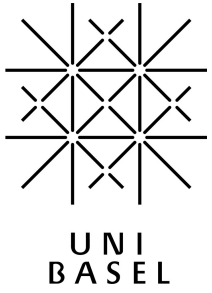
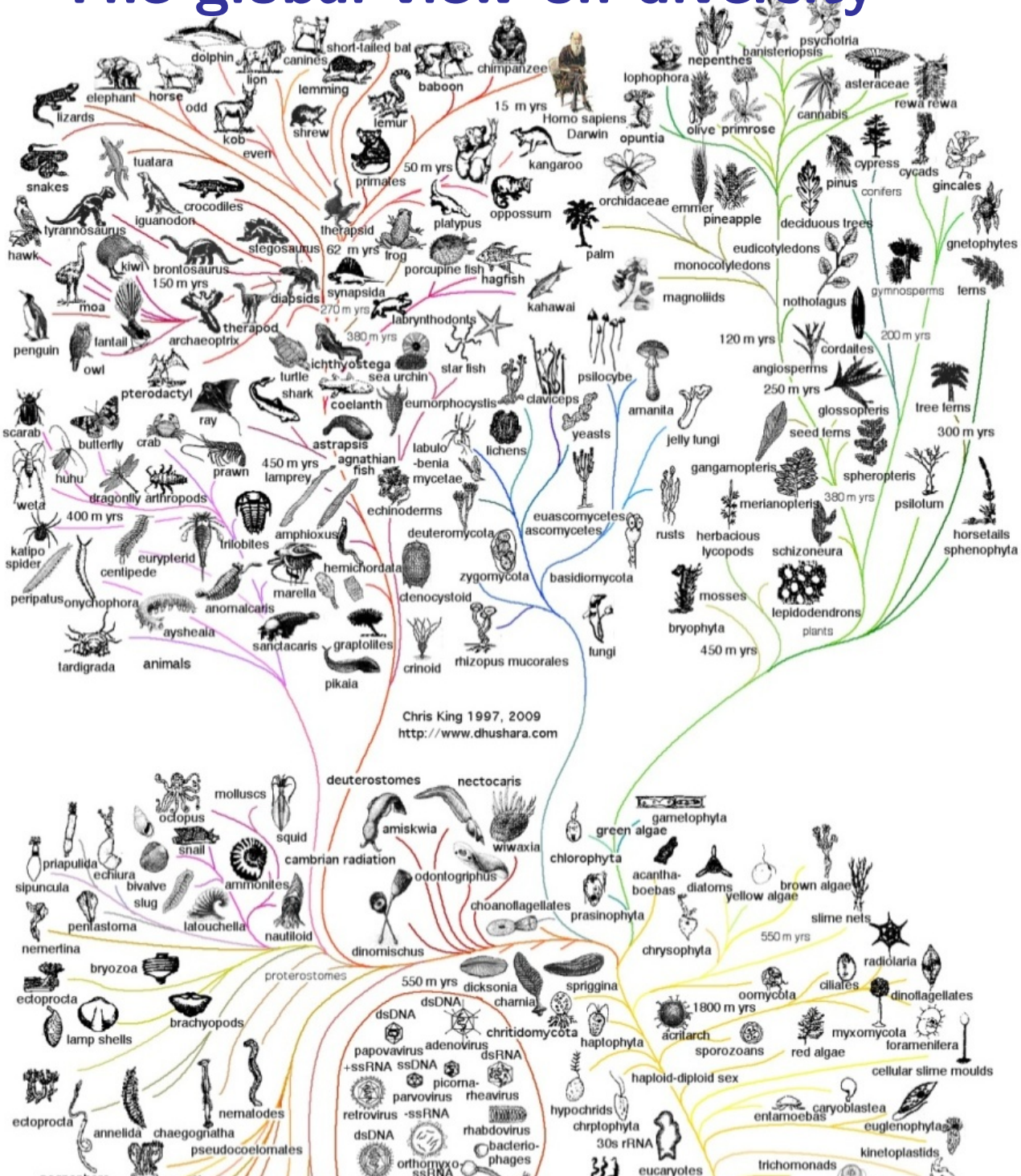
Inferring Genetic Diversity from Next-generation Sequencing Data

Slides adapted from ECCB 2012 Tutorial, together with Karin Metzner (UHZ) and Niko Beerenwinkel (ETHZ)

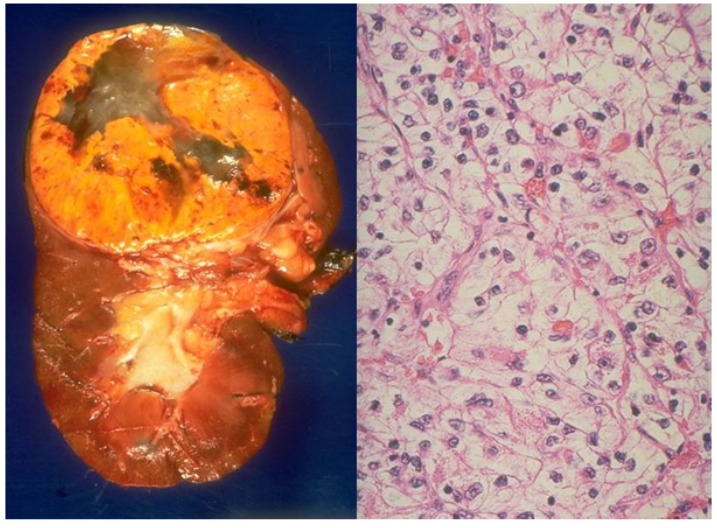


The global view on diversity

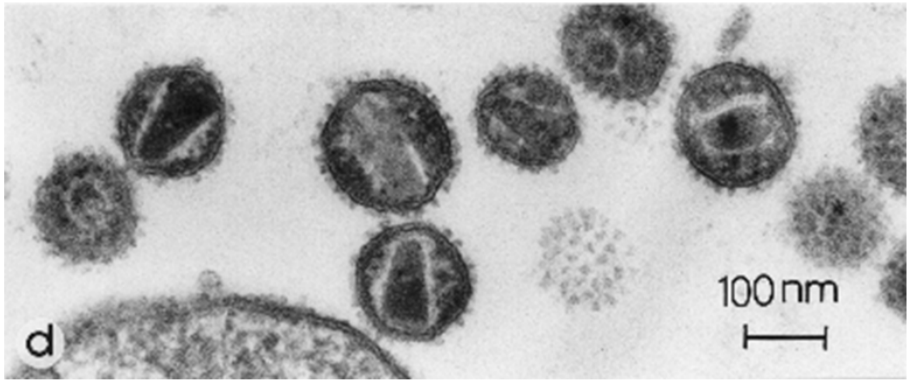


Zooming in: diversity within an organism

- Genetic diversity in an organism: **intra-tumour diversity**

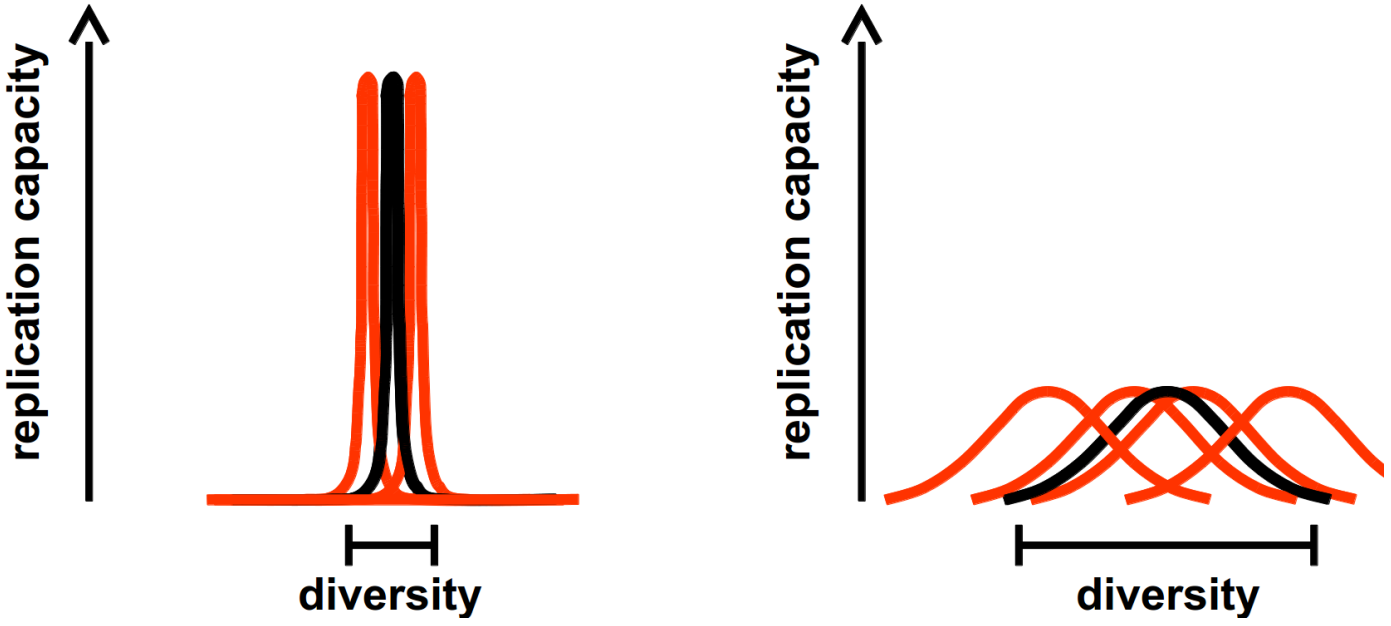


- Genetic diversity of a species in a defined environment: **intra-patient diversity of HIV-1**



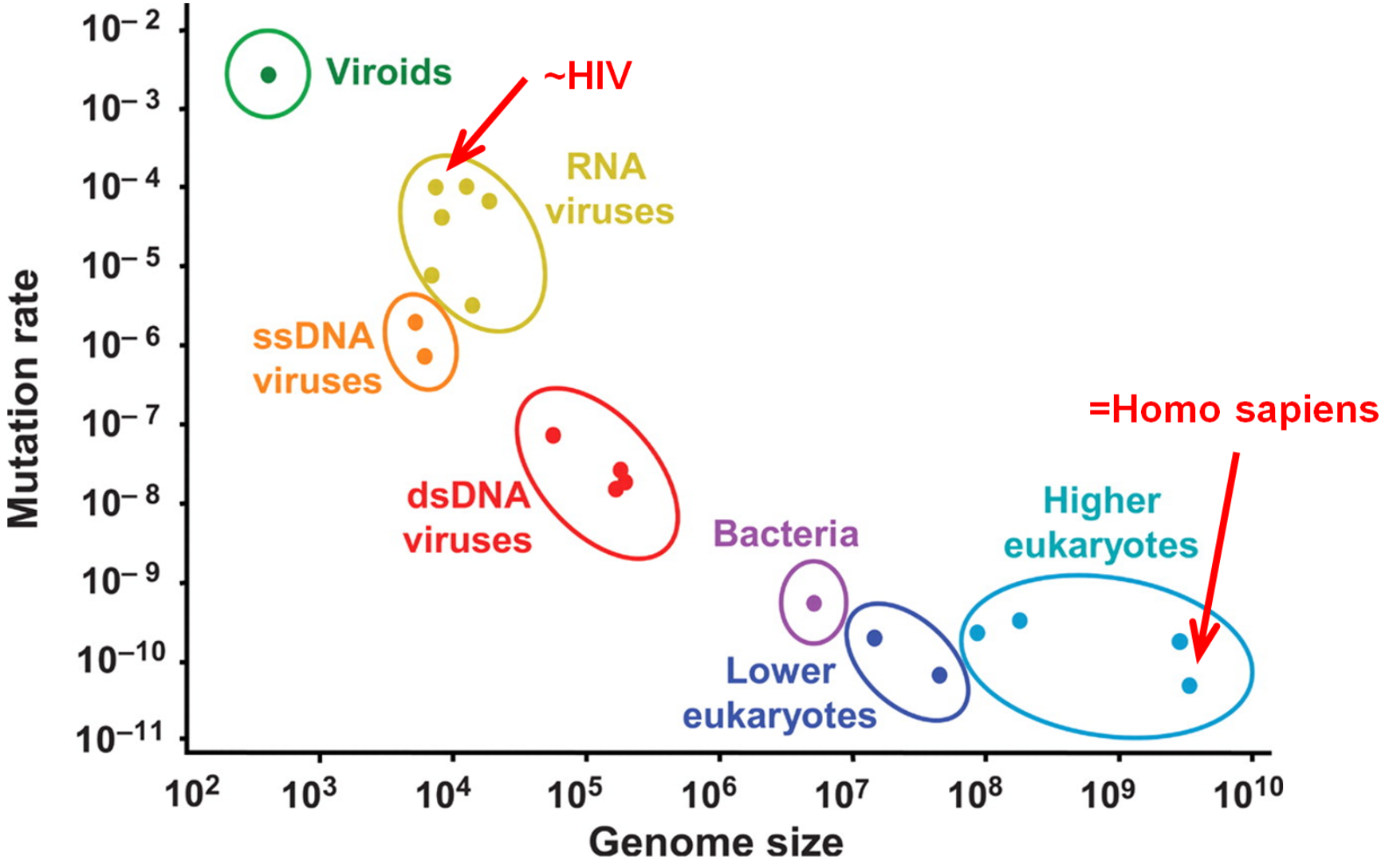
HR Gelderblom *et al.*, Virology 1987

Viral quasispecies



- Quasispecies model of molecular evolution (Eigen & Schuster, 1979)
- Selection pressure on the whole population rather than on single individual: generation of a broad quasispecies with members of approximately equal fitness might be a promising evolutionary strategy.
- Viral quasispecies = viral population = mutant cloud = swarm
  Virus variant = viral haplotype

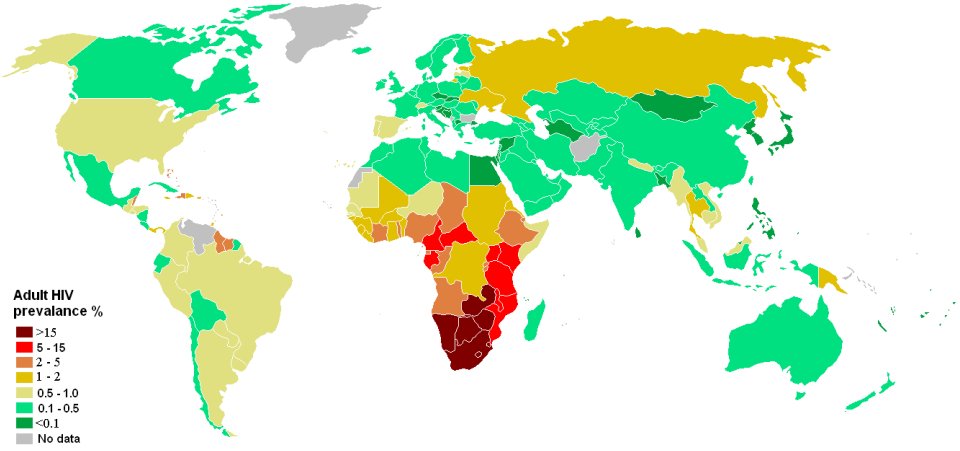
Mutation rate correlates with genome size



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

HIV and AIDS

- Acquired Immunodeficiency Syndrome (AIDS) is caused by Human Immunodeficiency Virus (HIV). More than 35 Million people are affected world-wide.



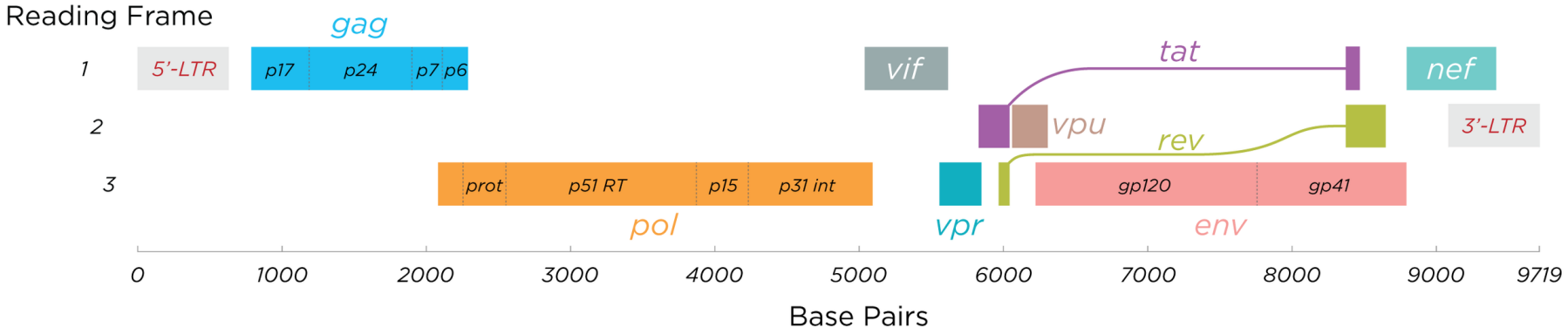
- HIV **quasispecies** are a set of genetically diverse (**haplotypes**).

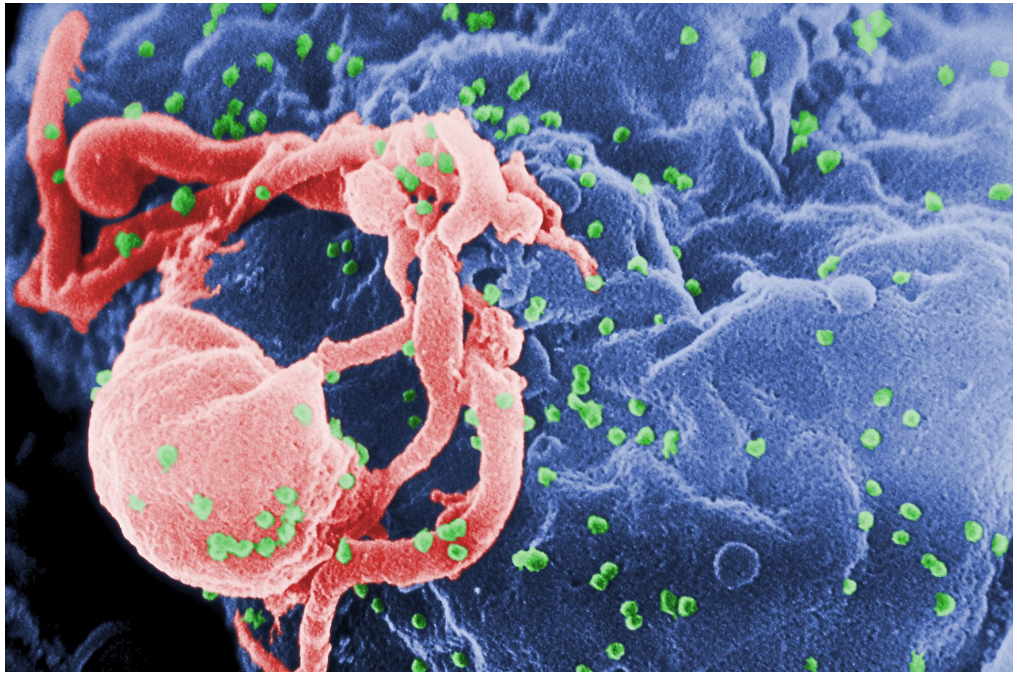
- HIV has **high mutagenicity**
  $\rightsquigarrow$ **evolutionary escape** from the host's immune response
  $\rightsquigarrow$ **drug resistance**.

- Identify this genetic diversity to enable **Personalized Medication**.

HIV-1 genome

Single-stranded RNA genome, organized in a highly sophisticated way

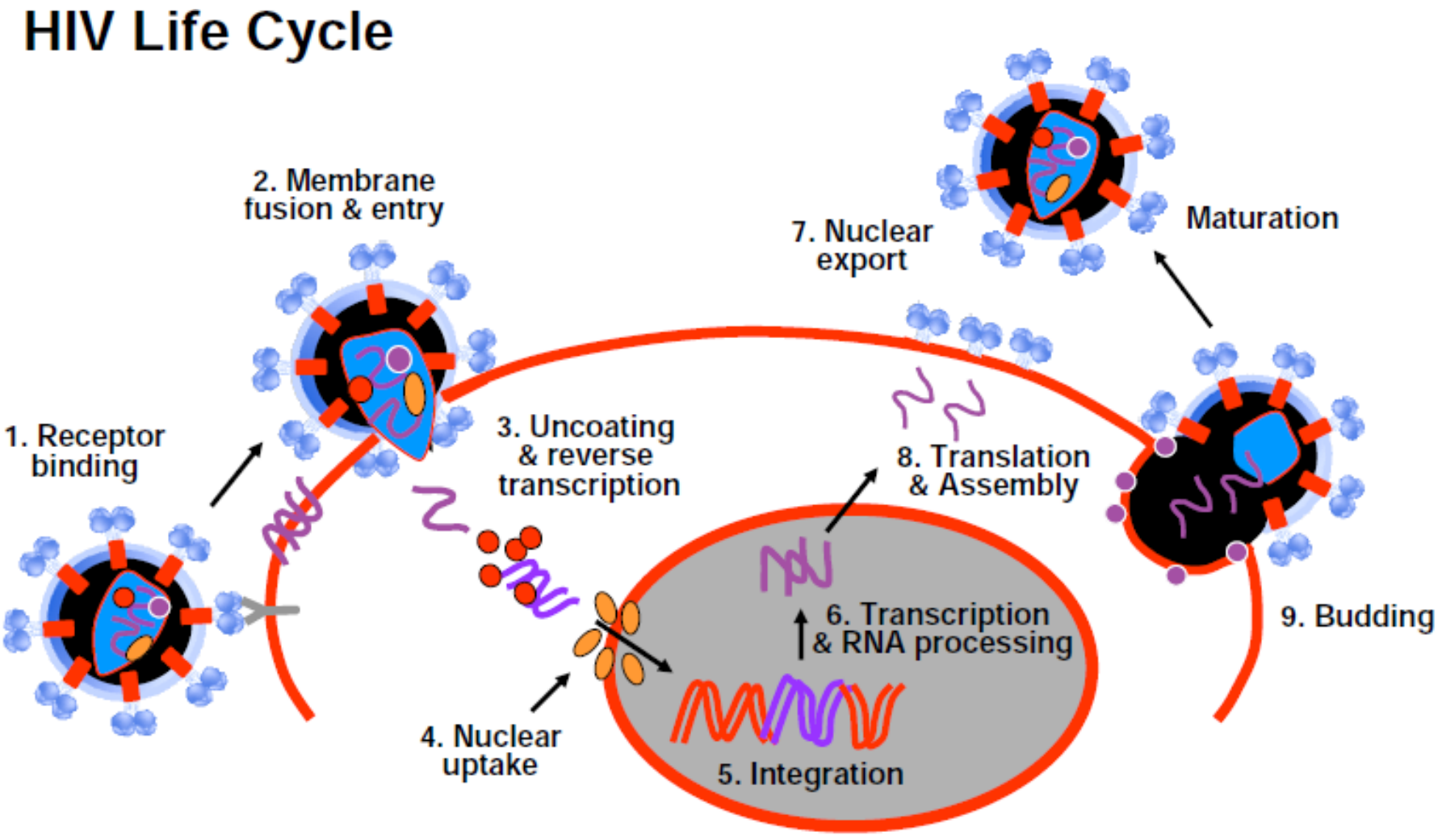




Wikipedia

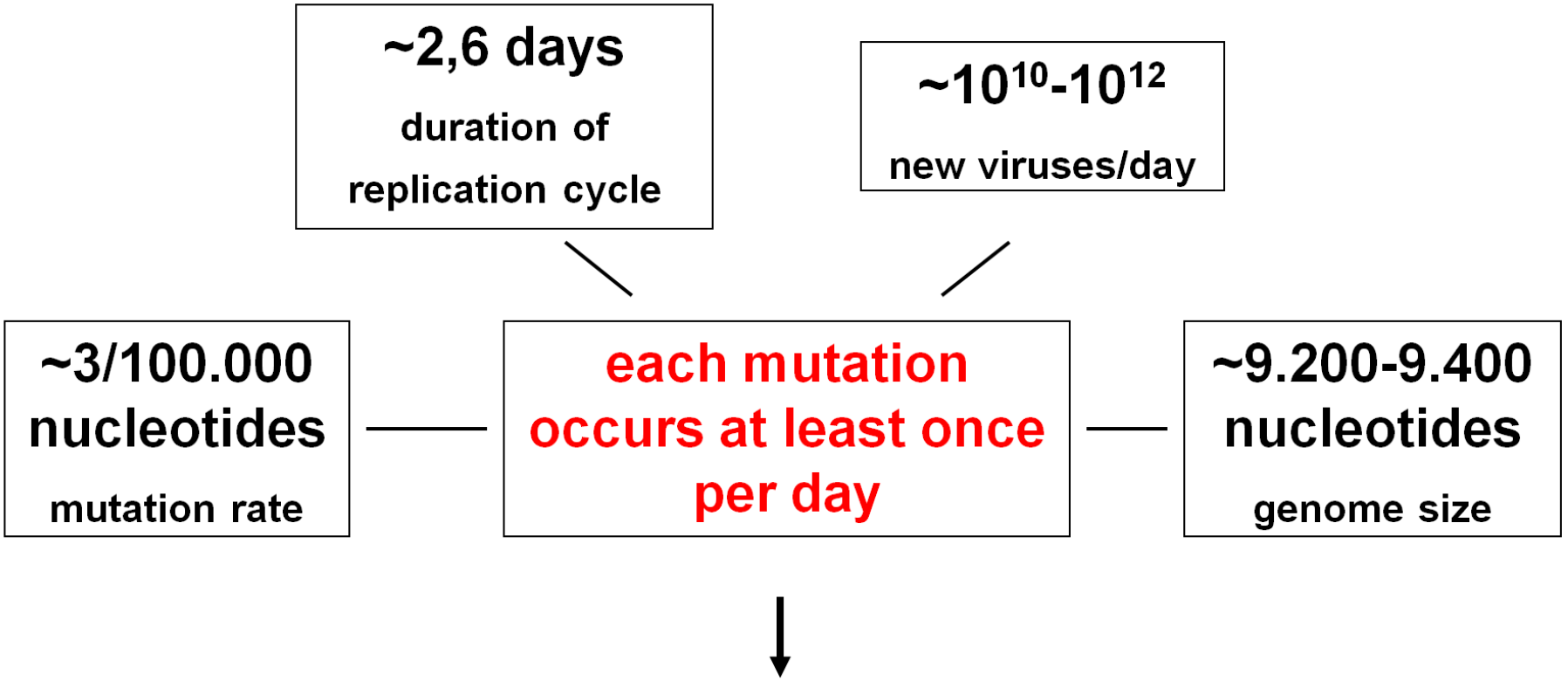
Scanning electron micrograph of HIV-1 (in green) budding from cultured lymphocyte.

The HIV Life Cycle

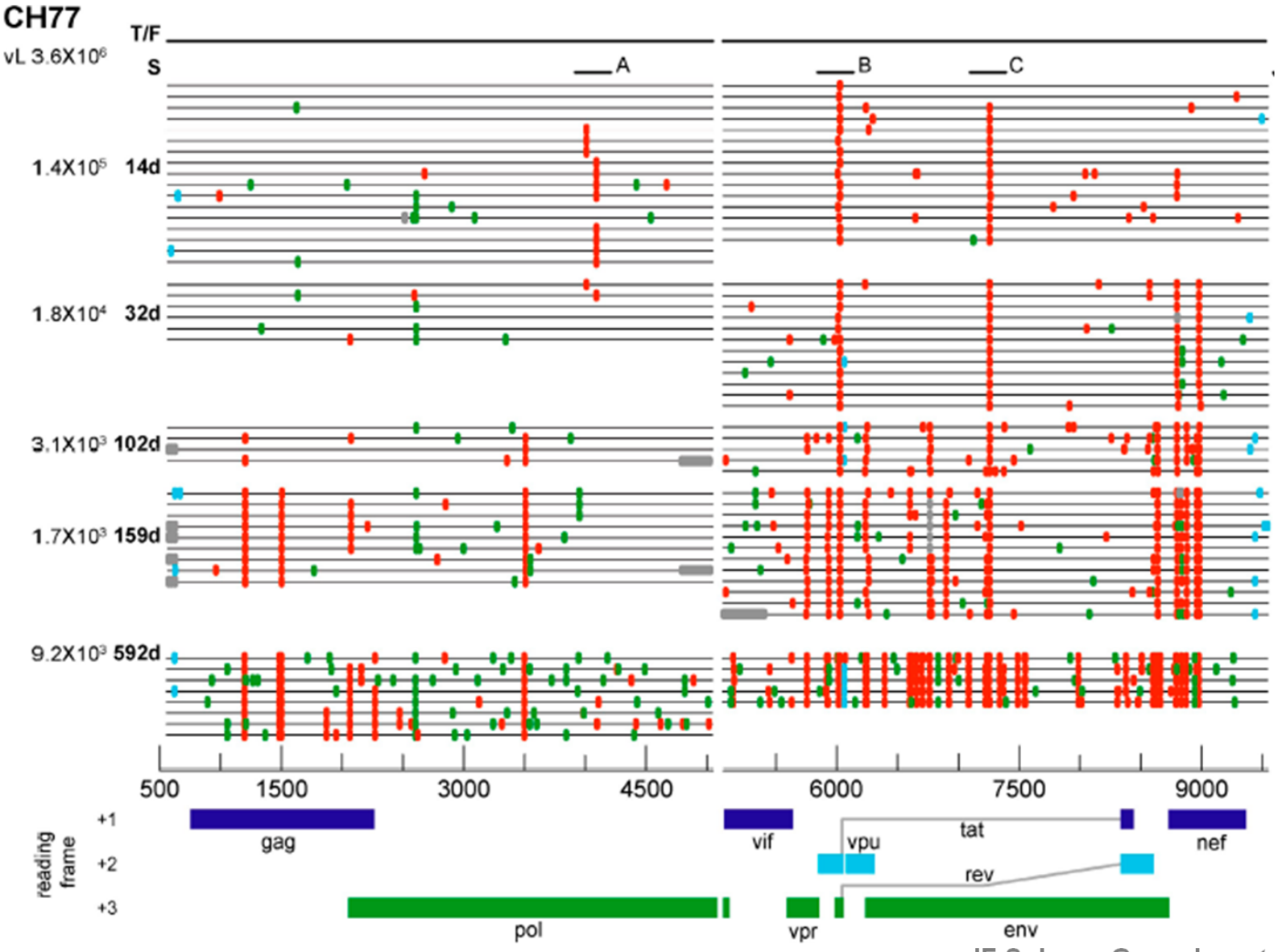


HIV "hijacks" T-helper cells $\rightsquigarrow$ reverse transcription $\rightsquigarrow$ integration in nucleus $\rightsquigarrow$ transcription $\rightsquigarrow$ translation & assembly $\rightsquigarrow$ new virus pushes out ("buds").

Evolution and diversity of HIV

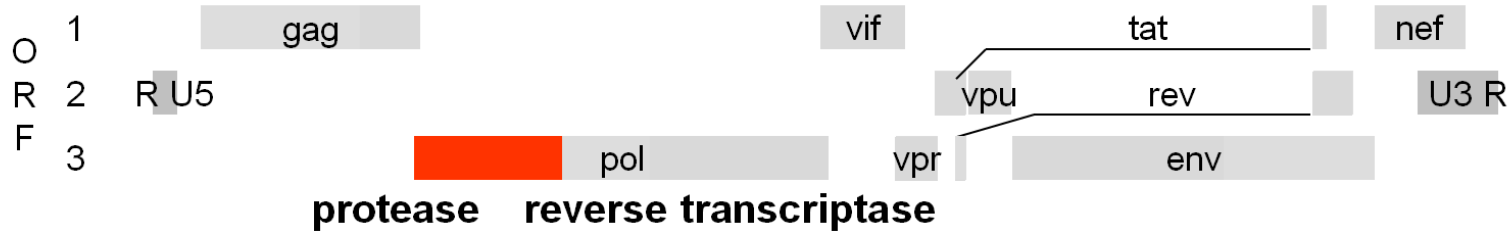


HIV-1 population dynamics within a host



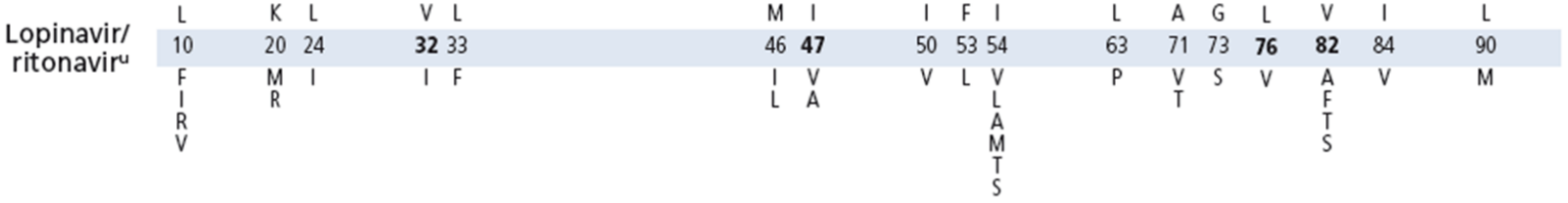
JF Salazar-Gonzalez *et al.*, JEM 2009

HIV drug resistance



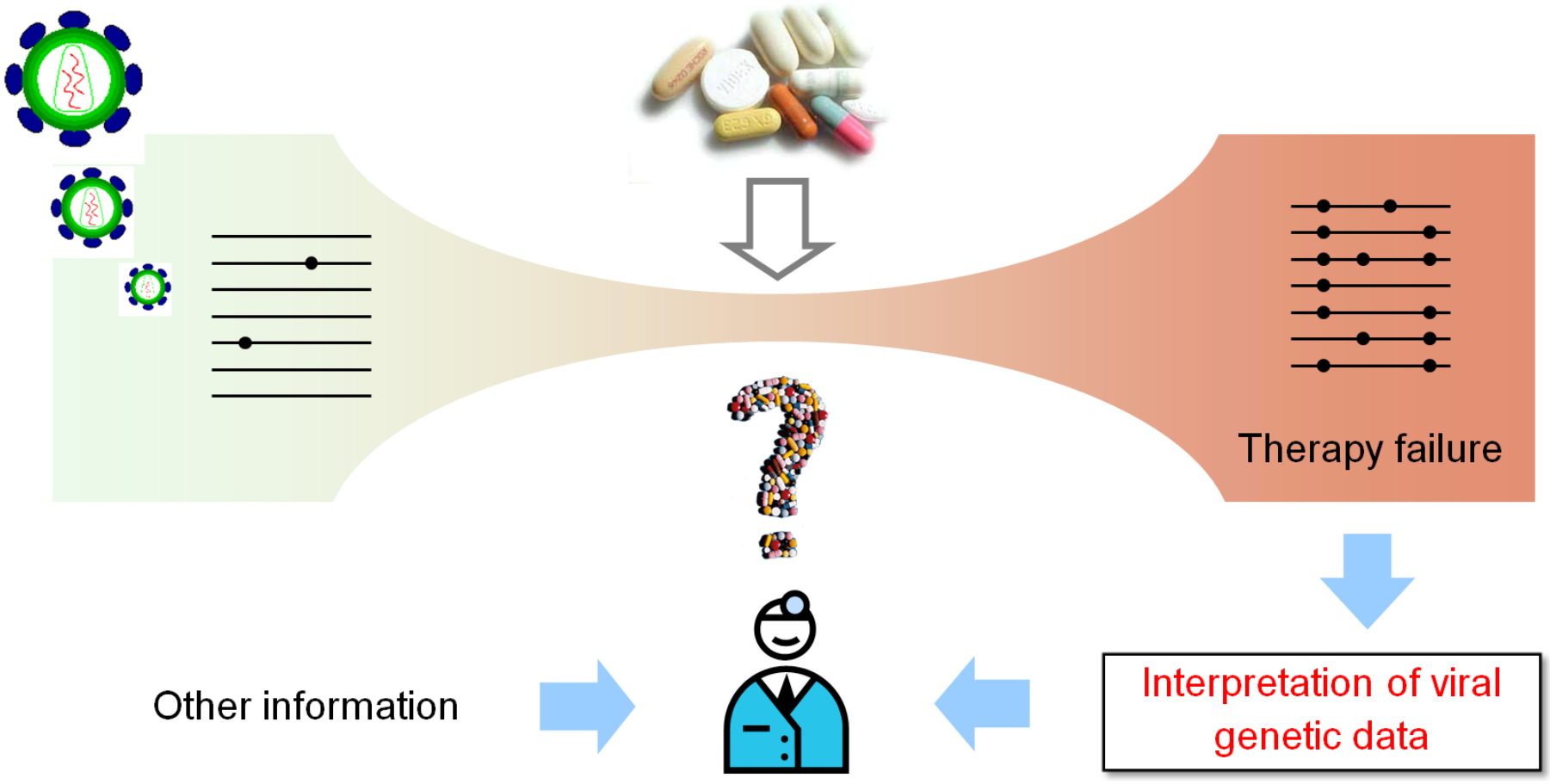
RT

PR



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

HIV drug resistance, individualized treatment



HIV evolution is **highly responsive to selection pressure** from drugs that are not fully suppressive

$\rightsquigarrow$ **rapid outgrowth of resistant variants**.

DNA Sequencing: Shotgun sequencing

Shotgun sequencing: Sequencing DNA strands by way of randomized fragmentation (analogy with quasi-random firing pattern of a shotgun).

Due to technical limitations, classical DNA sequencing (a.k.a. **Sanger sequencing**, starting in the 1980s) can only be used for short DNA strands of < 1000 base pairs.

Longer sequences are subdivided into smaller fragments ("reads")

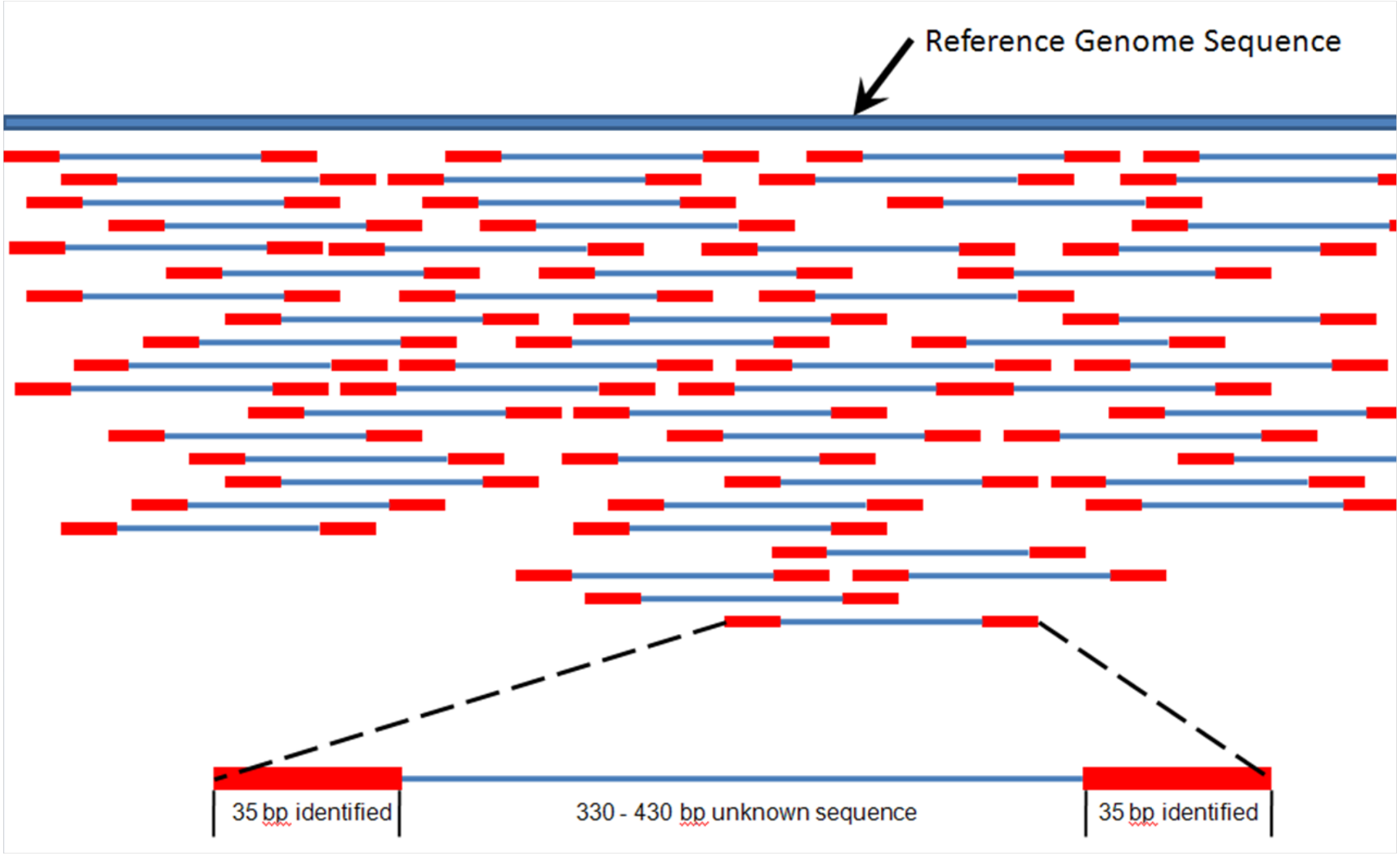
$\rightsquigarrow$ sequenced separately $\rightsquigarrow$ read assembly gives overall sequence.

```
Strand                    Sequence
Original                  AGCATGCTGCAGTCATGCTTAGGCTA
First shotgun sequence    AGCATGCTGCAGTCATGCT-------
                          ------------------TAGGCTA
Second shotgun sequence   AGCATG--------------------
                          ------CTGCAGTCATGCTTAGGCTA
Reconstruction            AGCATGCTGCAGTCATGCTTAGGCTA
```

DNA Sequencing: Shotgun sequencing

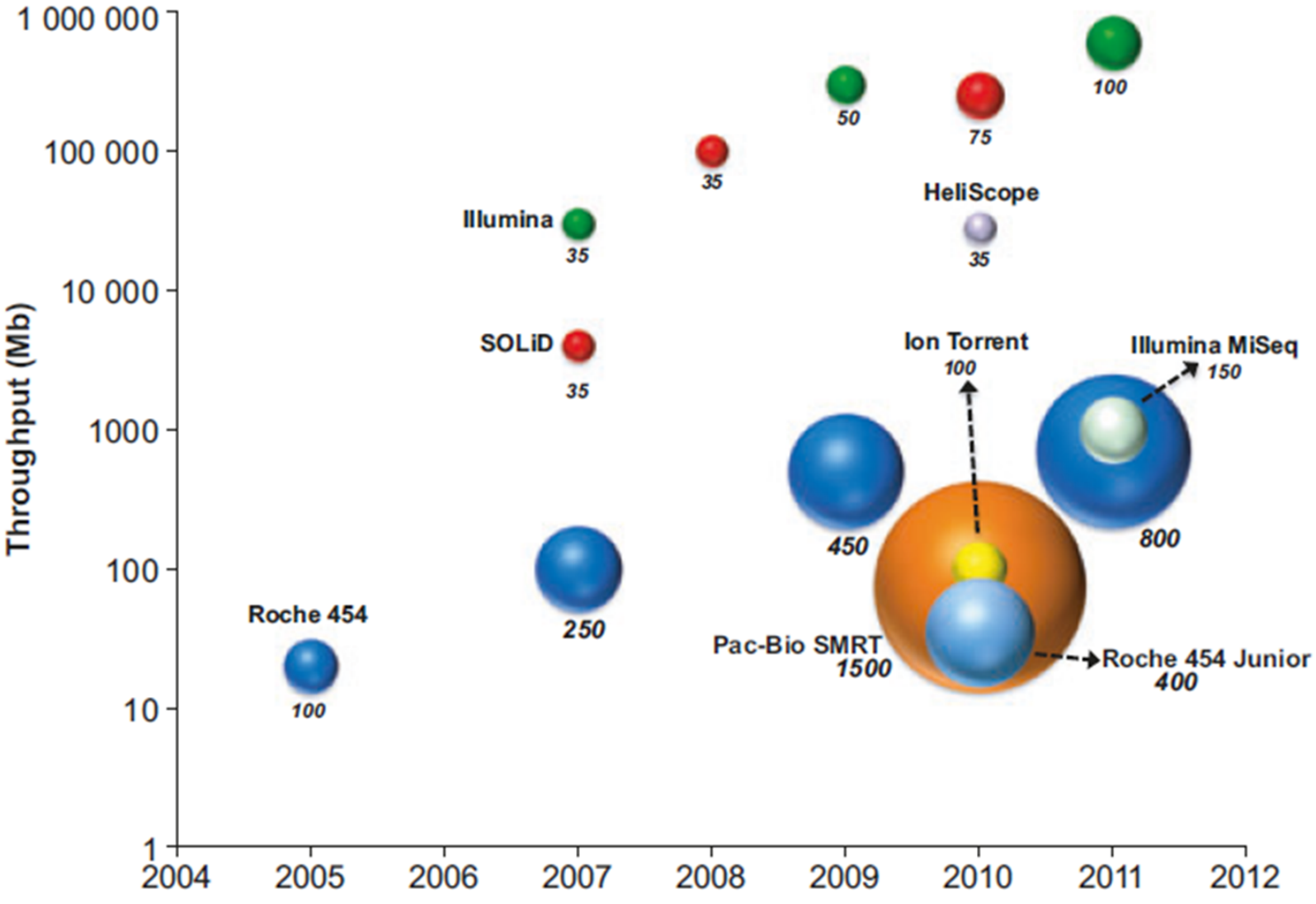
- Shotgun strategy still applied today, but using other sequencing technologies, such as **short-read** and **long-read sequencing.**
- Short-read or **next-generation** sequencing produces shorter reads (anywhere from 25-500 bp) but many hundreds of thousands or millions of reads in a relatively short time (on the order of a day).
- Vastly superior to Sanger sequencing: Fast, cheap...
  ...but for some applications, sequencing errors cannot be neglected, and the assembly of short and "noisy" reads can be difficut (computationally and statistically).
- **Third-generation sequencing:** Long reads, but usually higher error rate and much lower throughput.

Mapping short reads to reference genome



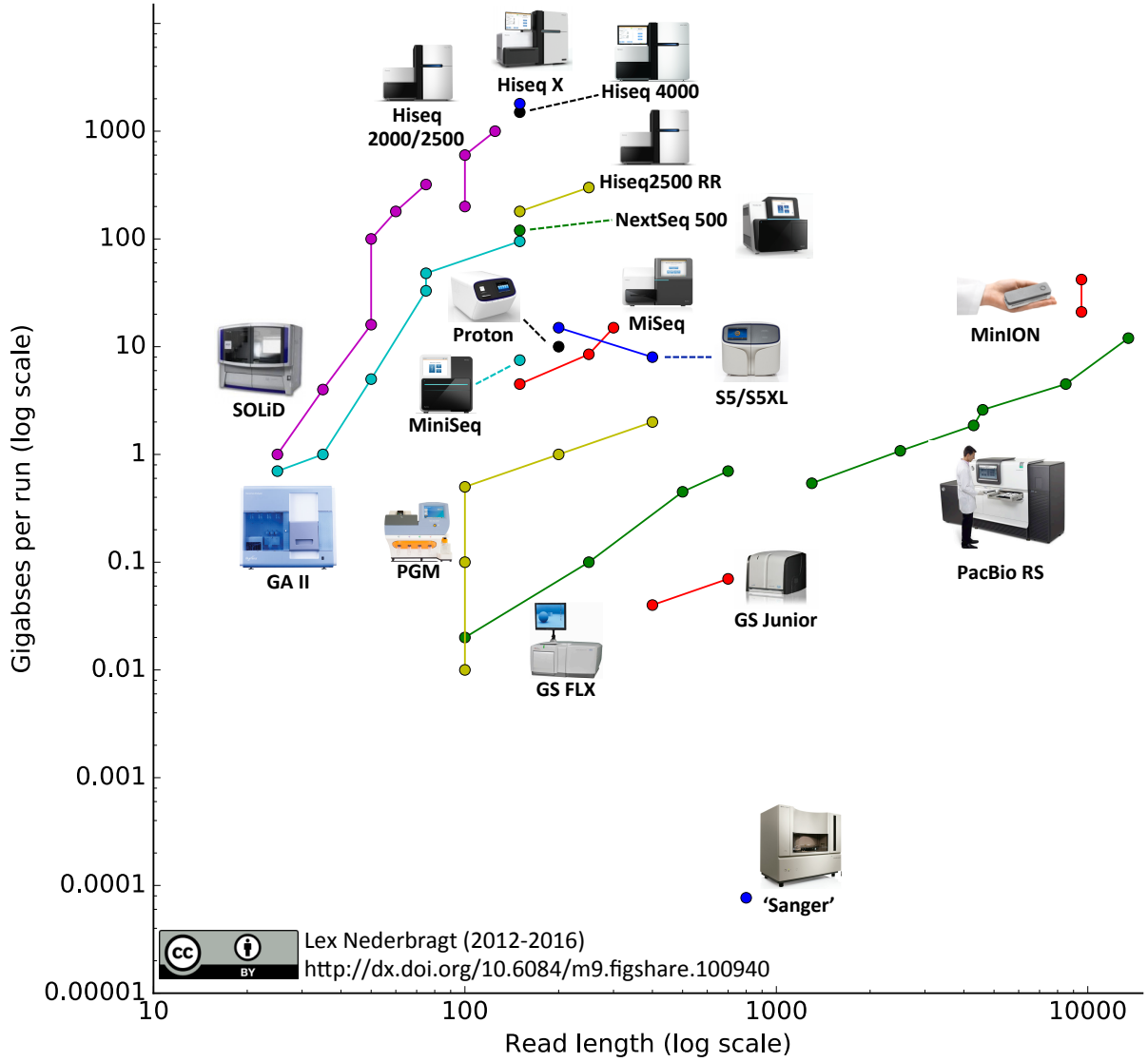
By Suspencewl - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=13764860

NGS technologies: Read length vs throughput



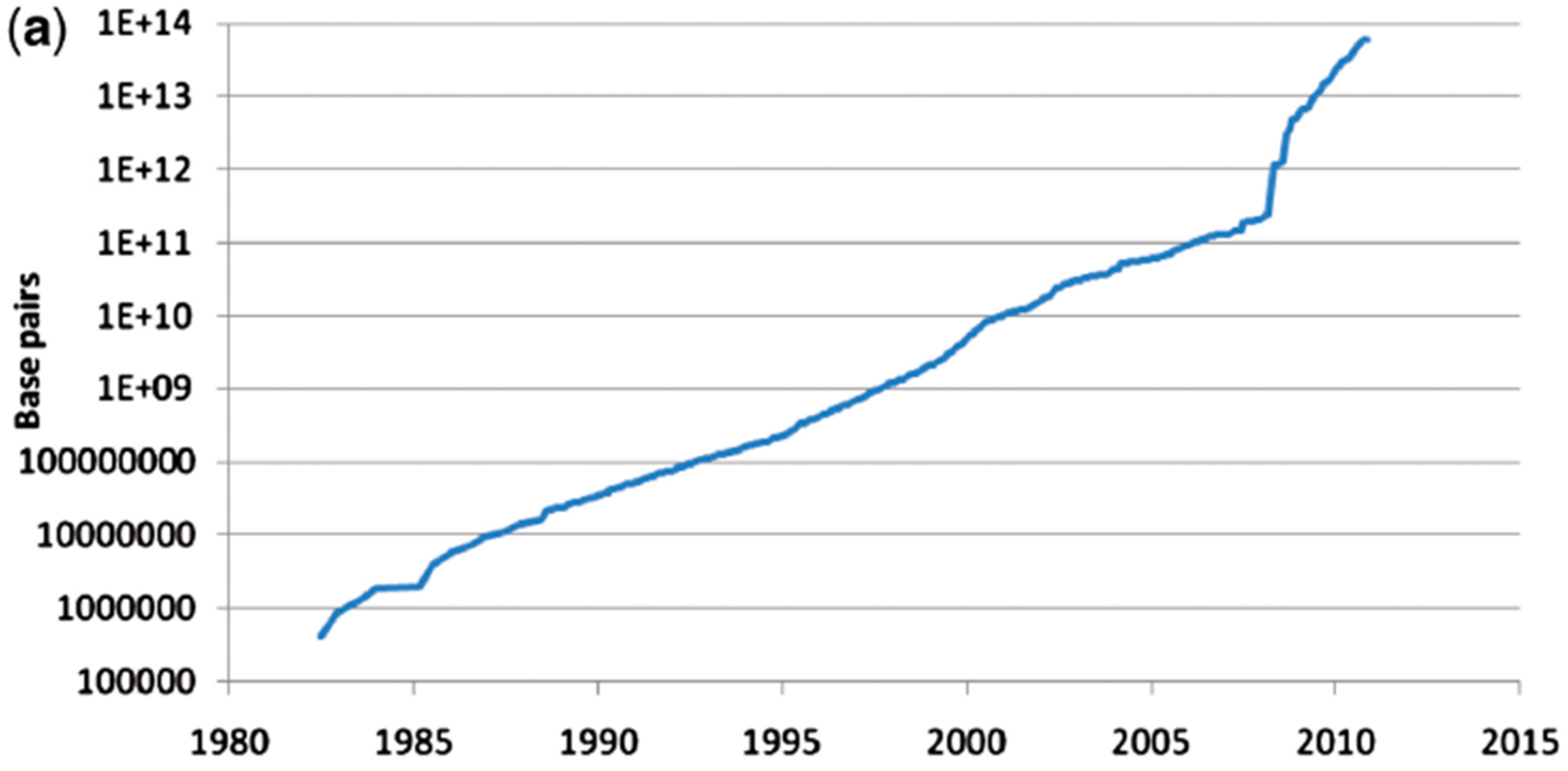
S Shokralla *et al.*, Mol Ecol 2012

NGS technologies: Read length vs throughput



Nederbragt, Lex (2016): developments in NGS. figshare. Dataset. https://doi.org/10.6084/m9.figshare.100940.v9

Cumulative data volume in base pairs over time



The International Nucleotide Sequence Database Collaboration, NAR 2010

Cumulative data volume in base pairs over time

Nucleic Acids Research, 2016, Vol. 44, Database issue ***D49***

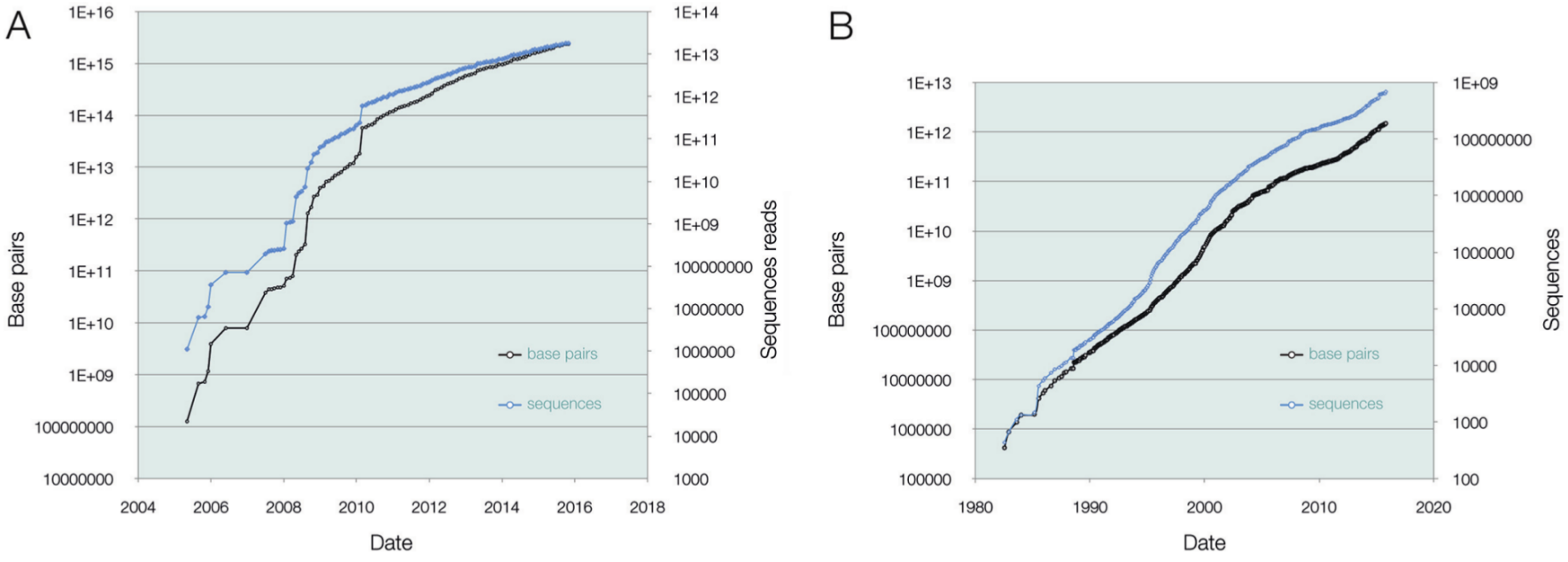
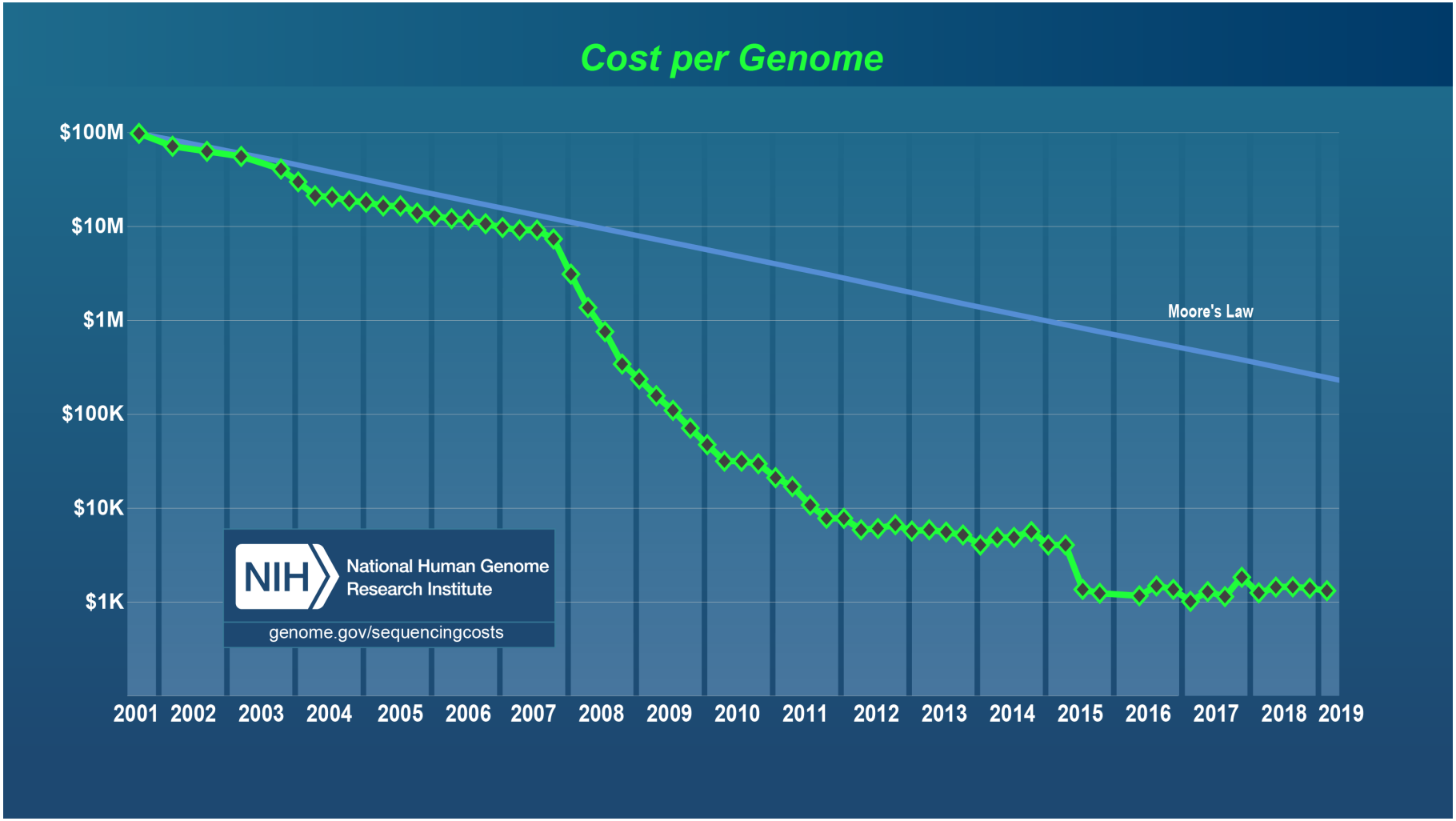


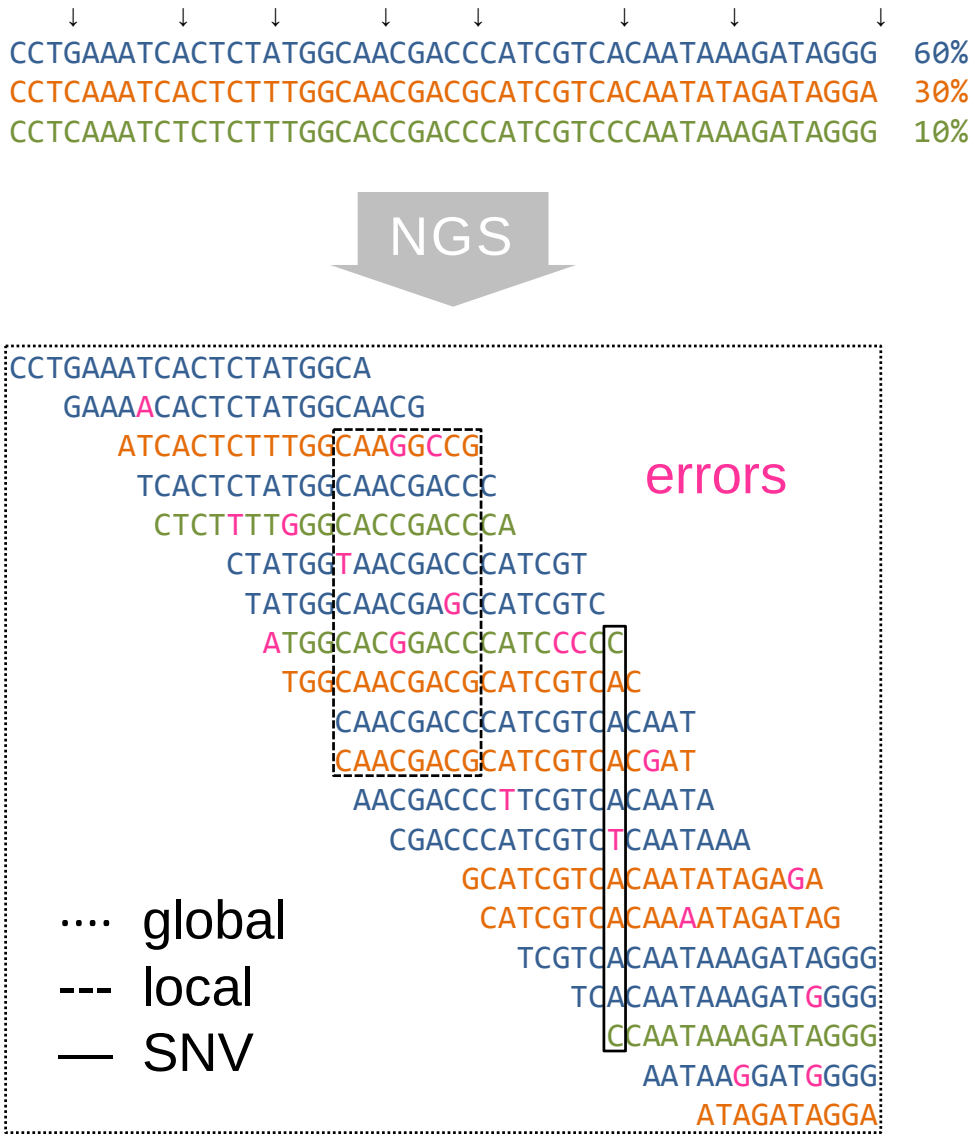
Figure 1. Cumulative growth in INSDC. (**A**) Base pairs (black, 2365.5 trillion) and sequence reads (blue, 17.8 trillion) for INSDC raw data. (**B**) Base pairs (black 1449 billion) and sequences (blue, 651.5 million) in INSDC assembled/annotated data.

Cochrane, Karsch-Mizrachi, Takagi, Nucleic Acids Research, 2016, Vol. 44, Database issue doi: 10.1093/nar/gkv1323

Costs per genome over time

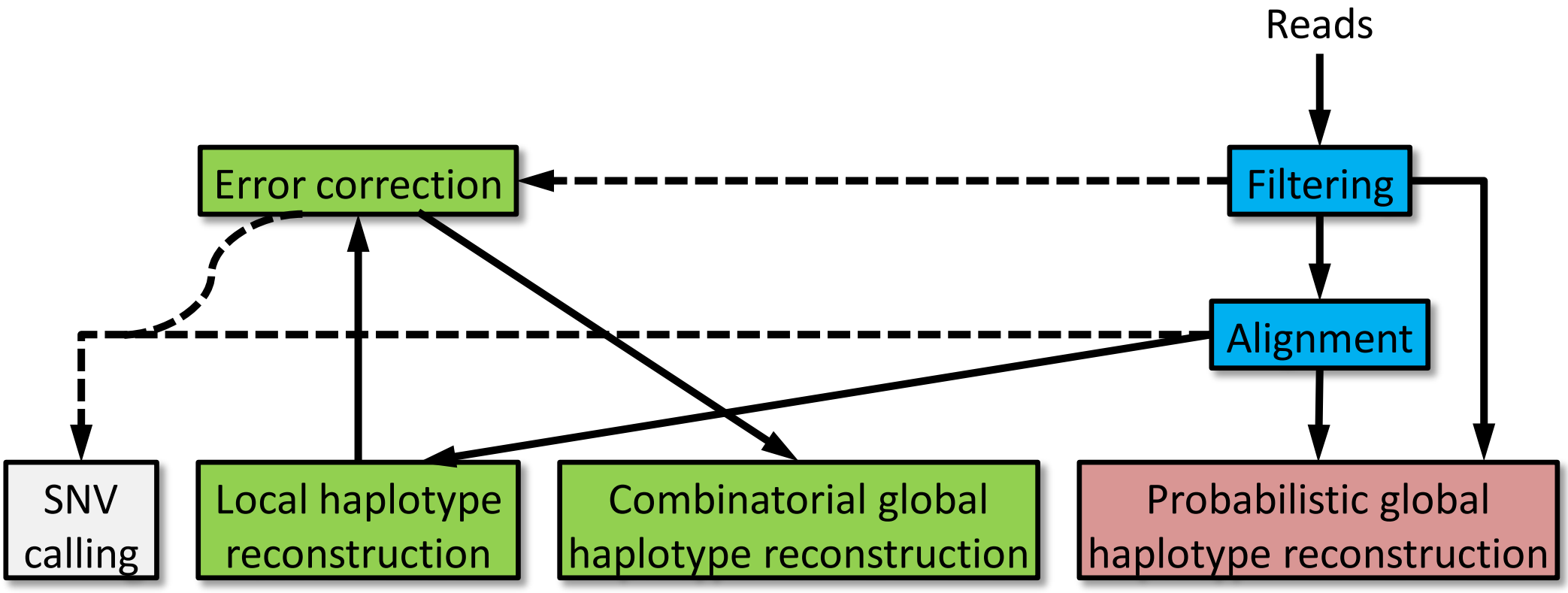


Global Haplotype Assembly



Beerenwinkel et al (Front Microbiol 2012)

Overview


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

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.**

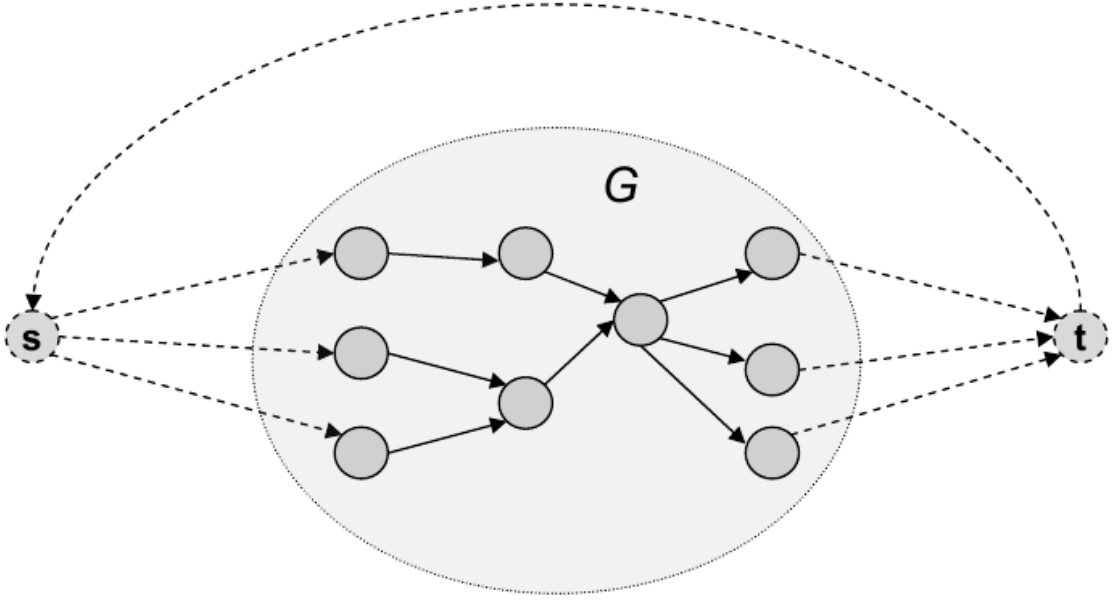
Probabilistic assembly

- 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.**

Combinatorial Assembly: Network Flow

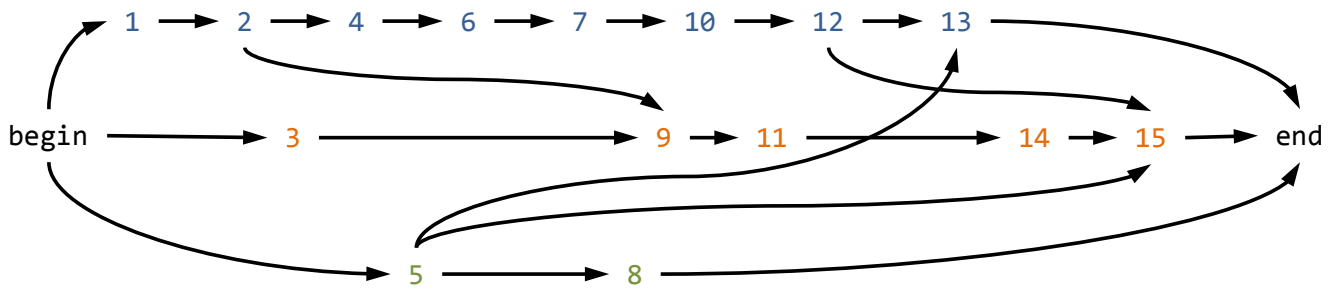


Westbrooks et al (ISBRA 2008)

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
```

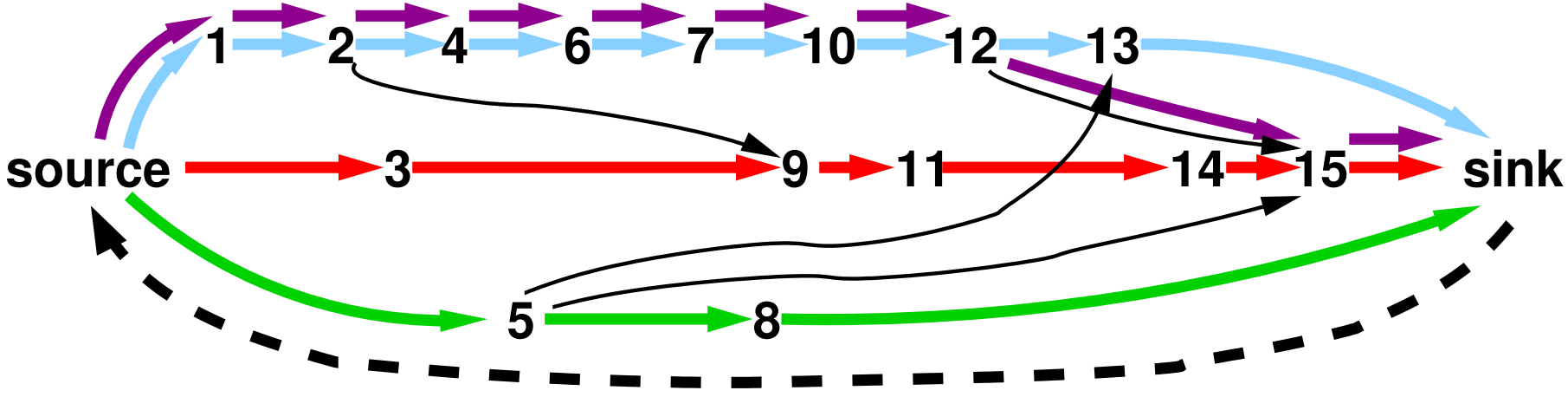


Each path is a potential haplotype

Beerenwinkel et al (Front Microbiol 2012)

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.



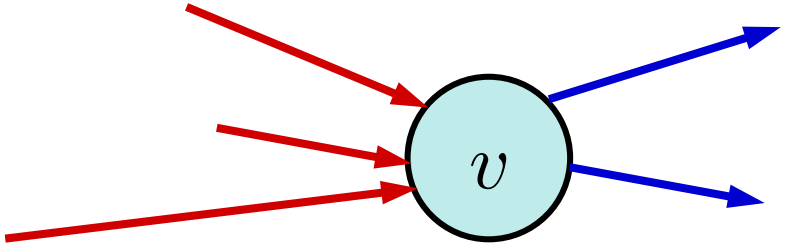
Quasispecies assembly via network flows

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:**

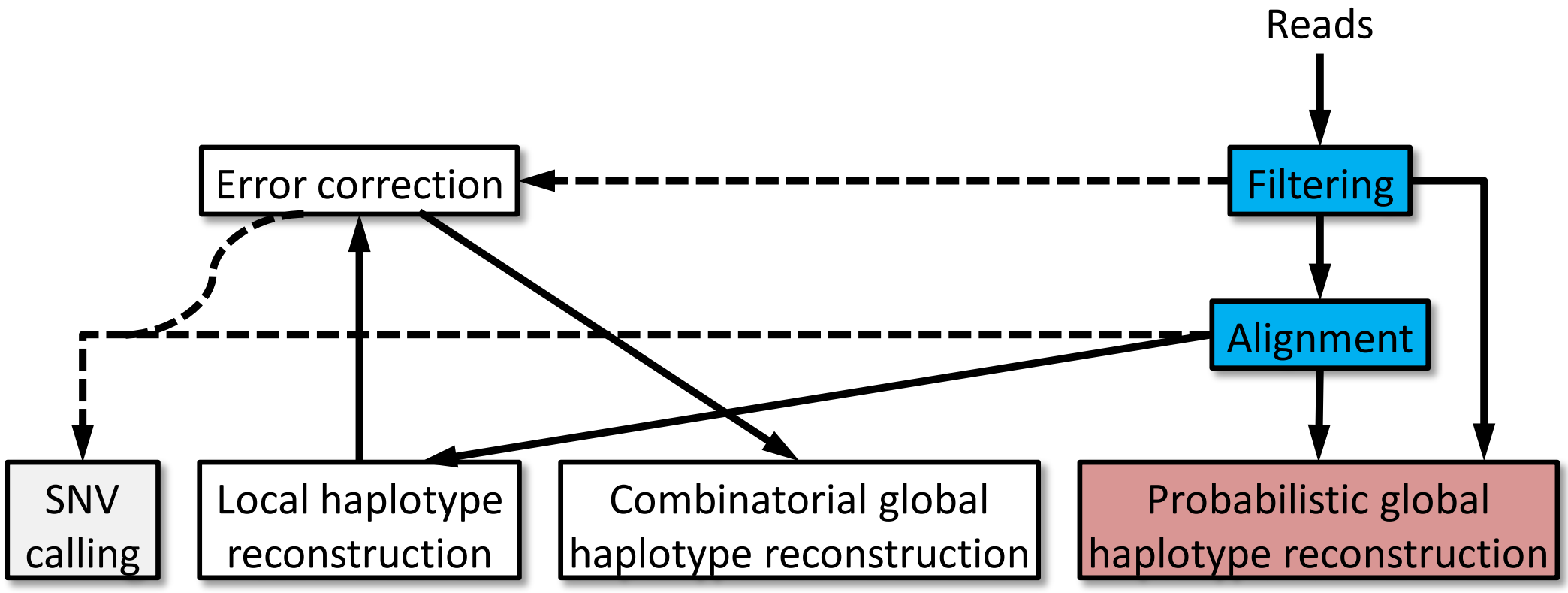


$$\sum_{e_{\text{in}}} f(e) = \sum_{e_{\text{out}}} f(e)$$

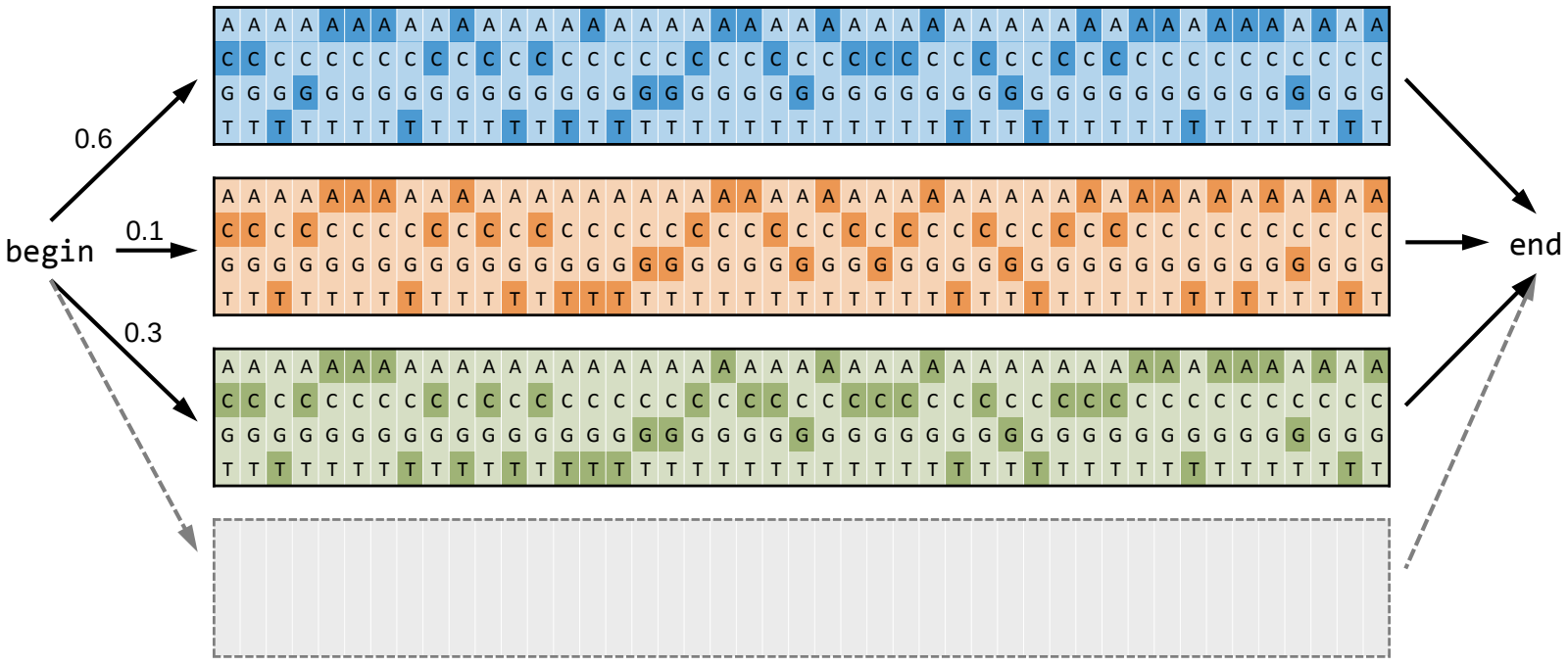
- **Each read covered by at least one haplotype**

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

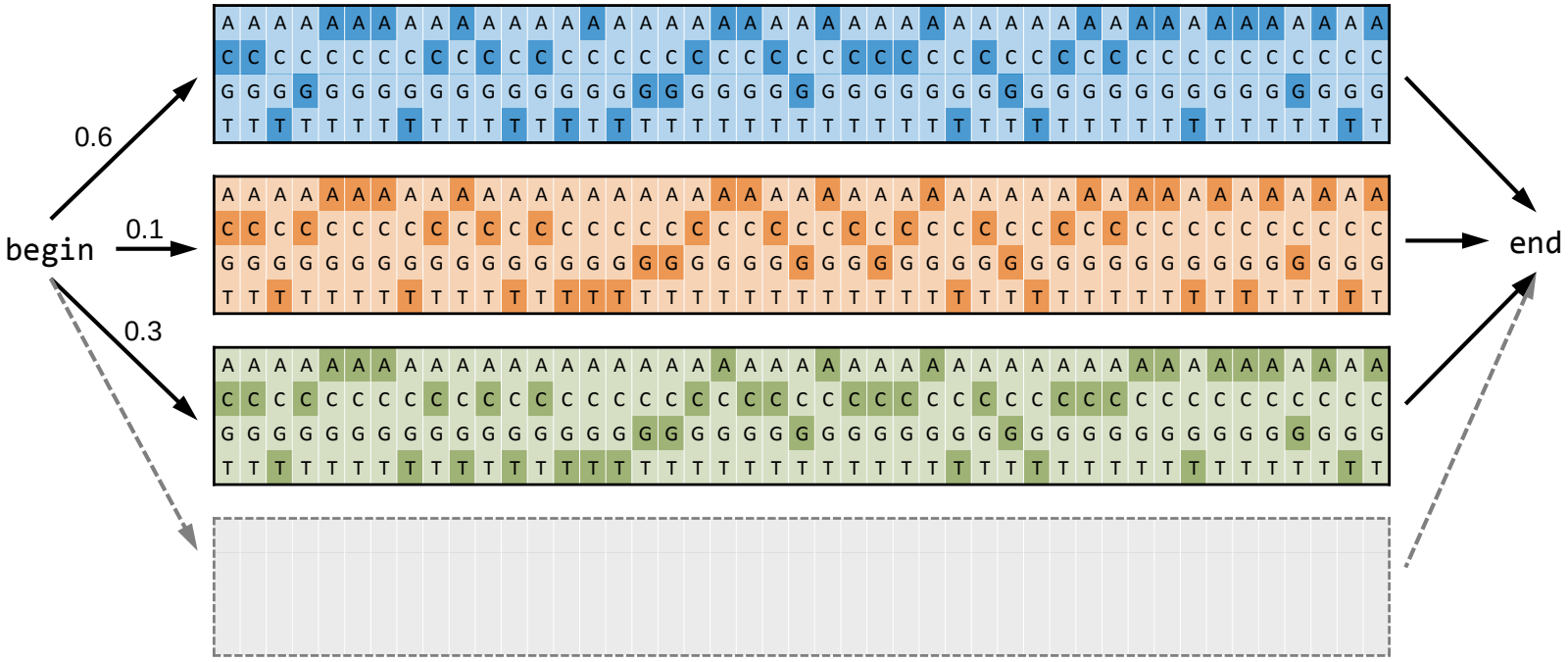
Overview


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

Probabilistic Assembly: Mixture Model



PredictHaplo: Generative Model for Reads

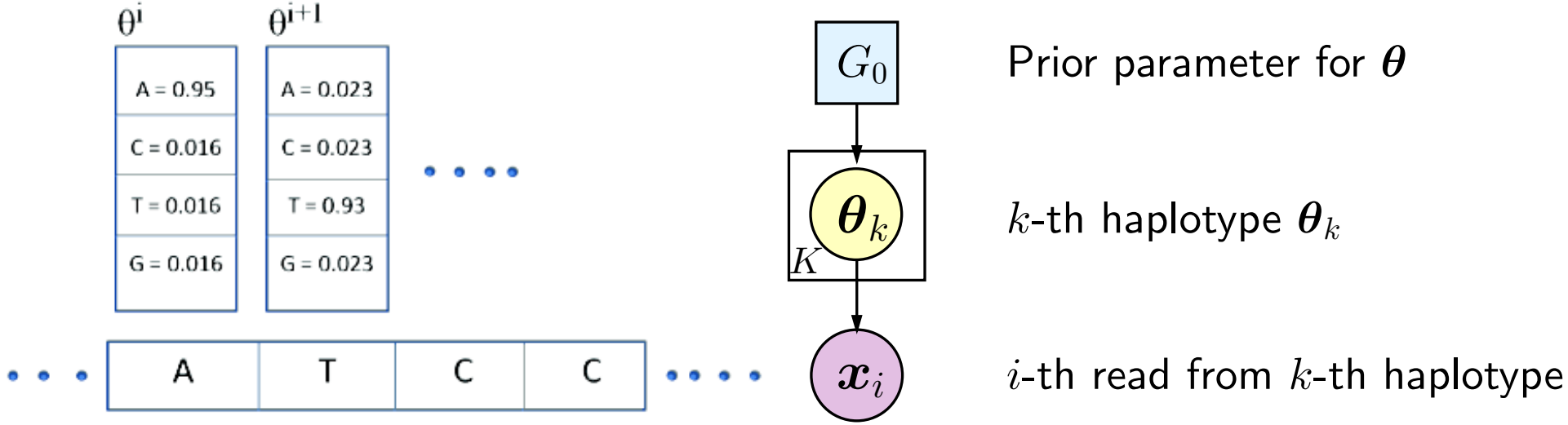


Bayesian mixture model: assign **priors** on class proportions and component distributions, integrate out latent variables.

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

Component Distributions

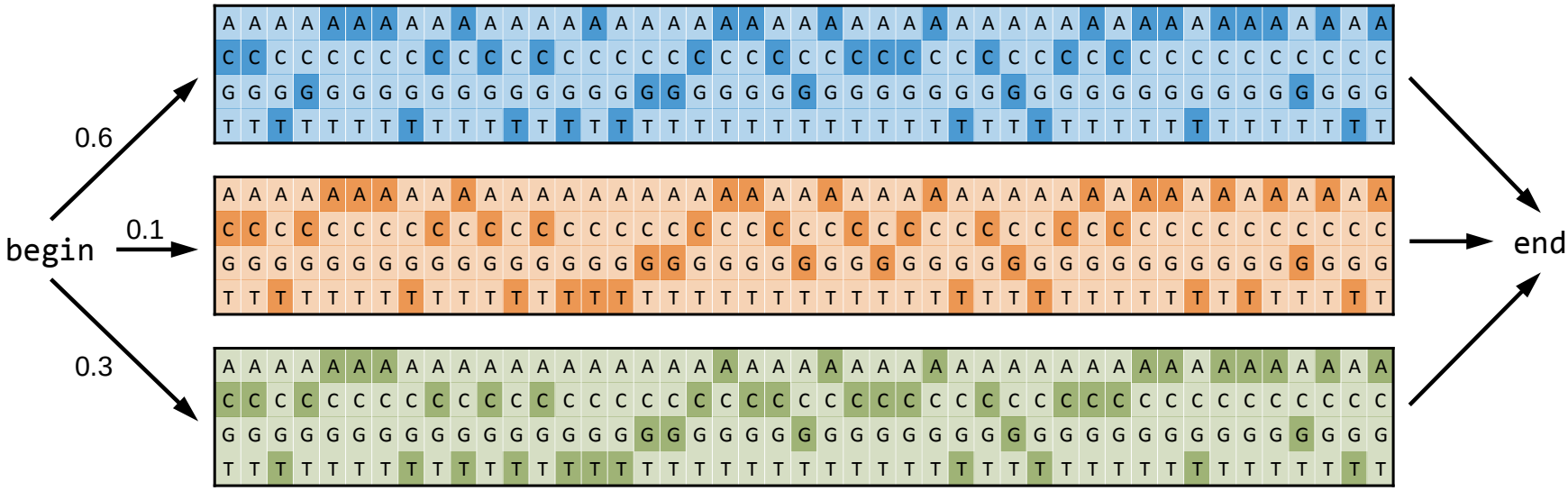
Haplotype: **position-wise multinomial probability tables θ**:

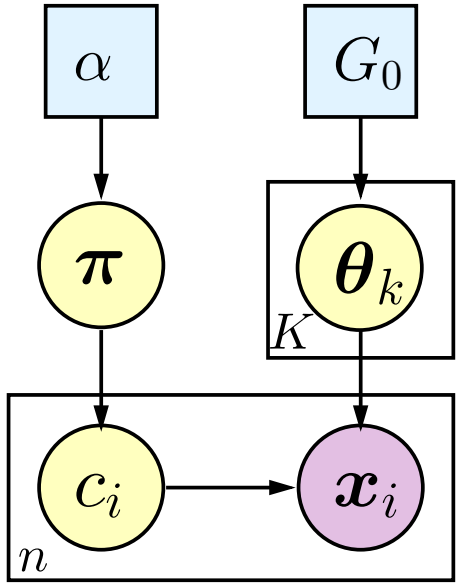


- Parameters for k-th haplotype: $\boldsymbol{\theta}_k = (\theta_k^1, \ldots, \theta_k^L)$
- Position-wise independence assumption: i-th read $\boldsymbol{x}_i$ ranging from position a to b drawn from k-th haplotype:

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

Finite Mixture of Haplotypes





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

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

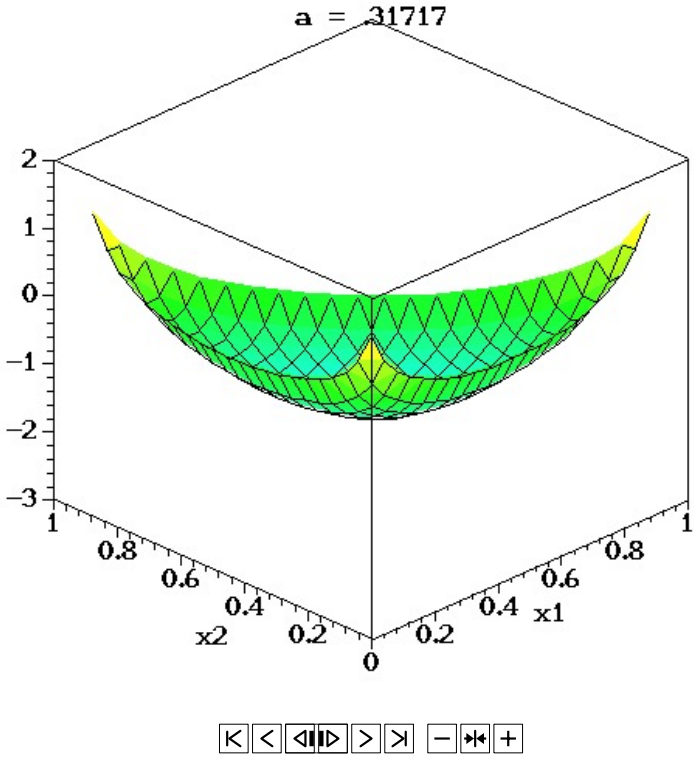
$$c_i \sim \text{Mult}(\boldsymbol{\pi})$$

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

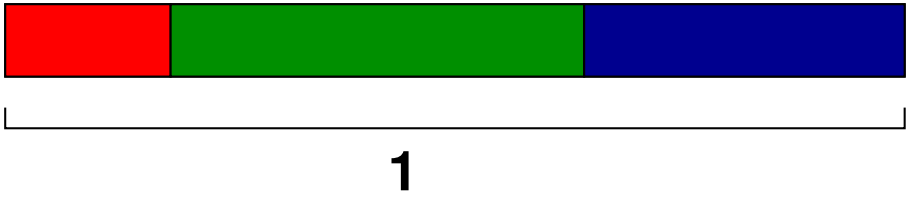
Beerenwinkel et al (Front Microbiol 2012)

Dirichlet Priors for Mixture Proportions

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



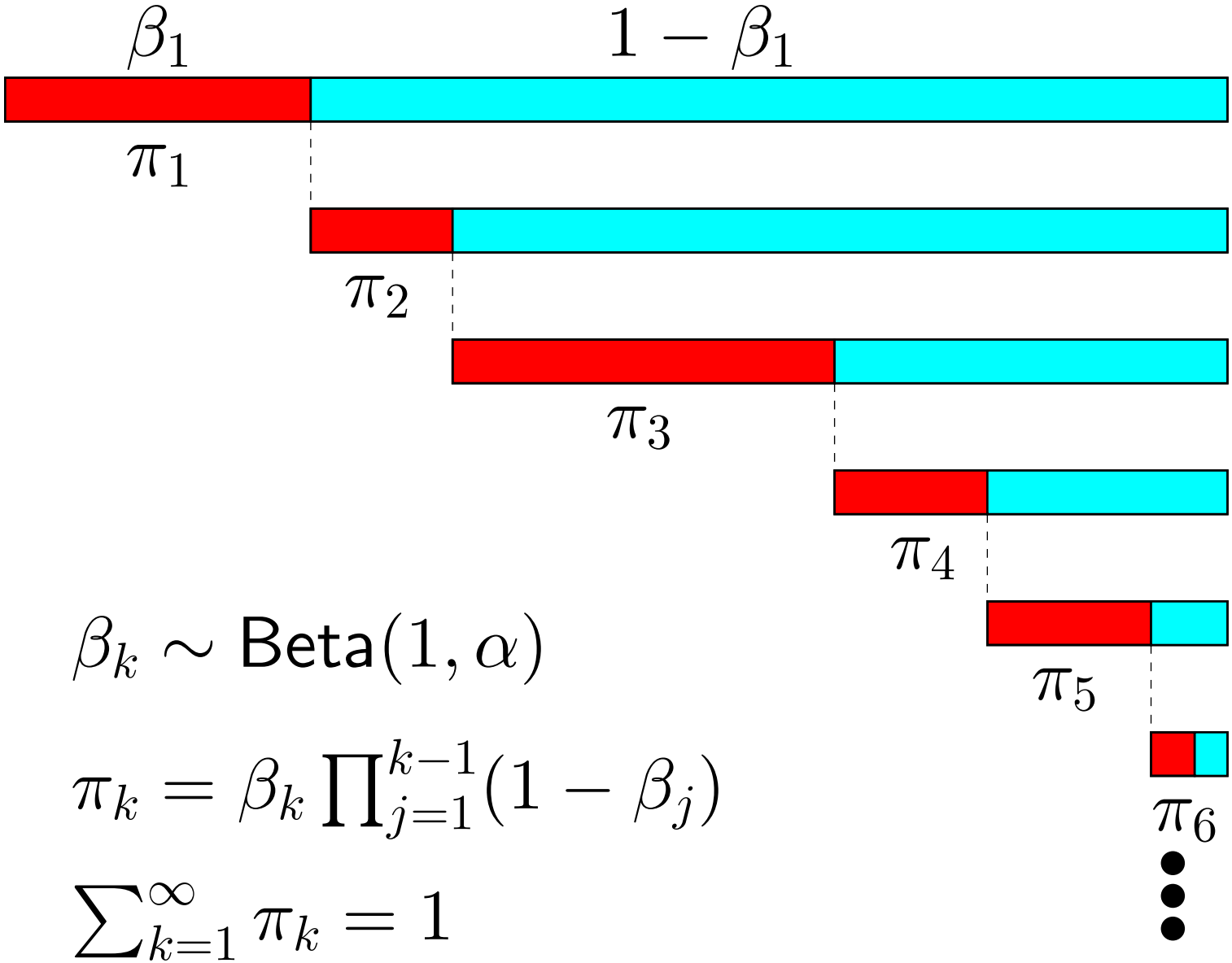
Interpretation: breaking a stick of length 1 into $K = 3$ parts



- **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

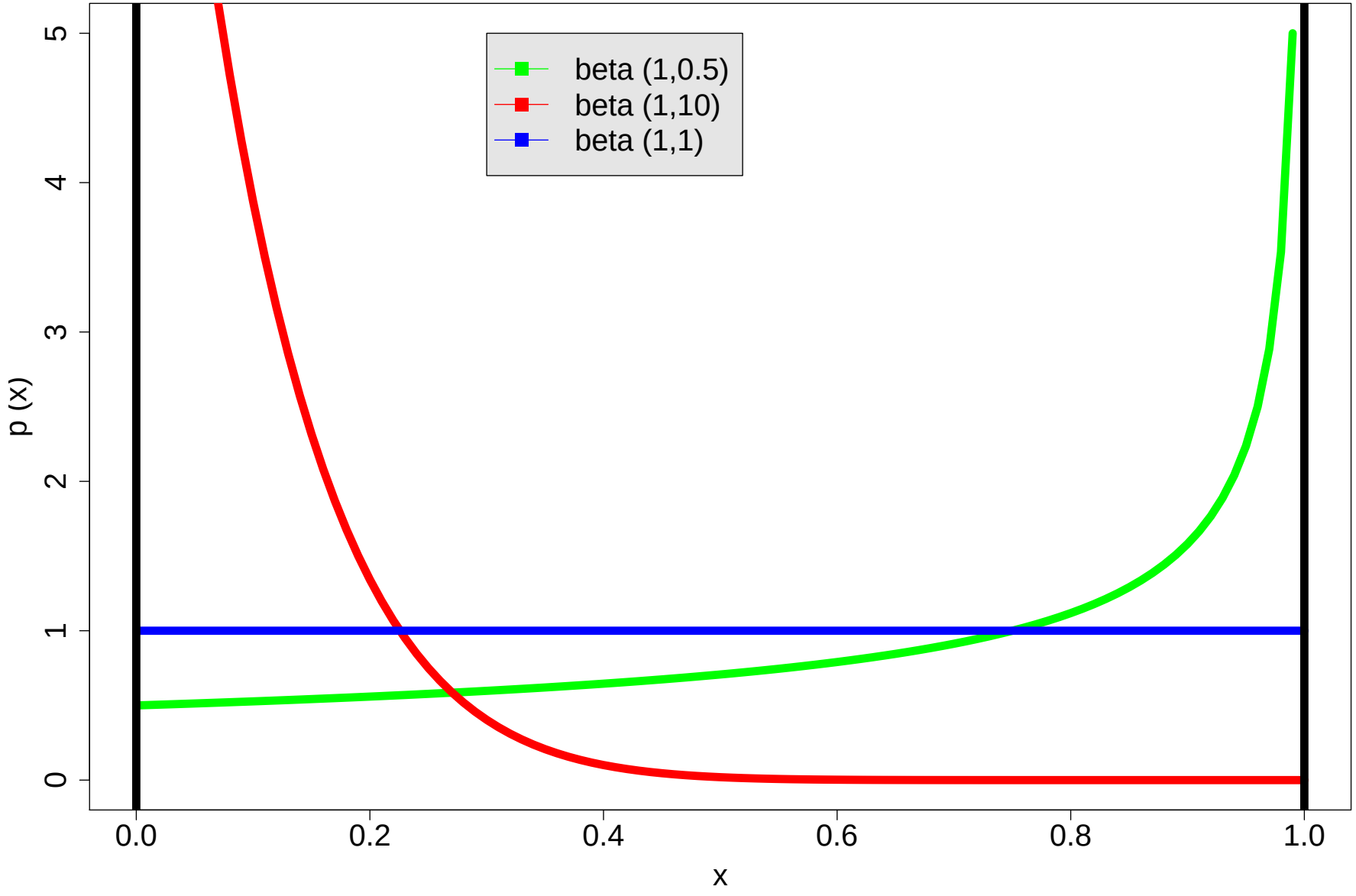


$$\beta_k \sim \mathsf{Beta}(1, \alpha)$$

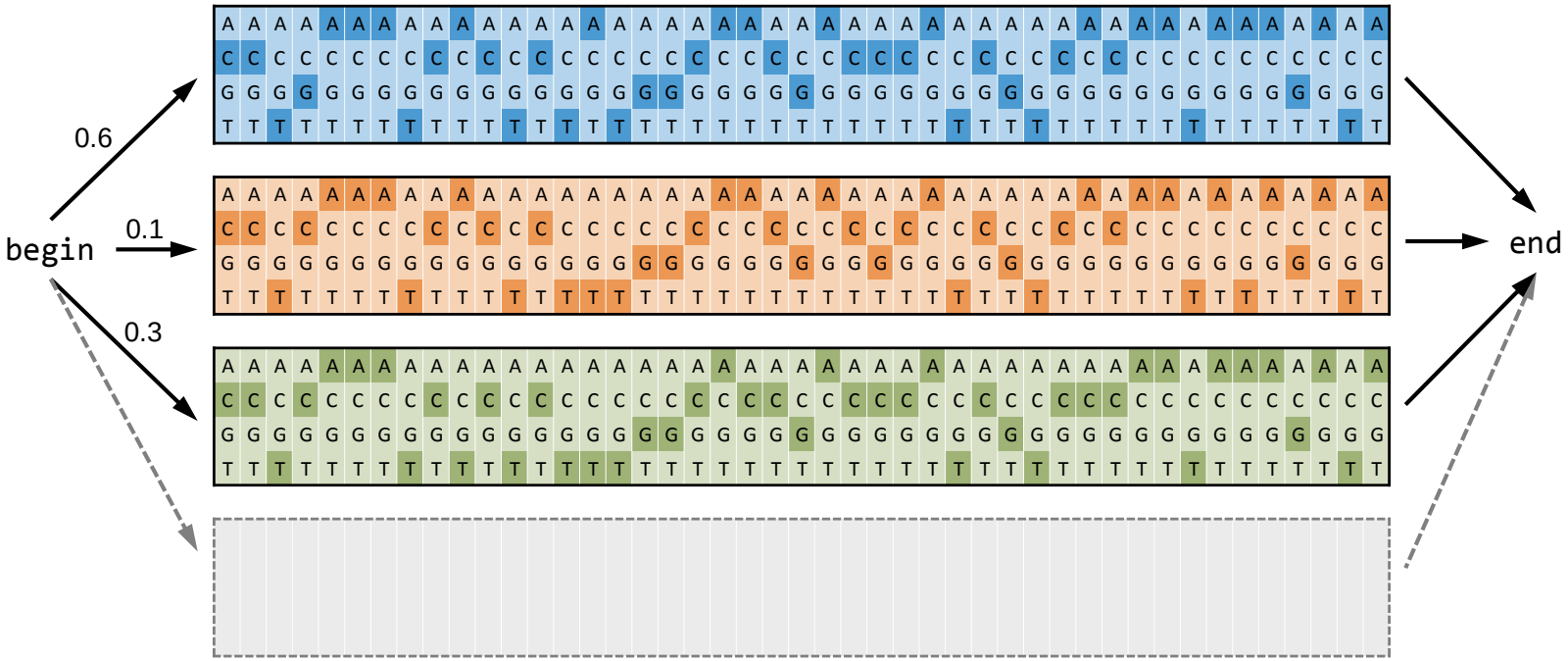
$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j)$$

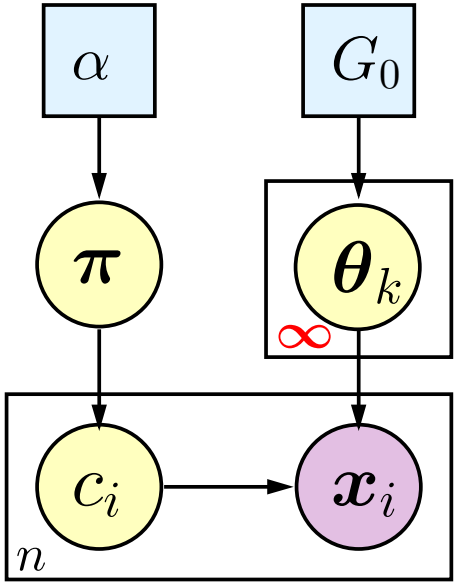
$$\sum_{k=1}^{\infty} \pi_k = 1$$

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

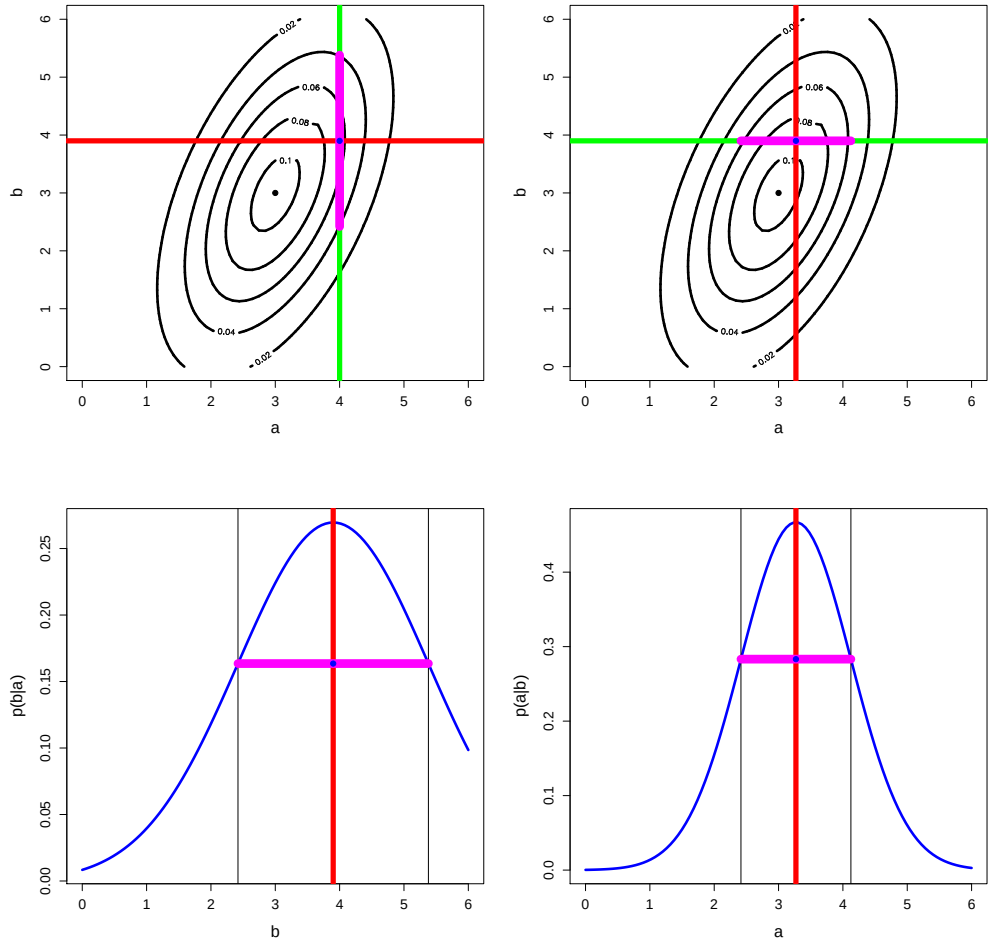
  1. draw $\boldsymbol{\theta}_k$ from $p(\boldsymbol{\theta}_k|\bullet)$ (all currently populated + 1 empty haplotype)
  2. draw c_i from $p(c_i|\bullet),\ i = 1, \ldots, n_{\text{reads}}$
  3. draw $\boldsymbol{\pi}$ from $p(\boldsymbol{\pi}|\bullet)$ (all currently populated + 1 empty haplotype)

- 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

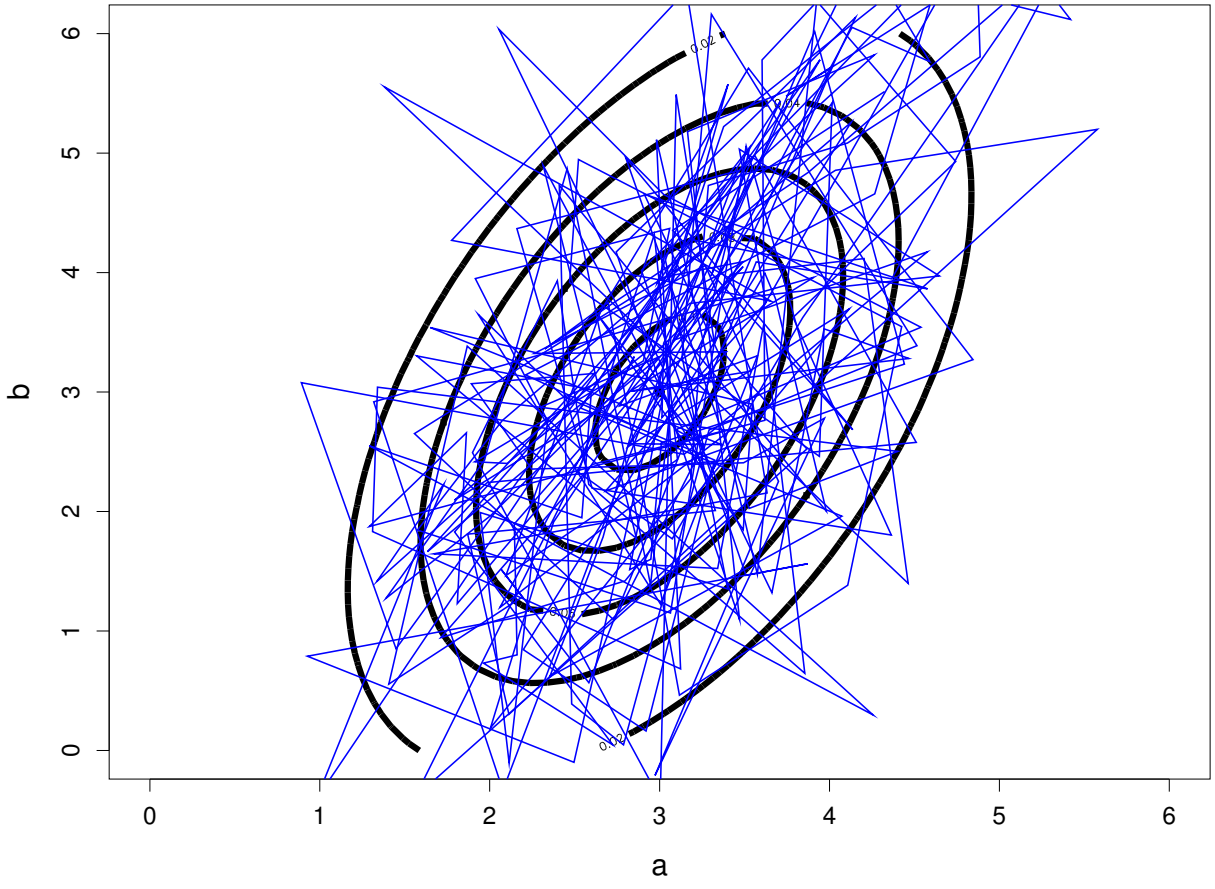
Assume you want to sample from a 2-dim Gaussian $p(a, b) \sim \mathcal{N}(a, b|\mu, \Sigma)$. You know that **conditionals of Gaussians are again Gaussians**, but you have forgotten how to sample from a 2-dim Gaussian.



Gibbs Sampling

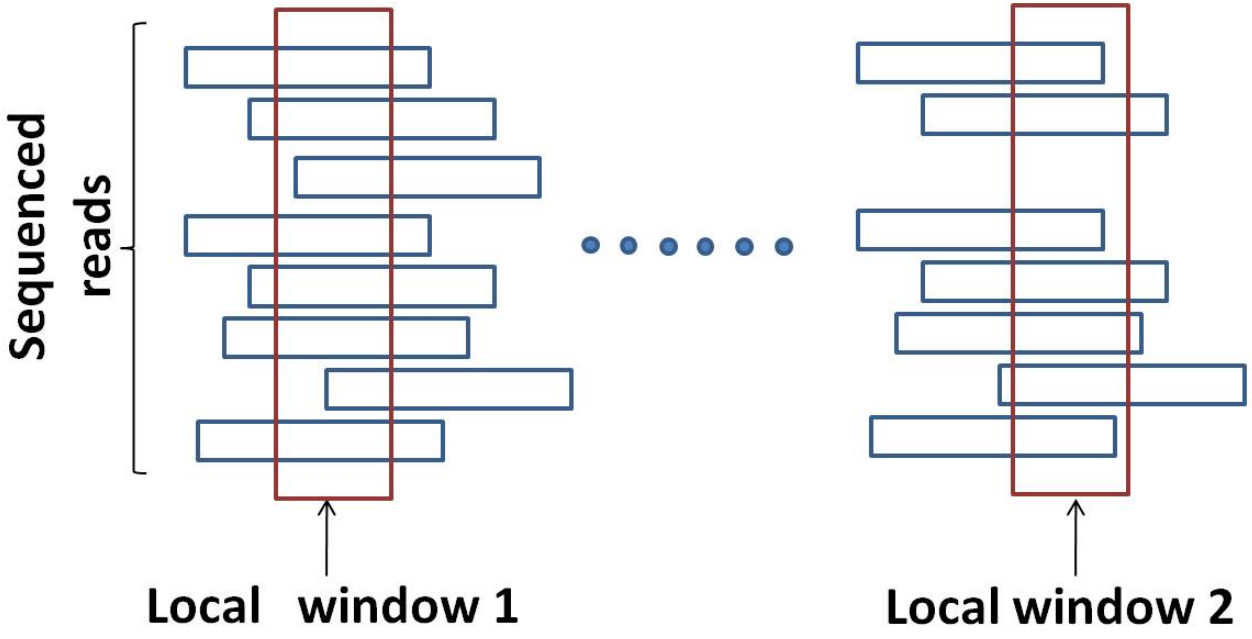
Solution: run a Gibbs sampler: Iterate:

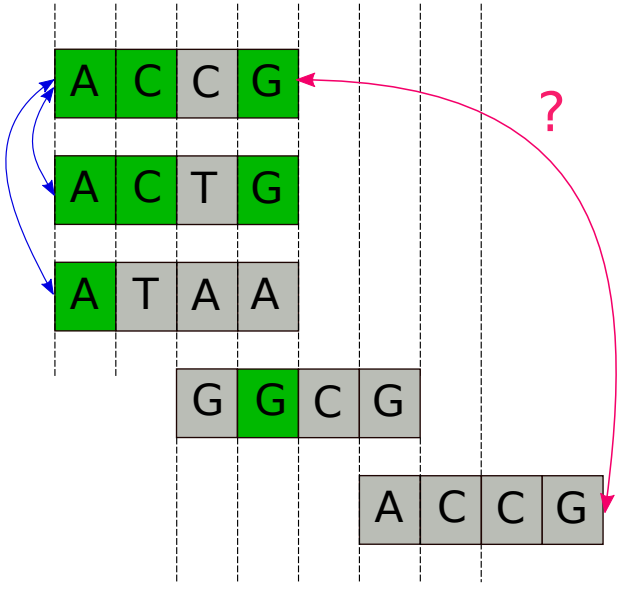
1. sample a from $p(a|b, \bullet) = \mathcal{N}(a|\mu', \Sigma')$
2. sample b from $p(b|a, \bullet) = \mathcal{N}(b|\mu'', \Sigma'')$



Local to Global

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

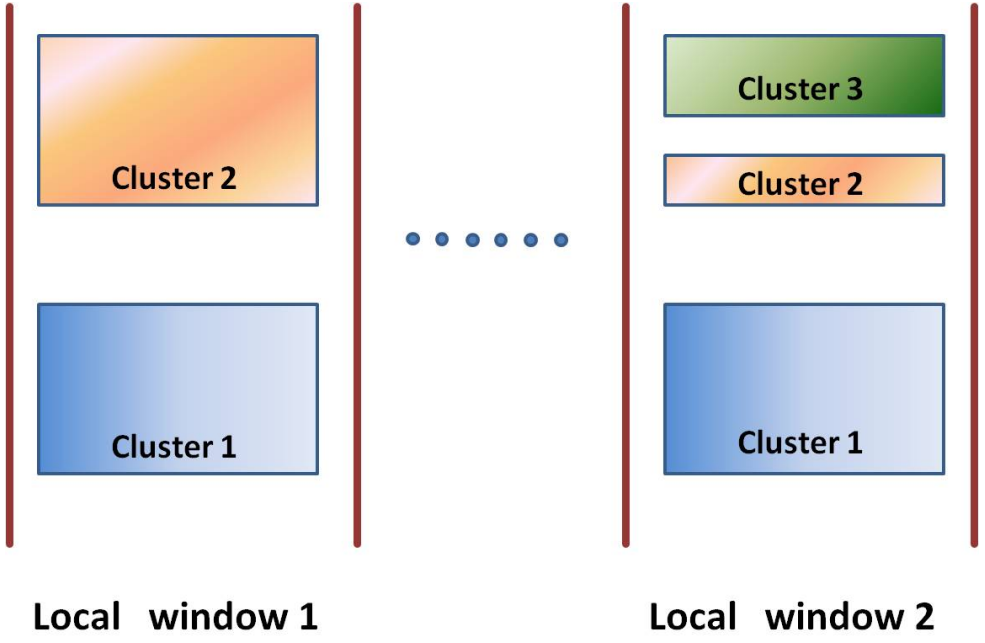




Global reconstruction

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

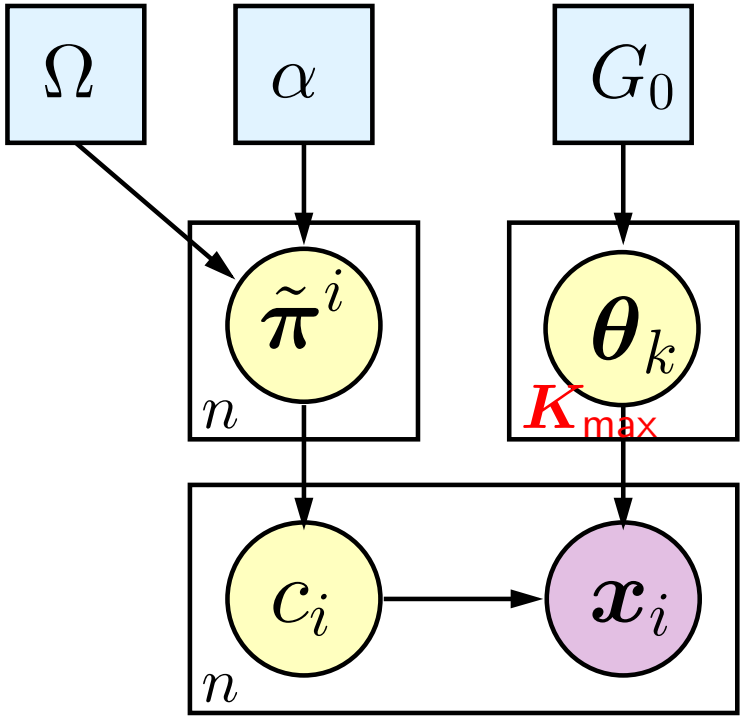
- **Distribution of the reads:**

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

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

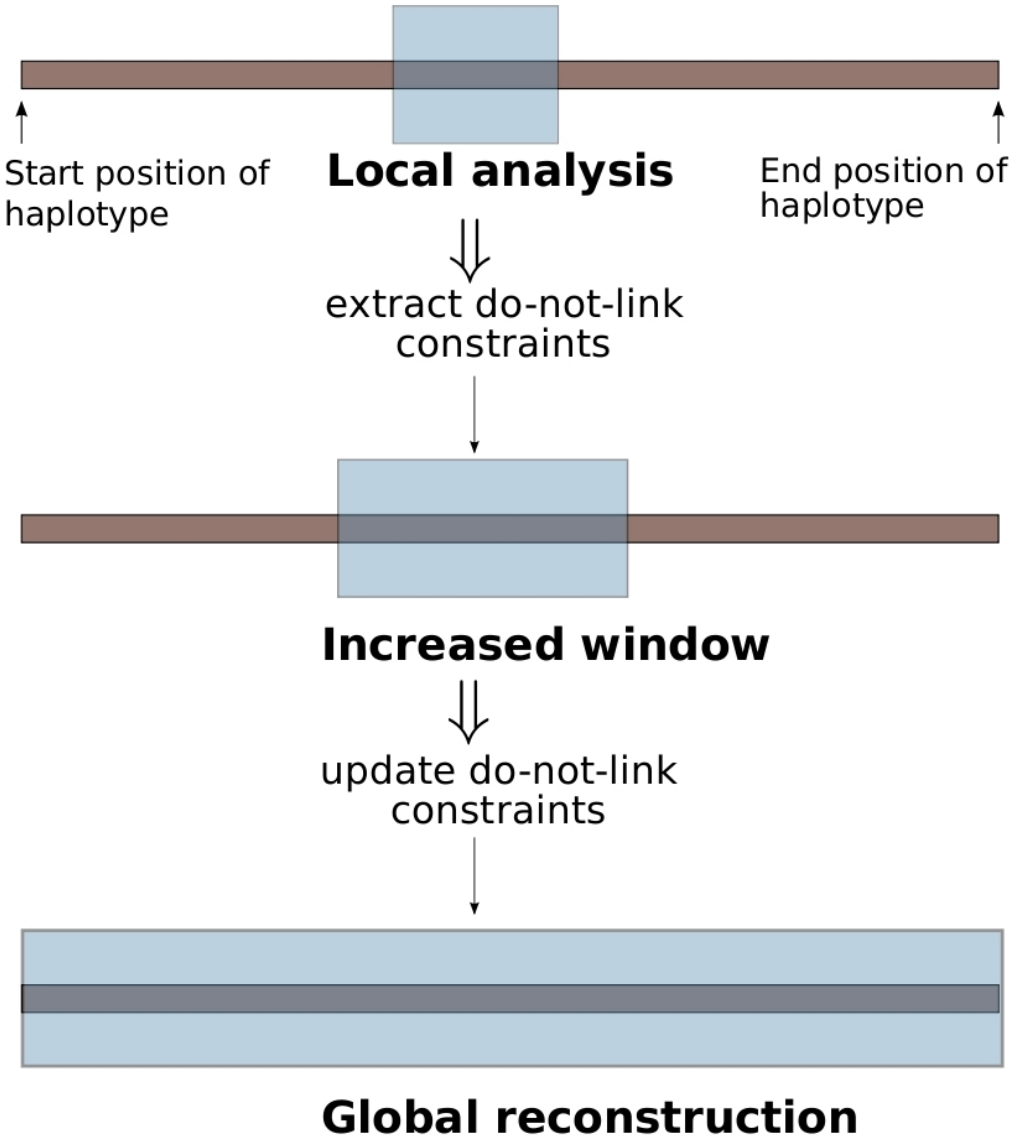
$$c_i|\tilde{\boldsymbol{\pi}}^i \sim \mathrm{Mult}\left(c_i|\tilde{\boldsymbol{\pi}}^i\right)$$

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



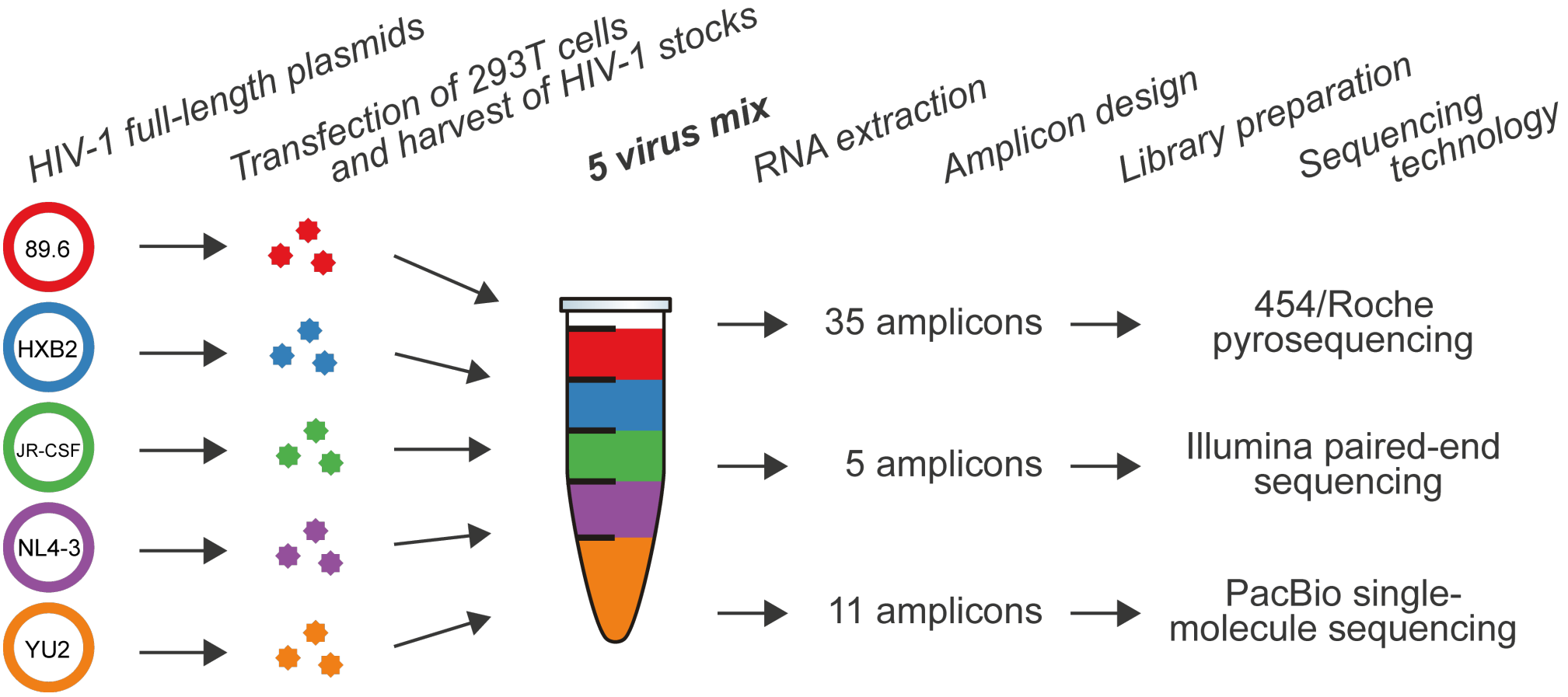
Prabhakaran et al., 2010

Constrained model: summary



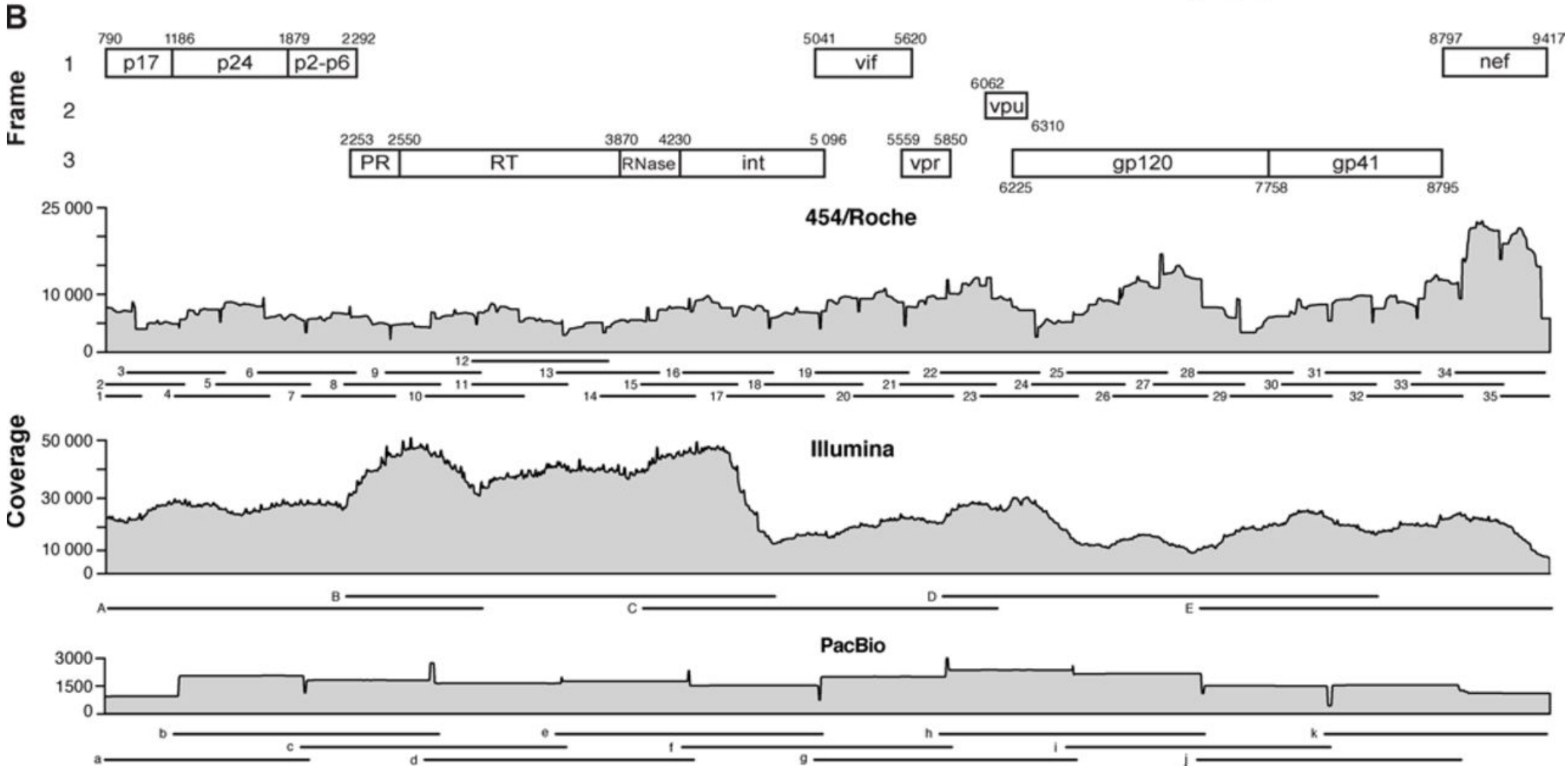
Beerenwinkel et al (Front Microbiol 2012)

Experiments: 5 Virus Mix



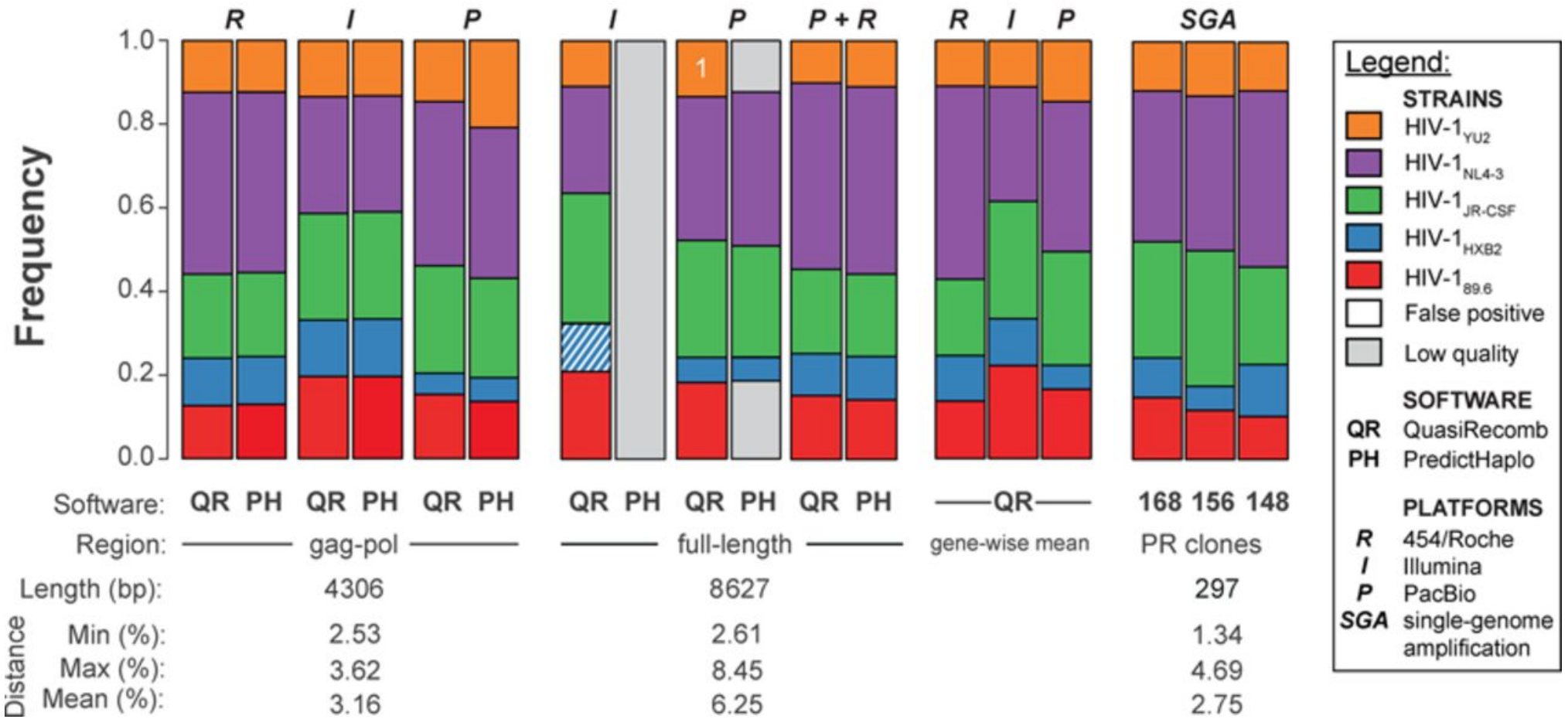
Di Giallonardo et al., Nucl Acids Res 2014

Experiments: 5 Virus Mix



Di Giallonardo et al., Nucl Acids Res 2014

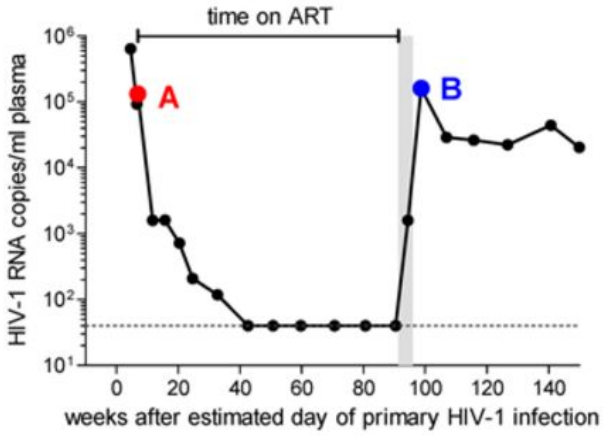
Experiments: 5 Virus Mix



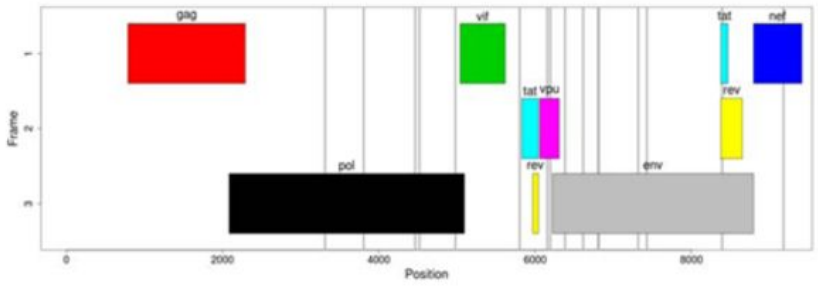
Di Giallonardo et al., Nucl Acids Res 2014

Experiments: Patient with Superinfection

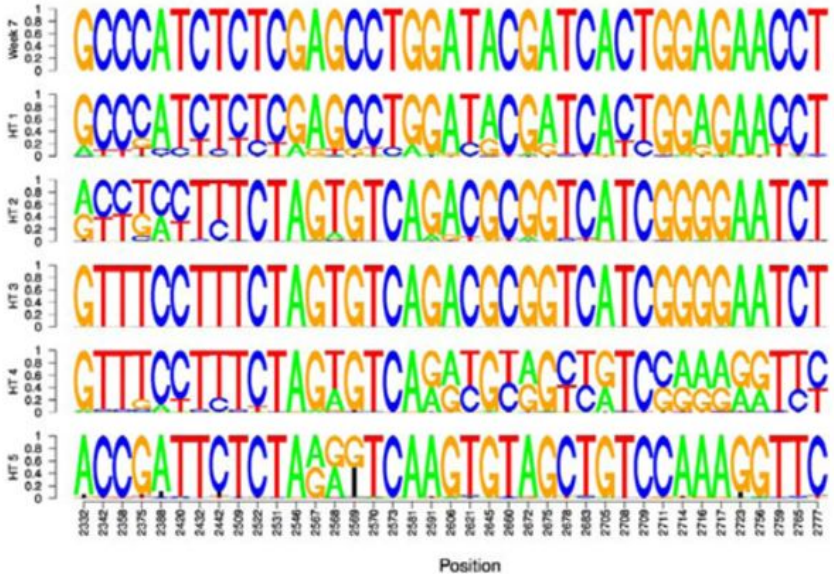
a



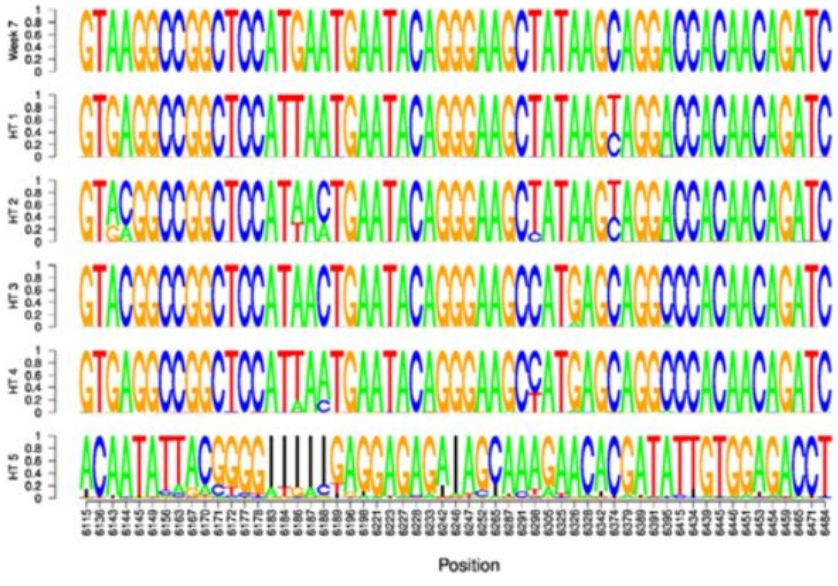
b



c



d



Di Giallonardo et al., Nucl Acids Res 2014

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

Acknowledgments

Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich: **Francesca Di Giallonardo, Yannick Duport, Christine Leemann, Stefan Schmutz, Nottania K. Campbell, Beda Joos, Huldrych F. Günthard, Karin J. Metzner**

Department of Biosystems Science and Engineering, ETH Zurich: **Armin Töpfer, Christian Beisel, Niko Beerenwinkel**

Functional Genomics Center Zurich: **Maria Rita Lecca, Andrea Patrignani**

Inst. Immunologie und Genetik, Kaiserslautern: **Martin Däumer**

Institute of Medical Virology, University of Zurich: **Peter Rusert, Alexandra Trkola**

Dept. Math. & Comp. Sci., U Basel: **Sandhya Prabhakaran, Sudhir Raman, Mélanie Rey, Sonali Parbhoo, Mario Wieser**