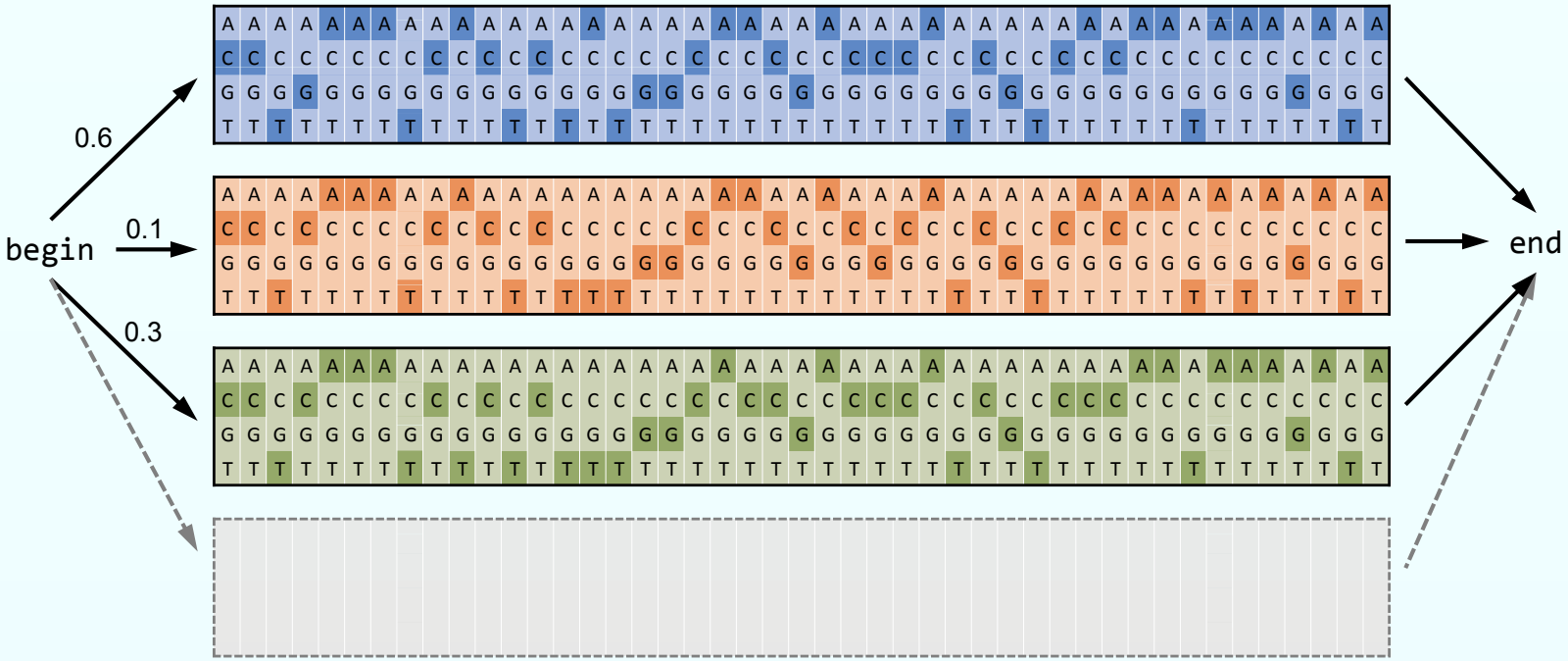
$$\pi_k = \beta_k \prod_{j=1}^{k-1}(1-\beta_j)$$

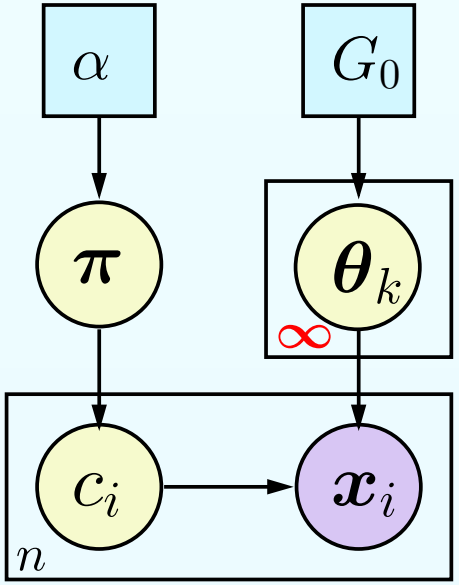
$$\sum_{k=1}^{\infty} \pi_k = 1$$

Beta distribution



Infinite Mixtures





$$\boldsymbol{\pi} \sim \text{Stick}(1, \alpha)$$

$$\boldsymbol{\theta}_k \sim G_0$$

$$c_i \sim \text{Mult}(\boldsymbol{\pi})$$

$$\boldsymbol{x}_i \sim P(\boldsymbol{x}_i | \boldsymbol{\theta}_{c_j})$$

Beerenwinkel et al (Front Microbiol 2012)

Infinite Mixtures: Inference

- **Making it fast: truncate the process.** Bound k from above by $K_{\text{max}} \gg$ "expected" K. Posterior estimates based on truncated process will be exponentially close to those based on the infinite process [Ishwaran & James, 2002].

- Use a **sampler**: Iterate

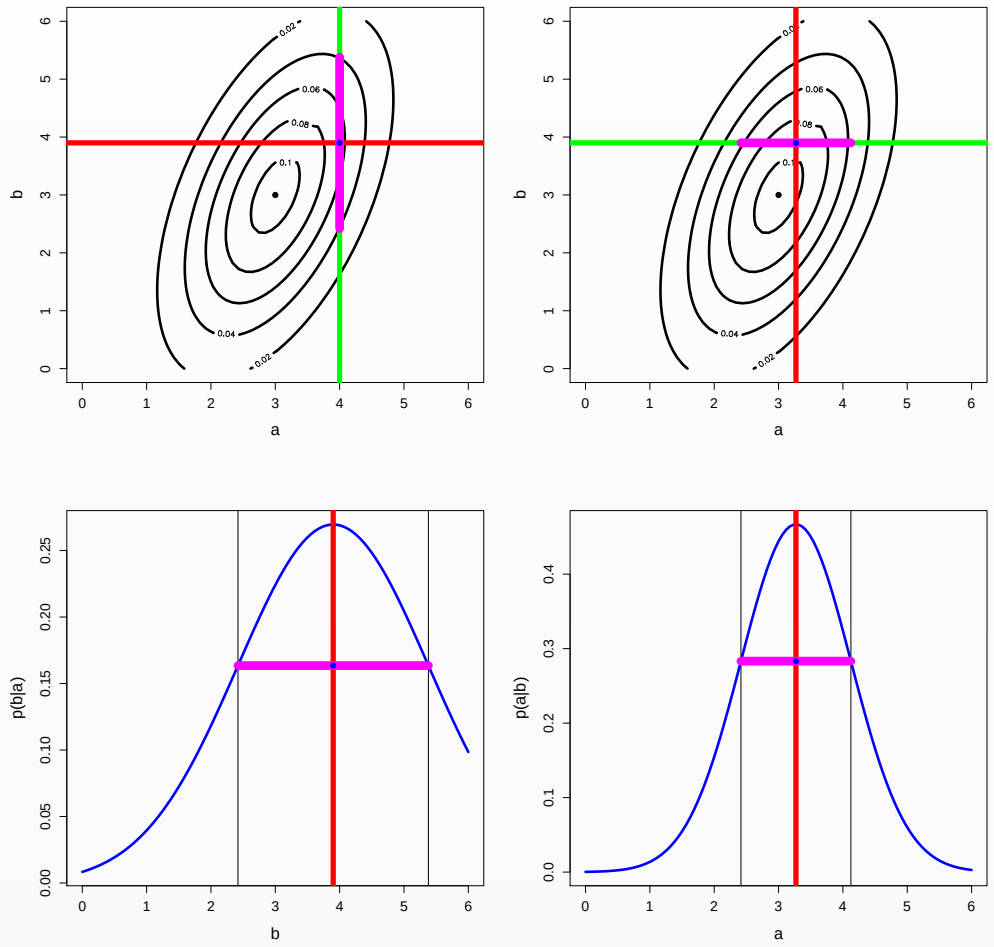
  1. draw $\boldsymbol{\theta}_k$ from $p(\boldsymbol{\theta}_k|\bullet)$ (all currently populated + 1 empty haplotype)
  2. draw c_i from $p(c_i|\bullet),\ i = 1, \ldots, n_{\text{reads}}$
  3. draw $\boldsymbol{\pi}$ from $p(\boldsymbol{\pi}|\bullet)$ (all currently populated + 1 empty haplotype)

- This is a **Gibbs sampler** (a MCMC method). The samples will converge to samples from the true posterior $p(\boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\pi}|\boldsymbol{x}, \bullet)$

- Here: all conditionals are in standard form $\rightsquigarrow$ sampling is easy.

Gibbs Sampling

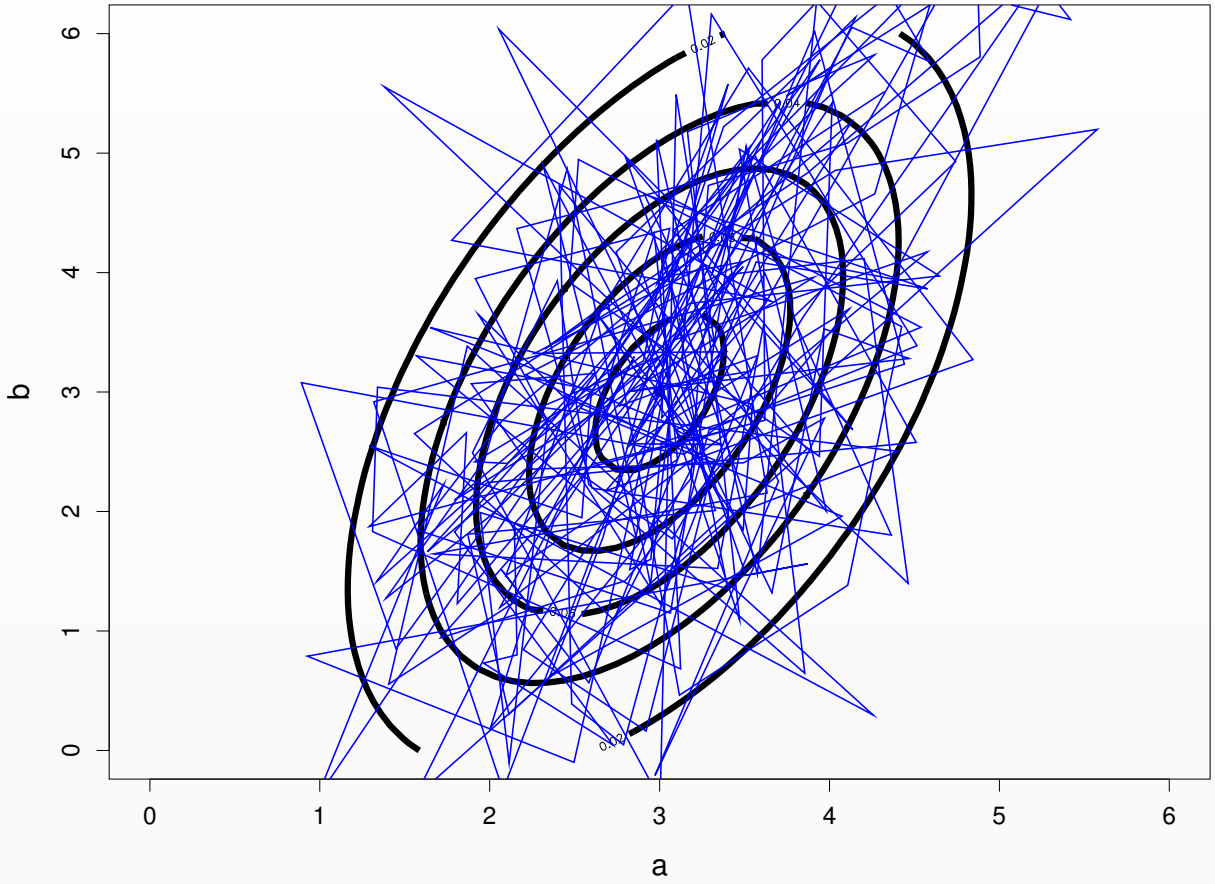
Assume you want to sample from a 2-dim Gaussian $p(a, b) \sim \mathcal{N}(a, b|\mu, \Sigma)$. You know that **conditionals of Gaussians are again Gaussians**, but you have forgotten how to sample from a 2-dim Gaussian.



Gibbs Sampling

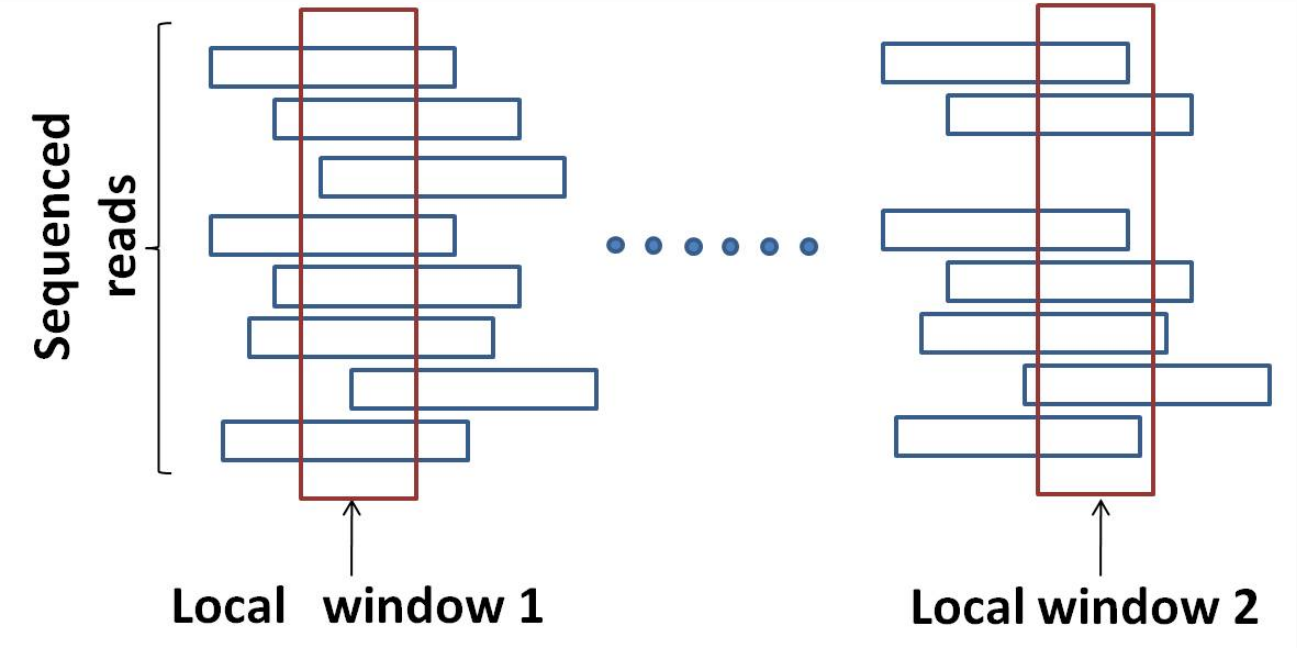
Solution: run a Gibbs sampler: Iterate:

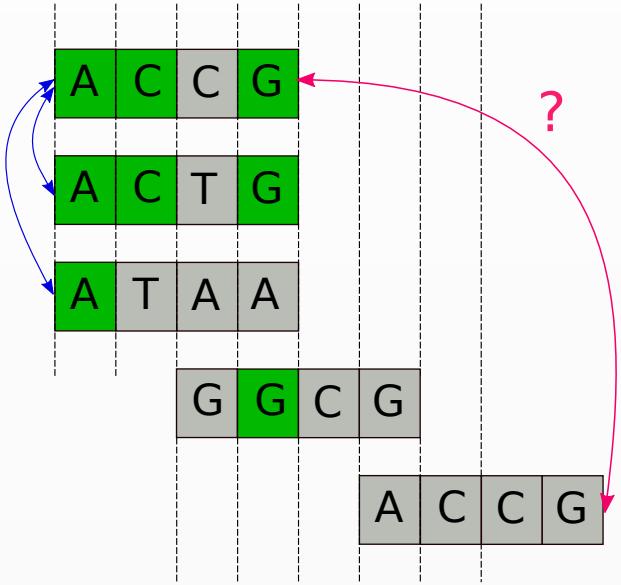
1. sample a from $p(a|b, \bullet) = \mathcal{N}(a|\mu', \Sigma')$
2. sample b from $p(b|a, \bullet) = \mathcal{N}(b|\mu'', \Sigma'')$



Local to Global

- Mixture model works for fully and partially overlapping reads...
- **...but not for global reconstruction!**



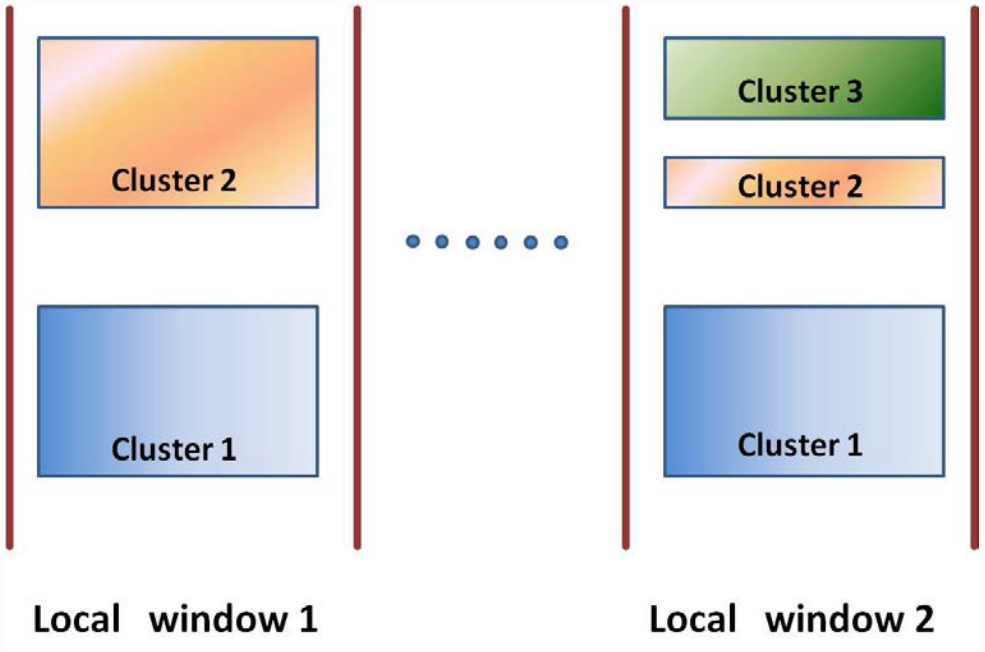


Global reconstruction

Prabhakaran et al., 2010

Extract do-not-link constraints

- **Idea:** Start with **local** inference
  $\rightsquigarrow$ extract local **do-not-link constraints** between reads:



- Local clusterings may be noisy $\rightsquigarrow$ **"soft" do-not-link constraints.**
- Include constraints in mixture model $\rightsquigarrow$ **global reconstruction.**

Prabhakaran et al., 2010

Constrained model

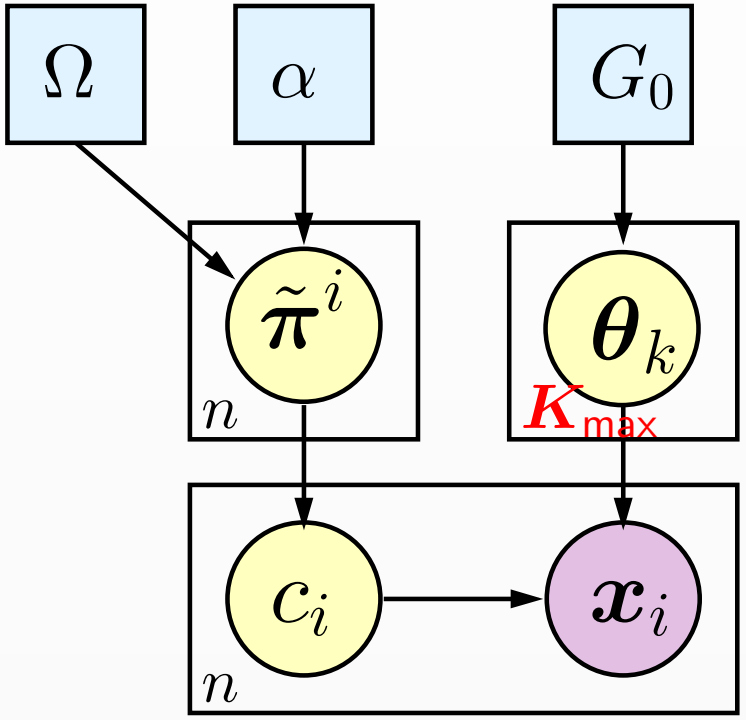
- **Distribution of the reads:**

$$p\left(\boldsymbol{x}_j|\boldsymbol{\pi},\boldsymbol{\theta}\right) = \sum_{k=1}^{K_{\text{max}}} \pi_k p\left(\boldsymbol{x}_j|\boldsymbol{\theta}_k\right).$$

- **Use constraints Ω to adjust parameters $\rightsquigarrow$ read-specific!**

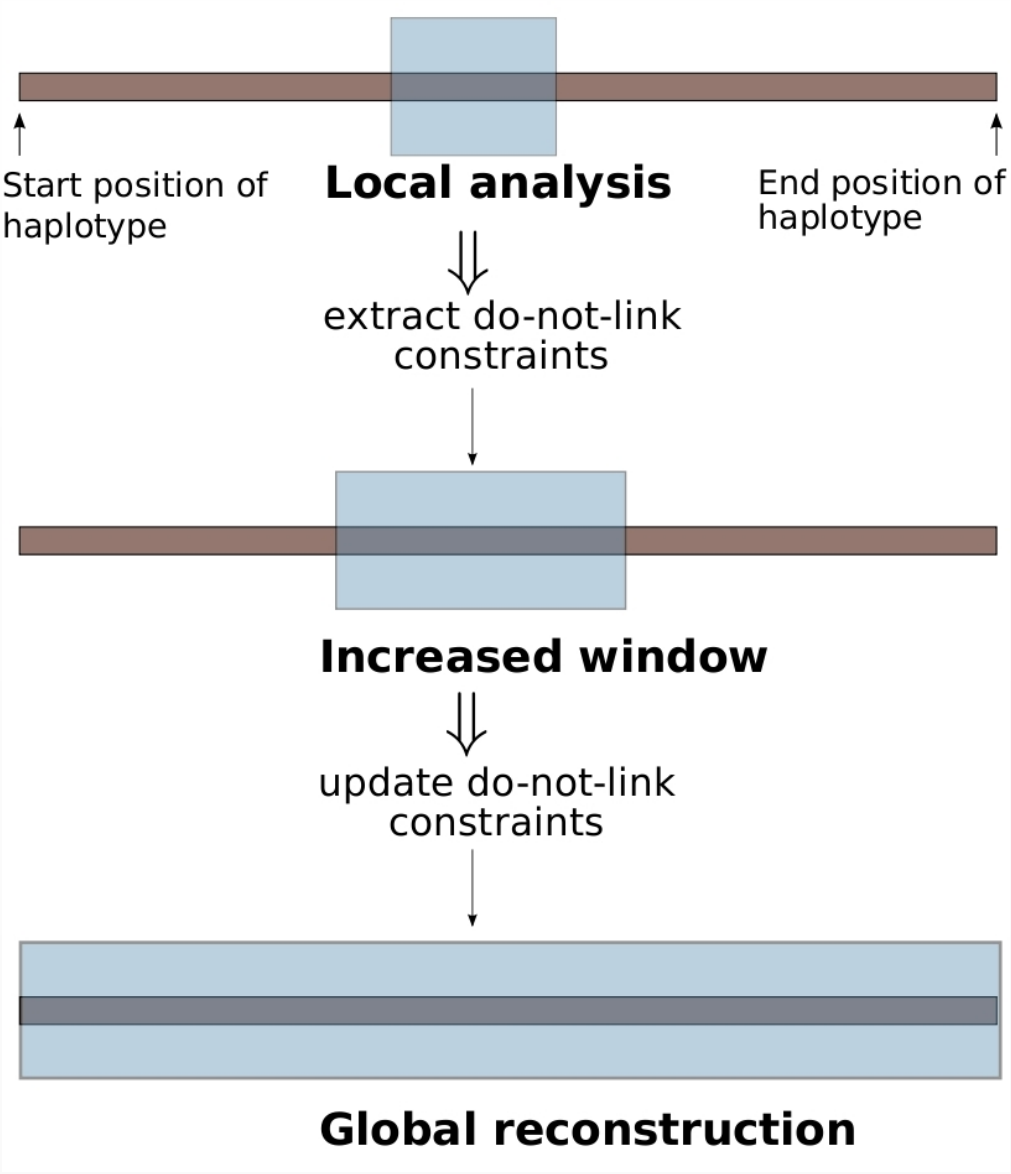
$$c_i|\tilde{\boldsymbol{\pi}}^i \sim \text{Mult}\left(c_i|\tilde{\boldsymbol{\pi}}^i\right)$$

where $\tilde{\boldsymbol{\pi}}^i$ are the **constraint-adjusted class probabilities** for the i-th read.



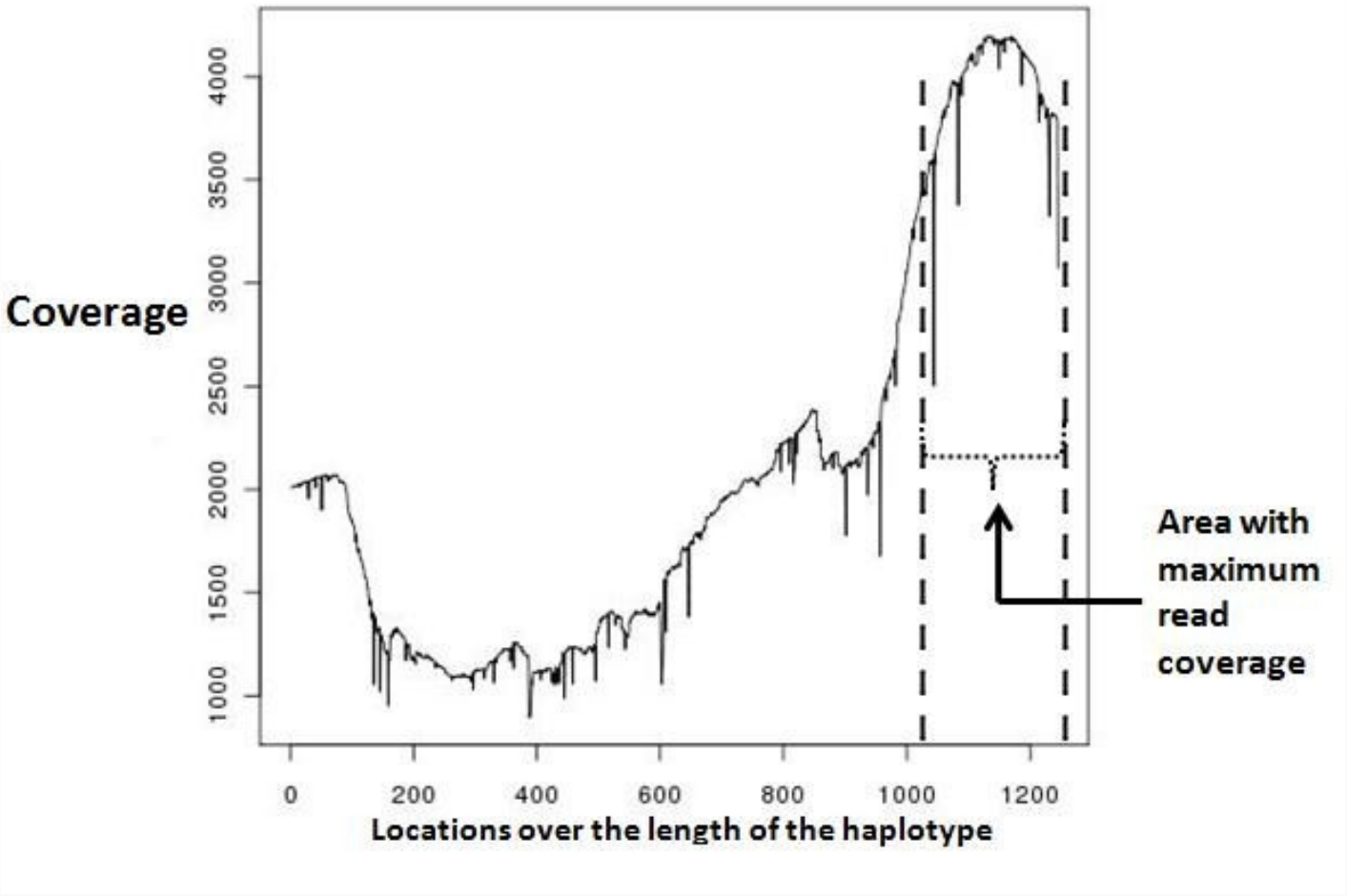
Prabhakaran et al., 2010

Constrained model: summary



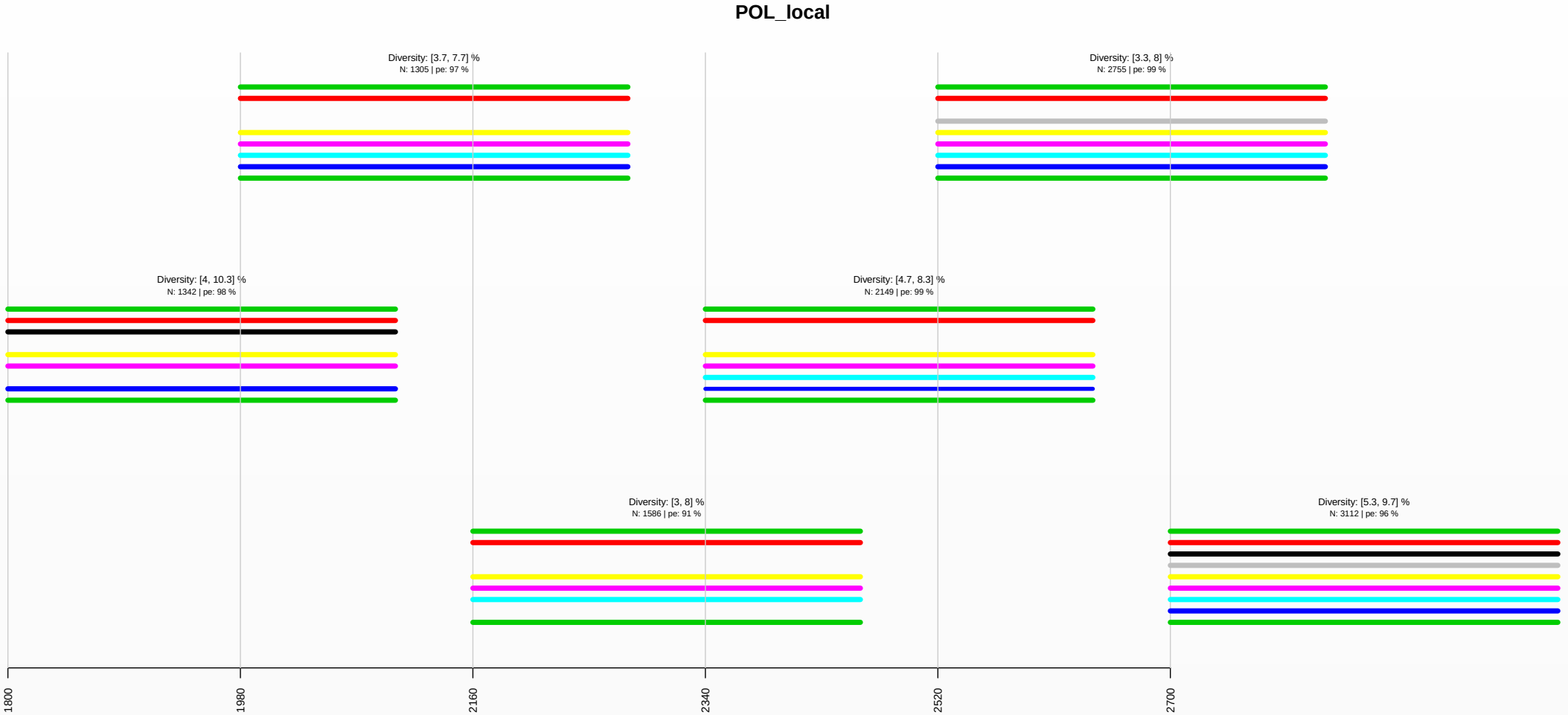
Prabhakaran et al., 2010

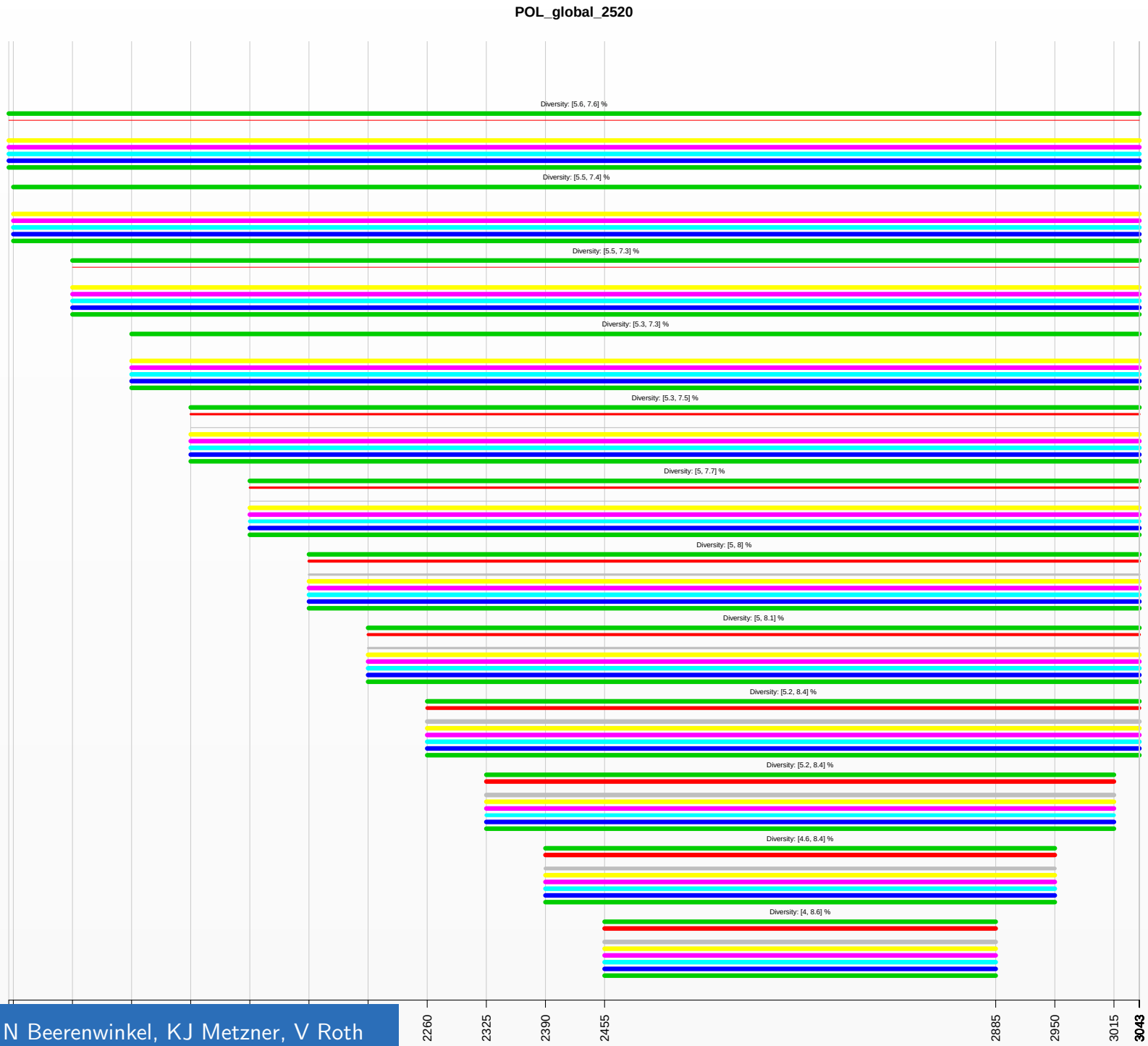
Experiments on 454/Roche data for pol gene



Prabhakaran et al., 2010

Pol gene, Local Reconstructions




POL_global_2520
Diversity: [5.6, 7.6] %
Diversity: [5.5, 7.4] %
Diversity: [5.5, 7.3] %
Diversity: [5.3, 7.3] %
Diversity: [5.3, 7.5] %
Diversity: [5, 7.7] %
Diversity: [5, 8] %
Diversity: [5, 8.1] %
Diversity: [5.2, 8.4] %
Diversity: [5.2, 8.4] %
Diversity: [4.6, 8.4] %
Diversity: [4, 8.6] %
2260
2325
2390
2455
2885
2950
3015
3043

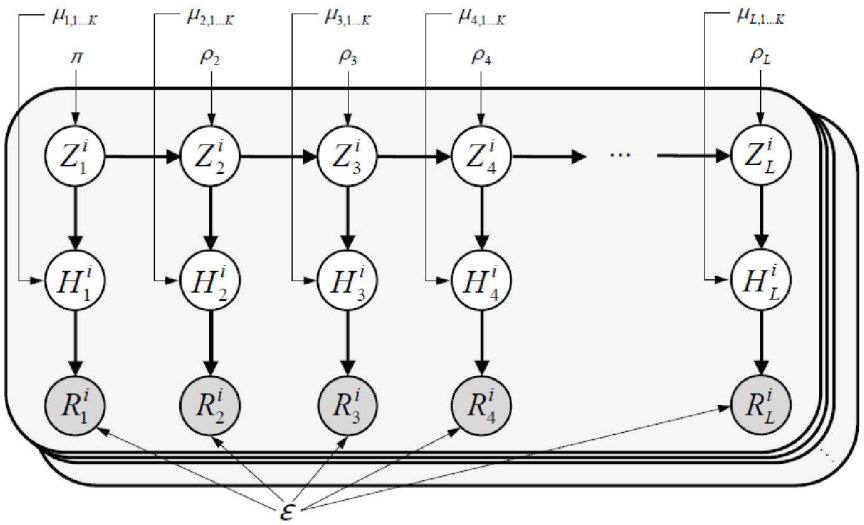
Pol gene, Results

Actual and Reconstructed haplotypes for HIV POL and simulated experiments on **454/Roche reads**.
All values are in %. X denotes 'undetected haplotype'.

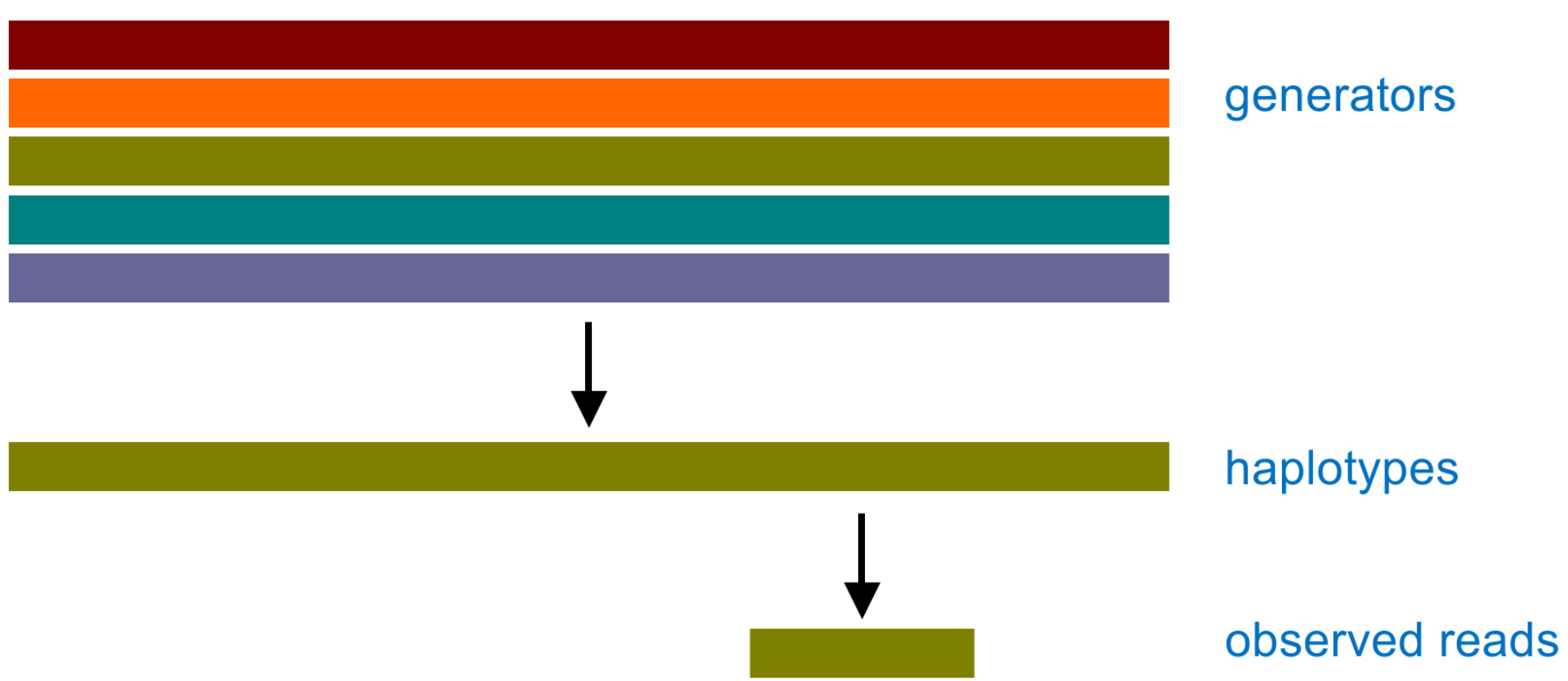
HIV POL	Actual	38.3	35.4	10.1	9.5	5.6	0.46	0.32	0.08	0.06	0.02
	Reconstructed	36.7	32.9	10.6	9.0	5.2	0.6	X	X	X	X
Simulation	Actual	50.8	24.3	12.6	6.3	3.0	1.6	0.8	0.4	0.25	0.06
	Reconstructed	50.4	24.3	12.5	6.3	3.0	2.0	0.8	X	X	X

- Can detect haplotypes **down to** 0.5% **frequency**.
- **Simulations agree perfectly** with real data!

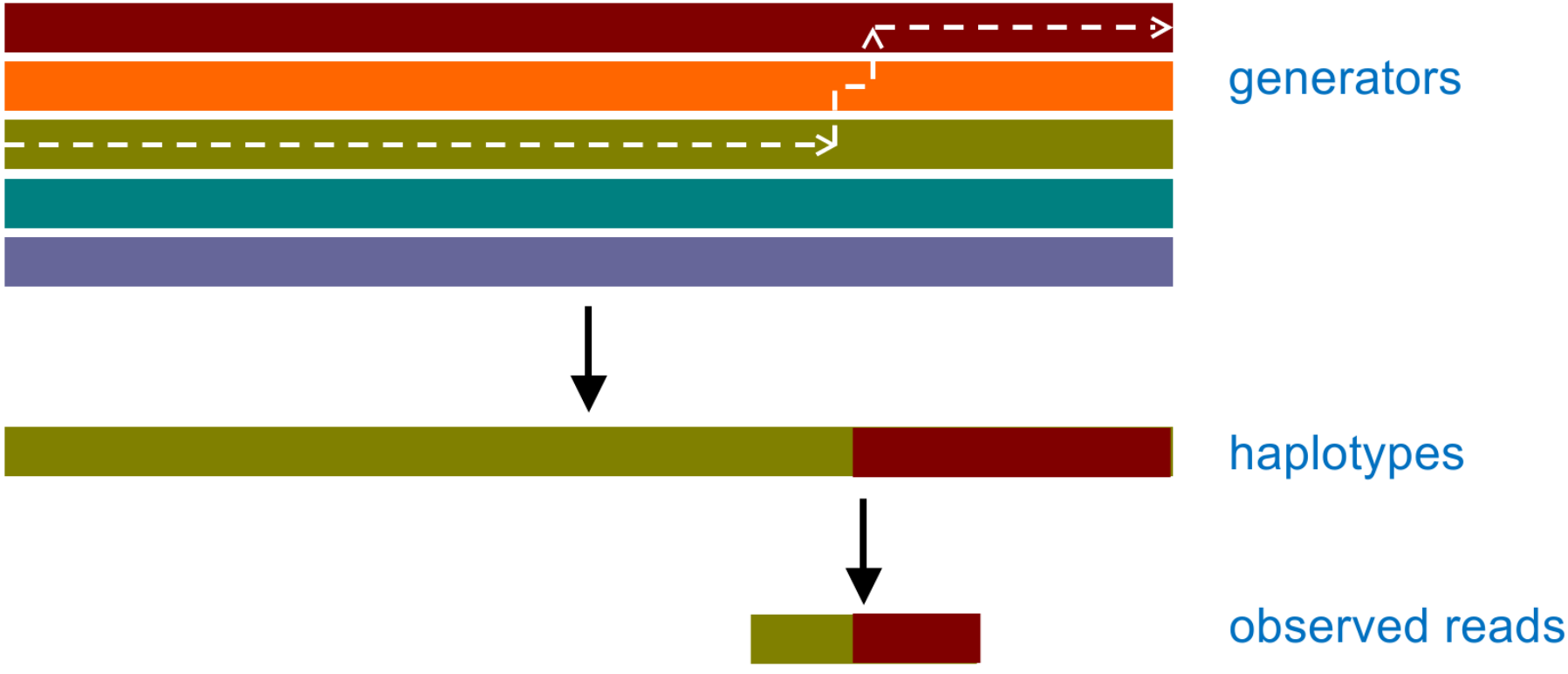
Probabilistic Assembly: Hidden Markov model



Hidden Markov model

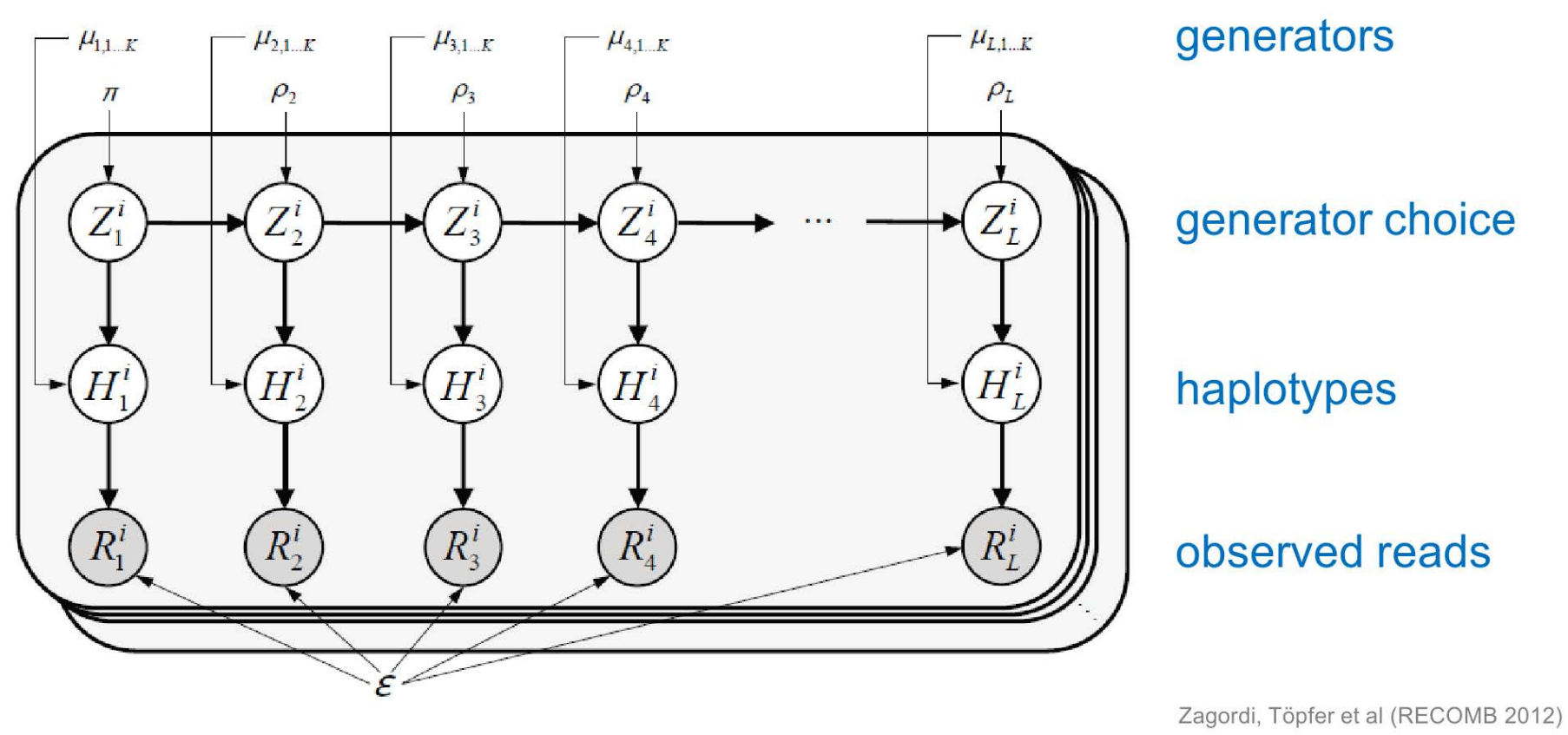


Recombination



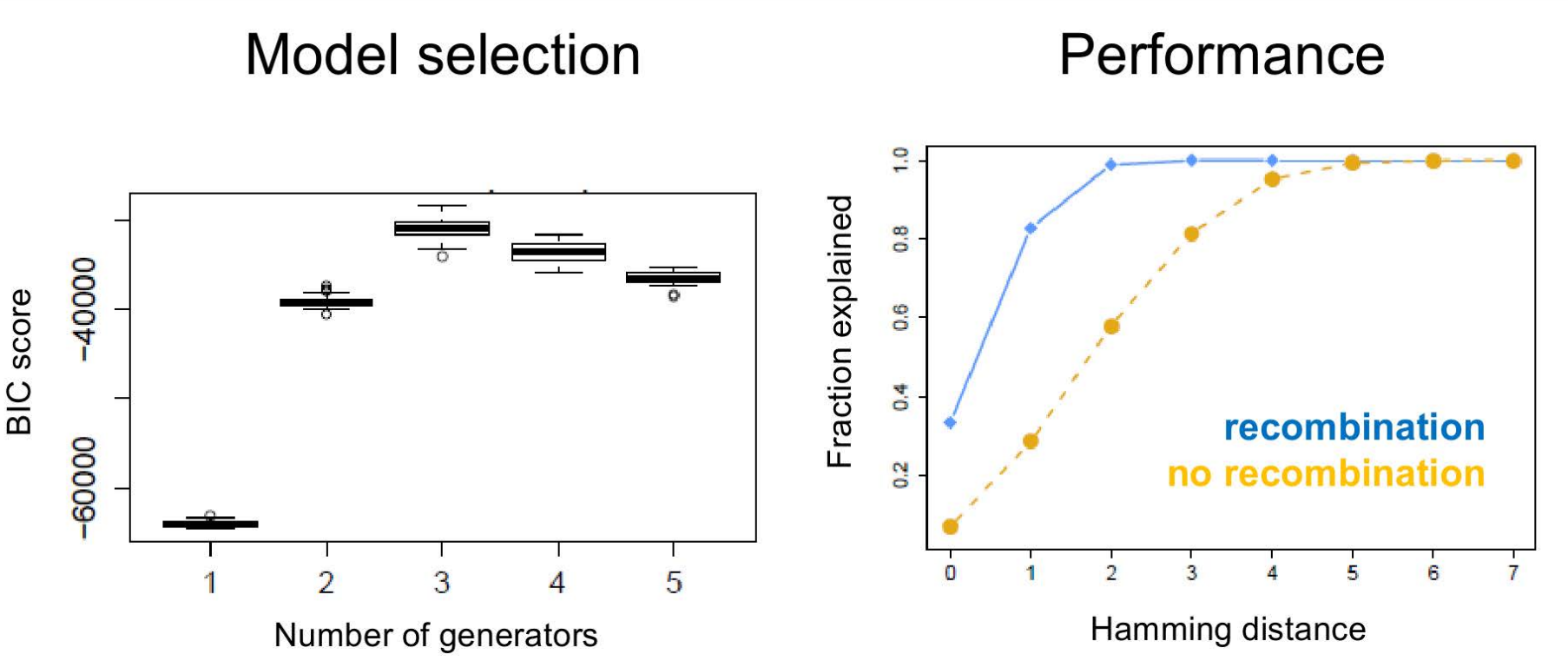
Jumping hidden Markov model

Many parameters $\rightsquigarrow$ requires strong regularization



Zagordi, Töpfer et al (RECOMB 2012)

Example: 3 generators + 24 recombinants



Zagordi, Töpfer et al (RECOMB 2012)

Global Haplotype Assembly: Summary

Two main classes:

Combinatorial assembly

$\rightsquigarrow$ *read graph, network flow, path cover, graph coloring* etc.

☺ well-studied **graph-theoretic background.**

☹ requires **error correction** prior to assembly.

Probabilistic assembly

$\rightsquigarrow$ *(infinite) mixture models, hidden Markov models* etc.

☺ **"integrated"**, no separate "hard" error correction.

☺ **flexible:** easy to include **constraints, recombinations**...

☹ **computational problems**, approximations needed.

General: **Reads must be long** enough to bridge conserved regions.
Missing length **cannot be compensated by higher coverage.**

Global Haplotype Assembly: Summary (2)

Software packages:

Program	Method	URL
QuRe	read graph	https://sourceforge.net/projects/qure/
ShoRAH	read graph	http://www.cbg.ethz.ch/software/shorah
ViSpA	read graph	http://alla.cs.gsu.edu/~software/VISPA/vispa.html
BIOA	read graph	https://bitbucket.org/nmancuso/bioa/
Hapler	read graph	http://nd.edu/~biocmp/hapler/
AmpliconNoise	probabilistic	http://code.google.com/p/ampliconnoise
PredictHaplo	probabilistic	http://bmda.cs.unibas.ch/HivHaploTyper/
QuasiRecomb	probabilistic	http://www.cbg.ethz.ch/software/quasirecomb

ECCB‘12 Tutorial 4
Inferring genetic diversity from next-generation sequencing data:
Computational methods and biomedical applications

Comparative Assessment of Methods, Demonstration of Case Studies

Niko Beerenwinkel
Volker Roth
Karin Metzner

Read length versus depth of coverage

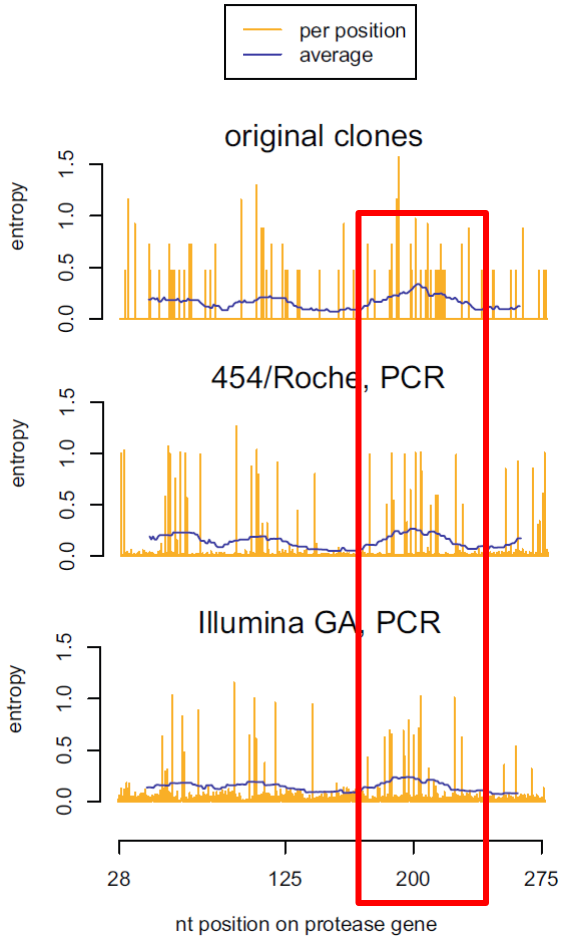
- Compare 10-clones mix between
  - 454/Roche (long reads, low coverage)
  - Illumina GA (short reads, high coverage)

Platform	PCR amplification	Total reads	Reads mapped to protease (10-93)	Mapped read length (mean ± sd)	Reads included in the analysis	Error rate [%] (mean ± sd)
454/Roche	No	16,540	668	232 ± 18	668	0.59 ± 0.02
454/Roche	Yes	45,973	4,331	236 ± 18	4,331	1.09 ± 0.01
Illumina GA	No	12,559,696	1,505,619	36	11,835	0.17 ± 0.01
Illumina GA	Yes	12,242,508	1,346,481	36	8,904	0.38 ± 0.01

Zagordi et al (PLOS ONE, in revision)

Local comparison

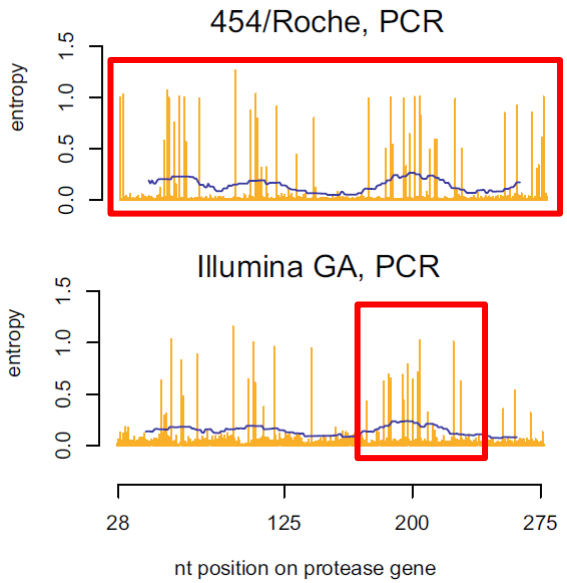
- Idea: Use region of highest diversity for Illumina-based local analysis.



Zagordi et al (PLOS ONE, in revision)

Local comparison

- Higher sensitivity and specificity with Illumina in detecting 10 clones.
- Low-frequency haplotypes go undetected with low coverage.

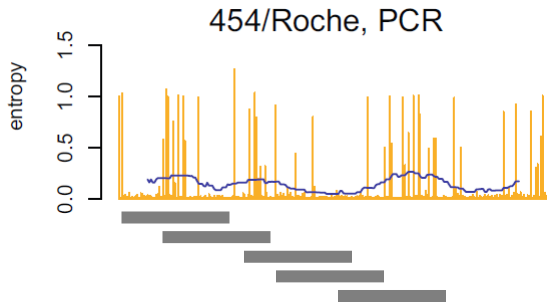


Platform	PCR amplification	Method	07-5 6681	07-548 25	07-569 51	08-597 12	08-041 34	08-013 15	08-026 59	08-578 81	08-045 12	Total
454/Roche	No	ShoRAH	10.6	14.1	14.1	13.9	4.9	—	—	—	—	57.6
454/Roche	No	Direct mapping	27.3	21.2	30.0	11.0	7.1	2.1	0.3	0.3	0.1	99.4
454/Roche	Yes	ShoRAH	3.6	15.7	22.0	11.4	7.0	0.3	—	—	—	60.0
454/Roche	Yes	Direct mapping	6.0	34.3	37.2	9.6	11.7	0.4	0.4	0.1	0.2	99.9
Illumina GA	No	ShoRAH	53.1	19.5	15.1	7.2	2.7	1.6	0.2	0.2	0.2	99.8
Illumina GA	No	Direct mapping	41.7	15.4	24.8	10.3	4.5	1.5	0.3	0.3	0.1	98.9
Illumina GA	Yes	ShoRAH	7.6	46.8	27.1	7.3	5.3	1.9	—	—	—	96.0
Illumina GA	Yes	Direct mapping	5.9	34.7	36.6	10.4	10.3	0.7	0.6	0.2	0.3	99.7

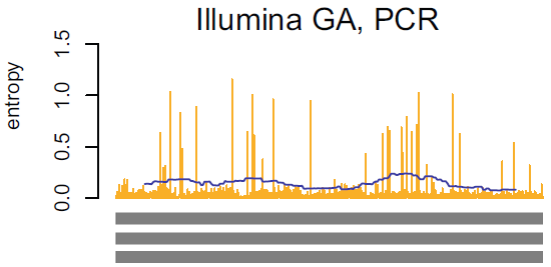
Zagordi et al (PLOS ONE, in revision)

Local versus global comparison

- Assemble short Illumina reads

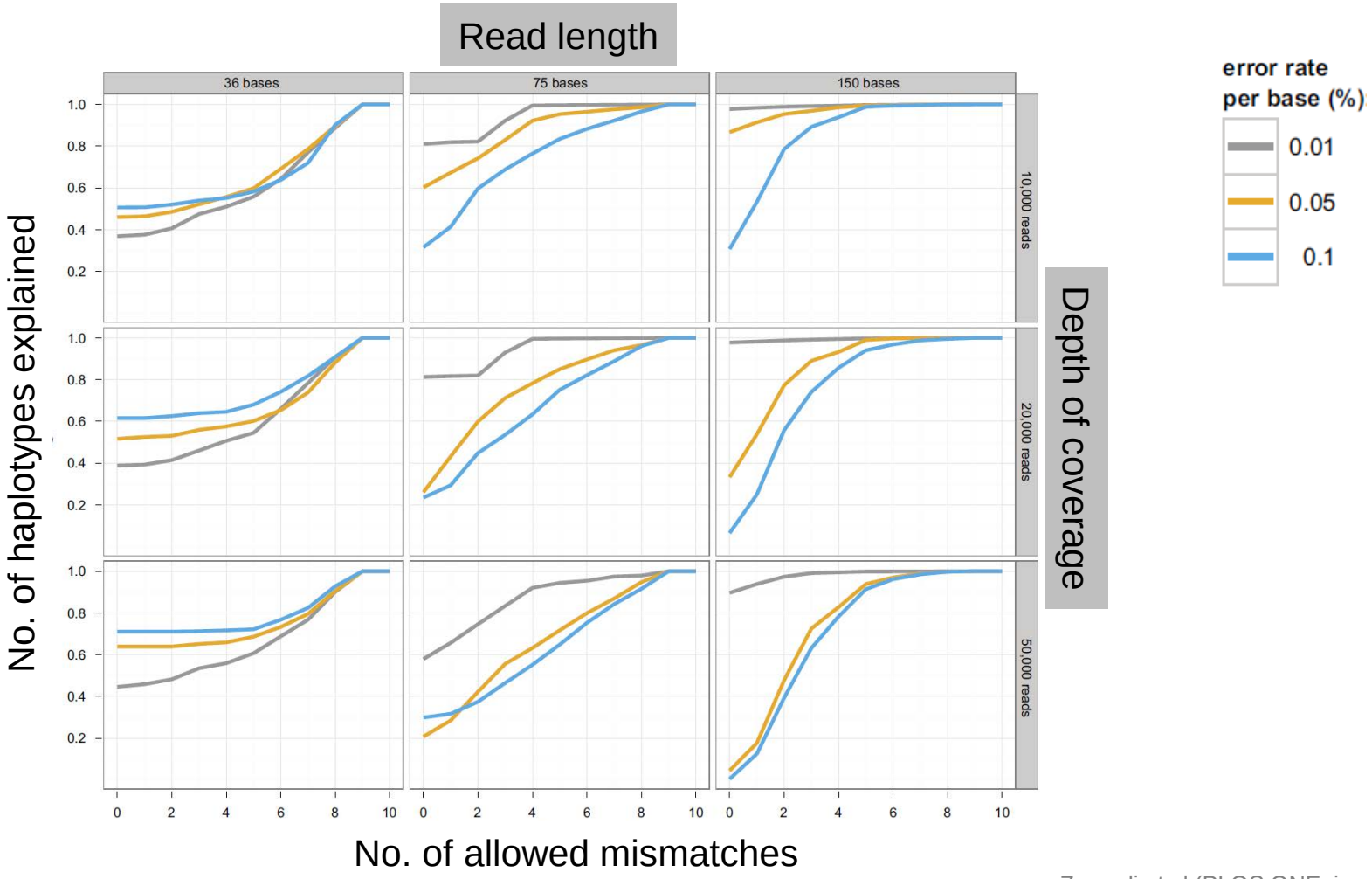


- Locally reconstruct long 454 reads



Zagordi et al (PLOS ONE, in revision)

Simulation study



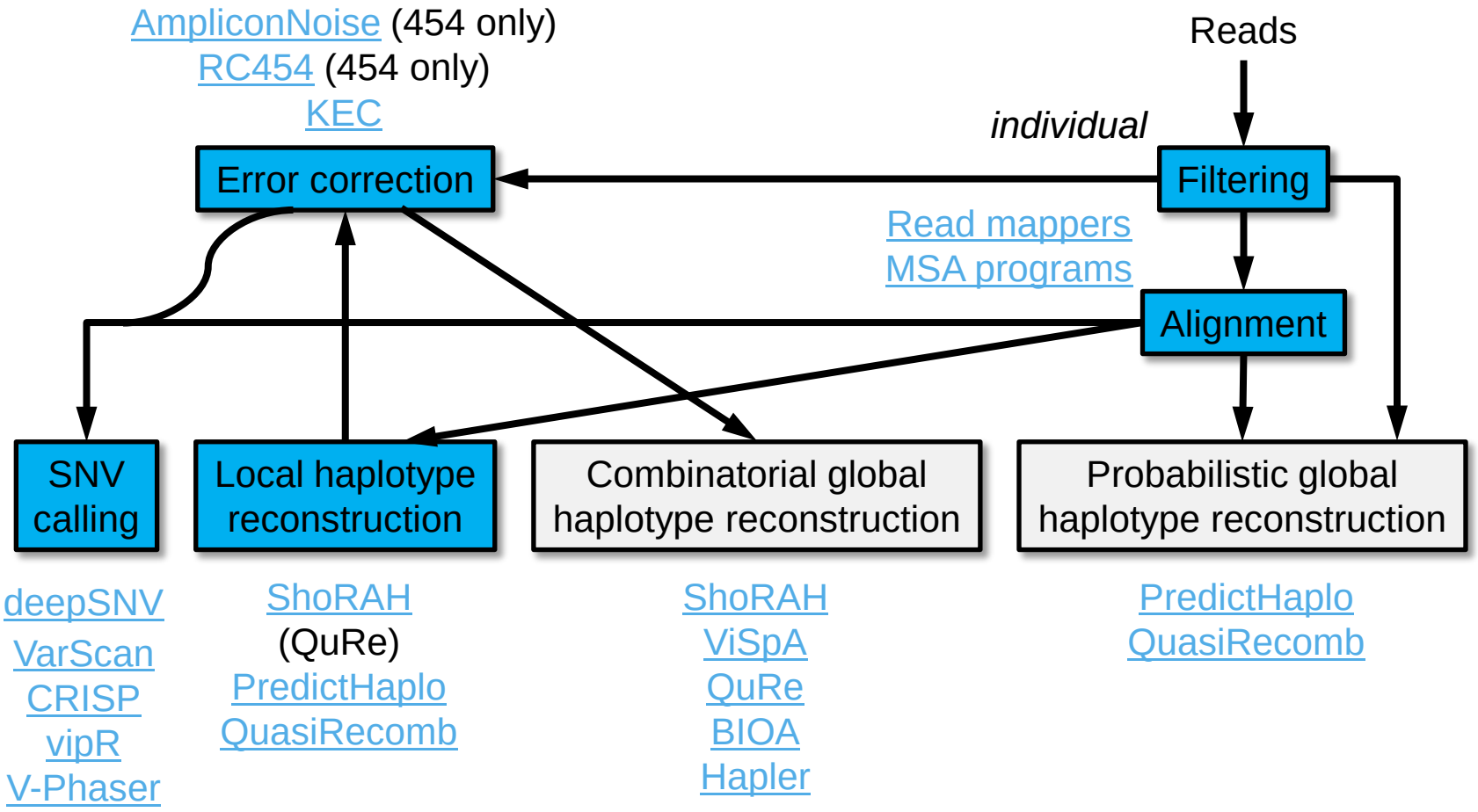
Zagordi et al (PLOS ONE, in revision)

Conclusions

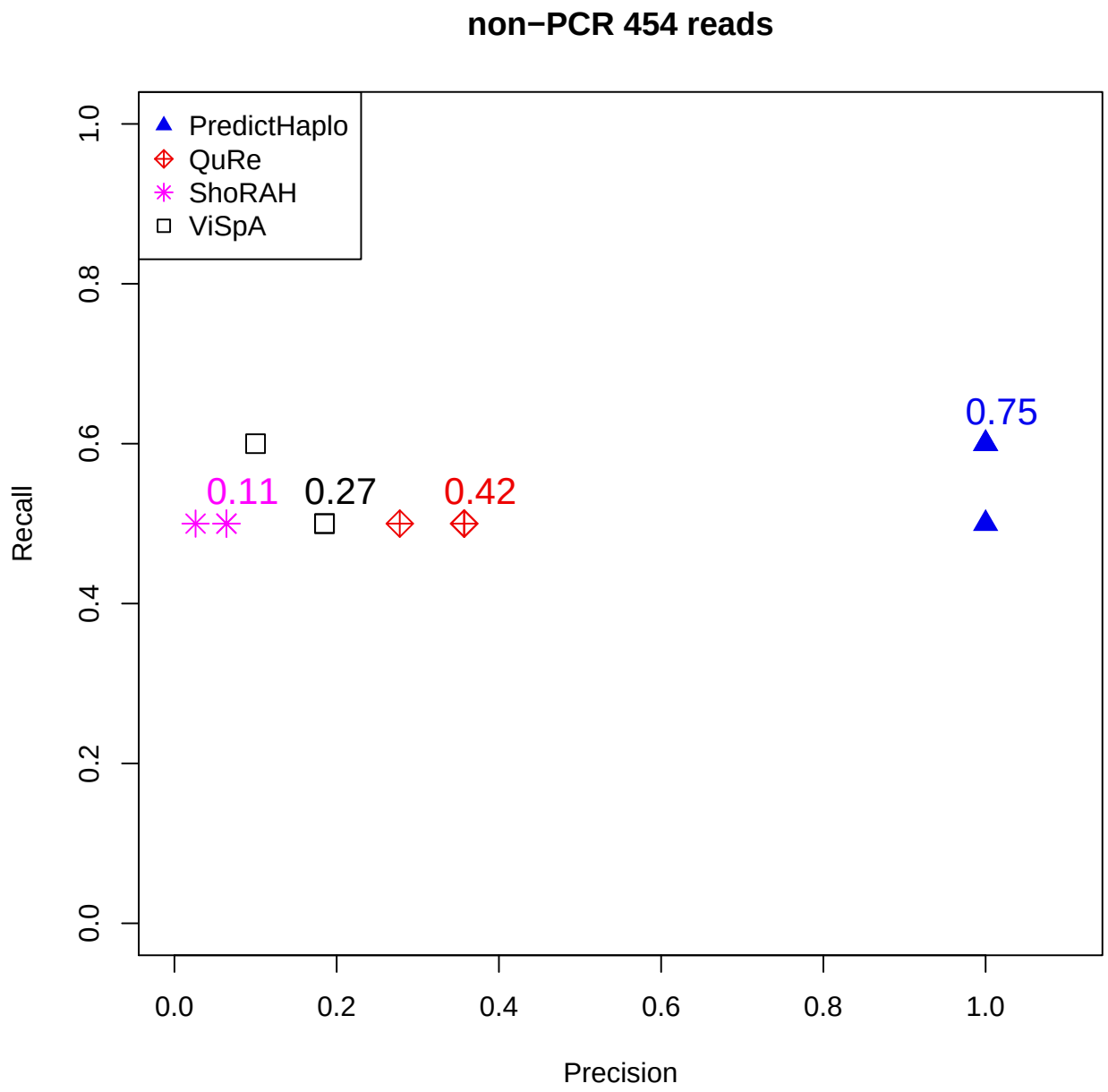
- Local reconstruction in high-diversity regions can provide most information about the number and frequency of clones (but not their full-length sequences).
- For detecting local variation, most notably SNVs, coverage is more critical than read length.
- For detecting global variation, read length is most critical:
  - If reads are too short, nothing helps.
  - If reads are long and errors are frequent, combinatorial reconstruction will generate too many false positives.
  - If reads are long and errors are rare, global reconstruction works.

Zagordi et al (PLOS ONE, in revision)

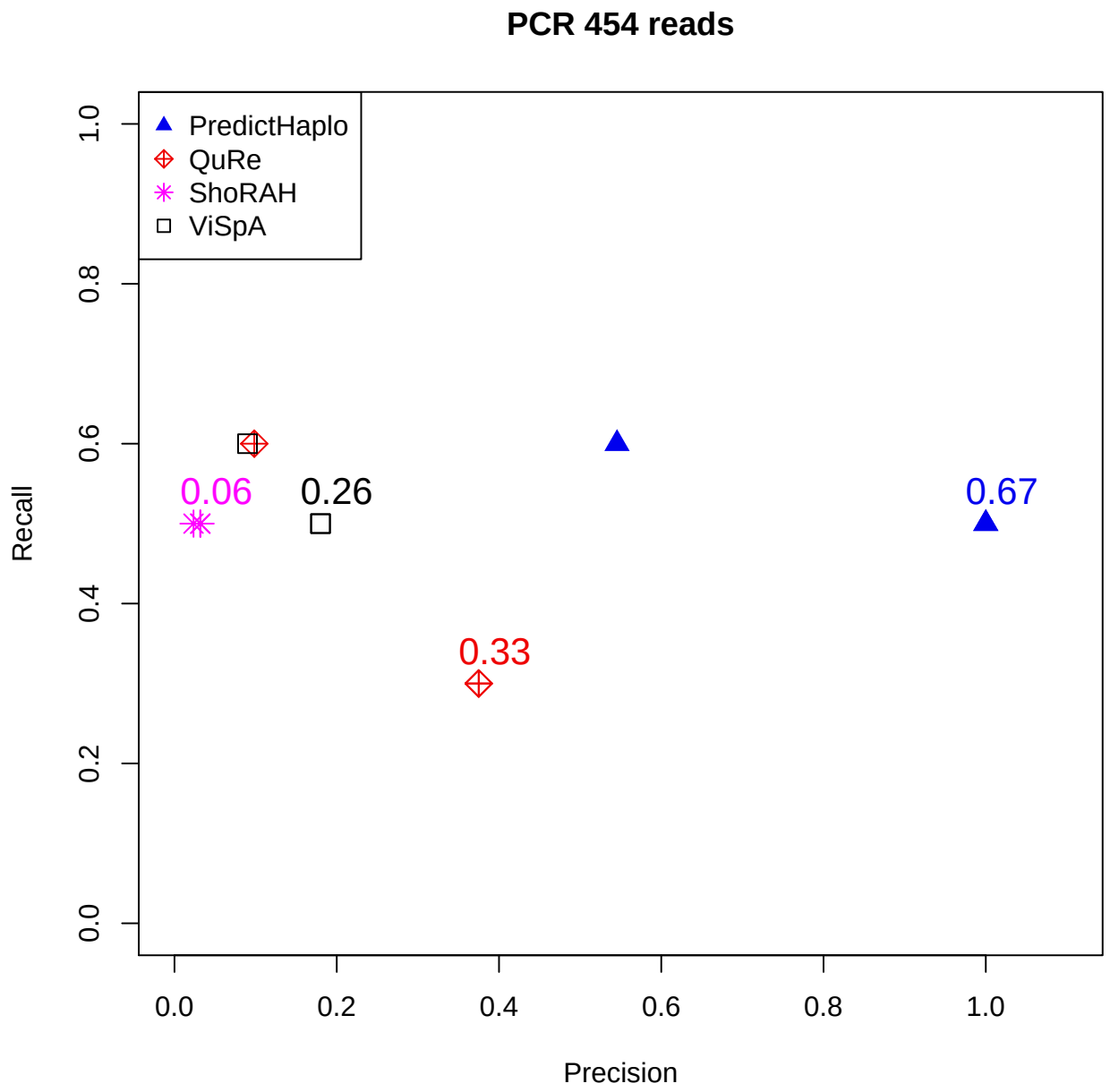
Software


AmpliconNoise (454 only)
RC454 (454 only)
KEC
Reads
individual
Error correction
Filtering
Read mappers
MSA programs
Alignment
SNV calling
Local haplotype reconstruction
Combinatorial global haplotype reconstruction
Probabilistic global haplotype reconstruction
deepSNV
VarScan
CRISP
vipR
V-Phaser
ShoRAH
(QuRe)
PredictHaplo
QuasiRecomb
ShoRAH
ViSpA
QuRe
BIOA
Hapler
PredictHaplo
QuasiRecomb

Comparing QuRe, ViSpA,ShoRAH,PredictHaplo



Comparing QuRe, ViSpA,ShoRAH,PredictHaplo

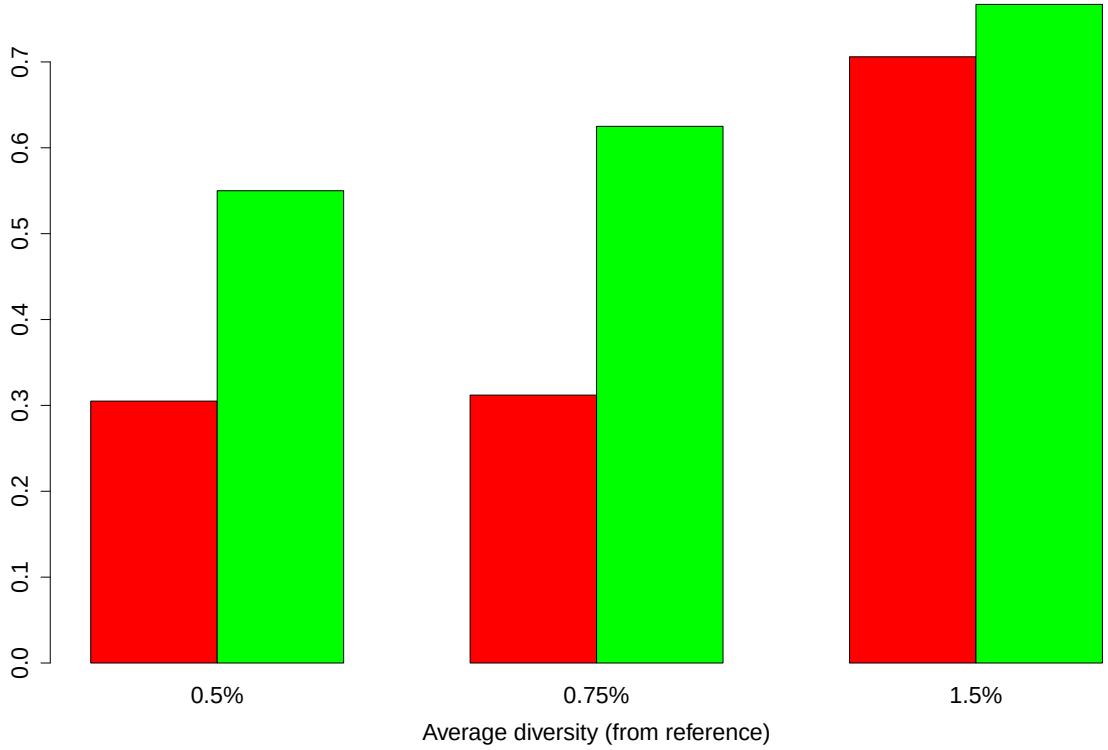


Comparing QuRe, ViSpA,ShoRAH,PredictHaplo

- Main performance difference between methods: **false positives**
- "Fuzziness" helps:
  - **Random path sampling** (QuRe) seems to be beter than strict path cover or network flow
  - **Fully probabilistic approach** seems to work best.
- ...you might forget about all the numbers presented and **run your own experiments:** all methods and the dataset can be freely downloaded from the web.

Evolution of the 454 Platform: Simulated Reads

Reads simulated with **MetaSim** at different read lengths and diversities.
Ten clone mix, frequencies decrease by factor of two from clone to clone.
Same number of reads in each experiments (200,000).
Red: 2008, avg read length ≈ 340
Green: 2012, avg read length ≈ 700.



Some Further Comments

- Length of reads is **the** important quantity.
- Preference for platforms producing long reads, like 454.
- Coverage is **important for detecting low-abundance haplotypes**.
- If read length is **too short to cover conserved regions**, increasing the coverage **will not help**.
- Next version of Solexa with 2×250 paired reads might become highly interesting, as well as alternative platforms like Pacific Biosciences etc.

ECCB 2012 Tutorial 4

Summary and clinical applications

Karin J. Metzner

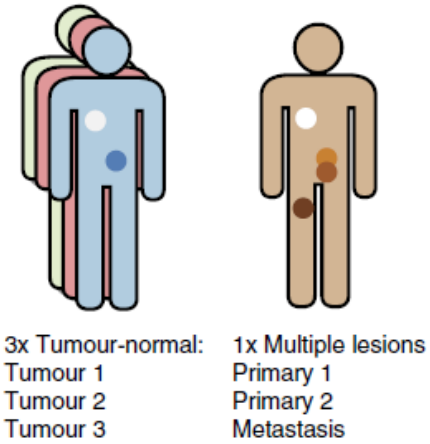
Division of Infectious Diseases and Hospital Epidemiology

University of Zurich[UZH]

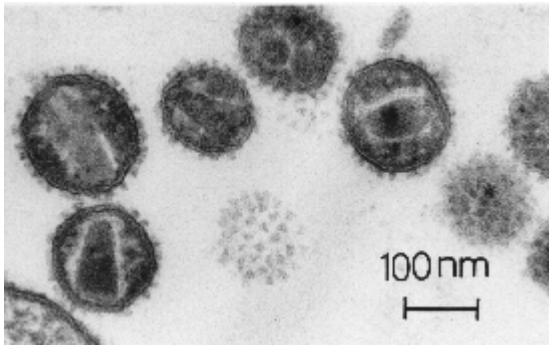
UniversityHospital Zurich

Case studies/examples

- Genetic diversity in tumors: Detecting low-frequency single-nucleotide variants (SNVs)



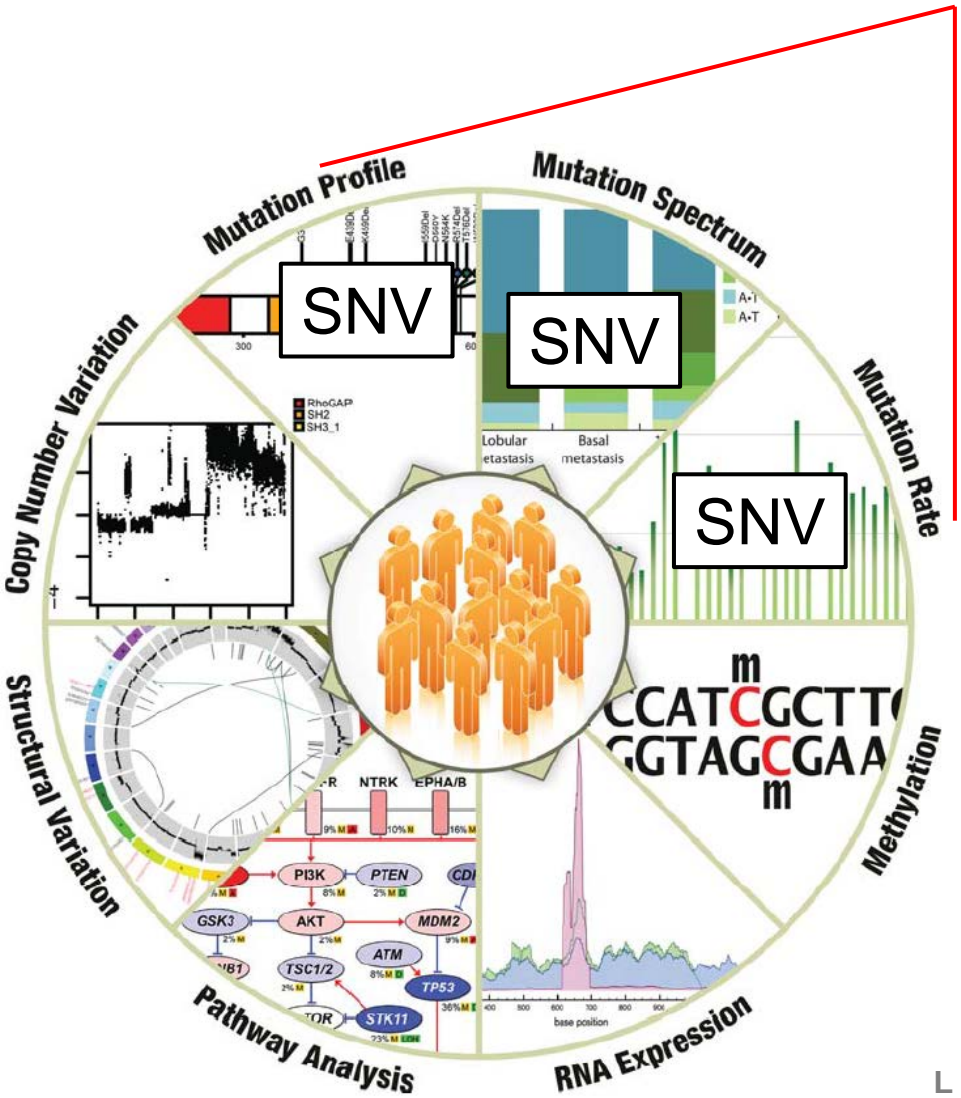
- Genetic diversity in virus populations: Local and global haplotype reconstruction



HR Gelderblom *et al*., Virology 1987

M Gerstung *et al*., Nat Comm 2012

Landscape of genomics analyses

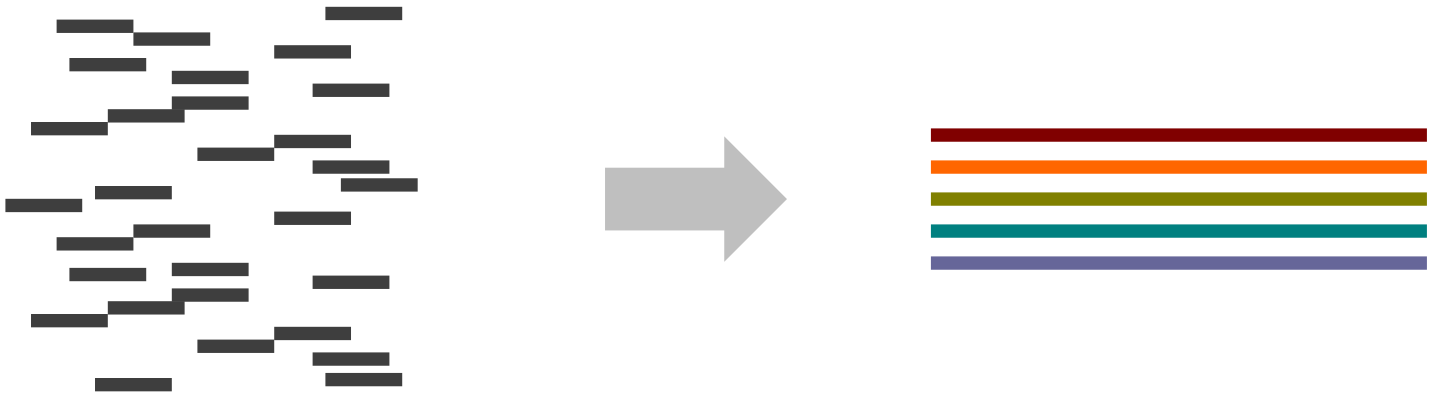


haplotype reconstruction

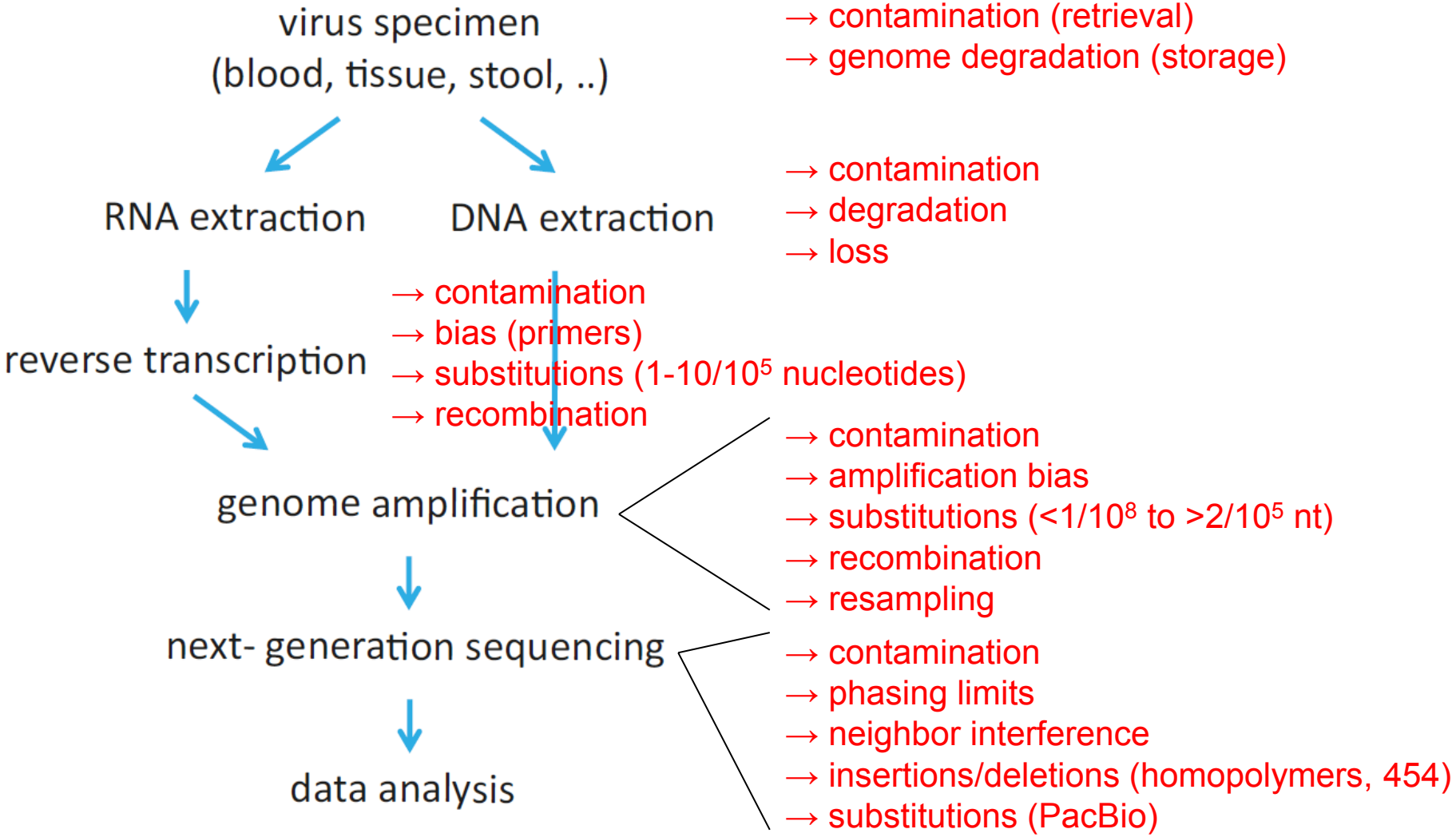
L Ding *et al.*, Hum Mol Genetics 2010

NGS-based diversity estimation: Main challenges

- Alignment (mapping) uncertainty
- Confounding sources of variation (errors) of multiple types
- Short read length relative to genomic region of interest



Error sources in next-generation sequencing



N Beerenwinkel *et al.*, Frontiers Microbiol 2012

Comparison of next-generation sequencing platforms

454/Roche GS-FLX:
up to 1’000’000 sequences/run
length: 400-700 bp/read

Illumina HiSeq 2000:
up to 1’500’000’000 seq./run
length: 2x100 bp/read

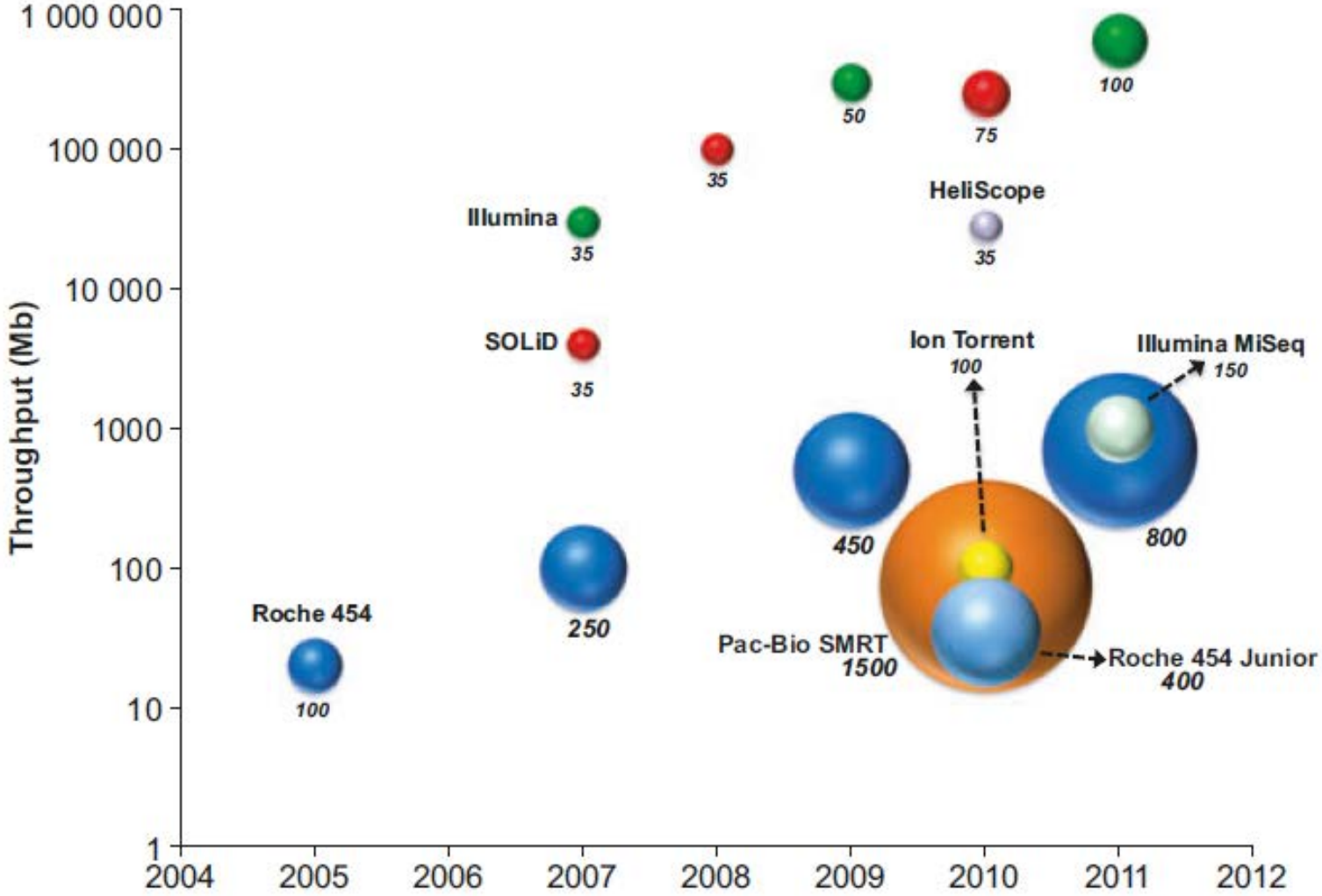
ABI 5500 SOLiD :
up to 900’000’000 seq./run
length: 50-75 bp/read

Pacific Biosciences RS:
up to 800’000 seq./run
length: ~1’500 bp/read

Ion Torrent PGM:
up to 5’000’000 seq./run
length: 35-200 bp/read

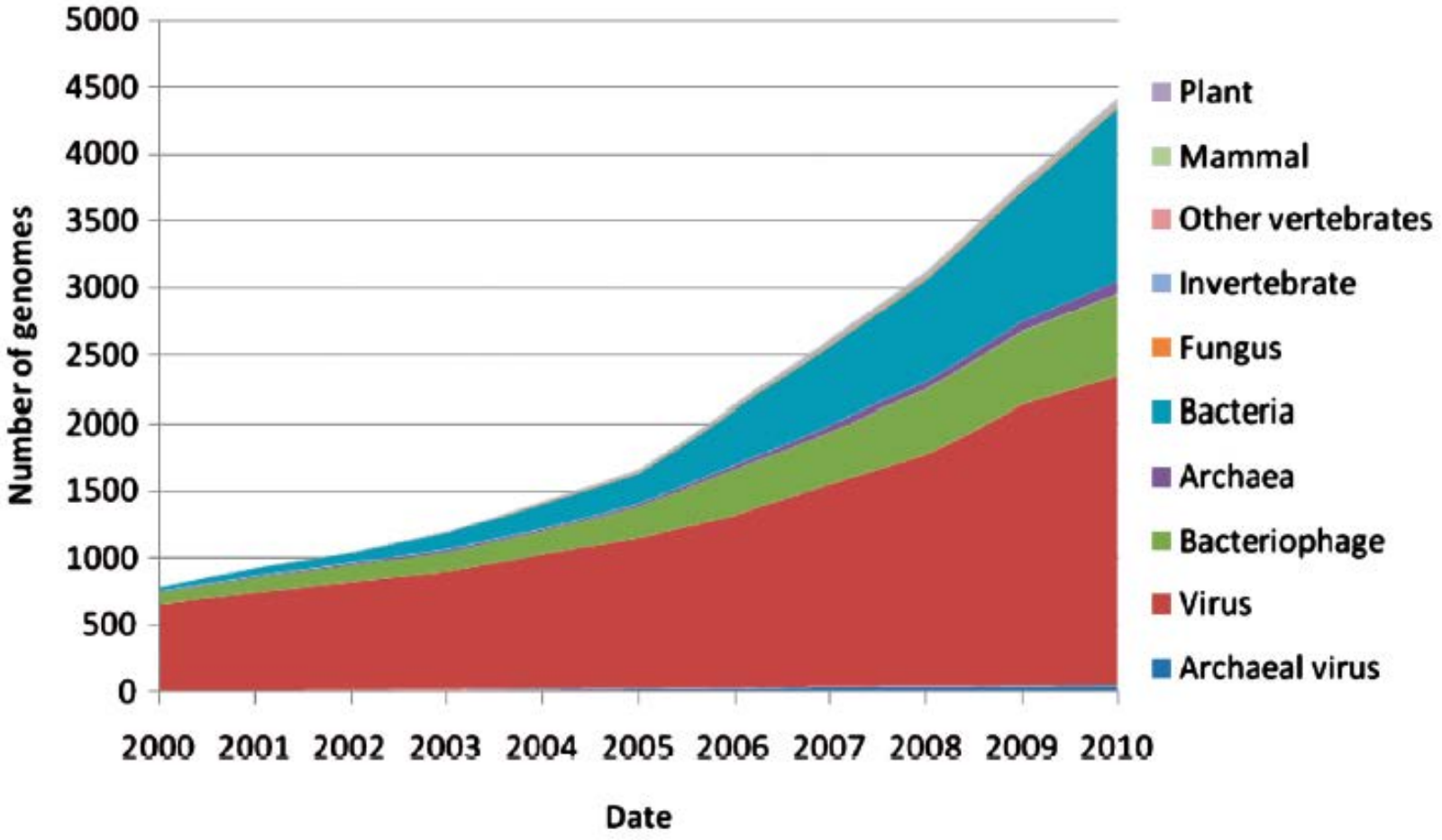
Helicos HeliScope:
up to 800’000’000 seq./run
length: 25-55 bp/read

Historical development of next-generation sequencing technologies



S Shokralla *et al*., Mol Ecol 2012

Growth in complete genomes



The International Nucleotide Sequence Database Collaboration, NAR 2010

Faster turn-around and dropping costs

Table 5 Sequencing statistics of six individual human genomes

platform	individual	No. of reads (millions)	read length (bases)	read coverage	genome coverage (%)	SNPs (millions)	No. of runs	estimated cost (US$)	references
Sanger	J. Craig Venter	31.9	800	7.5×	N/A	3.21	>340,000	70,000,000	Levy et al., 2007
Roche 454	James D. Watson	93.2	250	7.4×	95	3.32	234	1,000,000	Wheeler et al., 2008
SOLiD	James R. Lupski	238	35	29.6×	99.8	3.42	3	75,000	Lupski et al., 2010
Illumina Solexa	Yoruba male (NA18507)	3681	35	40.6×	99.9	4	40	250,000	Pushkarev 2009
	Han Chinese male (YH)	2950	35	36×	99.9	3.07	35	500,000	Wang et al., 2008
	Korean male (SJK)	1647	35, 74	29.0×	99.9	3.44	15	250,000	Ahn et al., 2009
	Korean male (AK1)	1910	36, 88, 106	27.8×	99.8	3.45	30	200,000	Kim et al., 2009
Helicos	Stephen R. Quake	2725	32	28×	90	2.81	4	48,000	Pushkarev et al., 2009

X Zhou *et al*., Protein Cell 2012

The future of next-generation sequencing: Data analysis

EasyGenomics Beta

Genome | INTERNATIONAL
DECODING FOR KNOWLEDGE DISCOVERY

GOLDEN HELIX
Accelerating the Quest for Significance

Era7 bioinformatics

GenomeQuest

INTREPID BIOINFORMATICS

SEQanswers
the next generation sequencing community

and more .. to come …

NGS and clinical trials

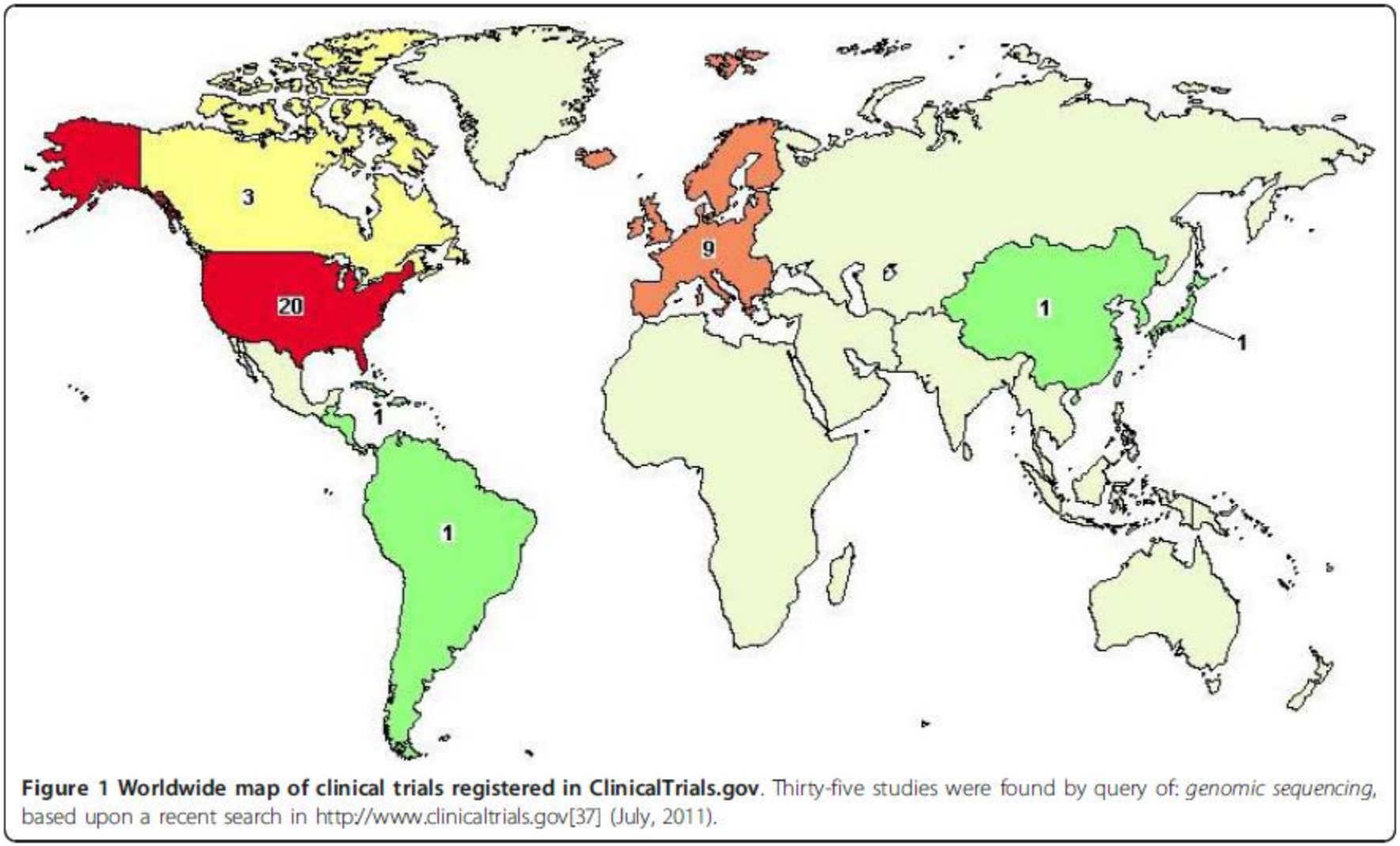


Figure 1 Worldwide map of clinical trials registered in ClinicalTrials.gov. Thirty-five studies were found by query of: *genomic sequencing*, based upon a recent search in http://www.clinicaltrials.gov[37] (July, 2011).

KK Mestan *et al*., J Trans Med 2011

Whole-genome sequencing studies in cancer

Study	Method	Cancer type	Number of samples sequenced	Aberration type
Ley *et al.*, 2008	Deep single-end whole-genome sequencing	AML	1	Point mutations, insertions, deletions
Campbell *et al.*, 2008	Shallow paired-end whole-genome sequencing	Lung	2	Deletions, amplifications, tandem duplications, interchromosomal rearrangements
Stephens *et al.*, 2009	Shallow paired-end whole-genome sequencing	Breast	24	Deletions, amplifications, tandem duplications, interchromosomal rearrangements, inversions
Pleasance *et al.*, 2010	Deep paired-end whole-genome sequencing	Melanoma	1	Point mutations, insertions, deletions, amplifications, interchromosomal rearrangements
Pleasance *et al.*, 2010	Deep paired-end whole-genome sequencing	Small-cell lung	1	Point mutations, insertions, deletions, amplifications, interchromosomal rearrangements
Mardis *et al.*, 2009	Deep paired-end whole-genome sequencing	AML	1	Point mutations, insertions, deletions, amplifications, interchromosomal rearrangements
Shah *et al.*, 2009	Deep paired-end whole-genome sequencing	Breast	1	Point mutations, insertions, deletions, amplifications, interchromosomal rearrangements
Ding *et al.*, 2010	Deep paired-end whole-genome sequencing	Breast	1	Point mutations, insertions, deletions, amplifications, interchromosomal rearrangements, inversions
Lee *et al.*, 2010	Deep paired-end whole-genome sequencing	Lung	1	Point mutations, insertions, deletions, amplifications, interchromosomal rearrangements, inversions

AML, acute myelogenous leukaemia.

M Meyerson *et al.*, Nat Rev Genetics 2010

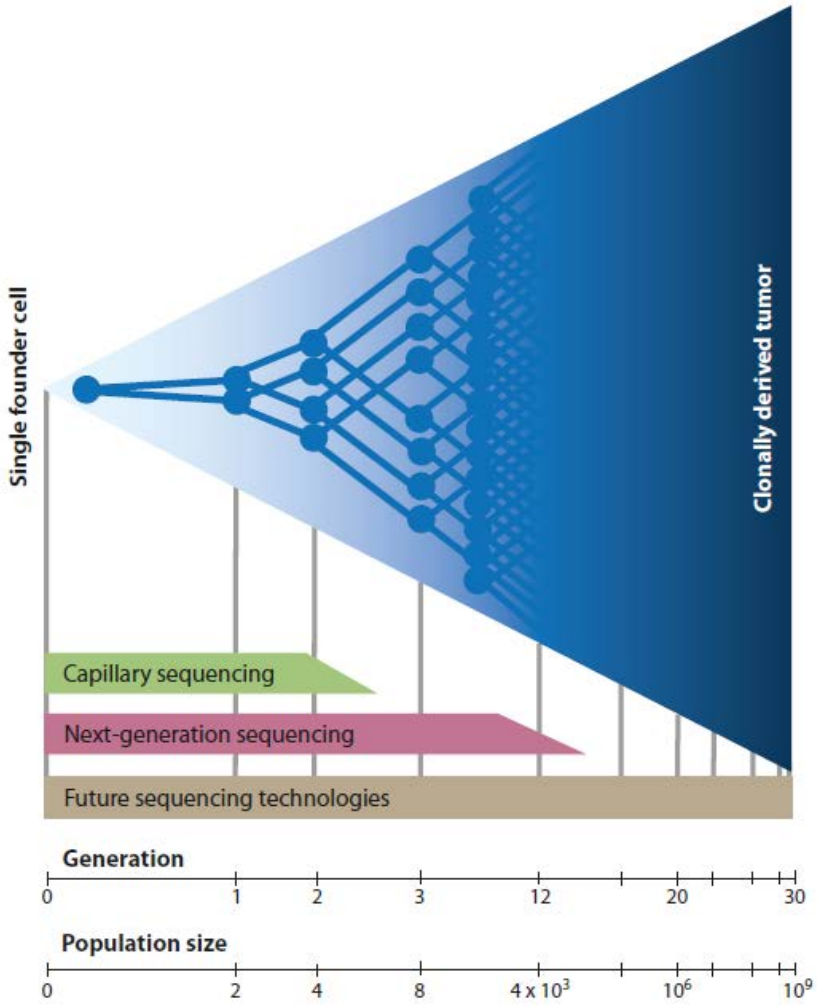
Cancer genome sequencing studies

Study type	Number of genes screened	Total number of mutations	Number of genes mutated	Average number of mutations per tumor	Estimated number of driver mutations	Reference(s)
Exomic						
Breast (n = 11)	18,191	1243	1137	84	140	87, 88
Colorectal (n = 11)	18,191	942	848	76	140	87, 88
Diverse (n = 210)	518	798	581	–	119	94
Pancreatic (n = 24)	20,661	1163	1007	48	160	98
Glioblastoma (n = 21)	20,661	748	685	47	155	102
Glioblastoma (n = 91)	601	453	223	–	8	103
Lung (n = 188)	623	1013	348	–	26	108
Genomic						
Acute myeloid leukemia (n = 1)	–	500–1000	10	Not applicable	10	82

JJ Salk *et al*., Ann Rev Pathol Mech Dis 2010

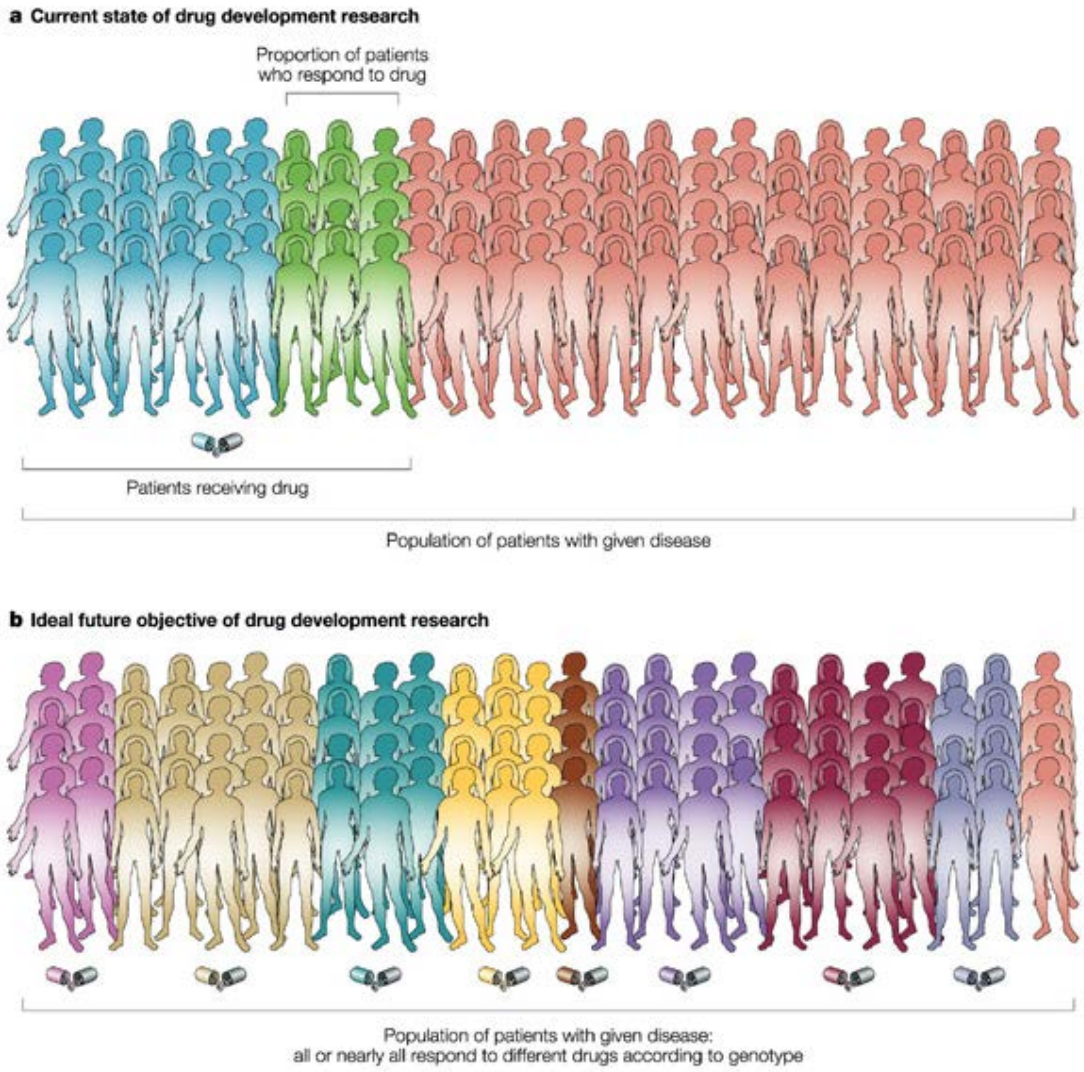
Limits of subclonal detection

- capillary sequencing
  - 25%
- NGS
  - 0.0002%
- future technologies
  - 0.?????x%



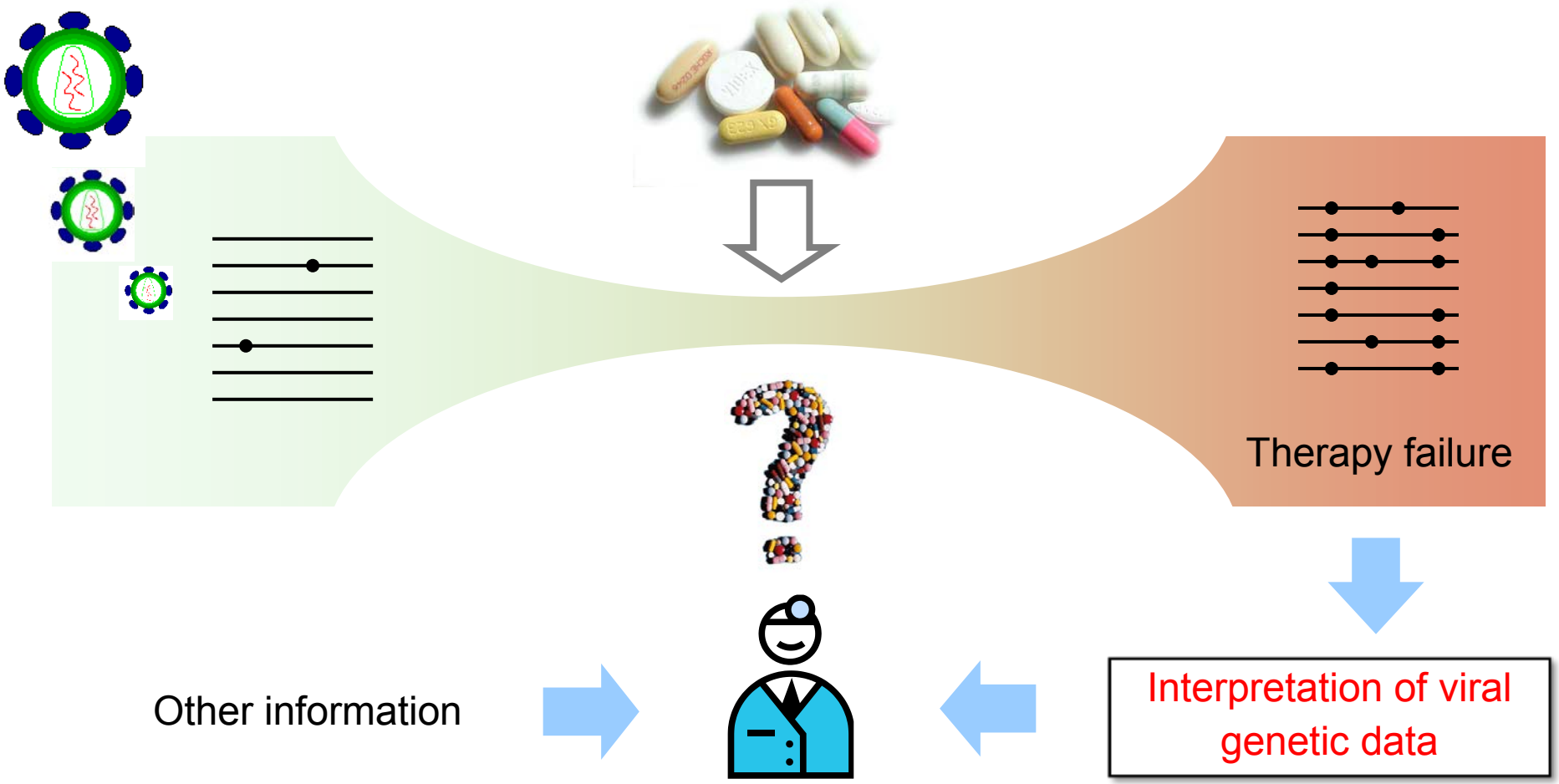
JJ Salk *et al*., Ann Rev Pathol Mech Dis 2010

The future is personalized medicine



AM Issa, Nat Rev Drug Dis 2002

HIV drug resistance, individualized treatment



Some of numerous open questions in basic research

- Origin of viruses
- Are virus populations changing within the next decades?
  - Can we predict virus evolution?
- Do mutation rates vary within and between infected individuals?
  - Do host factors influence virus evolution?
- …

Acknowledgements

- Christian Beisel (ETH Zurich)
- Rémy Bruggmann (FGCZ)
- Martin Däumer (SeqIT)
- Francesca Di Giallonardo (U Hospital Zurich)
- Yannick Duport (U Hospital Zurich)
- Nick Eriksson (23andMe)
- Moritz Gerstung (Sanger Institute)
- Huldrych Günthard (U Hospital Zurich)
- Eran Halperin (Tel Aviv U)
- Marzanna Künzli-Gontarczyk (FGCZ)
- Christine Leemann (U Hospital Zurich)
- Fabio Luciani (U New South Wales)
- Kerensa McElroy (U New South Wales)
- Holger Moch (U Hospital Zurich)
- Sandhya Prabhakaran (U Basel)
- Melanie Rey (U Basel)
- Bob Shafer (Stanford U)
- Armin Töpfer (ETH Zurich)
- Wan-Lin Yang (U Hospital Zurich)
- Osvaldo Zagordi (U Zurich)

Funding:
SNF
SHCS
SystemsX.ch
Vontobel Foundation