THESIS DEFENSE

Bayesian Assembly of Reads from High Throughput Sequencing

Jonathan Laserson

advised by

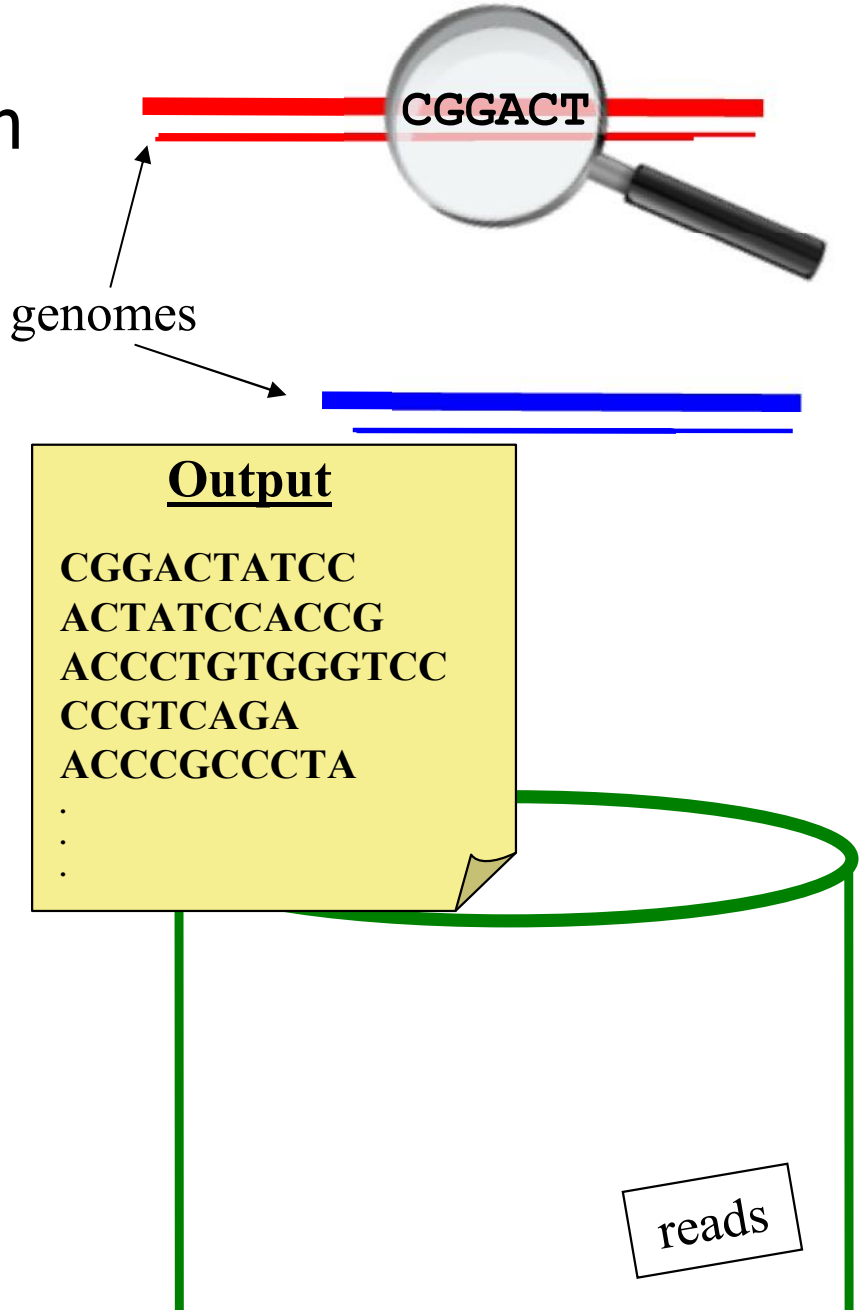
Daphne Koller

October 20th, 2011

High-Throughput Sequencing

- Millions of noisy short reads from a sample
- A **snapshot** of the genomic material in the sample

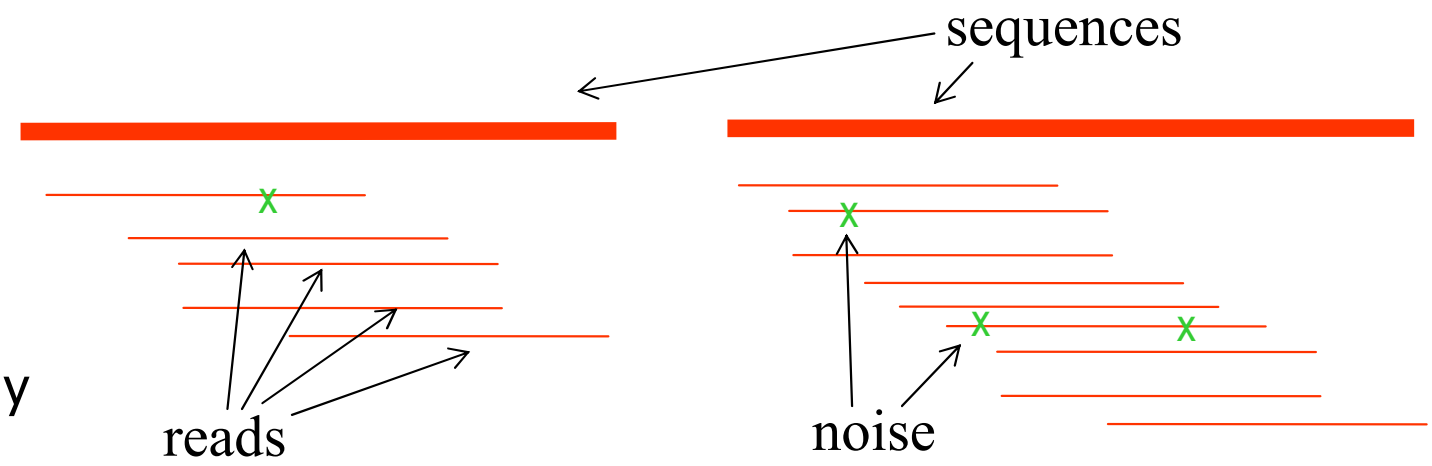
- Roche 454 GS-FLX Titanium
  - Avg read length 380b
  - reads in a run 500k
  - Base error rate 0.74% (84% indels)
  - Cost per base 0.05$/kb



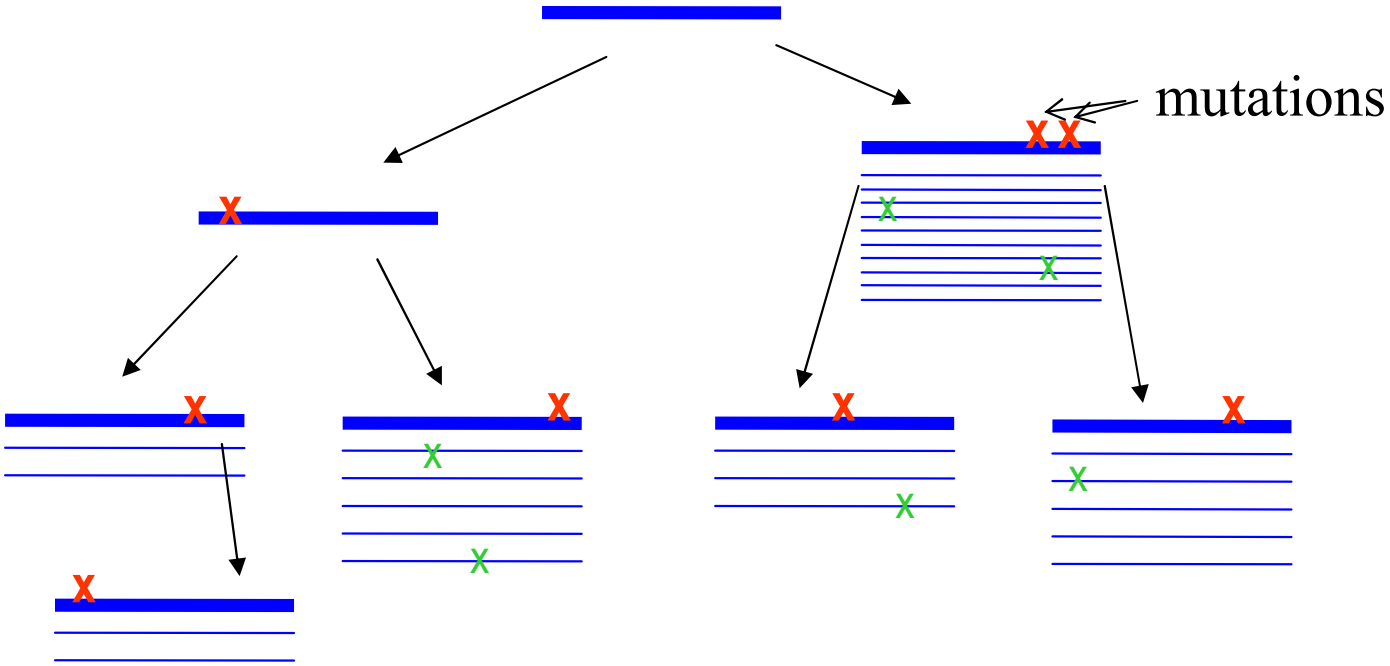
[*You* et al., 2011]

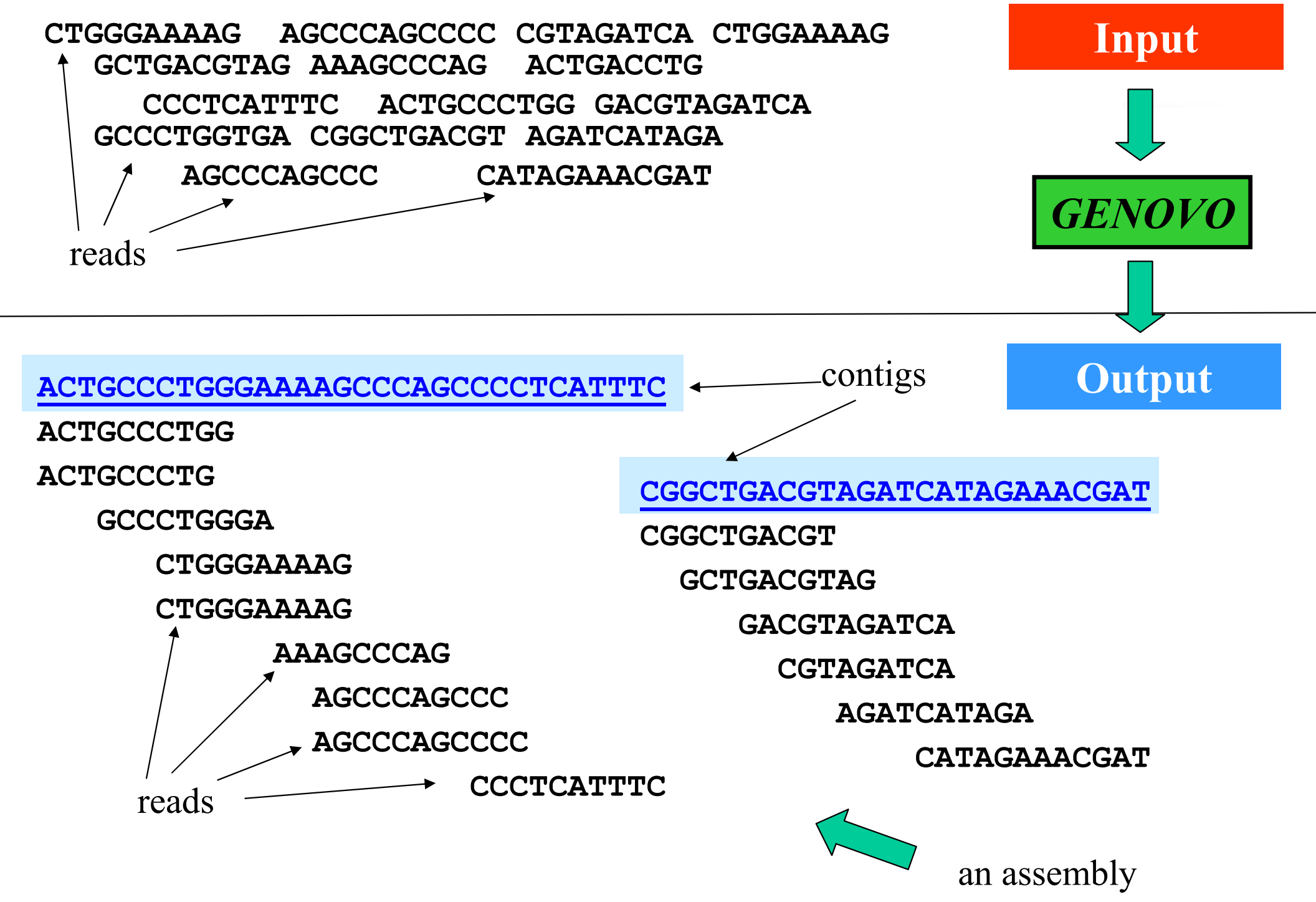
Population Sequencing

- **Breadth**
  - Metagenomics
  - *de novo* assembly



- **Depth**
  - Phylogenetic trees
  - Immune-seq




CTGGGAAAAG AGCCCAGCCCC CGTAGATCA CTGGAAAAG
GCTGACGTAG AAAGCCCAG ACTGACCTG
CCCTCATTTC ACTGCCCTGG GACGTAGATCA
GCCCTGGTGA CGGCTGACGT AGATCATAGA
AGCCCAGCCC CATAGAAACGAT
reads
Input
GENOVO
Output
ACTGCCCTGGGAAAAGCCCAGCCCCTCATTTC
contigs
ACTGCCCTGG
ACTGCCCTG
GCCCTGGGA
CTGGGAAAAG
CTGGGAAAAG
AAAGCCCAG
AGCCCAGCCC
AGCCCAGCCCC
CCCTCATTTC
reads
CGGCTGACGTAGATCATAGAAACGAT
CGGCTGACGT
GCTGACGTAG
GACGTAGATCA
CGTAGATCA
AGATCATAGA
CATAGAAACGAT
an assembly

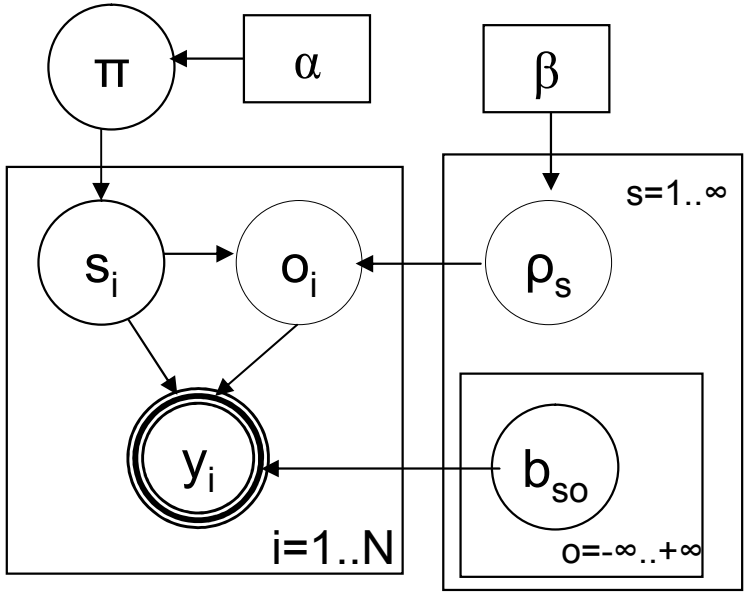
Previous Work

- Velvet, Euler:
  - Assume most reads have no error
  - Assume a single genome
  - Trash low-abundance reads
  - Fix reads with one error a-priori
- Newbler: No code / publication

Our Approach: Probabilistic Model

- **Fully Bayesian** model for high-throughput sequencing
- **Jointly** performing 3 tasks:
  - de-noising the reads
  - *de novo* sequence assembly
  - multiple alignment
- **We do not throw away reads**
- Can handle **unbalanced populations** of genomes

Our Approach: Probabilistic Model



$b_{so} \sim \mathrm{U}[\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}]$

$\mathbf{s} \sim \mathrm{CRP}(\alpha, N)$

$\rho_s \sim \mathrm{Beta}(1,1+\beta)$

$o_i \sim \mathcal{G}(\rho_{s\_i})$

$y_i \sim \mathrm{read\_model}(\mathbf{b}, s_i, o_i)$

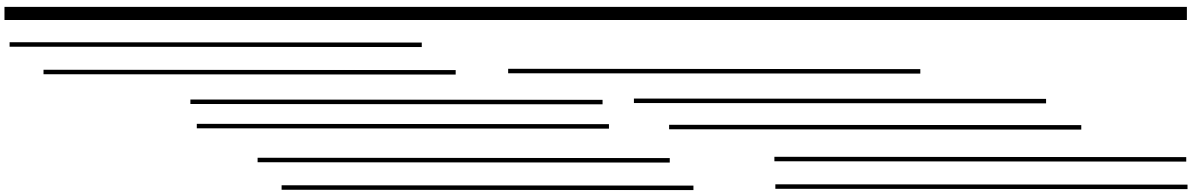
$$\log P(\mathbf{y},\mathbf{b},\mathbf{s},\mathbf{o}|\boldsymbol{\rho}) \approx \#\text{errors} \cdot \log(\varepsilon) - \#\text{bases} \cdot \log(4) + \#\text{seq} \cdot \log(\alpha)$$

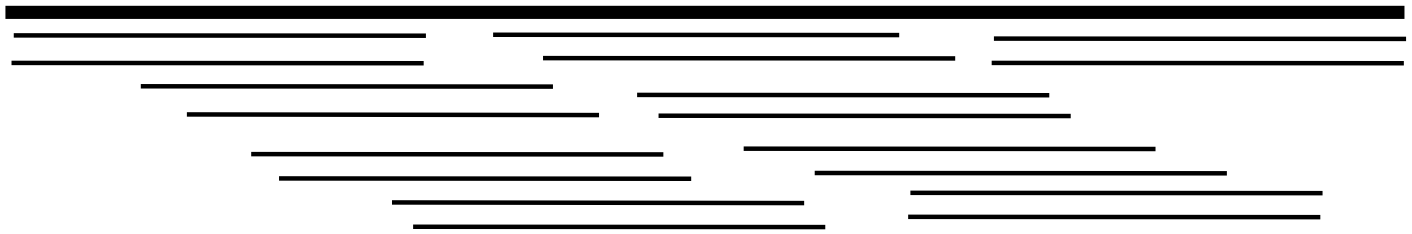
min read errors — min bases

we want to find

$$\underset{\mathbf{b},\mathbf{s},\mathbf{o}}{\operatorname{argmax}} P(\mathbf{b},\mathbf{s},\mathbf{o} \mid \mathbf{y})$$

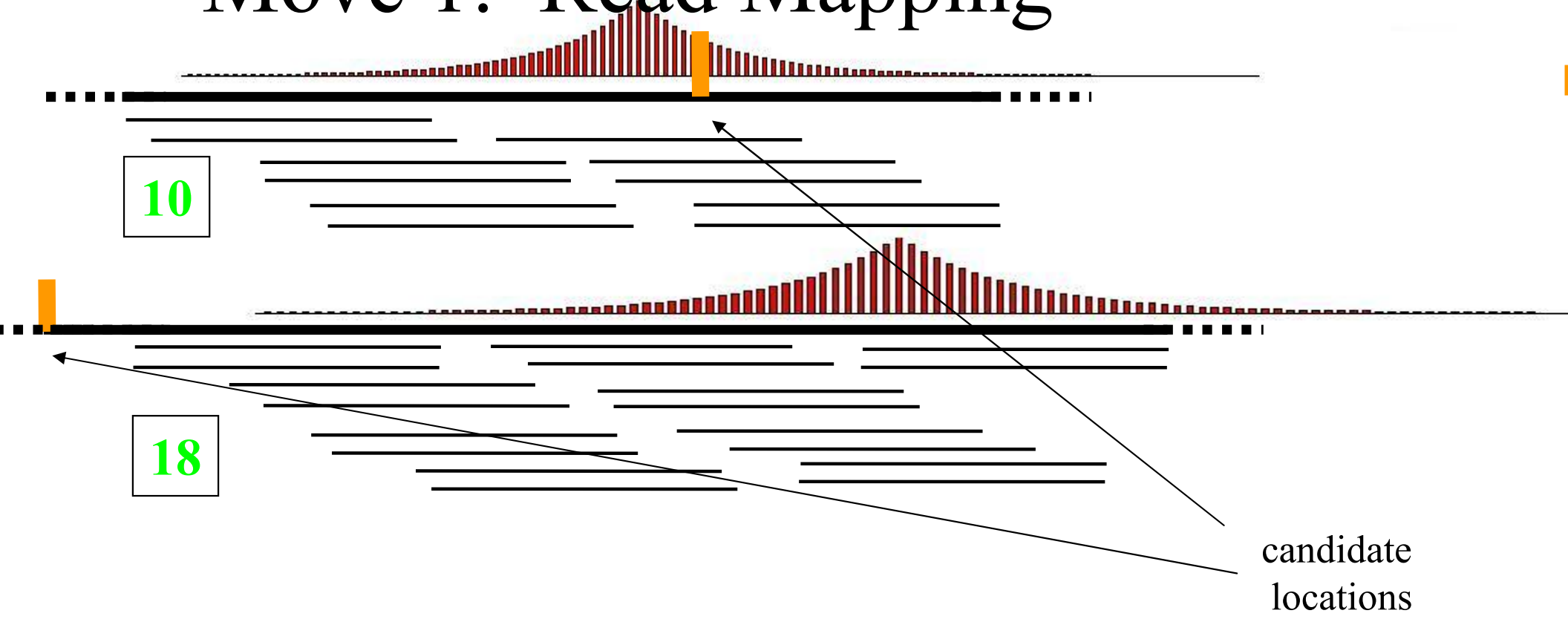
Inference





- MCMC
- Making moves in a probabilistic space
- Stochastic process
  - escape local optimum
  - multiple hypotheses

Move 1: Read Mapping

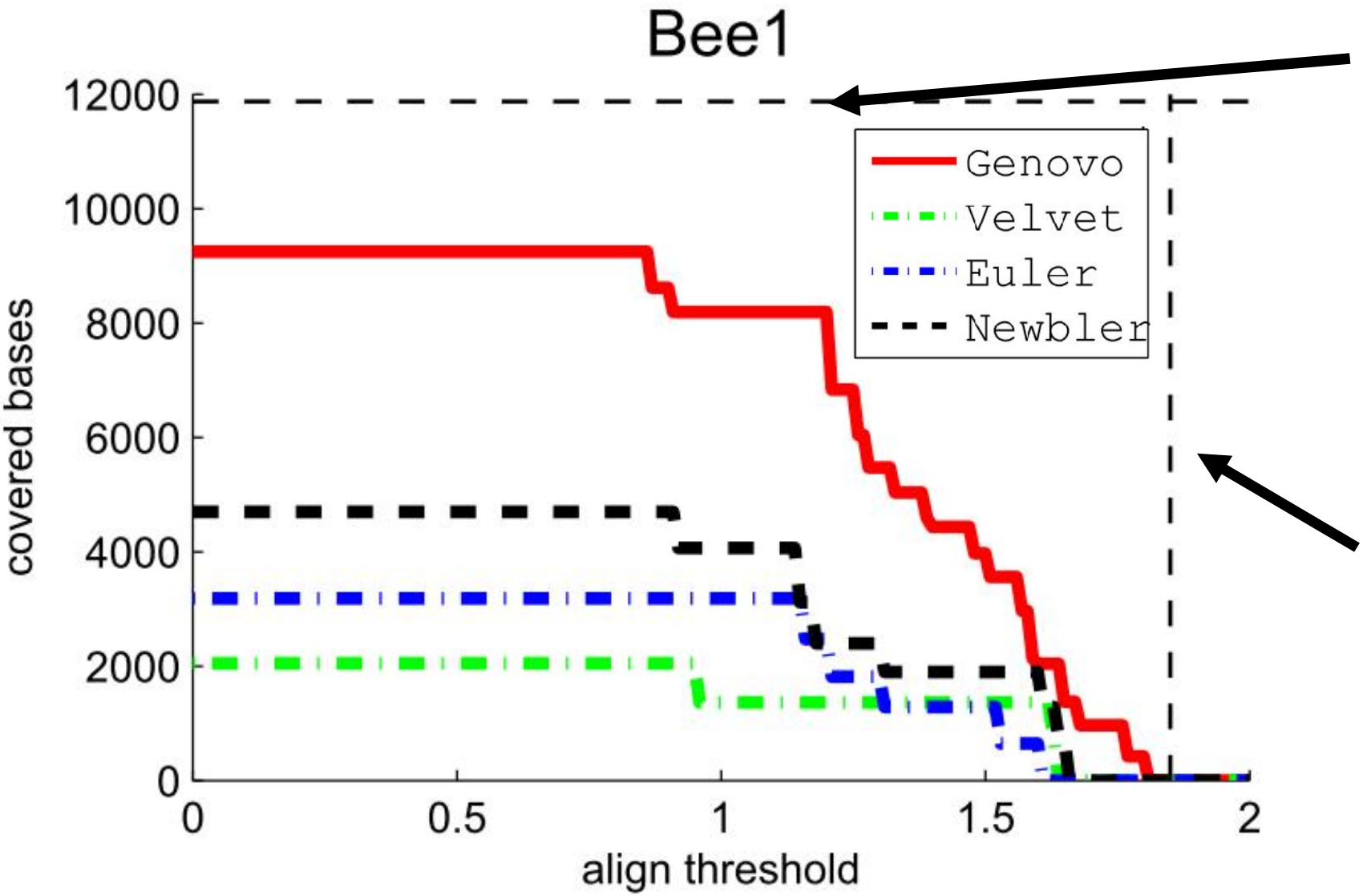


$$P(s_i{=}s, o_i{=}o \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\rho}, \mathbf{s}_{-i}, \mathbf{o}_{-i}) \propto$$ Likelihood Term | Stickiness Term | Geometric Term

A READ samples a *sequence* and an *offset* (inside the seq)

s_i o_i

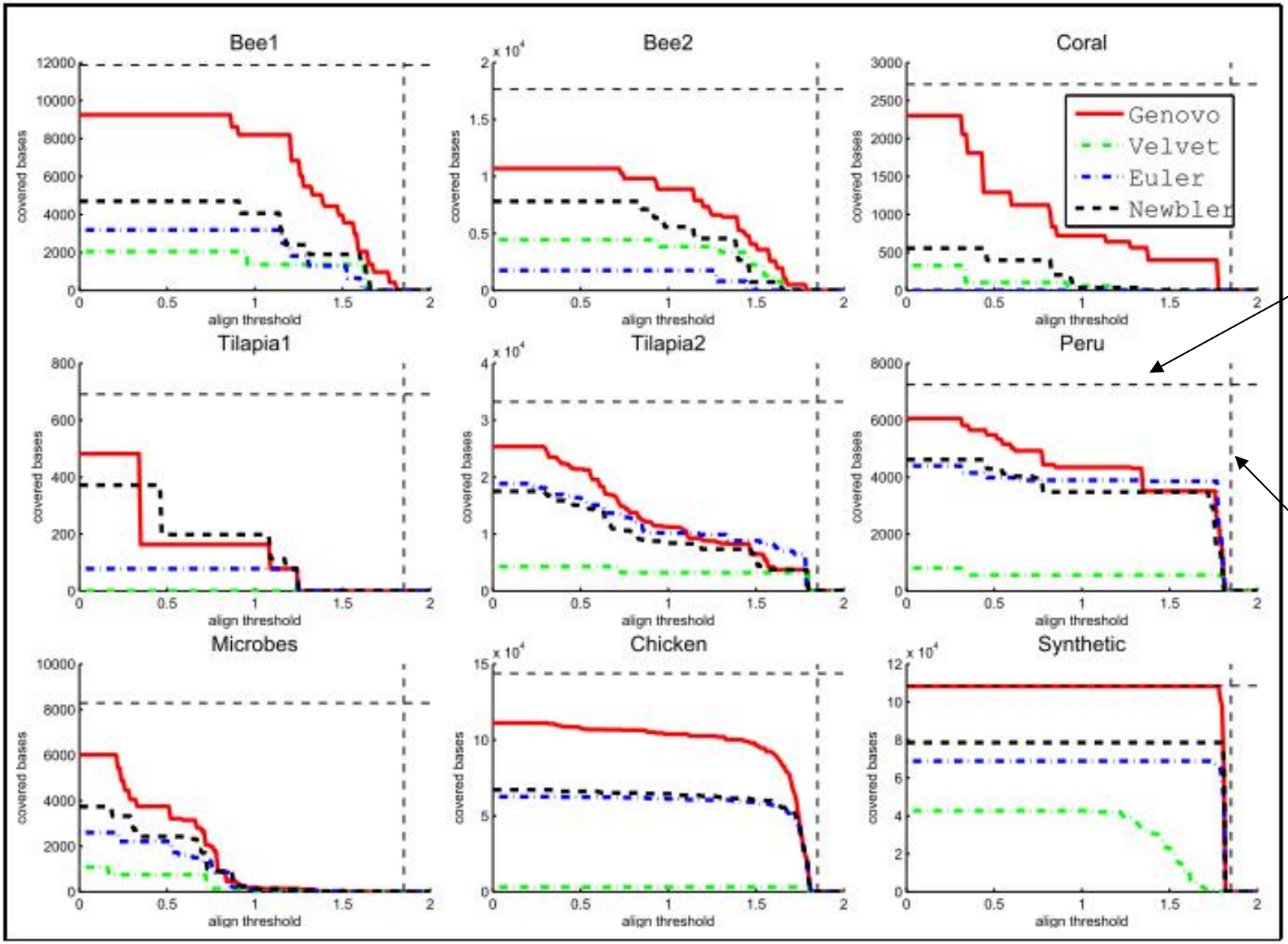
Comparison:
How many bases did you BLAST into?



horizontal line: total bases covered by all methods together

vertical line: threshold for a perfect match

Comparison:
How many bases did you BLAST into?



horizontal line: total bases covered by all methods together

vertical line: threshold for a perfect match

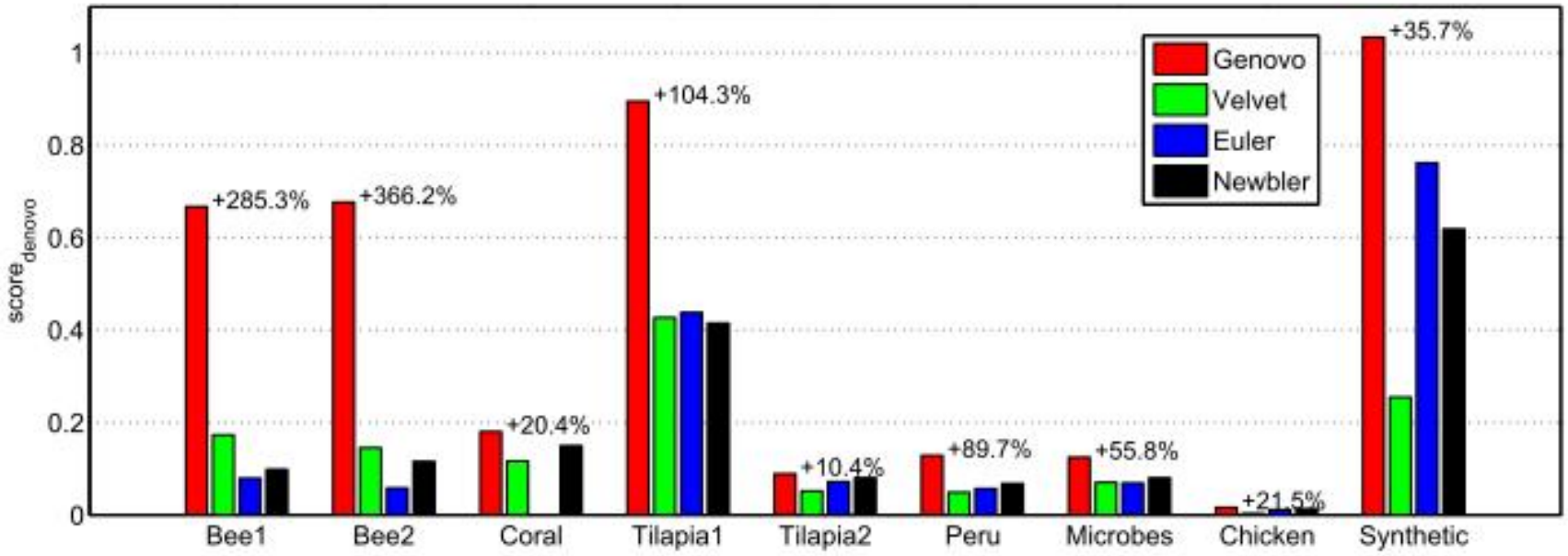
A Score for an Assembly

min read errors

min bases

α controls minimal overlap between adjacent reads

$$\text{score}_{denovo} = \#\text{errors} \cdot \log(\varepsilon) - \#\text{bases} \cdot \log(4) + \#\text{seq} \cdot \log(\alpha)$$



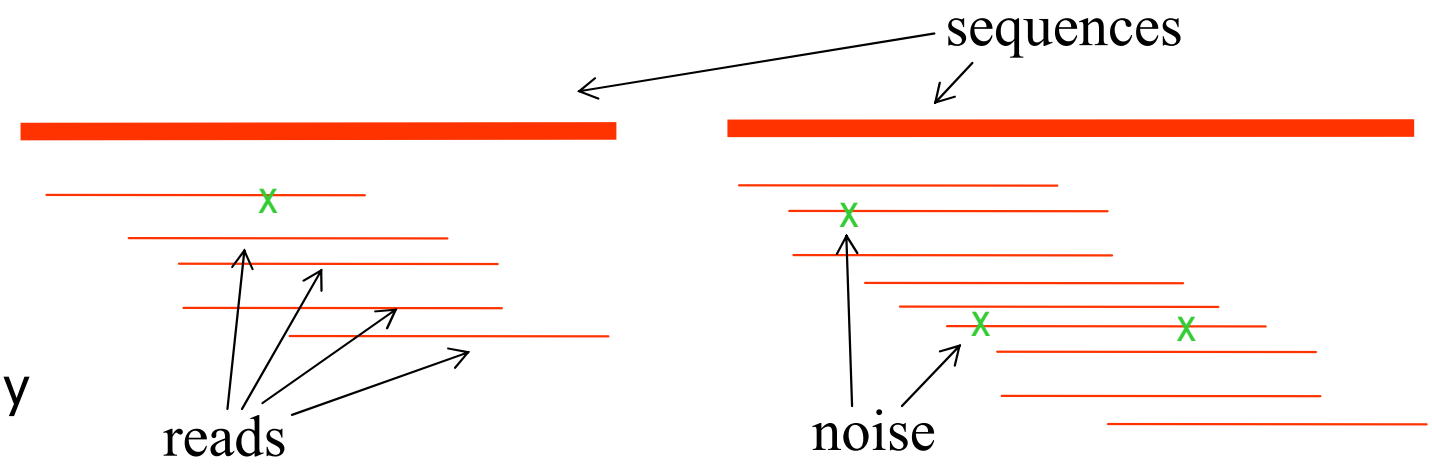
Contributions

- Fully Bayesian model for high-throughput sequencing
- An algorithm jointly performing 3 tasks:
  - de-noising the reads
  - *de novo* sequence Assembly
  - multiple alignment
- A score for *de novo* assembly
- Code freely available

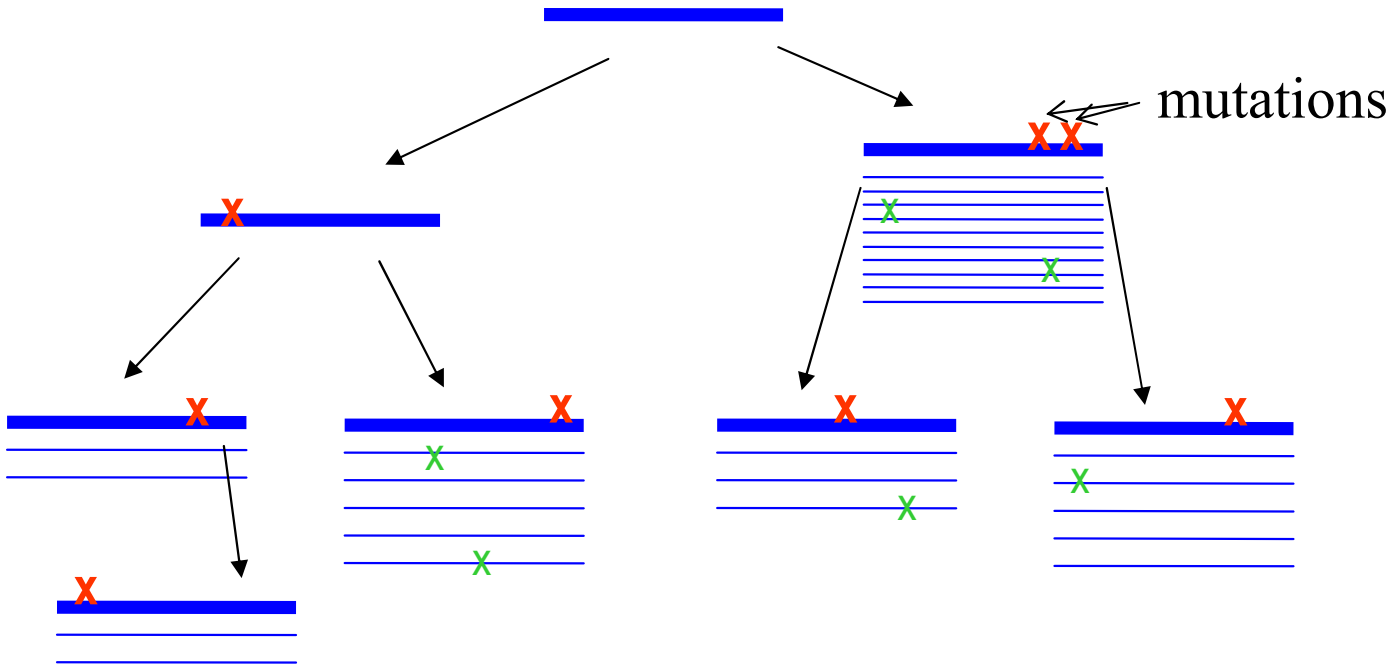
Laserson J, Jojic V, Koller D, **Genovo: *de novo* Assembly for Metagenomes,** *Journal of Computational Biology,* 2011.
RECOMB 2010.

Population Sequencing

- Breadth
  - Metagenomics
  - *de novo* assembly

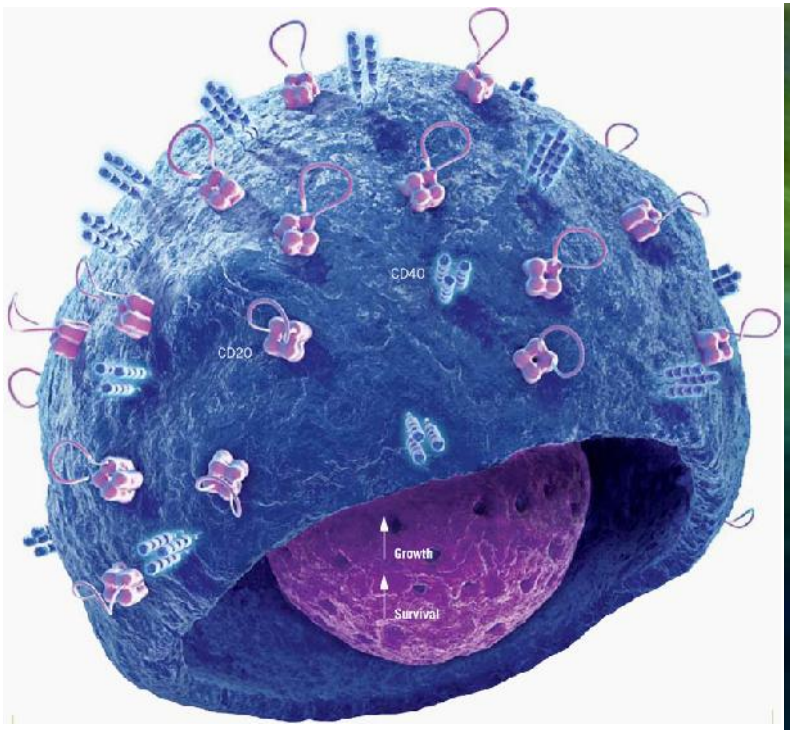


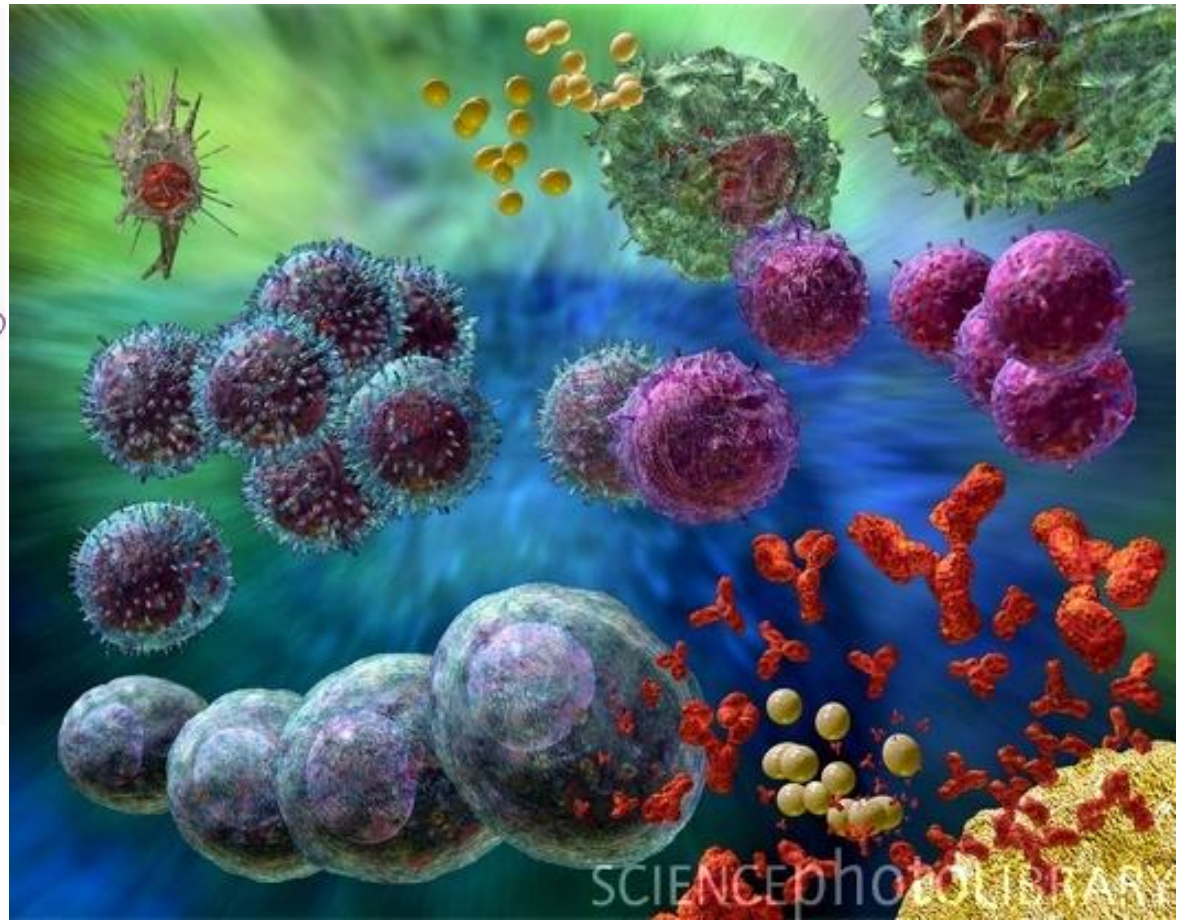
- Depth
  - Phylogenetic trees
  - Immune-seq

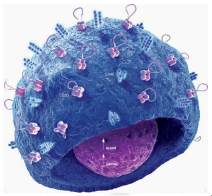


Immune System

- 10^{10} B-Cells in an individual



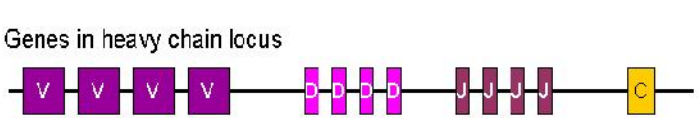


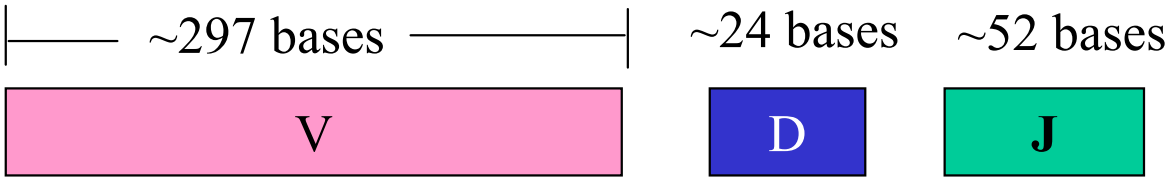


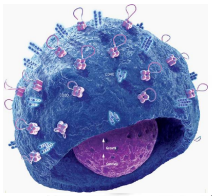
Immune System

- 10^{10} B-Cells in an individual
- Phase 1: generate basic seq
  - choose V,D,J from catalog

57 x 27 x 6

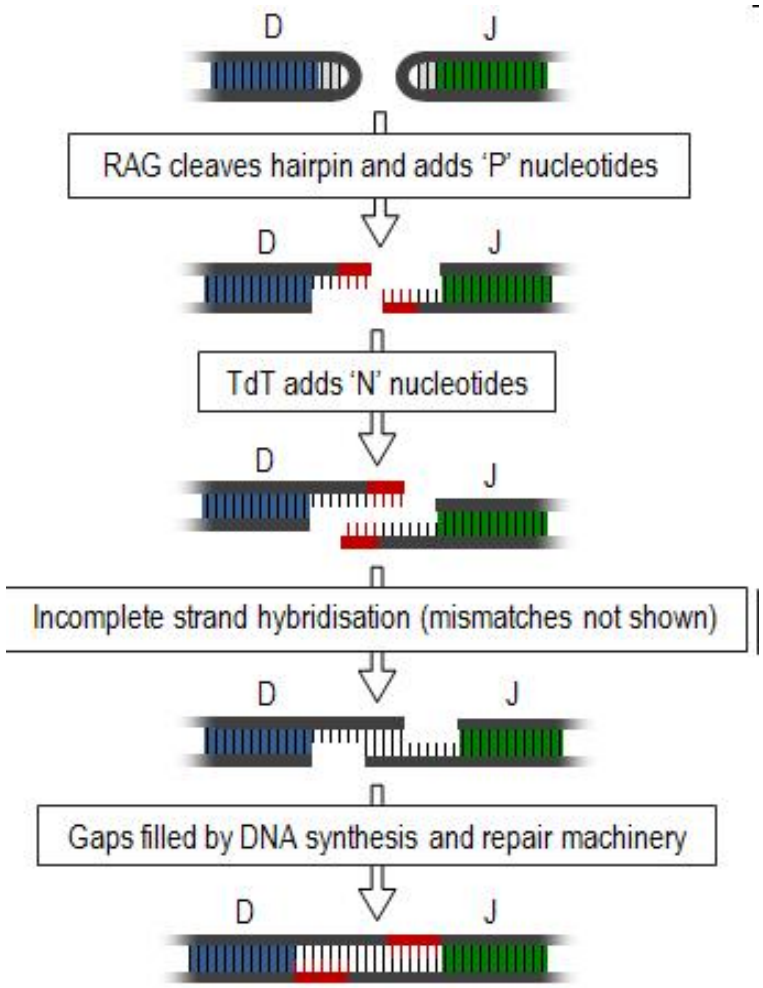




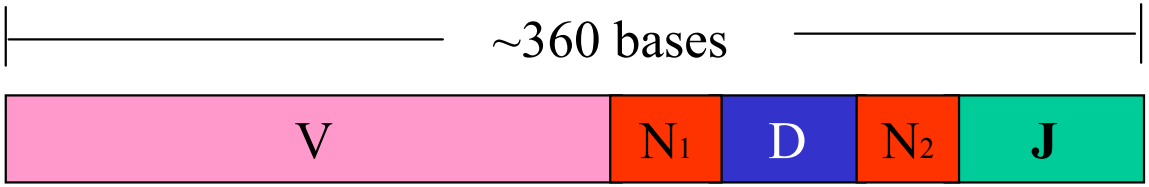


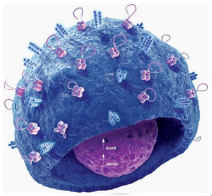
Immune System

- 10^{10} B-Cells in an individual
- Phase 1: generate basic seq
  - choose V,D,J
  - stitch V-D and D-J by:
    - eat up a few letters from ends
    - insert a few letters



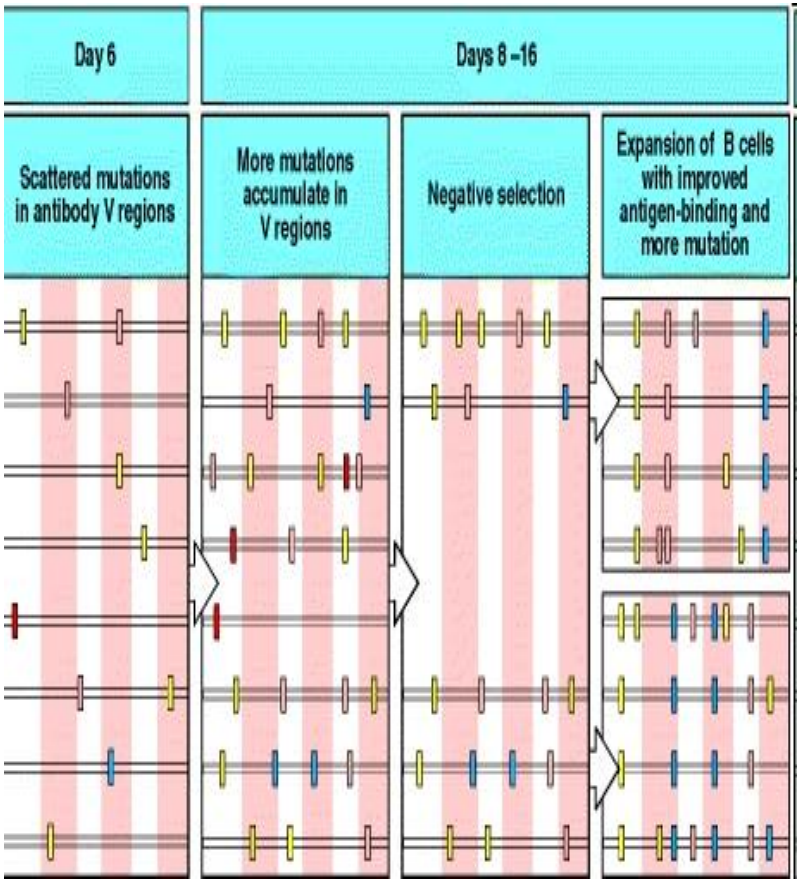
[Illustration: Wikipedia]



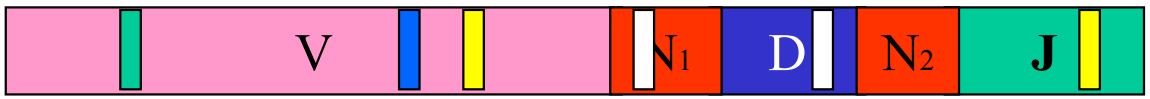


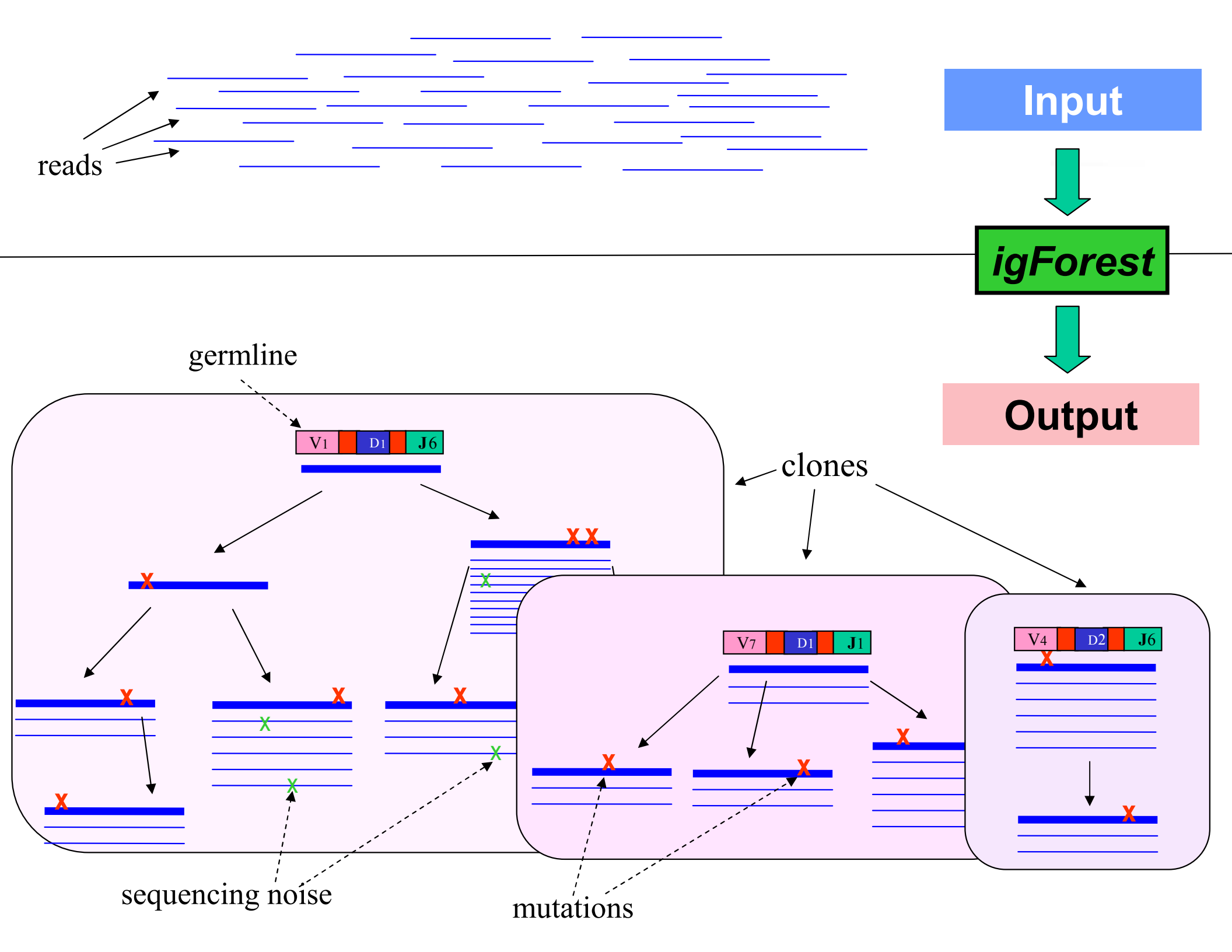
Immune System

- 10^{10} B-Cells in an individual
- Phase 1: generate basic seq
- Phase 2: hyper-mutate
  - If detecting an invader, **clone yourself with lots-of-mutations**
  - Even better detection? **keep cloning.**
  - **Remember the successful sequences** for future reference, and keep producing them.
  - This set of sequences is <u>one</u> **clone**.



[Illustration: Janeway, 2001]




reads
Input
igForest
Output
germline
V1
D1
J6
clones
V7
D1
J1
V4
D2
J6
sequencing noise
mutations

ReperTree

De Novo Annotation for B-Cells

- Phase 1: Divide reads into clones
  - 1M reads!
  - In a **healthy individual** almost every read is from a different clone.

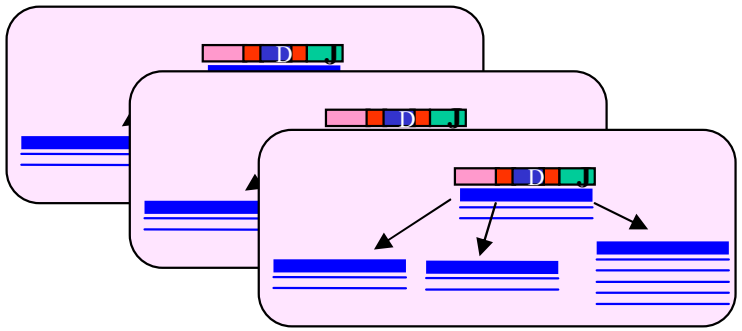
- Phase 2: Construct a mutation tree for each clone.
  - How to tell a **mutation** from a **sequencing error**?

Overview

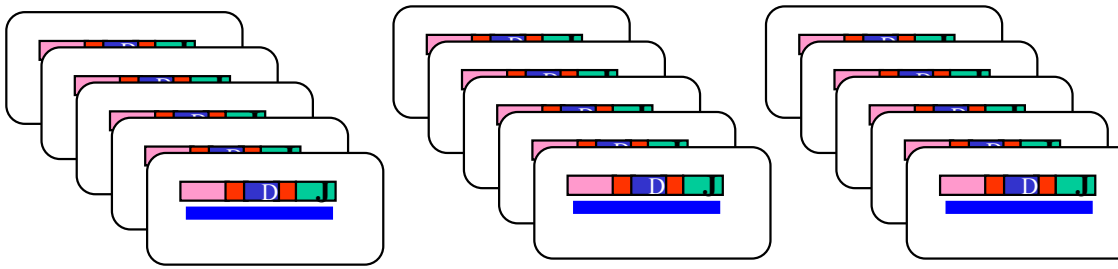
- 10 (healthy) Organ donors
- 4 samples from each individual
  - Blood, spleen, two lymph nodes
- 380k reads total

igForest

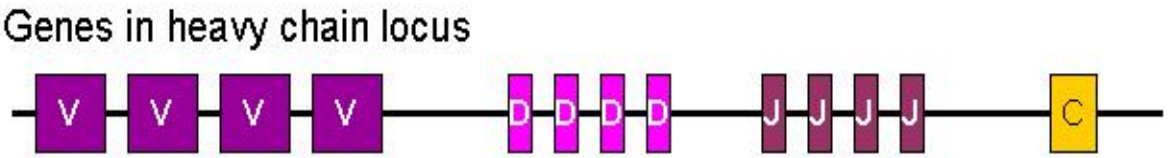
- 8,000 clones



- 165k other B-cells variants (not clones)

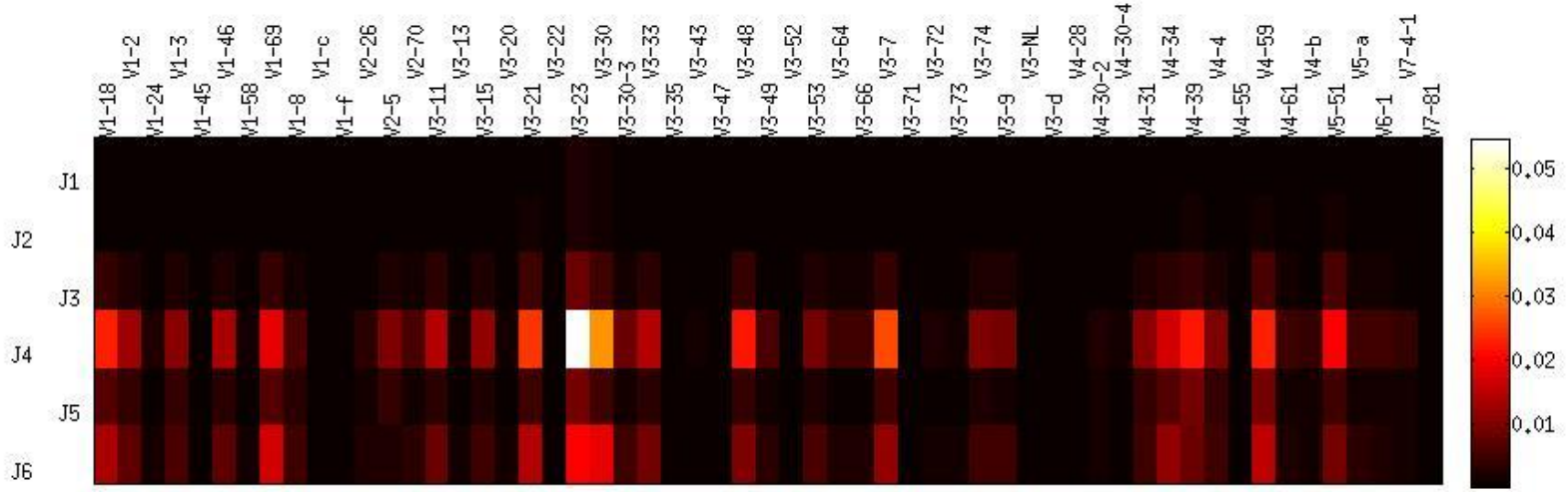


A Repertoire

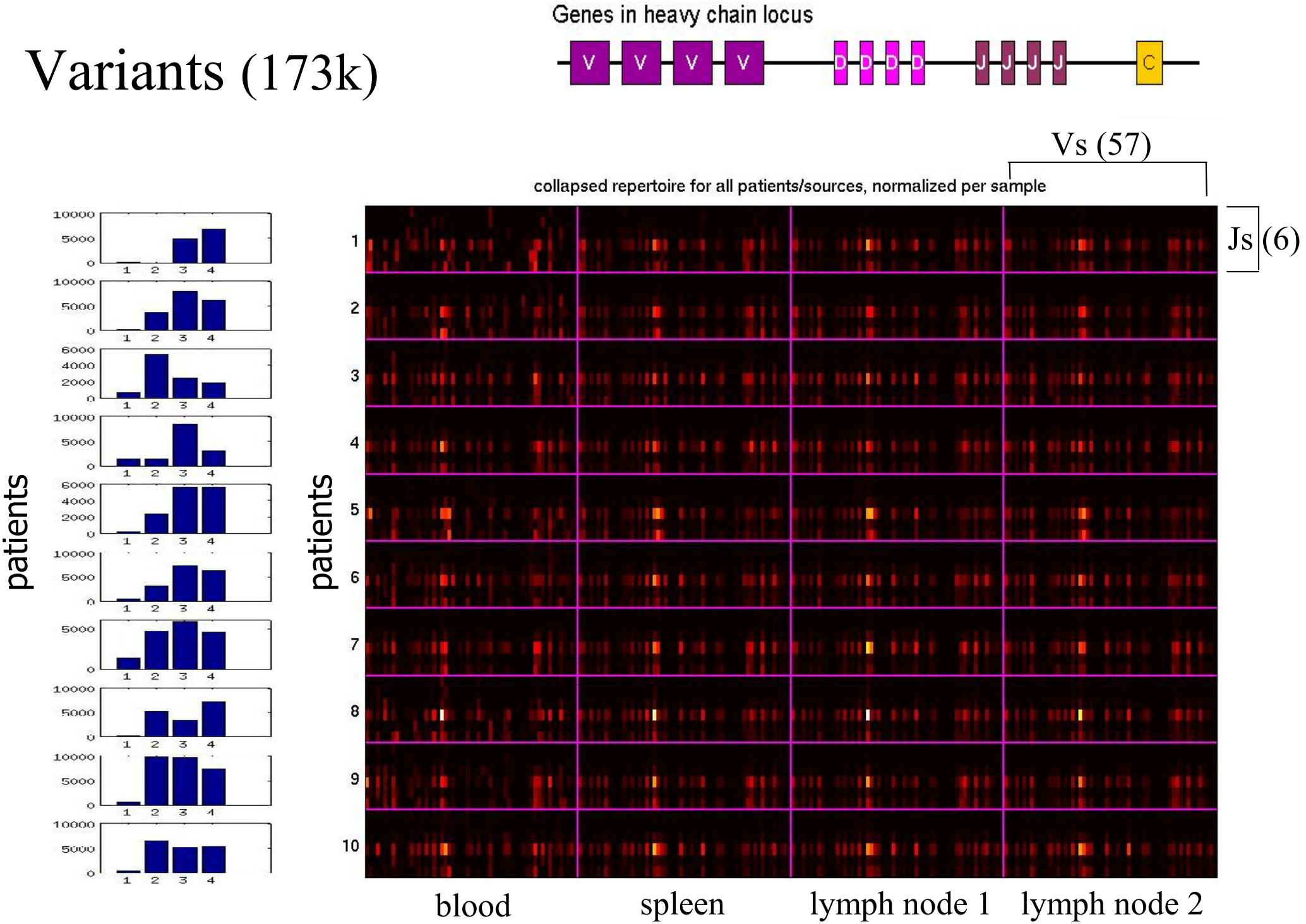


Vs (57)

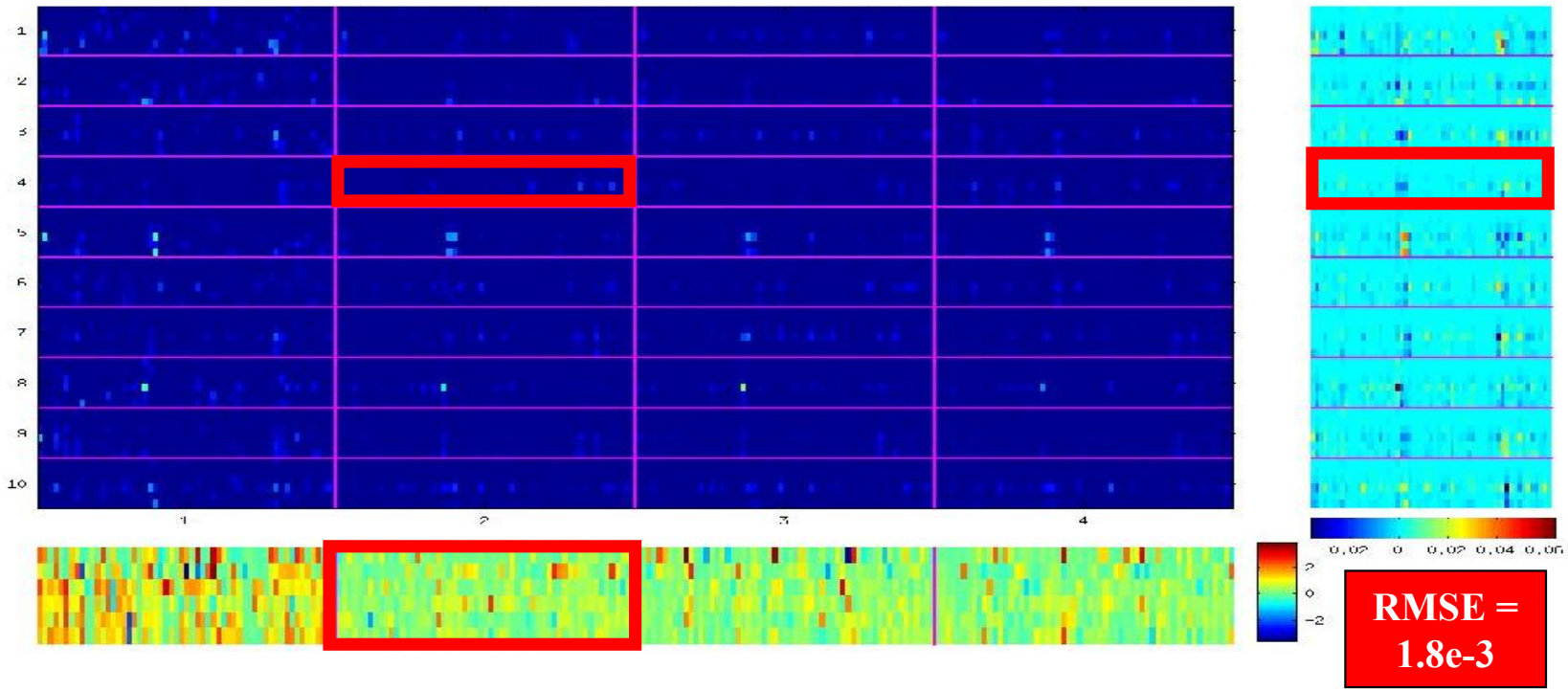
Js (6)



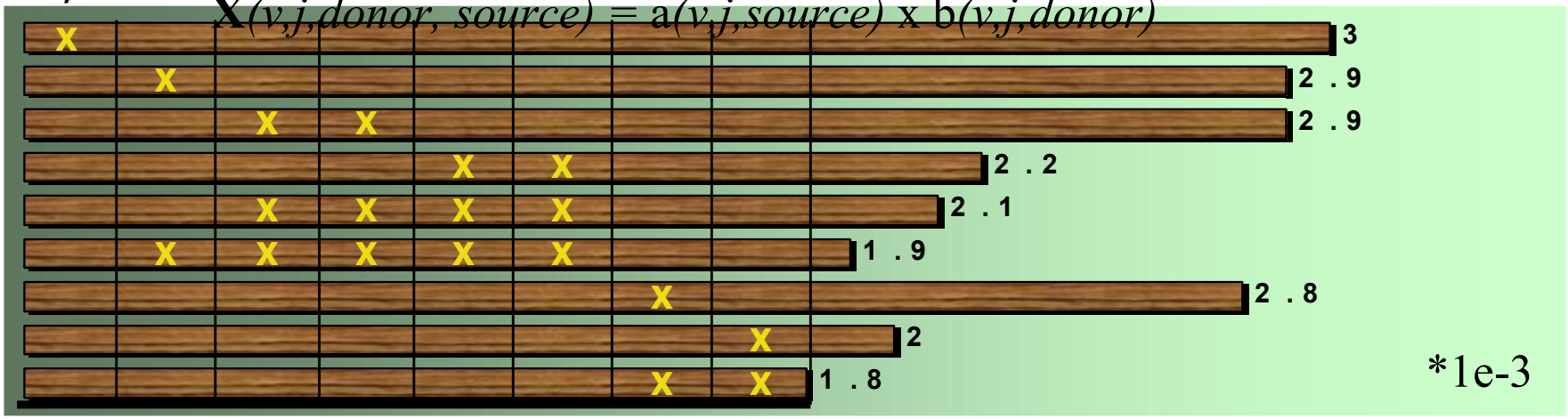
Variants (173k)



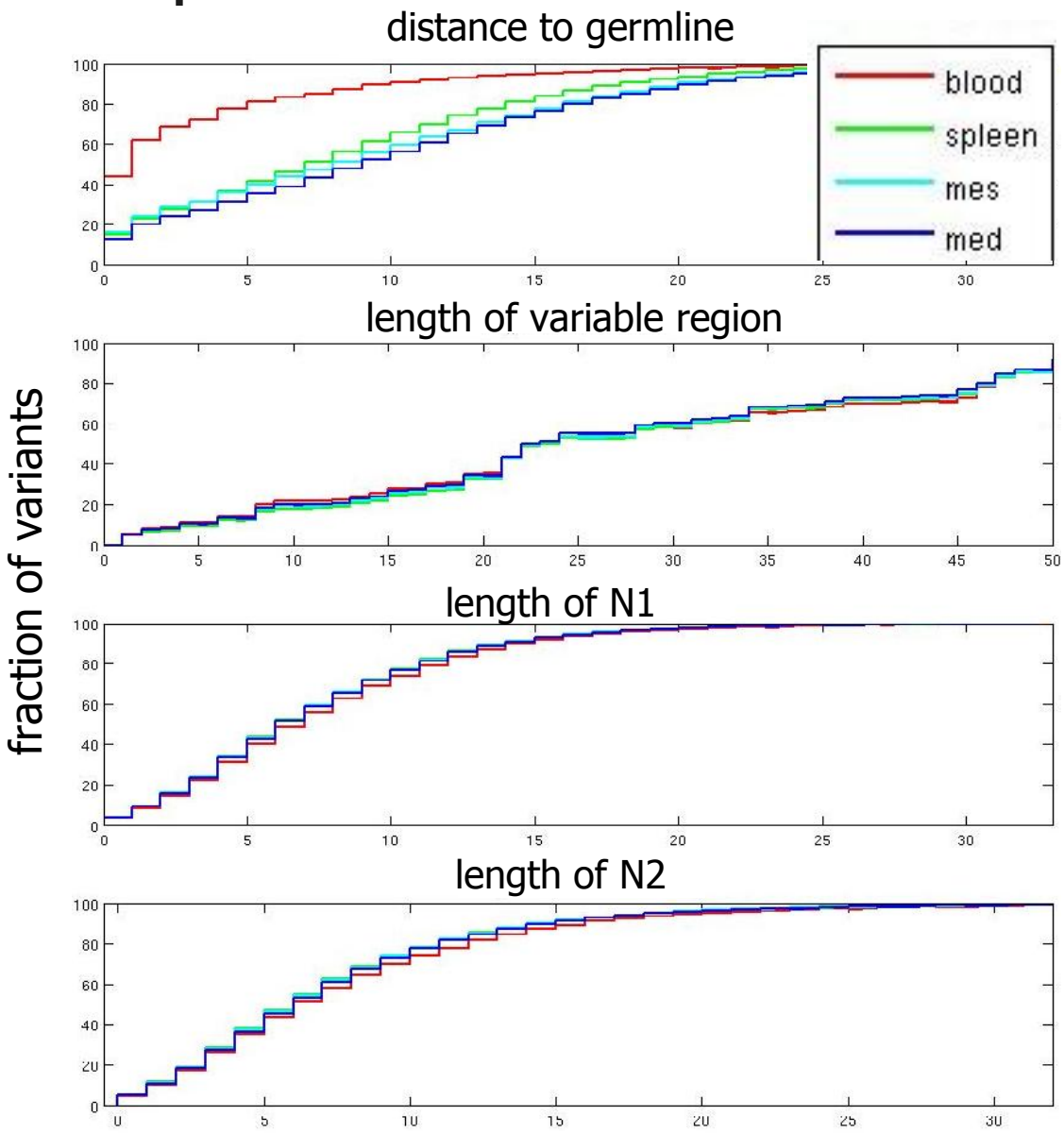
Repertoire Factorization

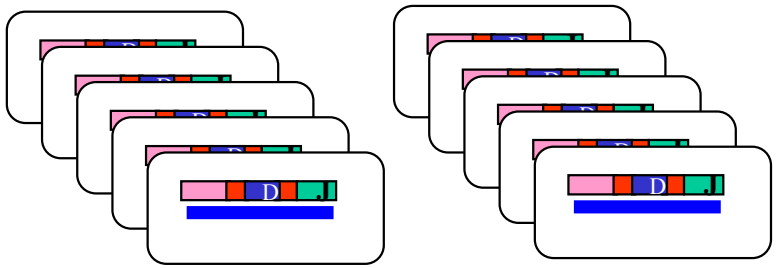


$$X(v,j,donor,\ source) = a(v,j,source) \times b(v,j,donor)$$

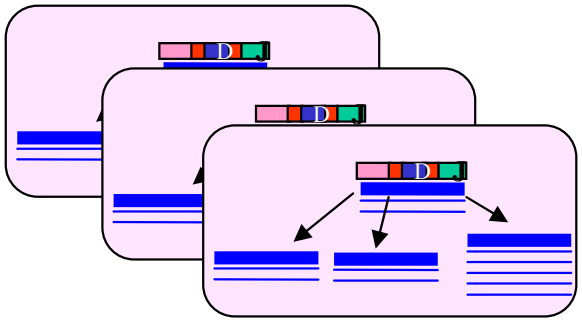


Properties Of Variants



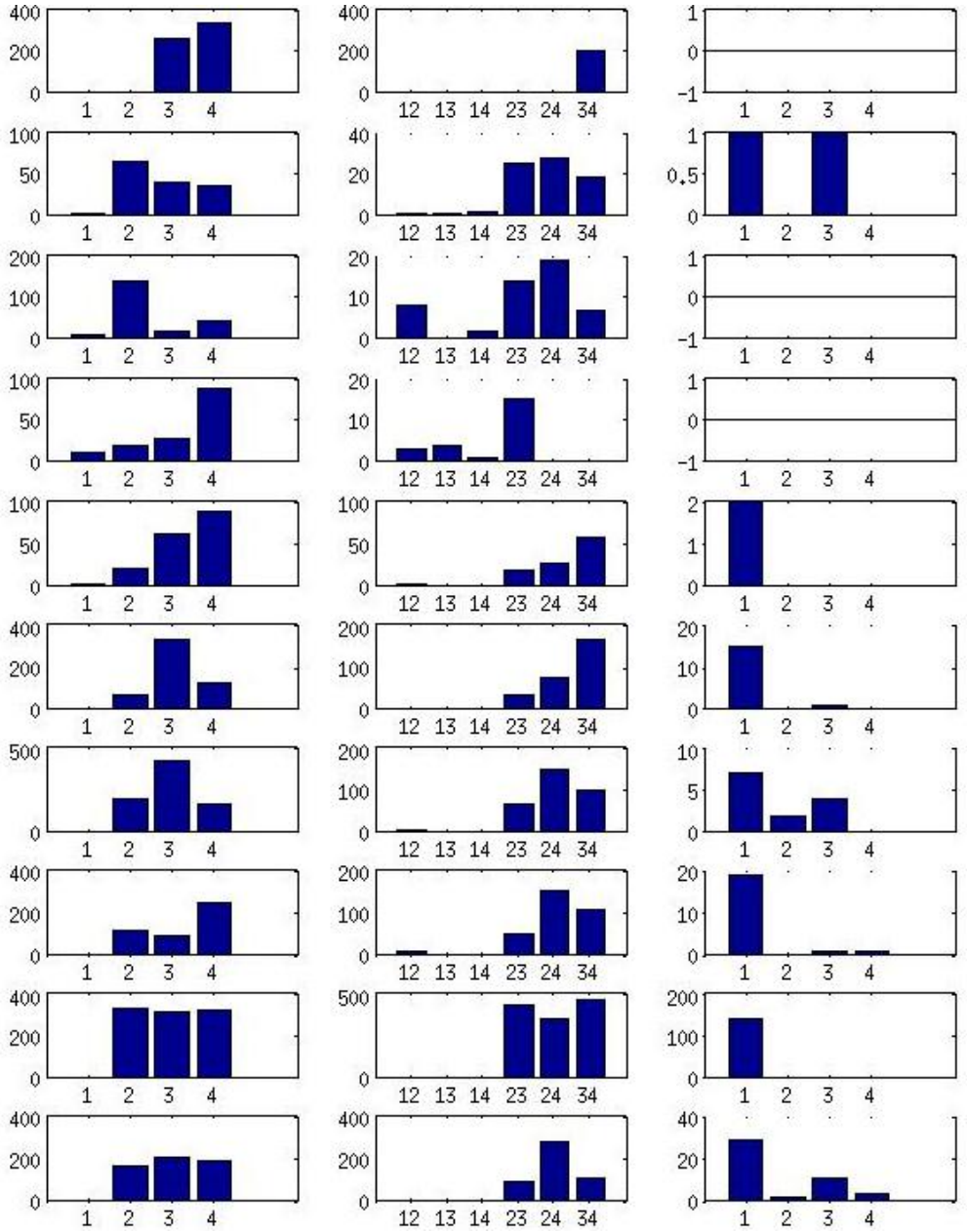


Clones (8,000)

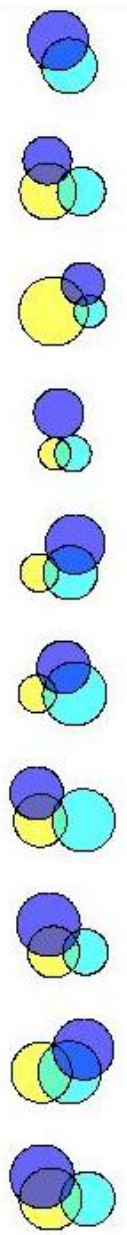


pure clones | shared 2 sources | shared 3 sources

patients



blood
spleen
Mesenteric
Mediastinal



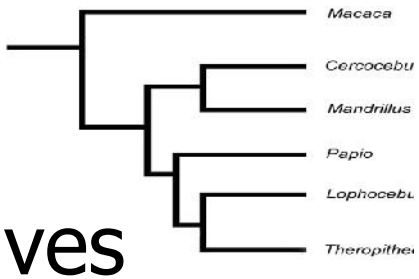
Evidence of clone migration

How do the trees look like?

Current Phylogenetic Trees

traditional phylo trees

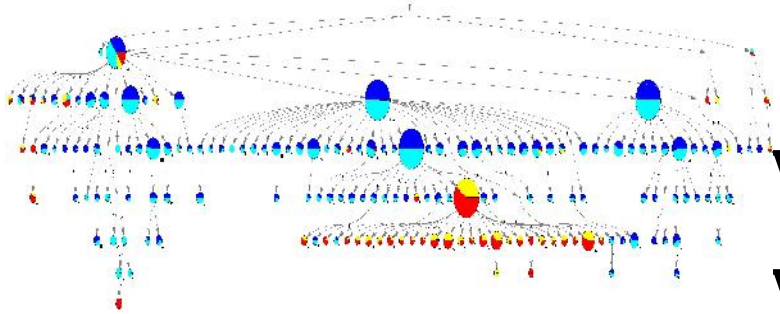
- X*100 sequences
- Weeks to run
- No sequencing/PCR errors
- Single solution



- Binary tree
- Data at the leaves
- Continuous mutation model

igForest

- X*10,000 √
- Hours √
- Noise model √
- Multiple/statistics √



√

√

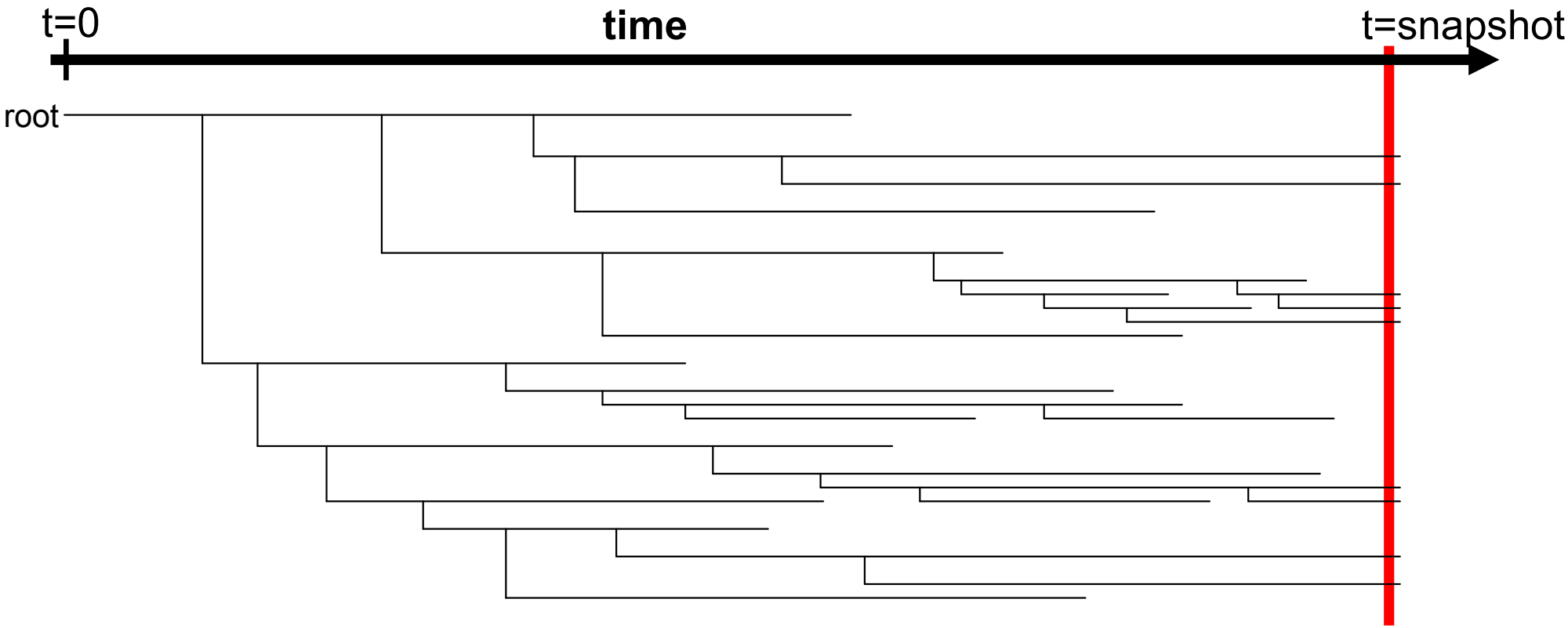
- Discrete √

Previous Work on B Cells

- ihmmune-align [Gaëta et al. 2007],
  SoDA2 [Munshaw and Kepler 2010]
  - examine each read independently, no tree
- igTree [Barak et al. 2008]
  - heuristic search, no high-throughput support
- Our Approach: ***igForest***
  - Joint **probabilistic model** for hypermutation *and* sequencing noise

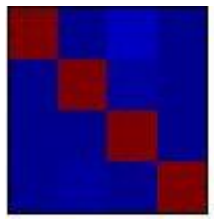
Generate Tree (Birth/death process)

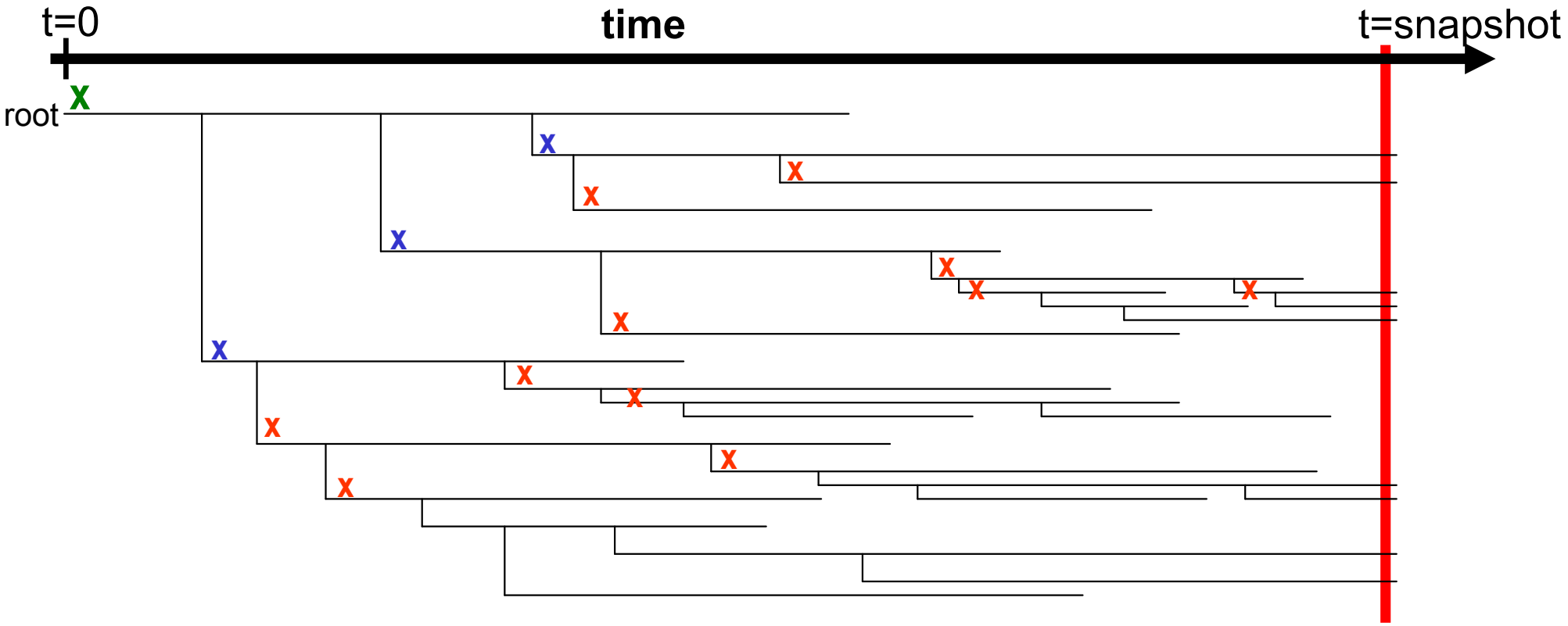
- **Start with one** live cell at *t=0*
- **Give birth** at rate λ
- **Die** at rate δ
- **Stop** when *t=* "snapshot time".



Generate Sequences

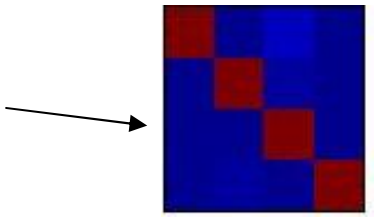
- **Start with** the inferred germline at the root
- **Generate each child sequence** from the sequence of its parent
  - Using a **Mutation Model**
- Continue so, **recursively**.

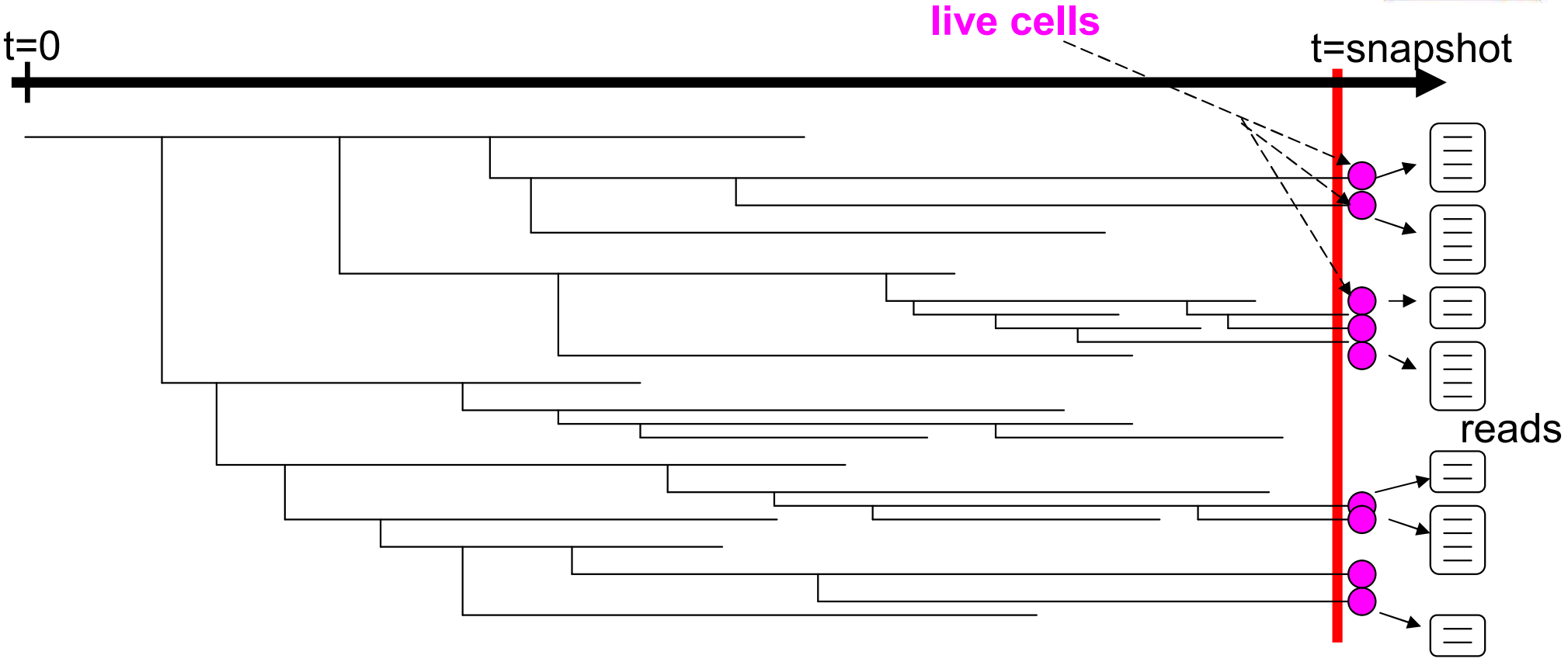




Generate Reads

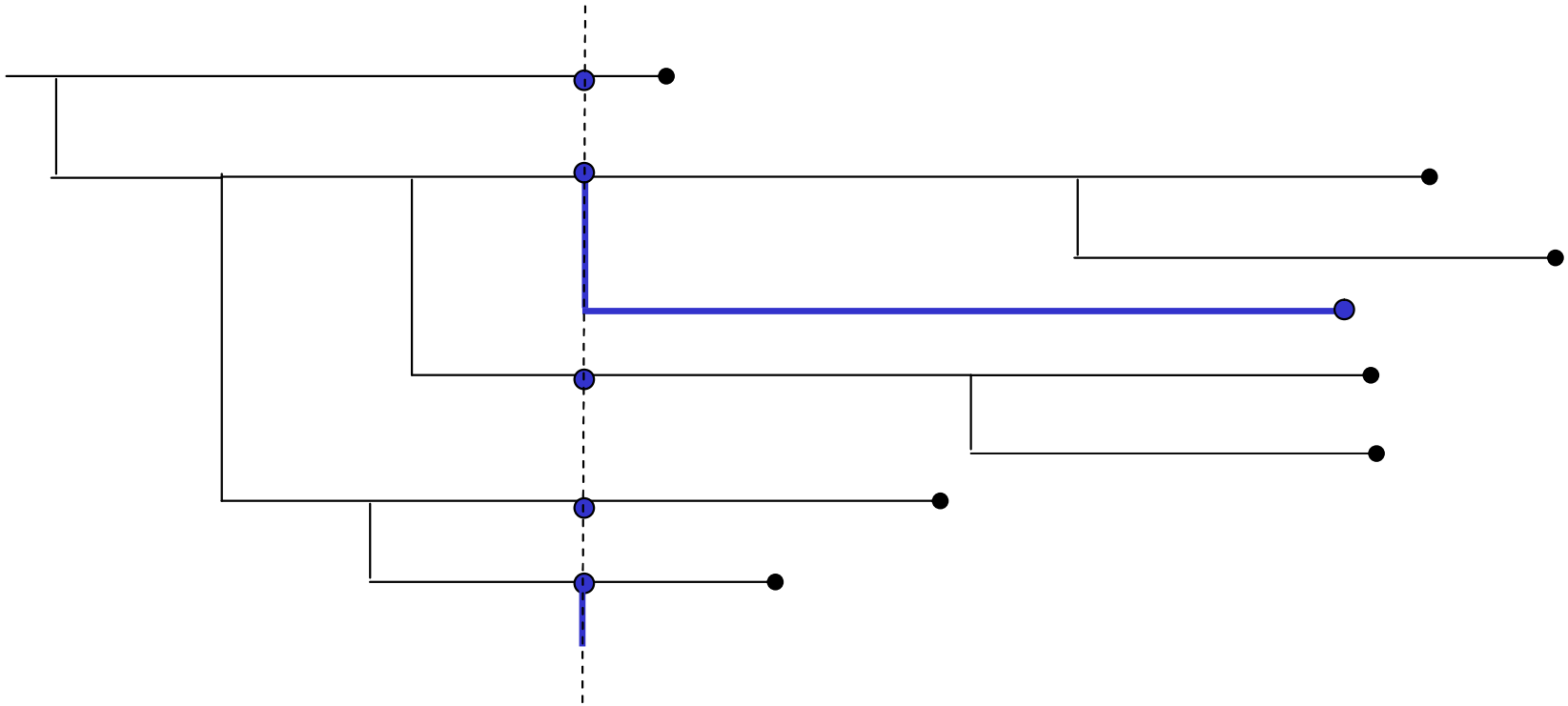
- **Total** N reads.
- **Assign** reads to **live cells** (uniformly).
- Copy read sequence from cell, based on **read model**.



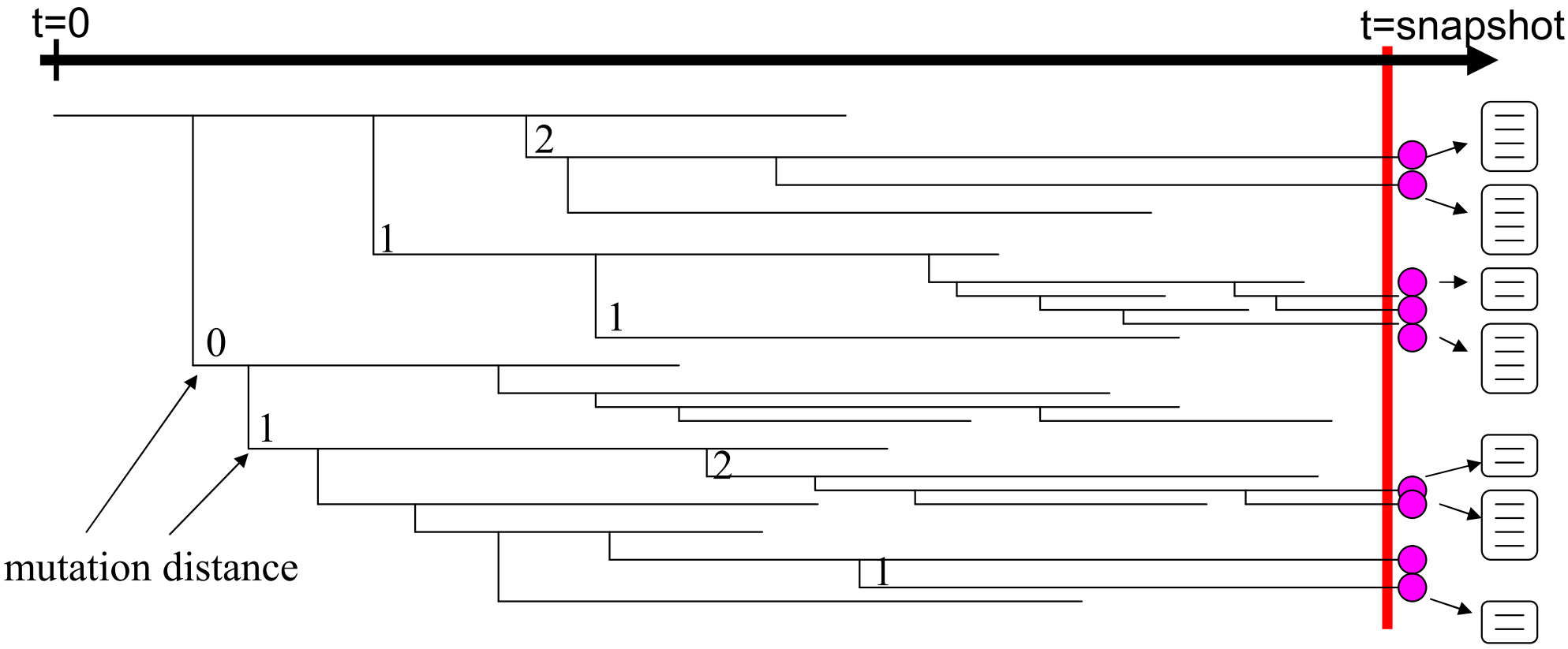


Inference

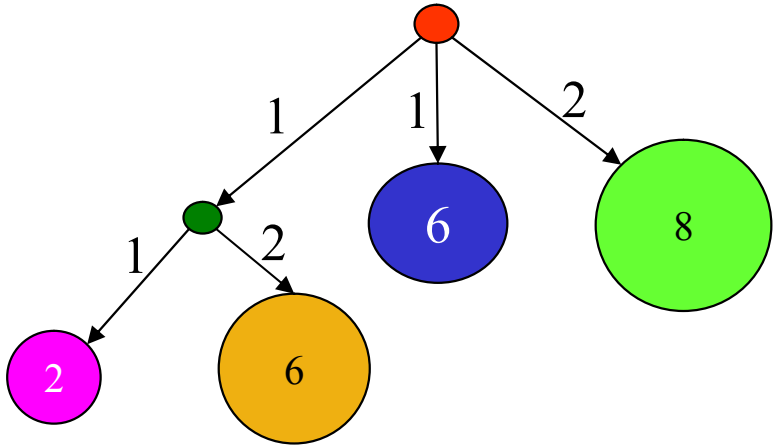
- MCMC
- Moves to **manipulate tree** and read assignments
- Generating samples of trees along the way

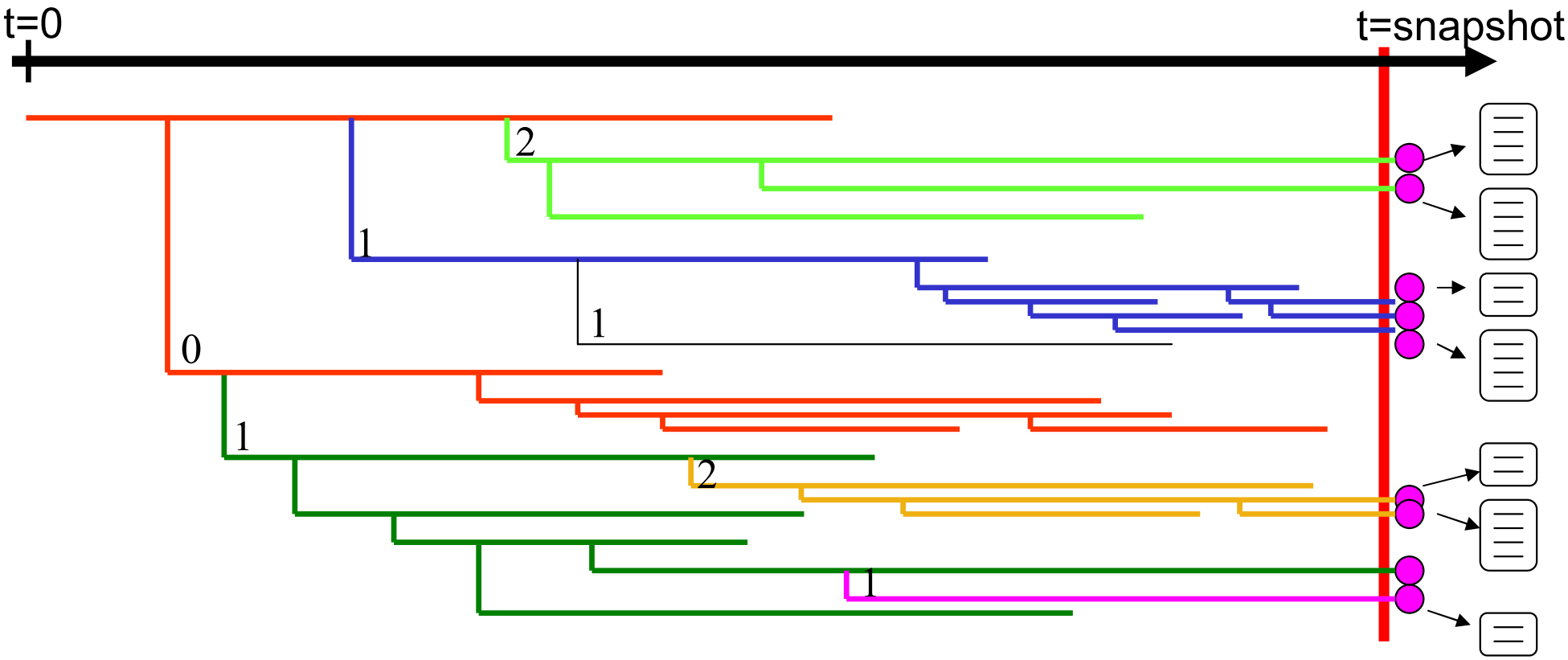


Output Tree

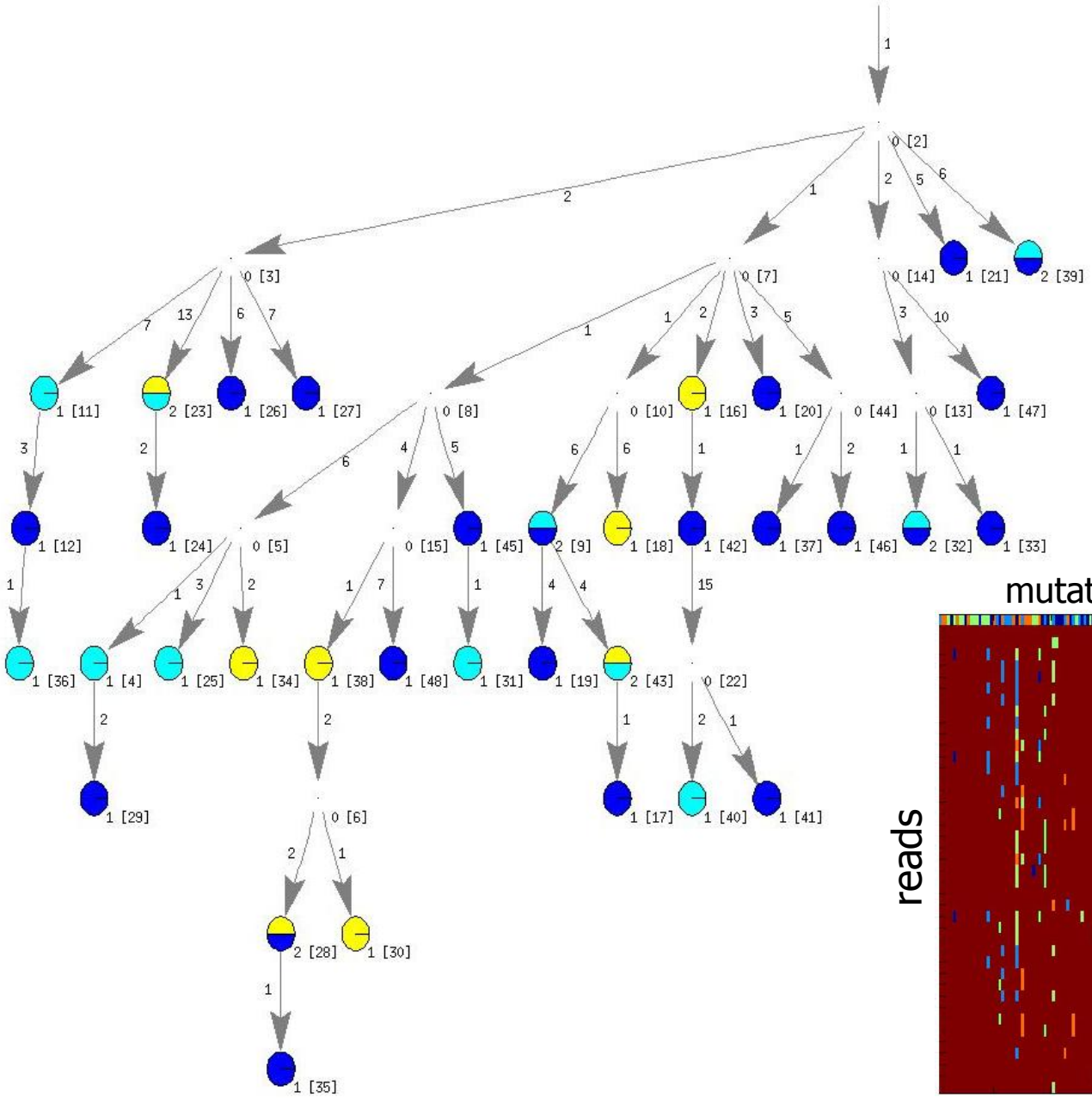


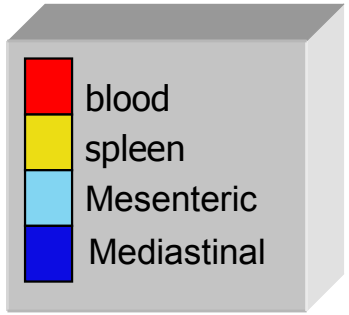
Output Tree



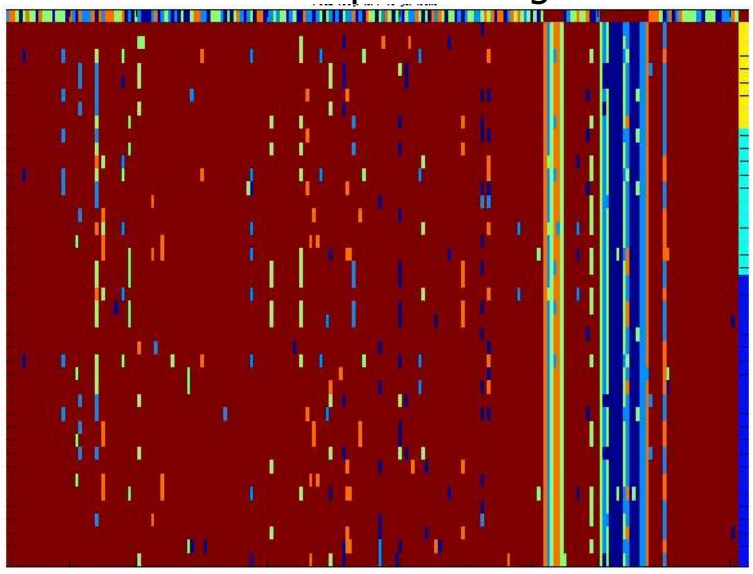


Sample Trees

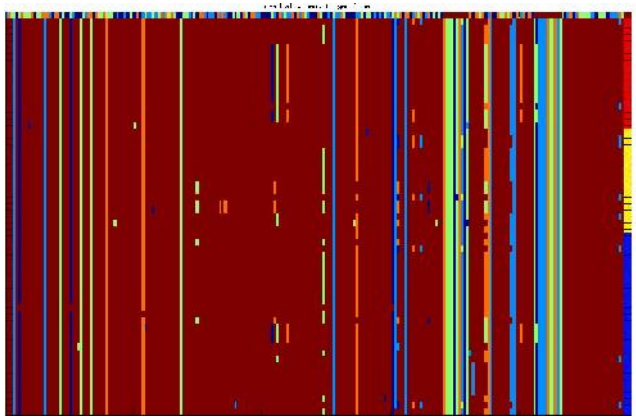


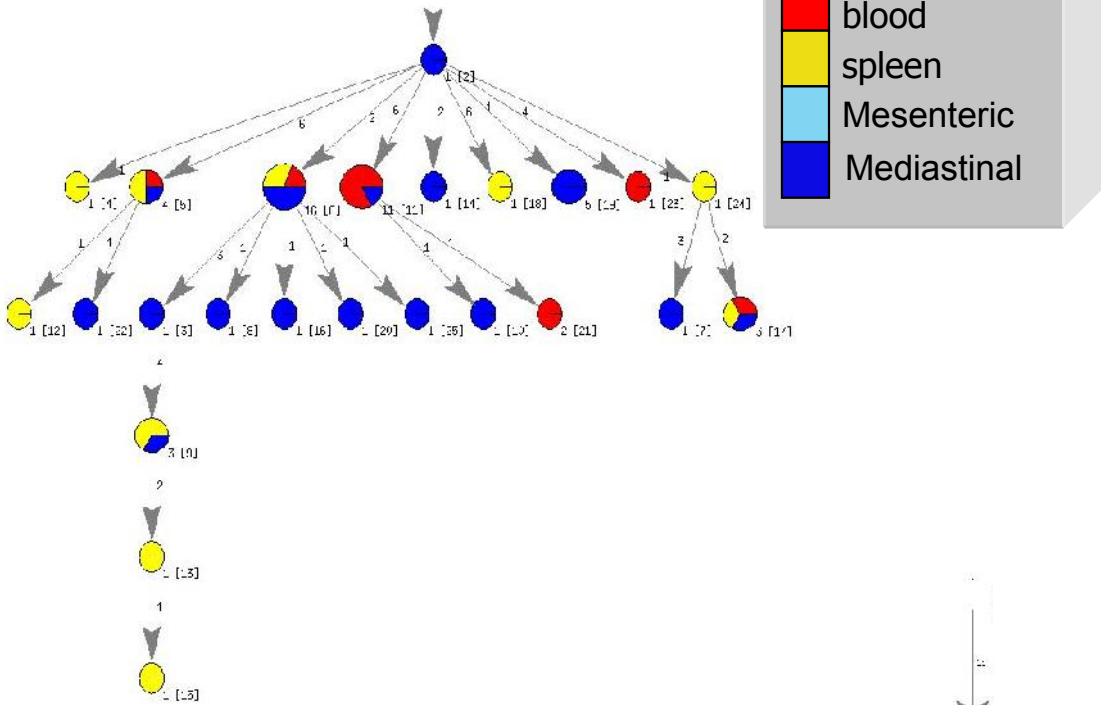


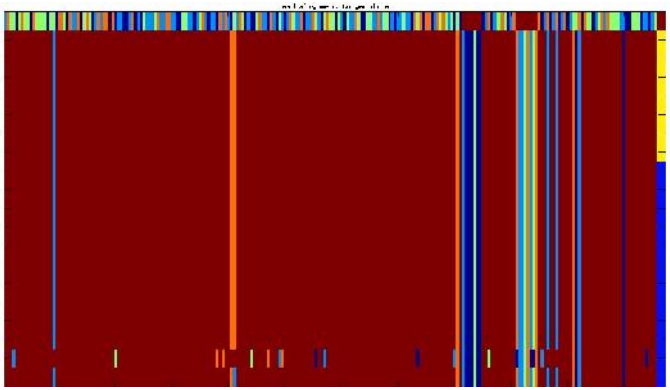
mutations compared to germline

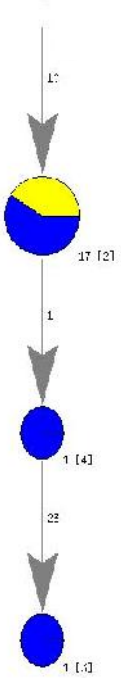


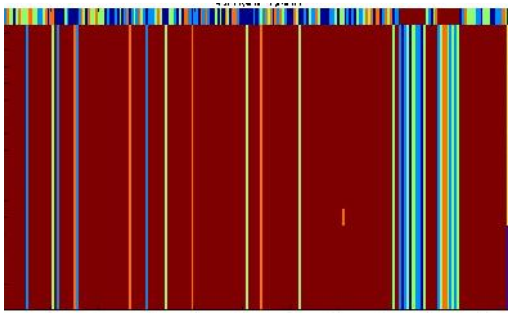
Sample Trees

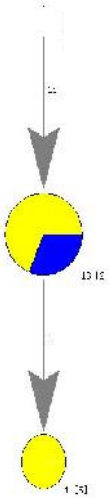


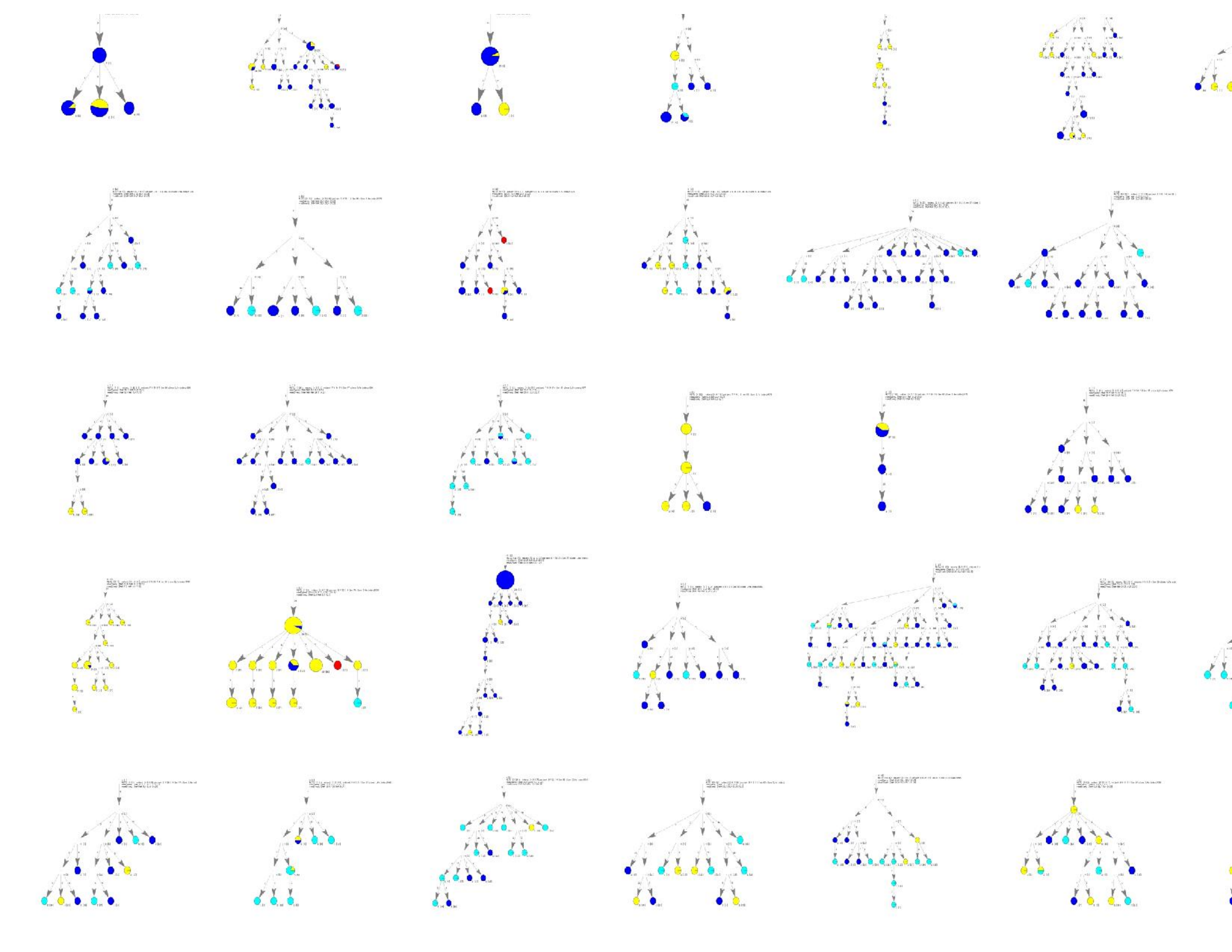




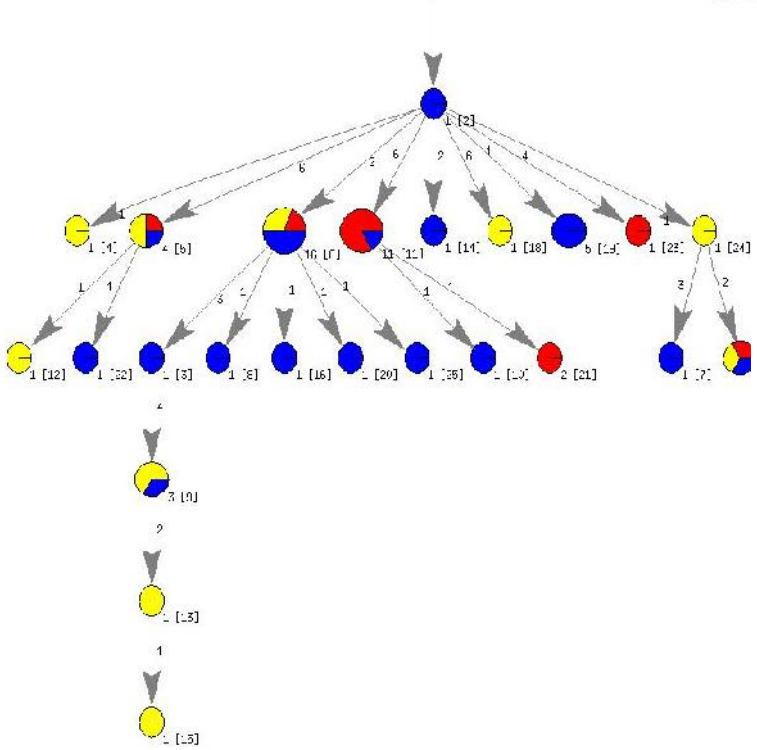


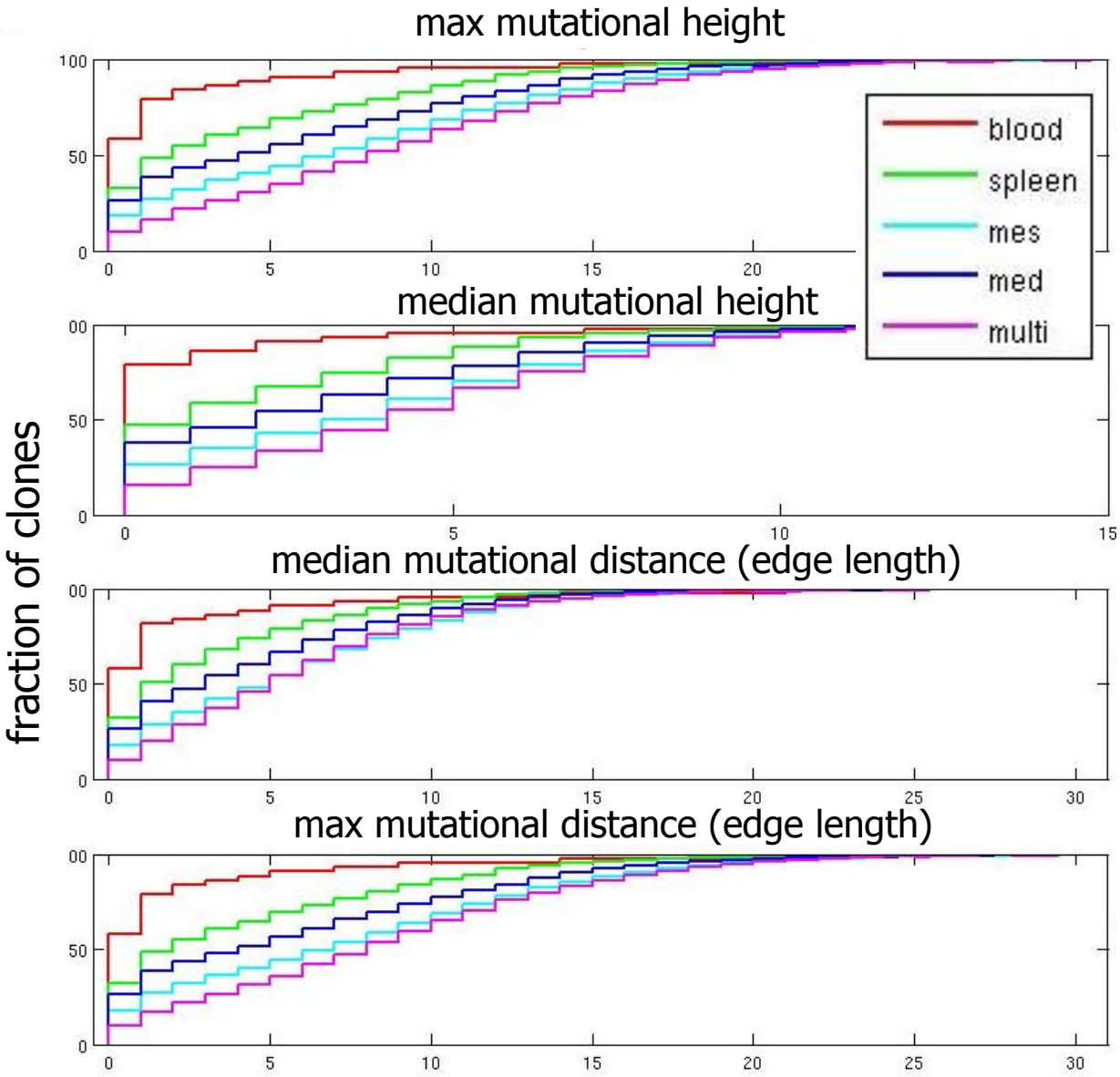






Properties Of Trees

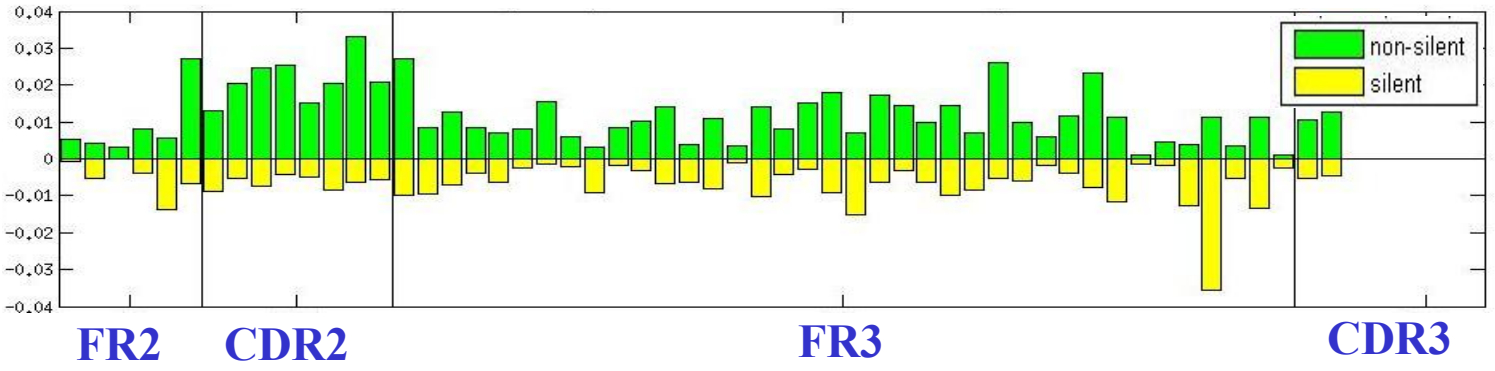




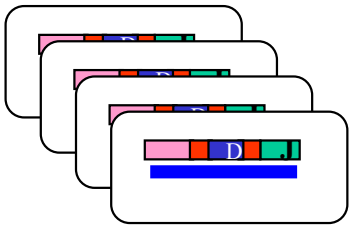
Mutation Analysis – V region

mutations binned by their aligned codon position

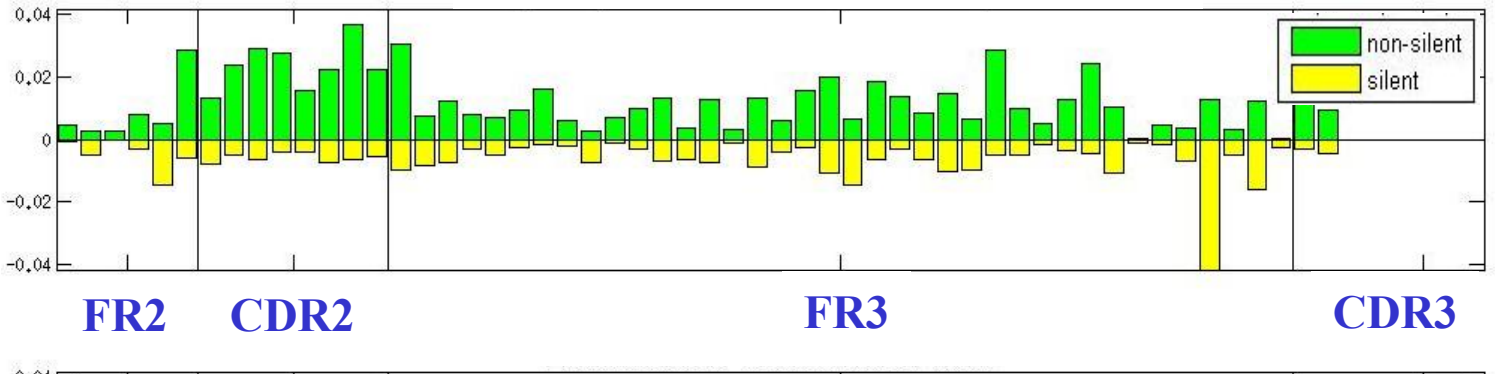
fraction of mutations



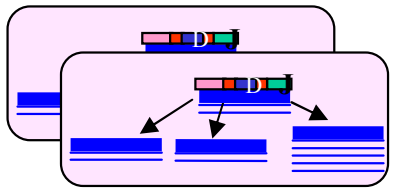
variants



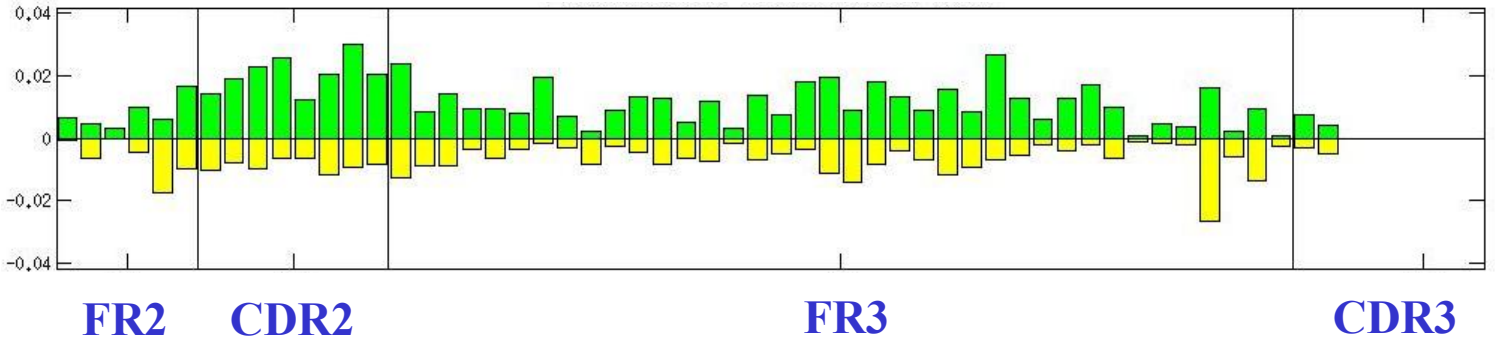
fraction of mutations



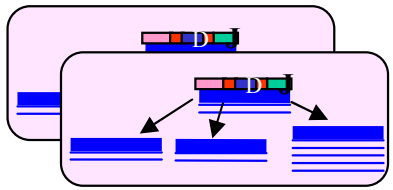
Clones



germline to root

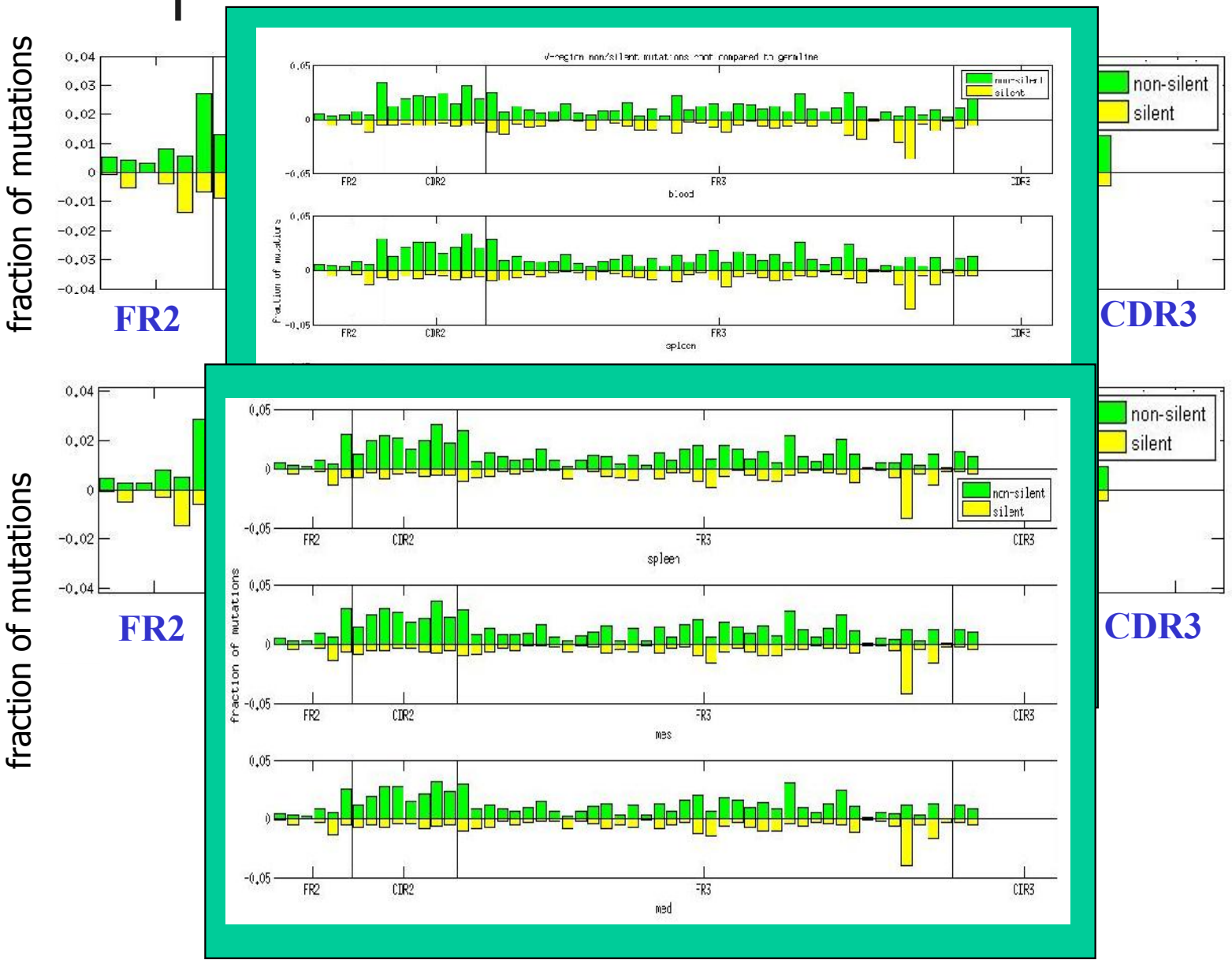


Clones

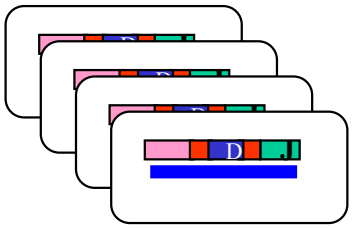


root down the tree

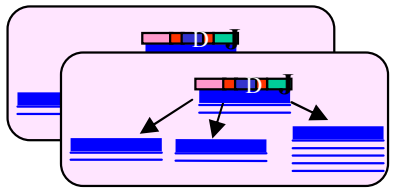
Mutation Analysis – V region



variants



Clones



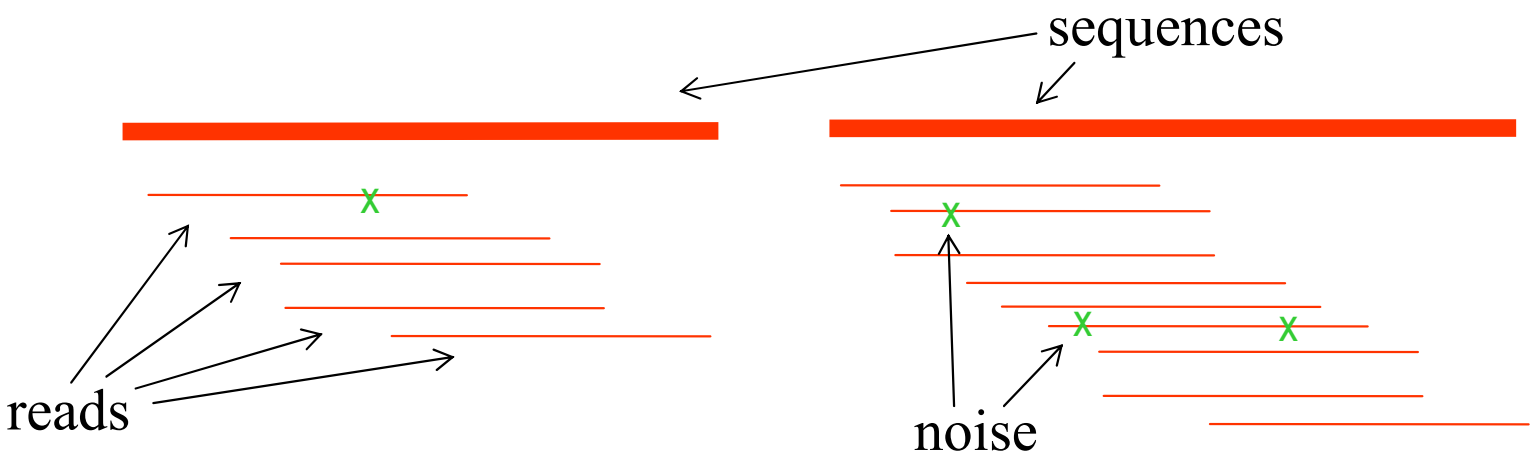
germline to root

Contributions

- Fully Bayesian phylogenetic analysis
  - O(10k) reads
  - Handles sequencing noise
- Big Picture view of the B cell repertoire
  - Intuitive clone description
- Insight on the Immune System
  - Evidence of clone migration
  - Differences between tissues
  - Deep hypermutation analysis
- Code (will be) freely available

Thank you!

- Breadth



- Depth