# Model selection in Bioinformatics: three short stories

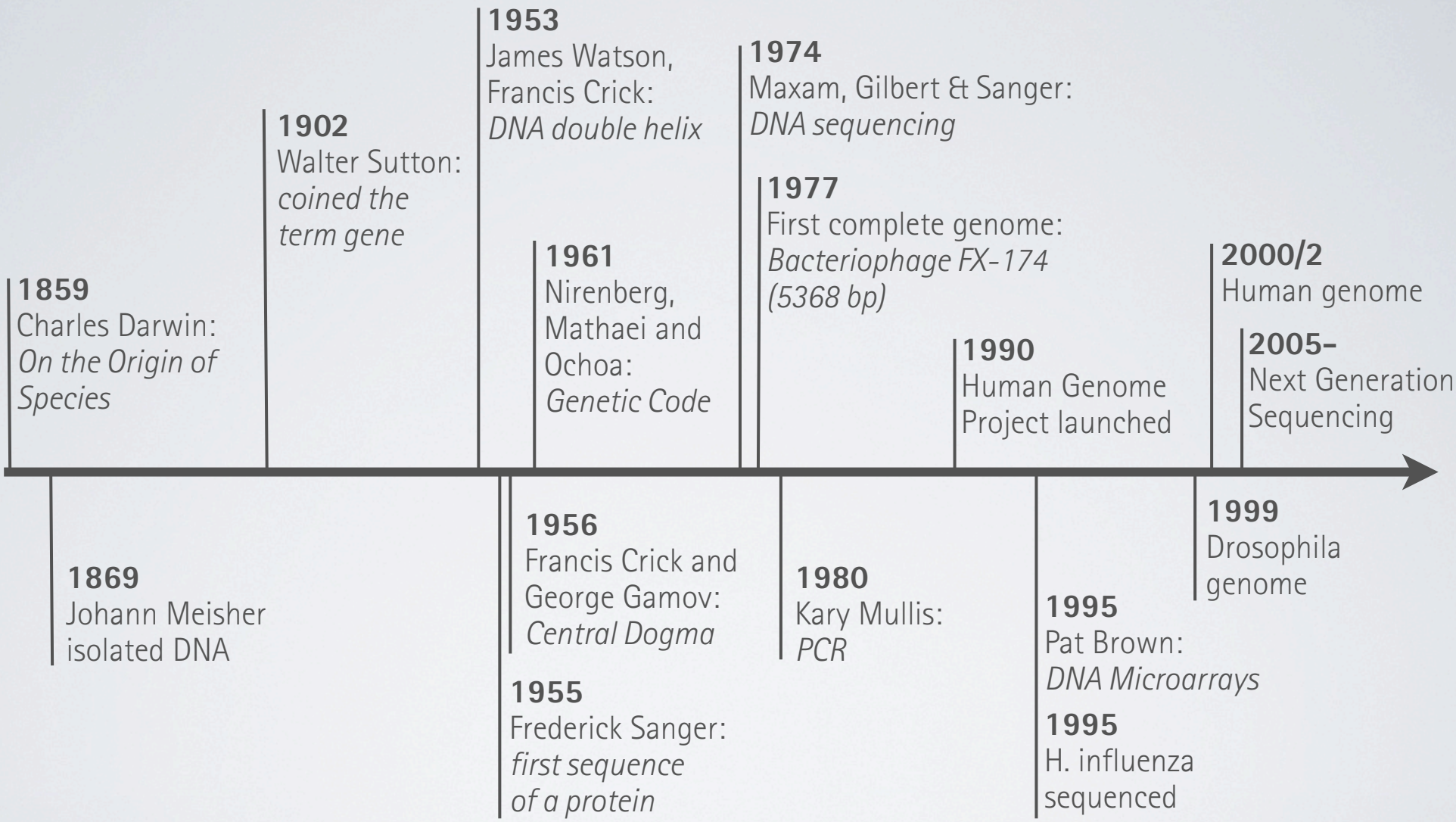
Alexander Schliep

Department of Computer Science
BioMaPS Institute for Quantitative Biology
Rutgers

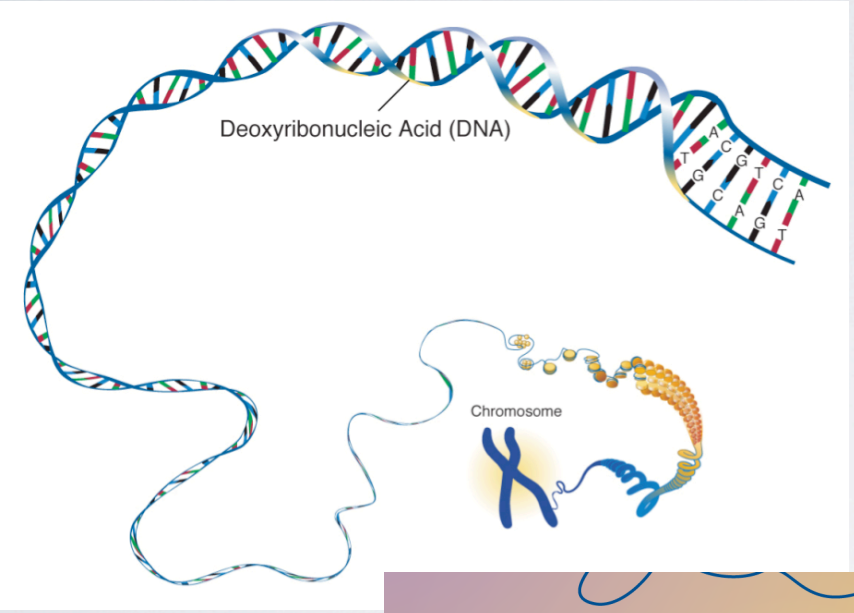
# Molecular Biology

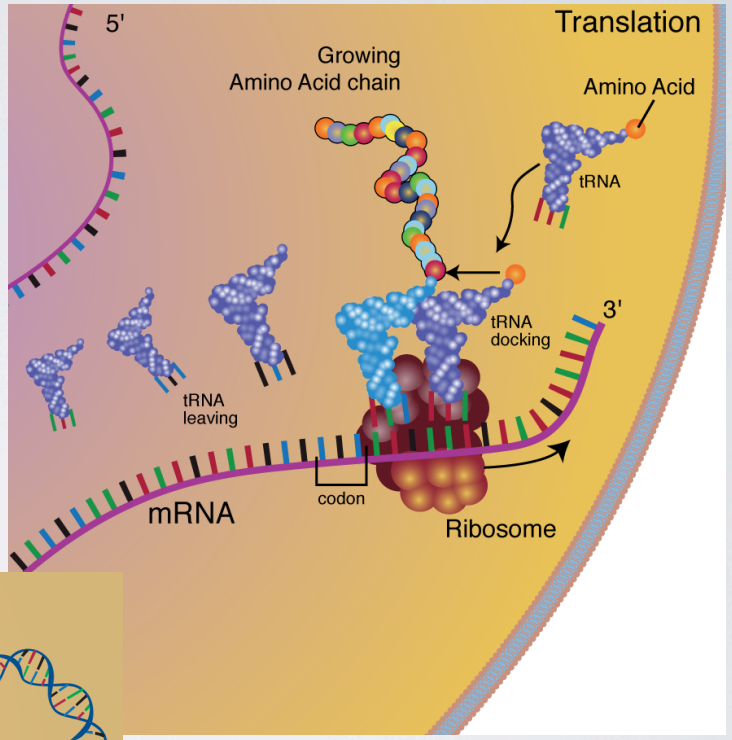
*Prolog*

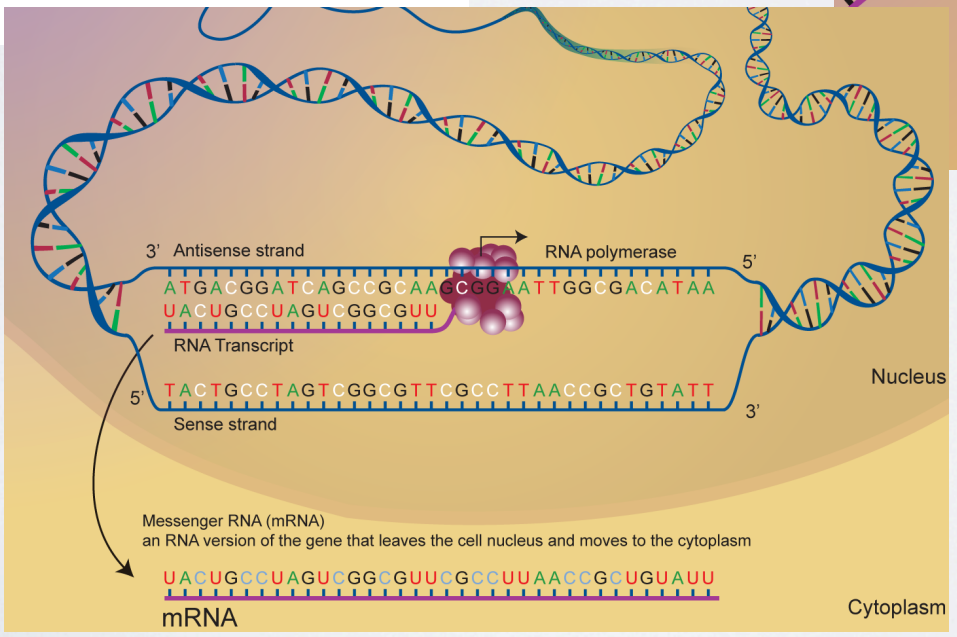
# Timeline of Molecular Biology



# Key players: DNA, RNA & Proteins




Illustrations from http://genome.gov

# Central Dogma (1956)

DNA ⇒ messenger RNA ⇒ Protein

**one gene** **one function**

# Success story: BRCA1/2

- Early onset, aggressive breast cancer (60% vs. 12% life-time risk for breast, 15-40% vs. 1.4% for ovarian cancer)
- Linkage analysis on pedigrees: genetic variations & approximate chromosomal positions

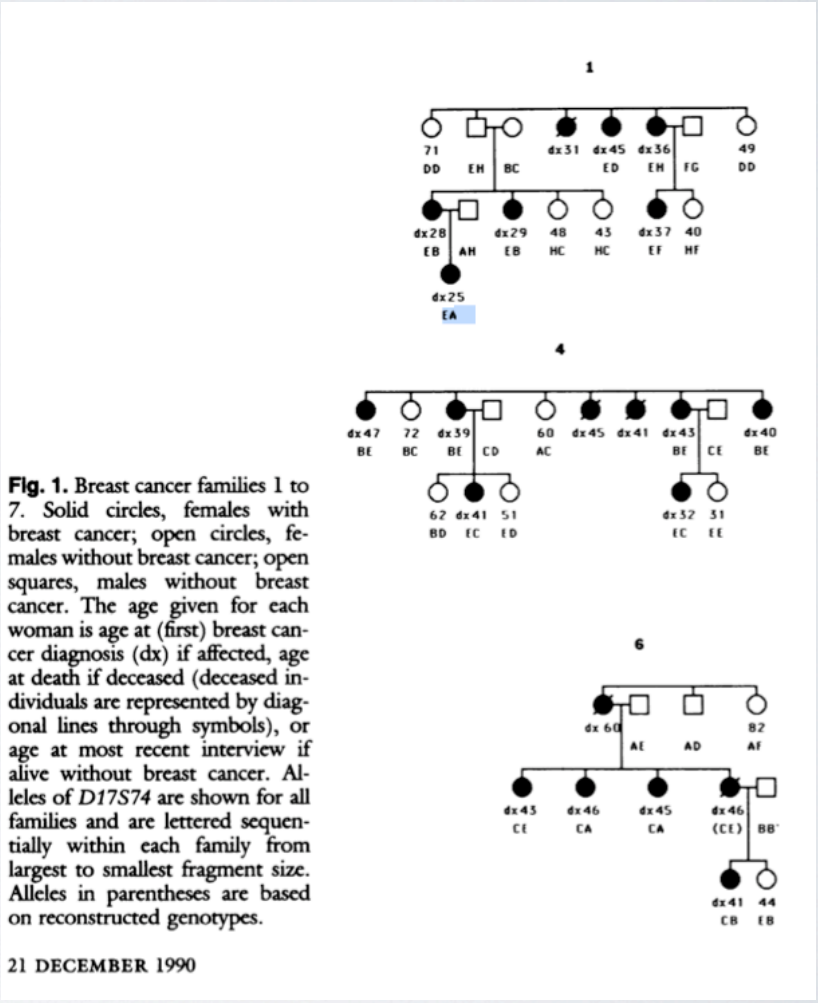


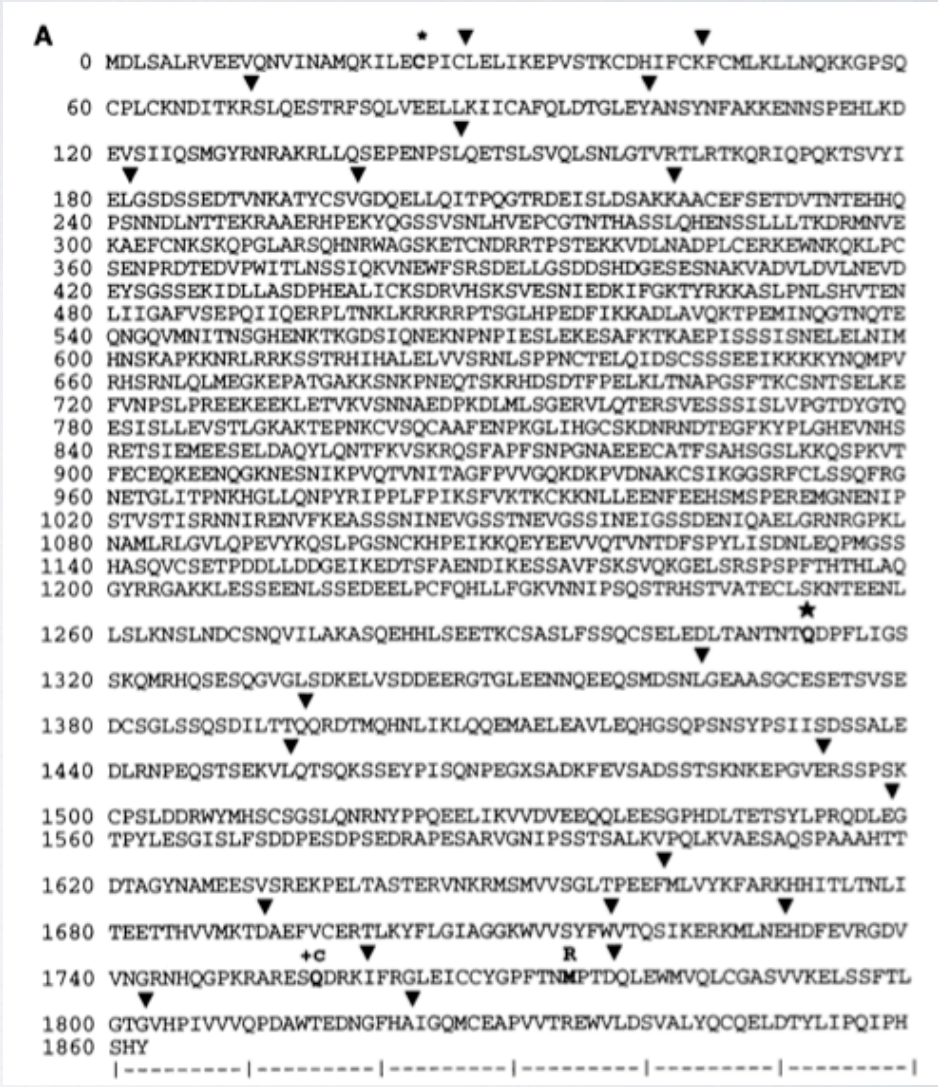

**Fig. 1.** Breast cancer families 1 to 7. Solid circles, females with breast cancer; open circles, females without breast cancer; open squares, males without breast cancer. The age given for each woman is age at (first) breast cancer diagnosis (dx) if affected, age at death if deceased (deceased individuals are represented by diagonal lines through symbols), or age at most recent interview if alive without breast cancer. Alleles of *D17S74* are shown for all families and are lettered sequentially within each family from largest to smallest fragment size. Alleles in parentheses are based on reconstructed genotypes.

21 DECEMBER 1990

Hall et al., Science 1990; Miki et al., Science 1994

# Success story: BRCA1/2

- Sequence of the Breast Cancer type 1 susceptibility protein
- BRCA 1/2 alert cells to DNA damage, initiating DNA repair or cell death



# Hope for Human Genome Project

Understand disease (and biology) by

- identifying all genes,
- investigating the proteins each gene codes for, and
- relating genetic variations to phenotypes.

# Number of Genes?

| Organism | Genome Size | #Genes |
|---|---|---|
| C. elegans | 100 Mbp | 19,000 |
| Drosophila | 123 Mbp | 17,000 |
| H. sapiens (estimates) | 3,080 Mbp | >> 100,000 |
| H. sapiens | | ~21,000 |

# Today's Picture

DNA modifications

transcriptional regulation

DNA ⇒ messenger RNA ⇒ Protein

multiple genes

post-transcriptional regulation

alternative splicing

multiple proteins

protein complexes

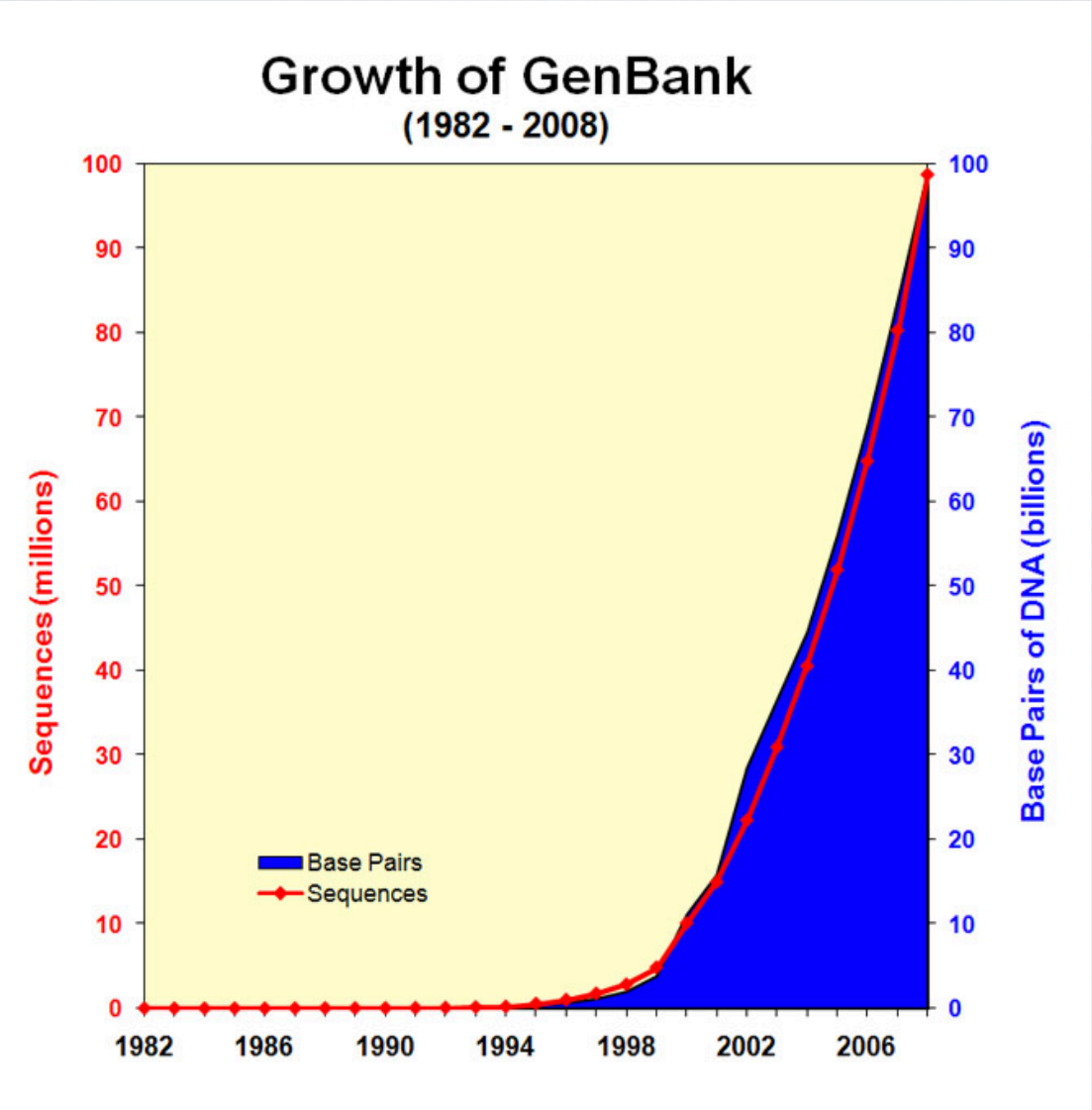
... a specific function in a specific context

# Regulation of Gene Expression

Where, when, why, are which proteins

produced from a gene?

# Amount of Data




http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html

# Types of Data



- Genbank: DNA Sequences
  started 1979 at LANL; today at NCBI/NIH
- PDB: Protein Structures
  started 1971 at Brookhaven; today at Rutgers
- Today: thousands of other data bases for specific types of data, organisms, diseases etc.
- NAR publishes ~100 novel data bases in its yearly database issue

# Workflow

- Data mining produces *interesting* observations
- Formulate hypothesis
- Biologists test hypothesis

# Workflow

- Data mining produces *interesting* observations
- Formulate hypothesis
- Biologists test hypothesis

**The selected model is (part of) the hypothesis**

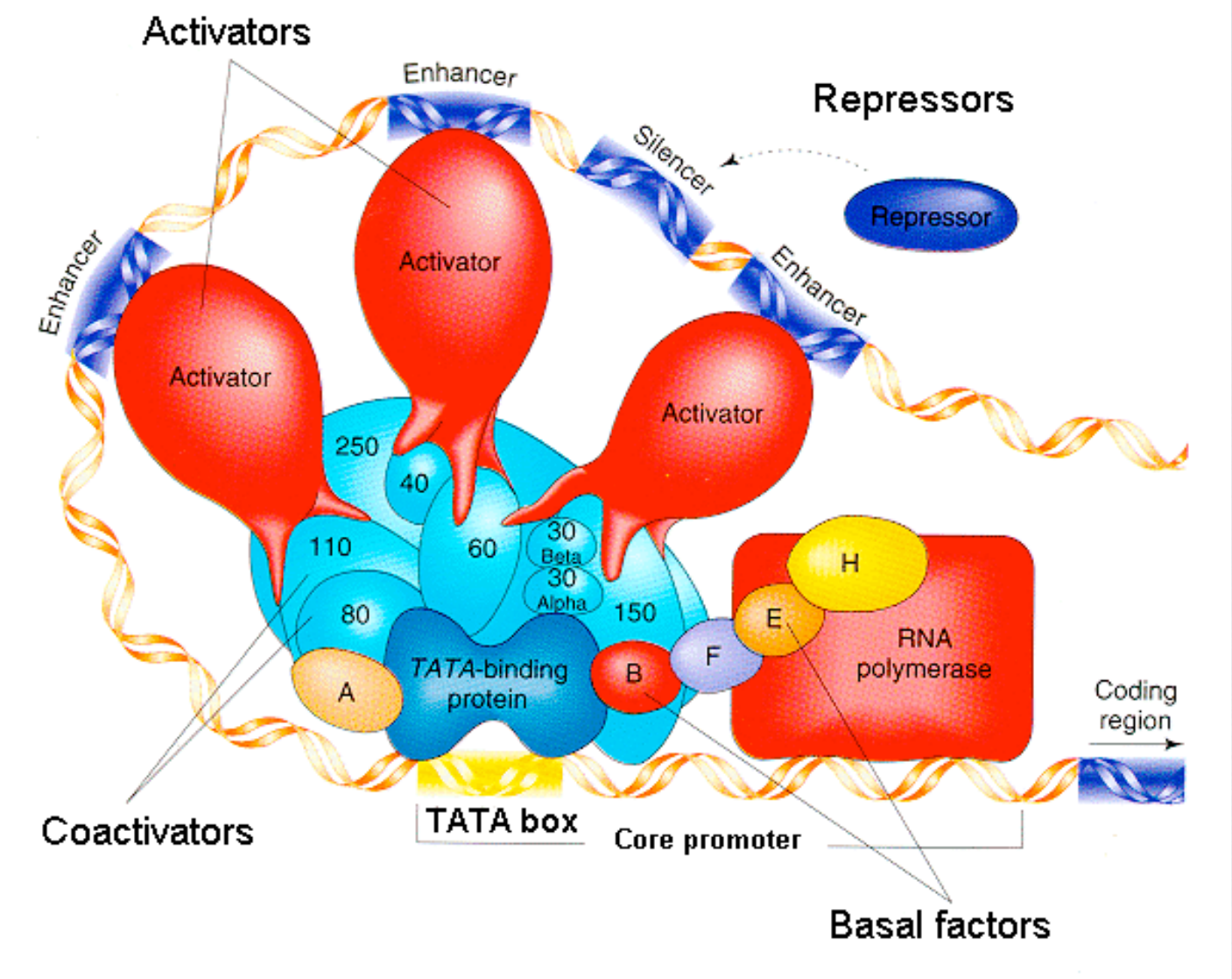
# Problem scale

- Individual problems are large (dimension and/or samples)
- Many problems
  - E.g. Databases with HMM for protein families
  - Personalized Medicine
- Frequent updates

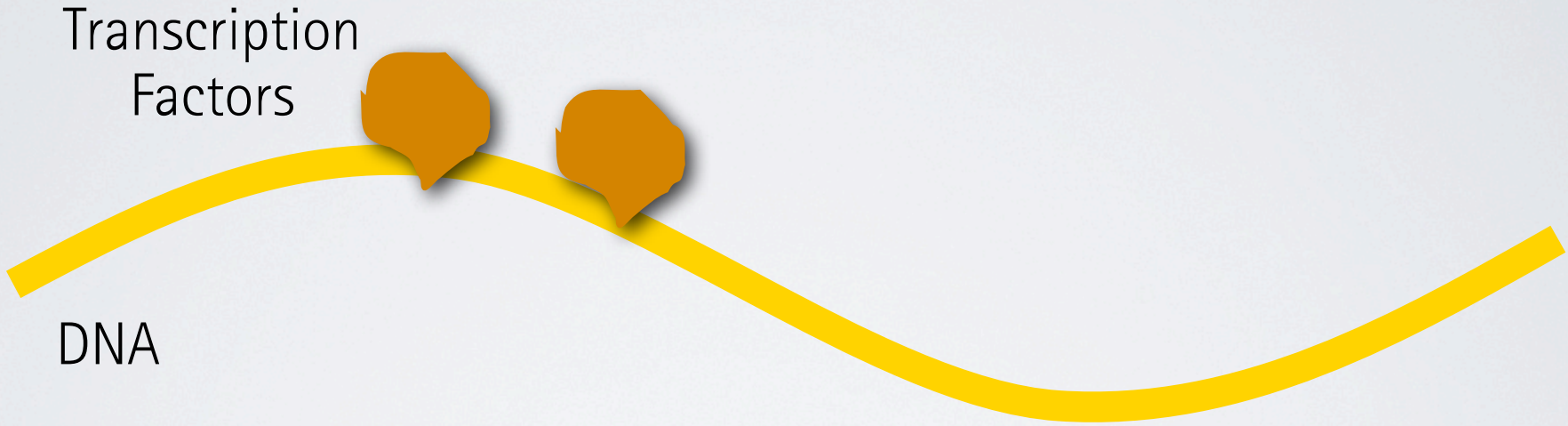
# Transcription factors & CSI models

*Story #1*

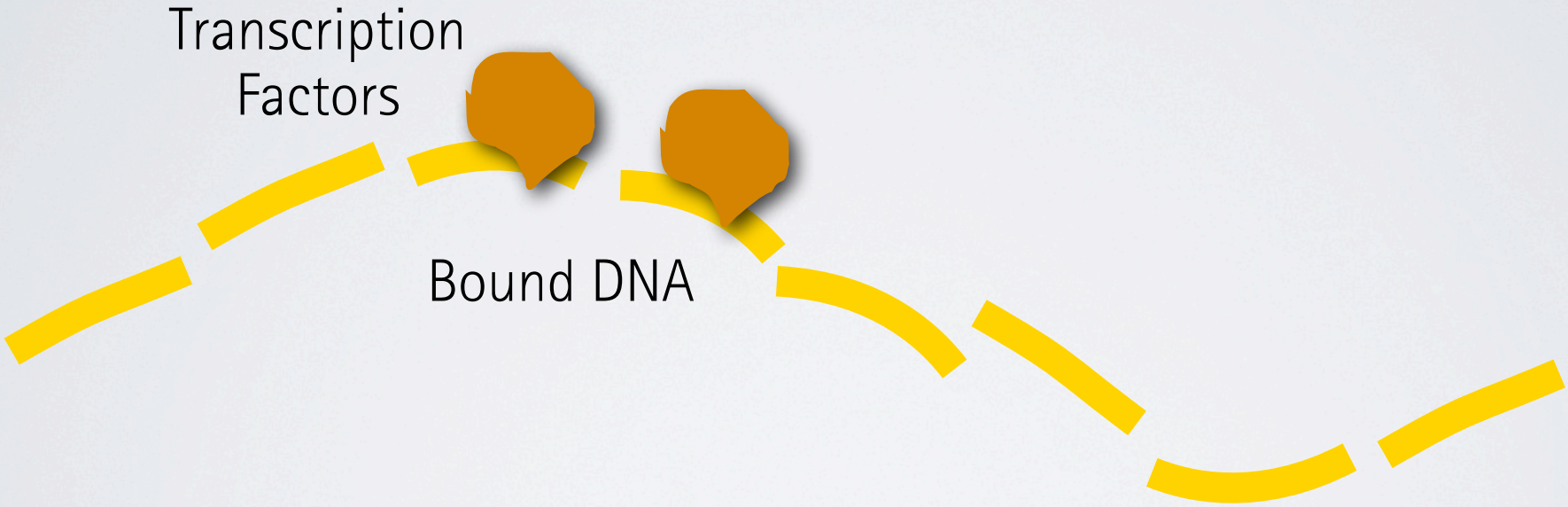
# Transcription factors



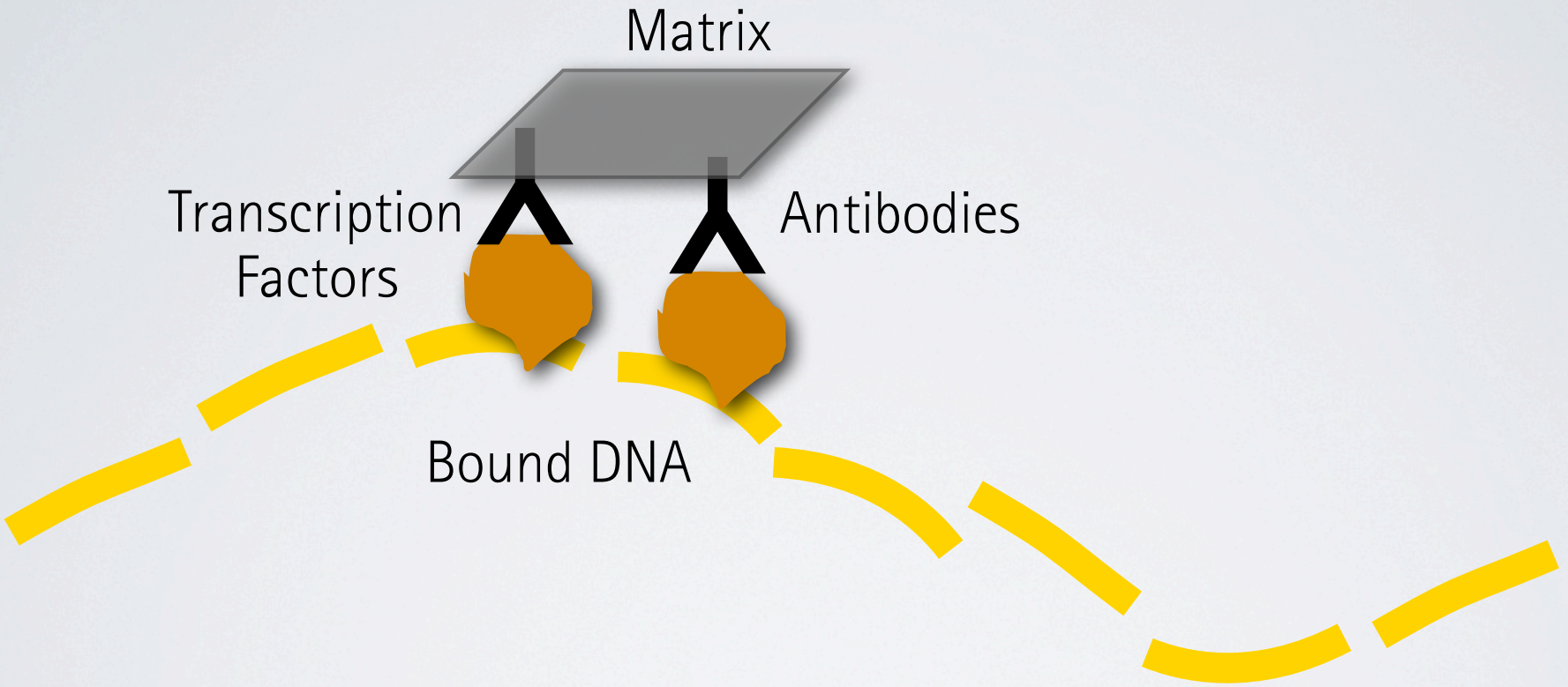
# Experimental setup



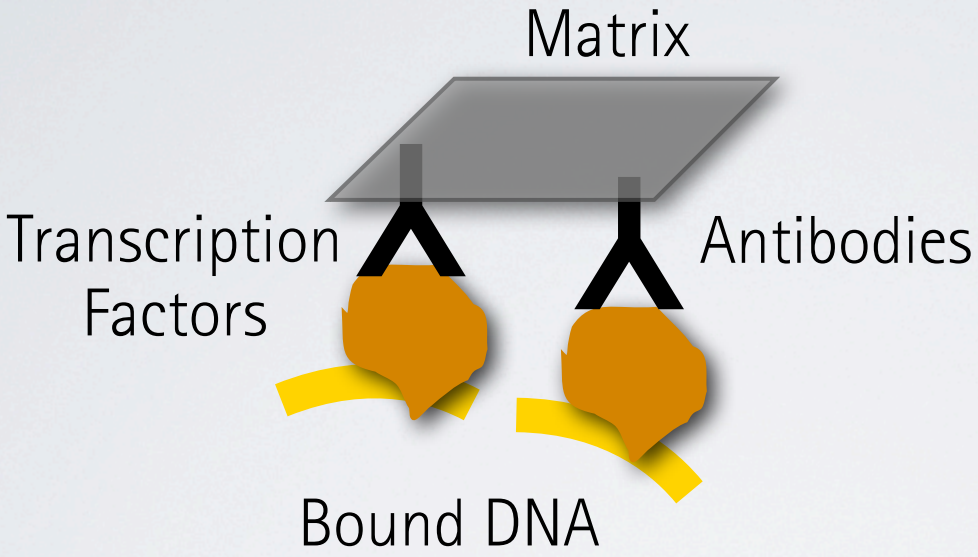
# Experimental setup



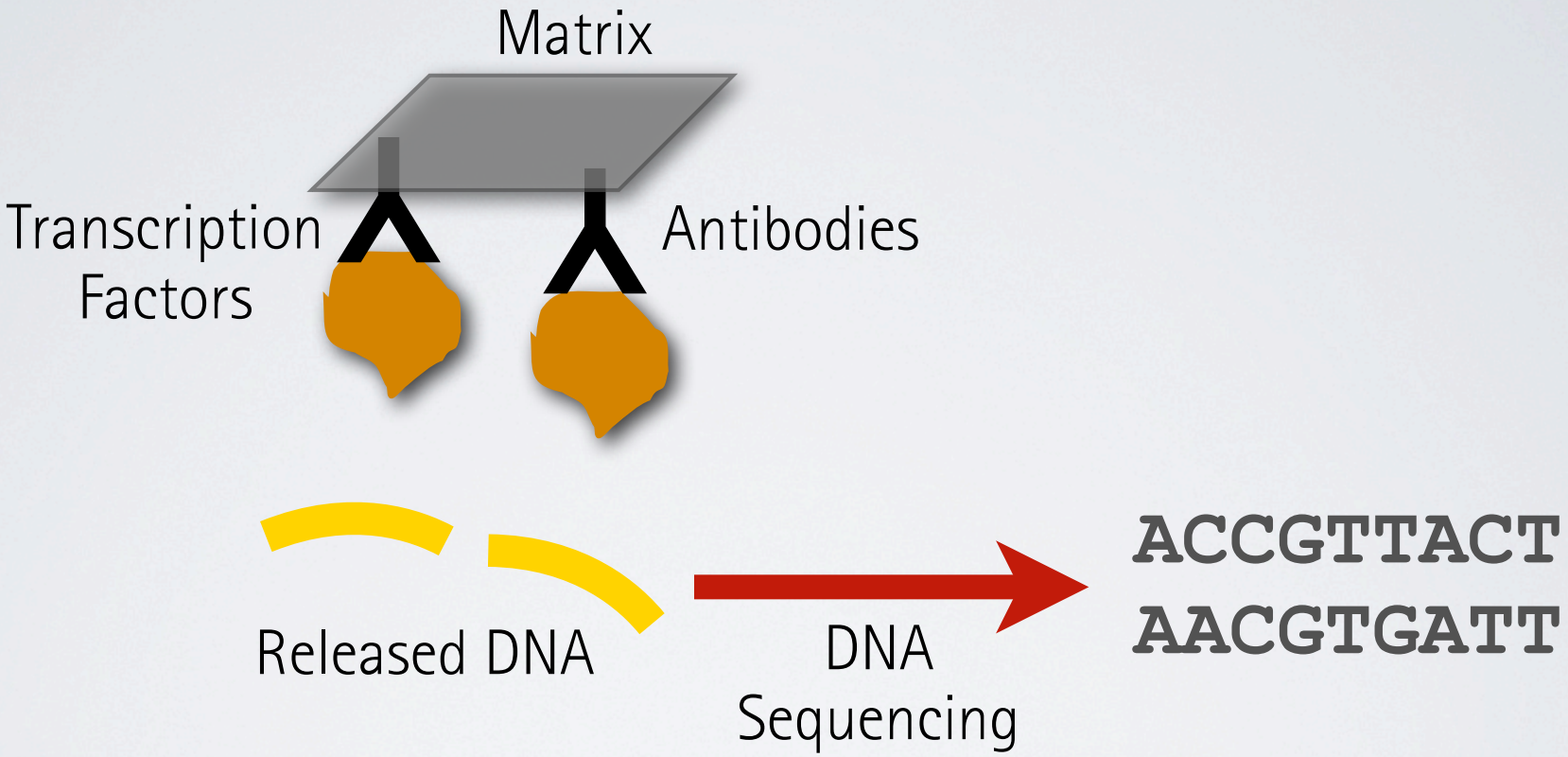
# Experimental setup



# Experimental setup



# Experimental setup

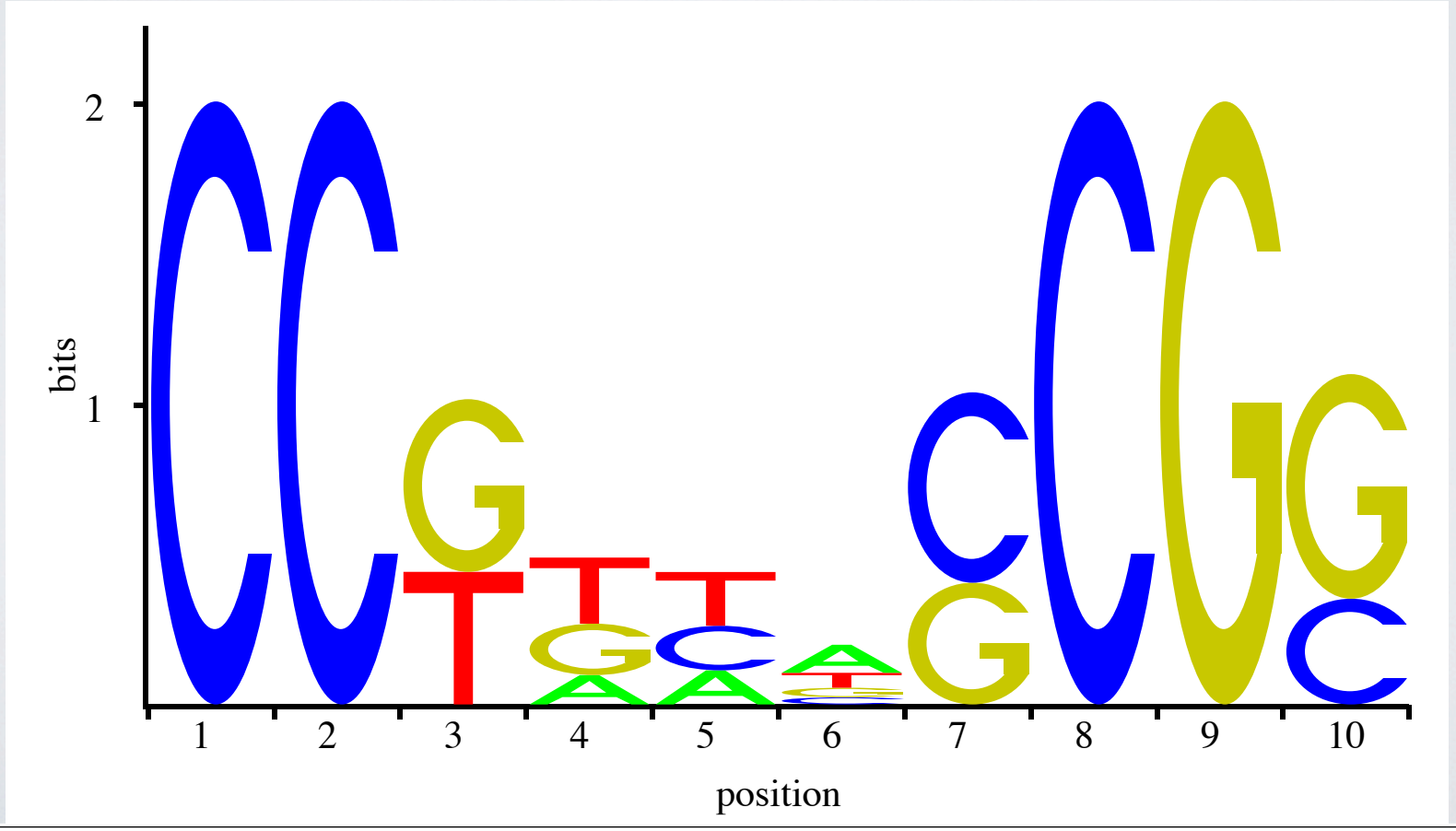


# Example Leu3

| C | C | G | G | T | A | C | C | G | G | Consensus |
|---|---|---|---|---|---|---|---|---|---|---|
| G | – | – | – | A | – | A | – | – | – | |
| – | – | – | T | C | – | A | – | – | – | |
| G | – | T | – | A | – | – | – | – | – | |
| – | – | – | C | C | – | – | – | – | – | |
| – | – | – | T | C | – | – | – | – | C | |
| – | – | – | – | G | G | – | – | – | – | |
| – | – | T | A | C | – | – | – | – | – | |
| – | – | – | A | A | – | T | – | – | – | |
| – | – | T | – | A | – | A | – | – | – | |
| ... | | | | | | | | | | |
| – | – | – | C | G | G | G | – | – | – | |
| G | – | – | – | A | T | G | – | – | – | |

# Leu3 Sequence Logo

```
A  [  0   0   0  20  26  47   0   0   0   0 ]
C  [100 100   0   0  34  12  61 100   0  32 ]
G  [  0   0  56  35   0  16  39   0 100  68 ]
T  [  0   0  43  45  41  25   0   0   0   0 ]
```



# Naive Bayes mixture

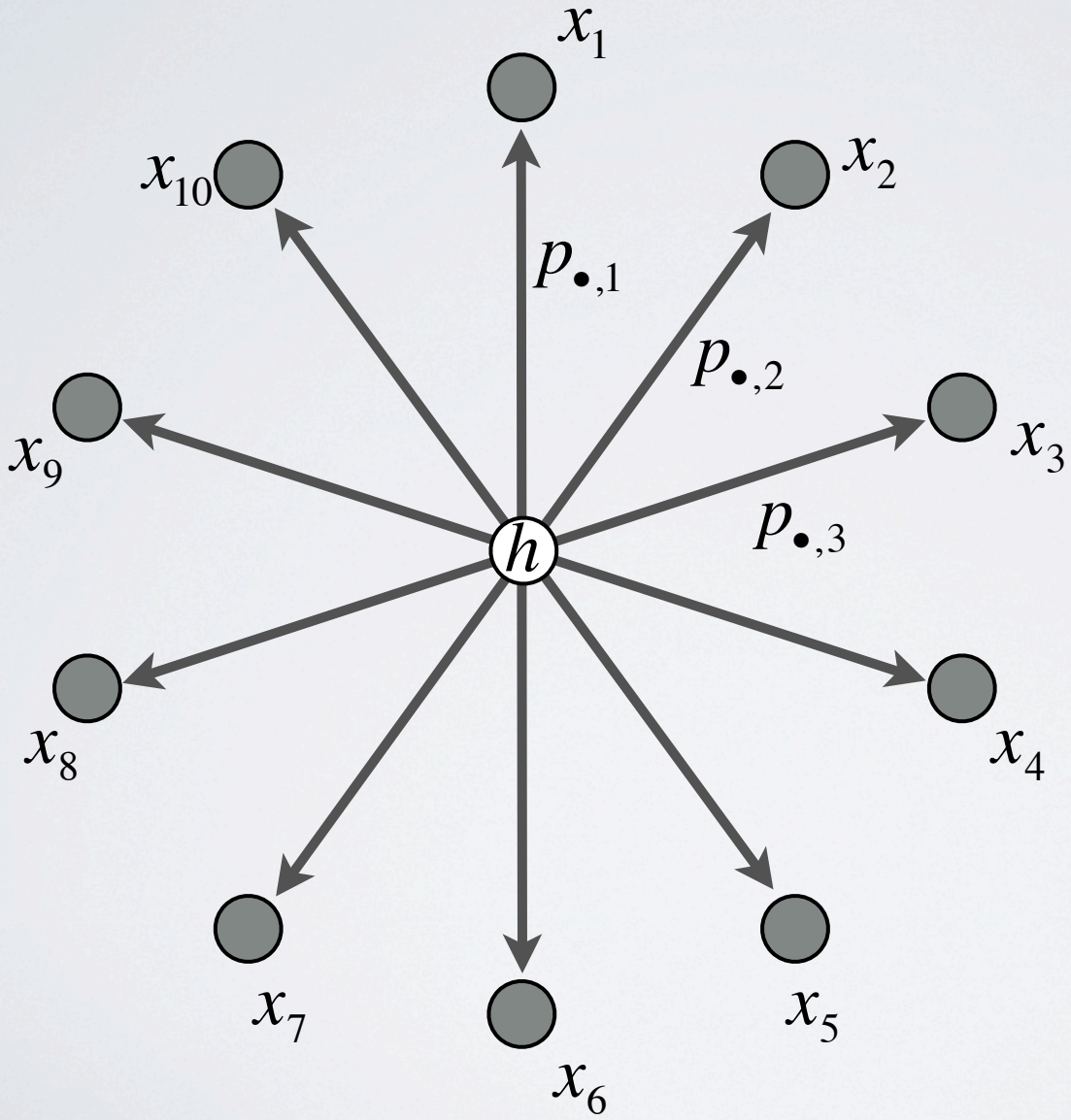
$$x = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}) \quad x_i \in \{A, C, G, T\}$$

$$P(x \mid \omega) = \alpha \prod_{j=1}^{10} p_{1,j}(x_j) + (1-\alpha) \prod_{j=1}^{10} p_{2,j}(x_j)$$

$$p_{i,j} = (p_{i,j}(A), p_{i,j}(C), p_{i,j}(G), p_{i,j}(T))$$

$$\omega = (\alpha, p_{1,1}, \ldots, p_{2,10})$$

# Naive Bayes mixture




$$P(h = 1) = \alpha$$

# Context-specific independence CSI

Mixture of Naive Bayes *and* relations between parameters

$$P(x \mid \omega) = \alpha \prod_{j=1}^{10} p_{1,j}(x_j) + (1-\alpha) \prod_{j=1}^{10} p_{2,j}(x_j)$$

$$R = \{\}$$

# Context-specific independence CSI

Mixture of Naive Bayes *and* relations between parameters

$$P(x \mid \omega) = \alpha \prod_{j=1}^{10} p_{1,j}(x_j) + (1-\alpha) \prod_{j=1}^{10} p_{2,j}(x_j)$$

$$R = \{p_{1,2} = p_{2,2}\}$$

# Context-specific independence CSI

Mixture of Naive Bayes *and* relations between parameters

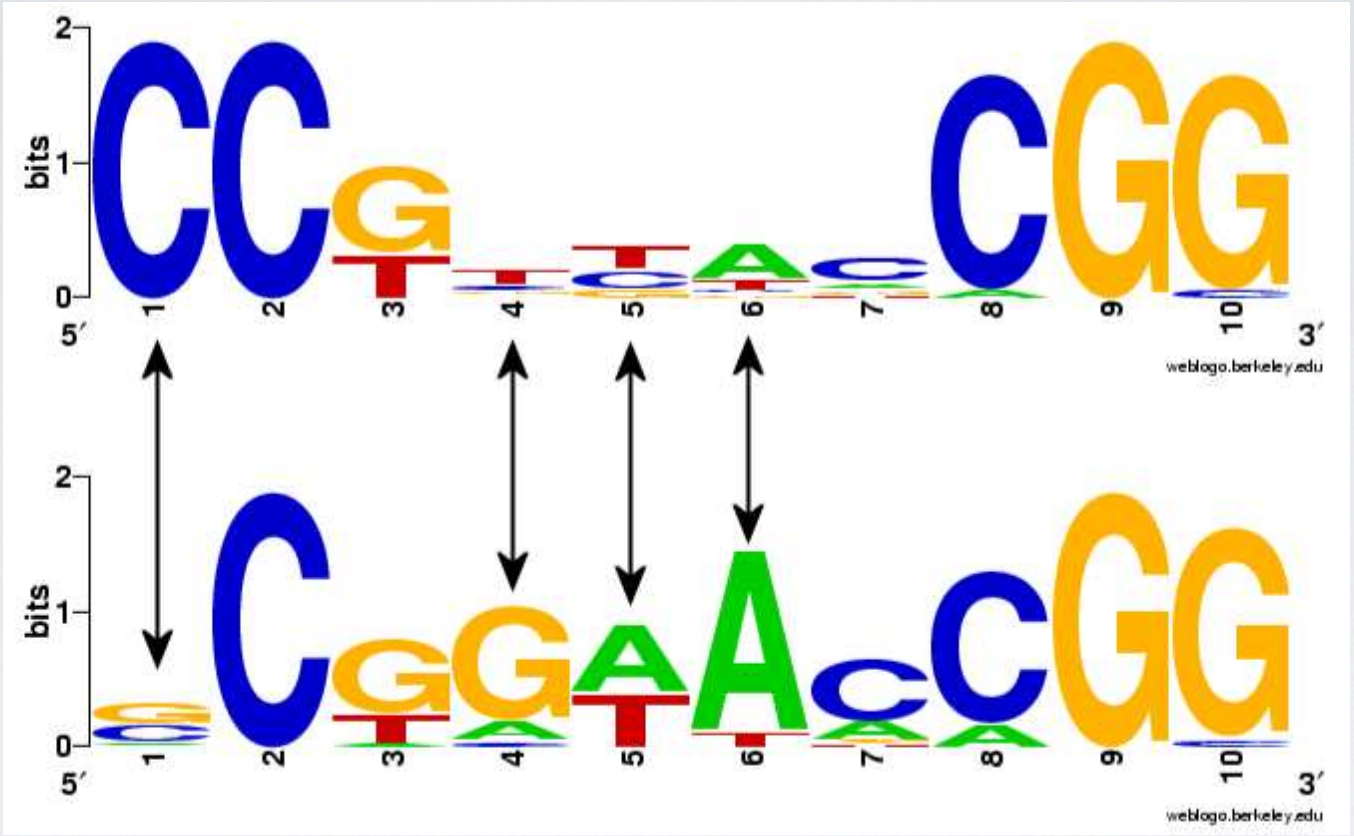
$$P(x \mid \omega) = \alpha \prod_{j=1}^{10} p_{1,j}(x_j) + (1-\alpha) \prod_{j=1}^{10} p_{2,j}(x_j)$$

$$R = \{p_{1,2} = p_{2,2}, \ldots, p_{1,7} = p_{2,7}, p_{1,8} = p_{2,8},$$
$$p_{1,9} = p_{2,9}, p_{1,10} = p_{2,10}\}$$

# MAP Estimation

- Priors:
  - Parameters (pseudocounts)
  - Structure (#components,#relations)
- Structural EM from full mixture (BIC for #components)
  - Given structure, reestimate parameters
  - Coordinate-wise greedy introduction of relations maximizing posterior

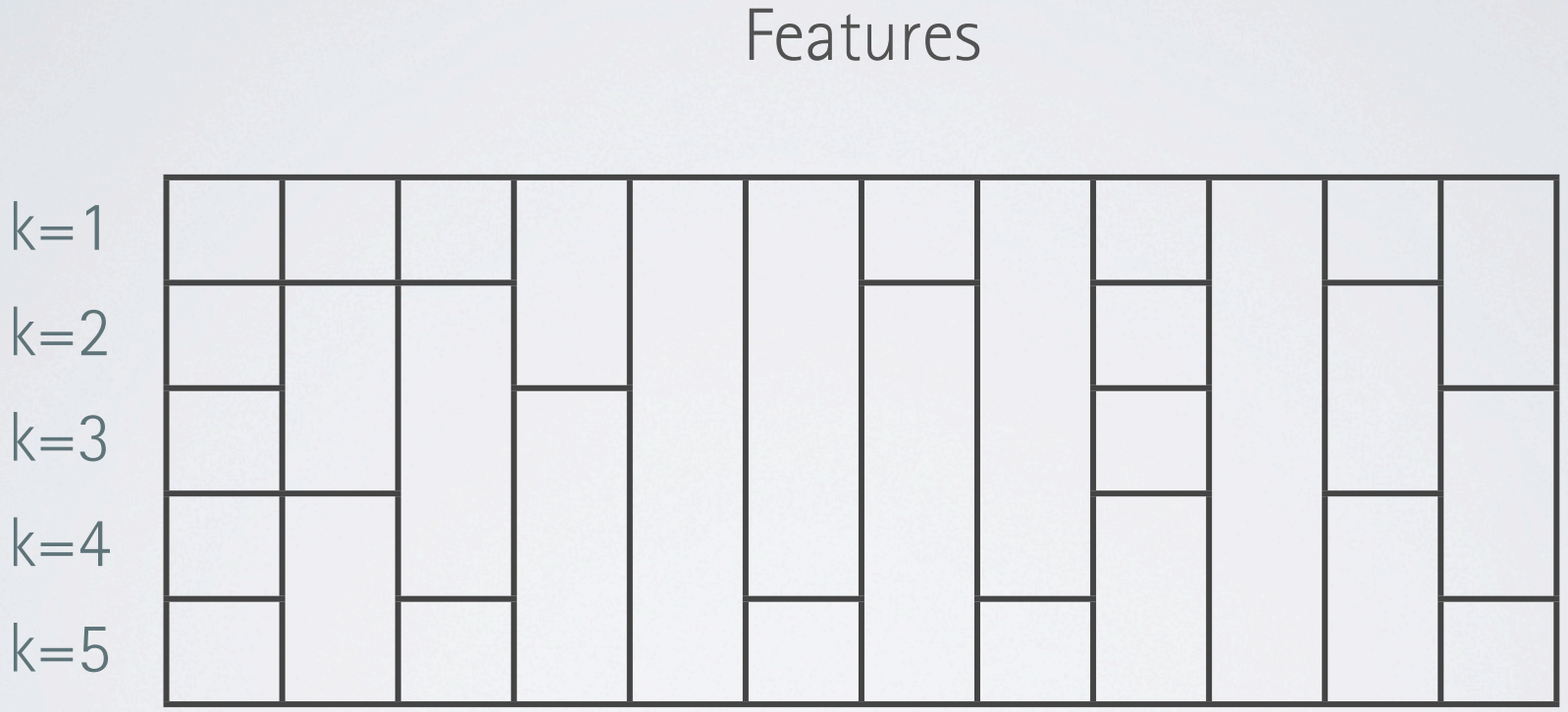
# Results Leu3




# Numbers

- Several databases
- # transcription factors > 400
- # sequences per factor ~ 20-50
- Yearly updates
- Identical problem for other regulatory elements

# CSI & Feature selection



# Copy Number Variants & Hidden Markov models

*Story #2*

# Genetic variants beyond mutations

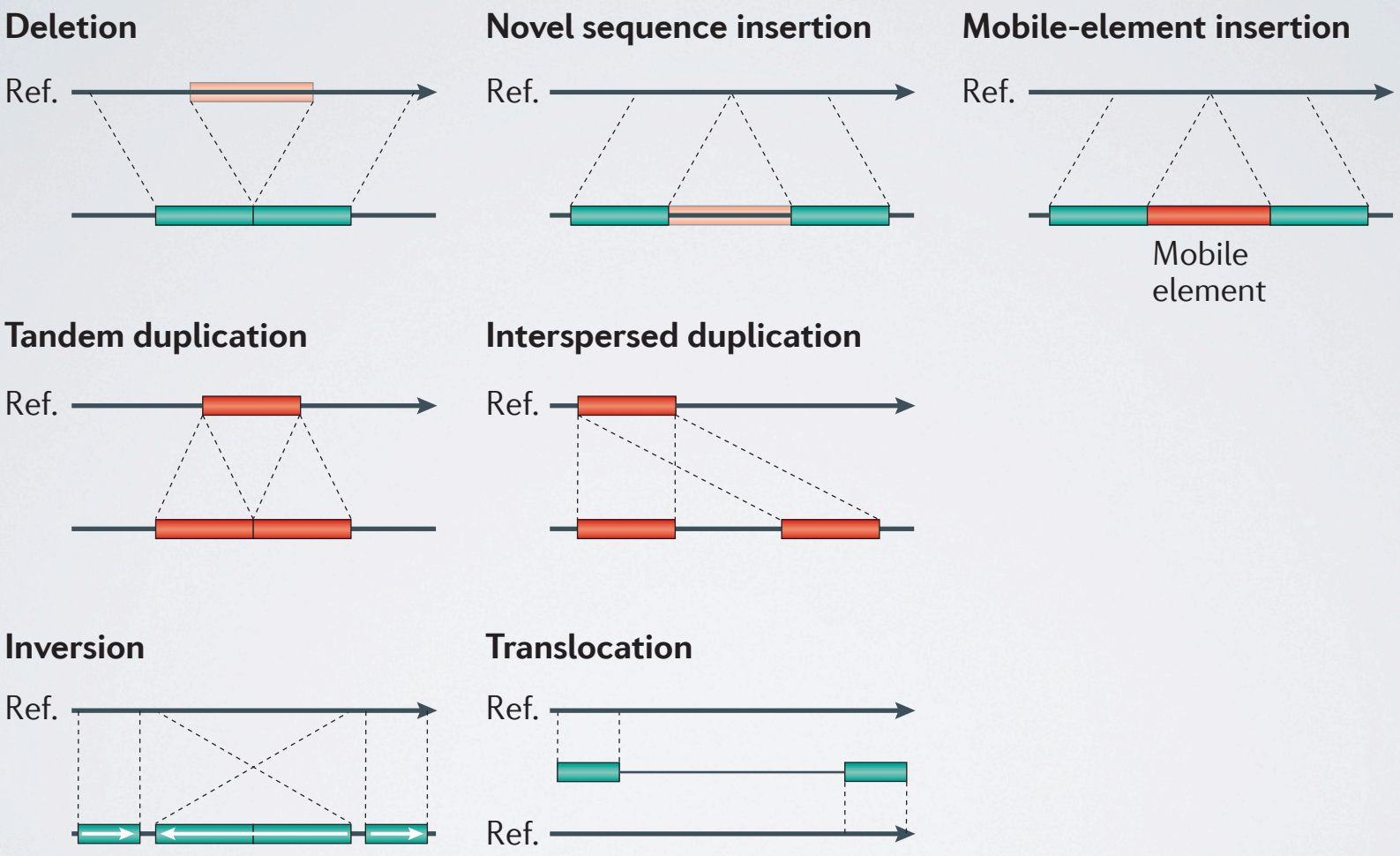



Figure from Alkan et al. Nature Reviews Genetics, 2011

# Genetic variants beyond mutations

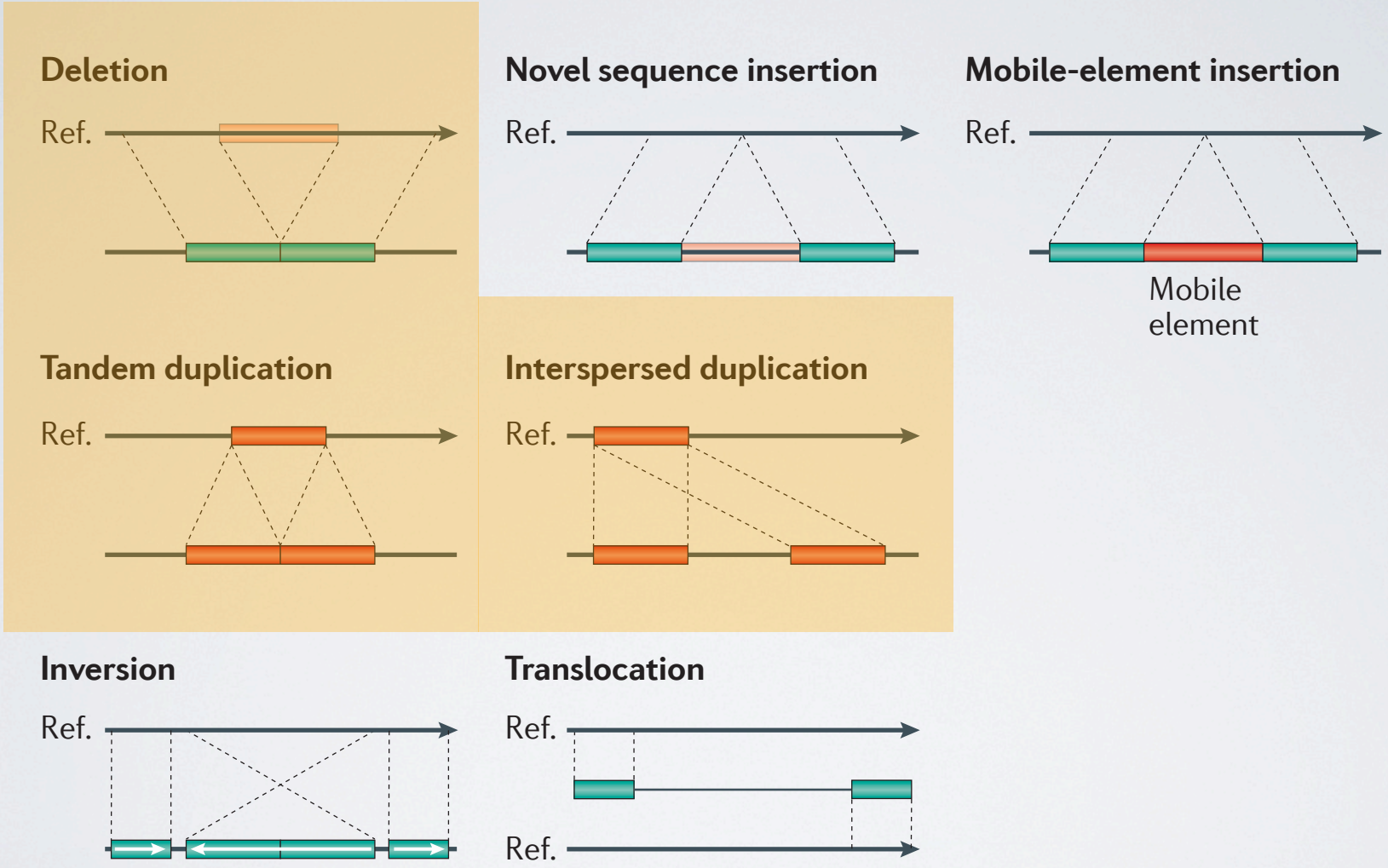
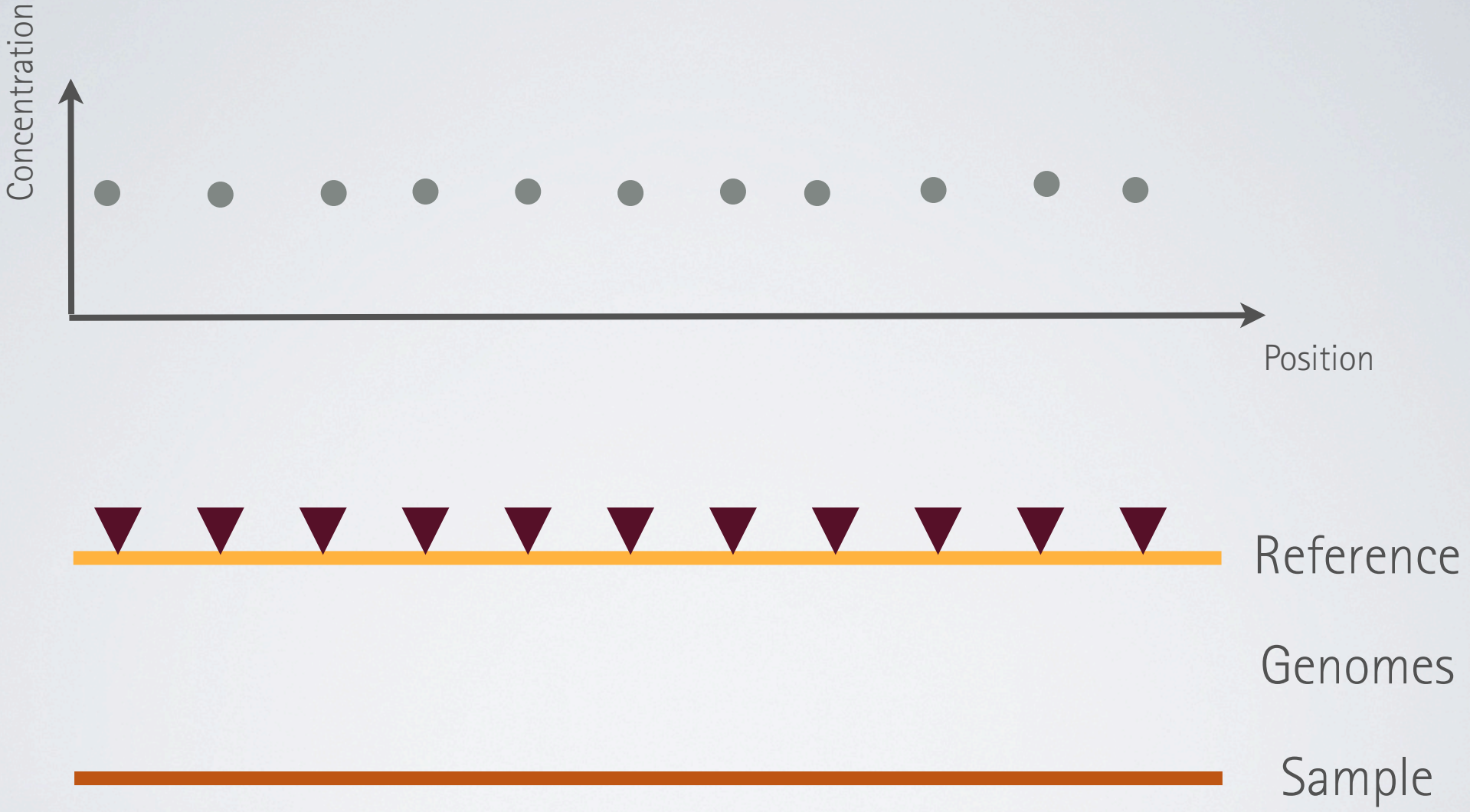
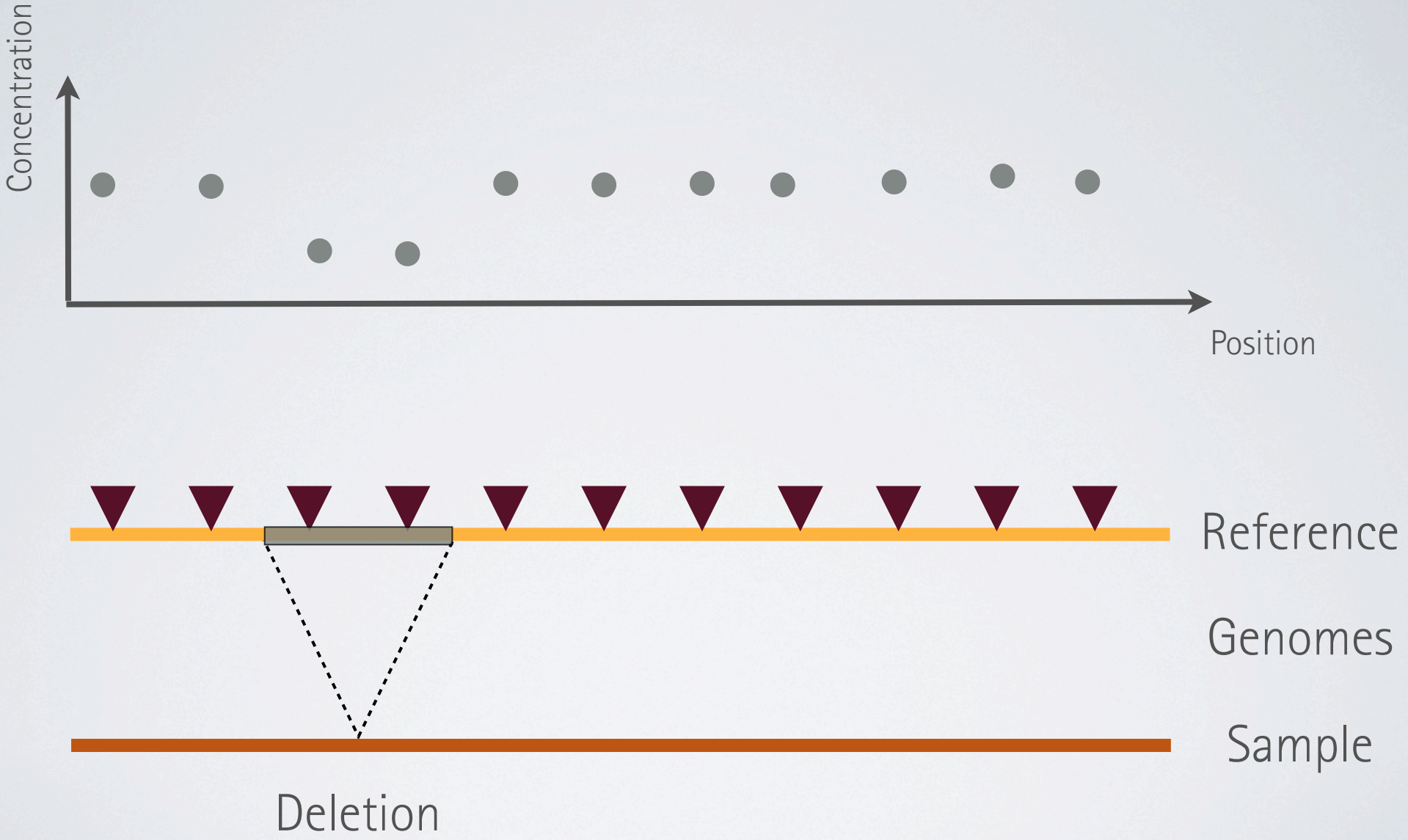



Figure from Alkan et al. Nature Reviews Genetics, 2011

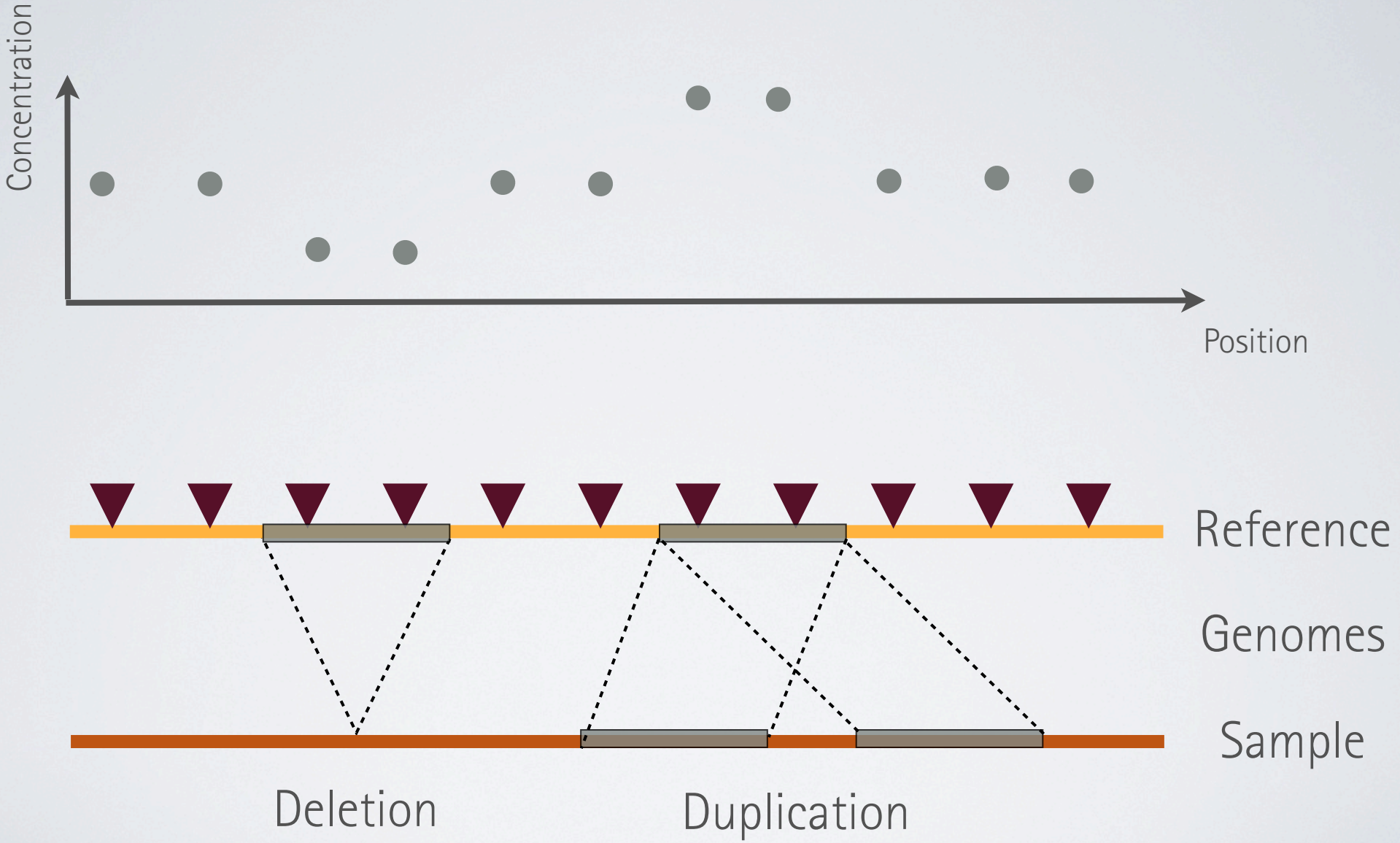
# Experiments for detecting CNV



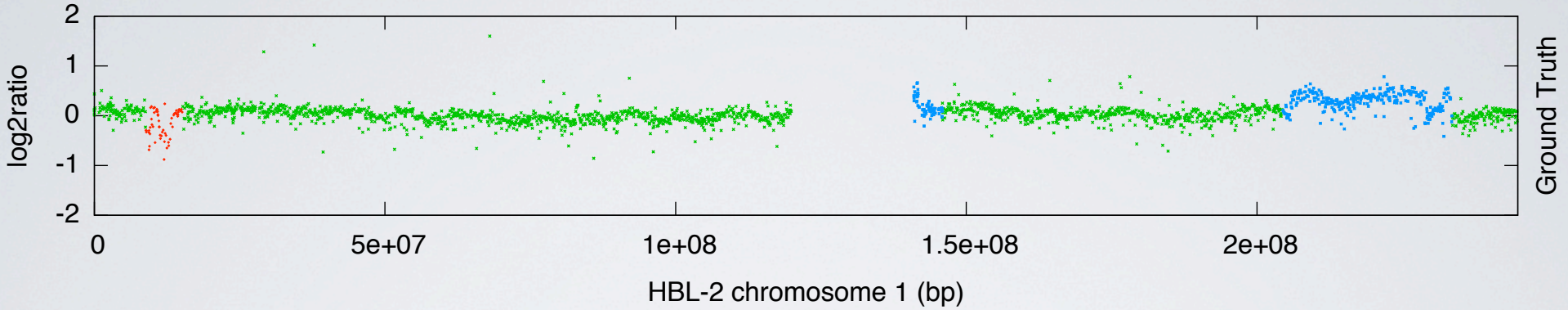
# Experiments for detecting CNV



# Experiments for detecting CNV

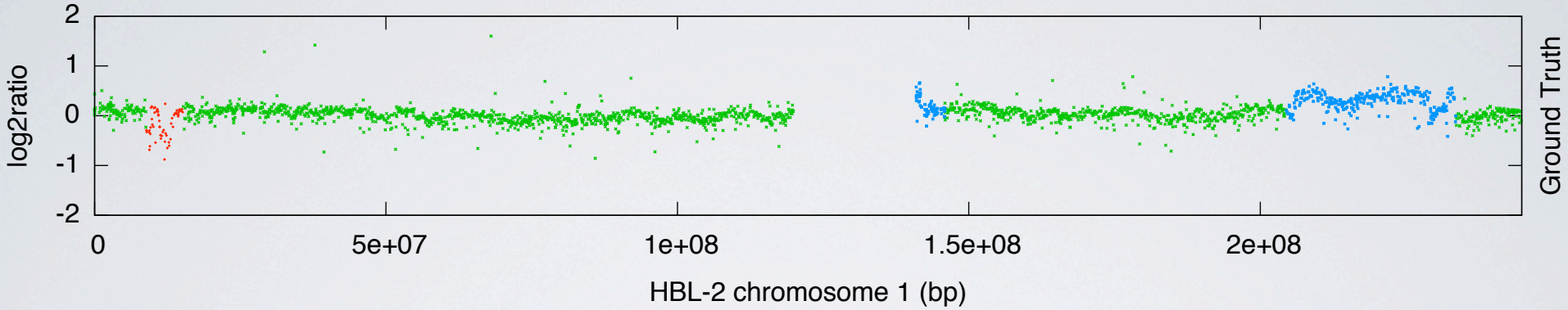


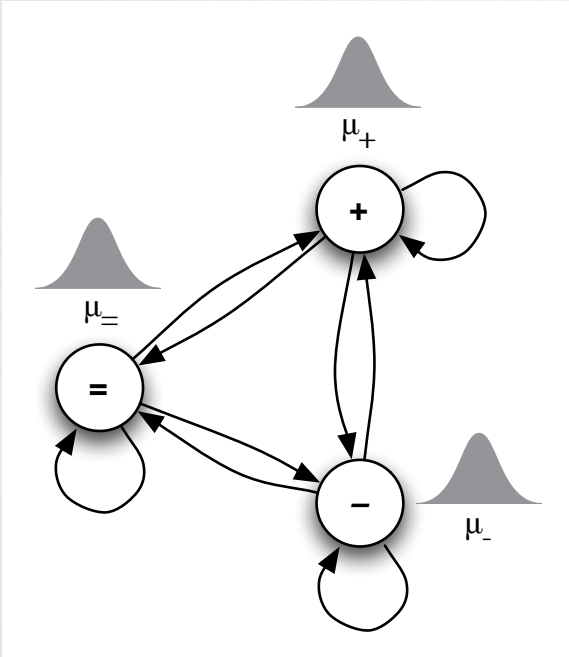
# Segmenting observation sequences




- Duplications and deletions affect contiguous regions
- Assume: piece-wise constant + noise
- Copy numbers are discrete

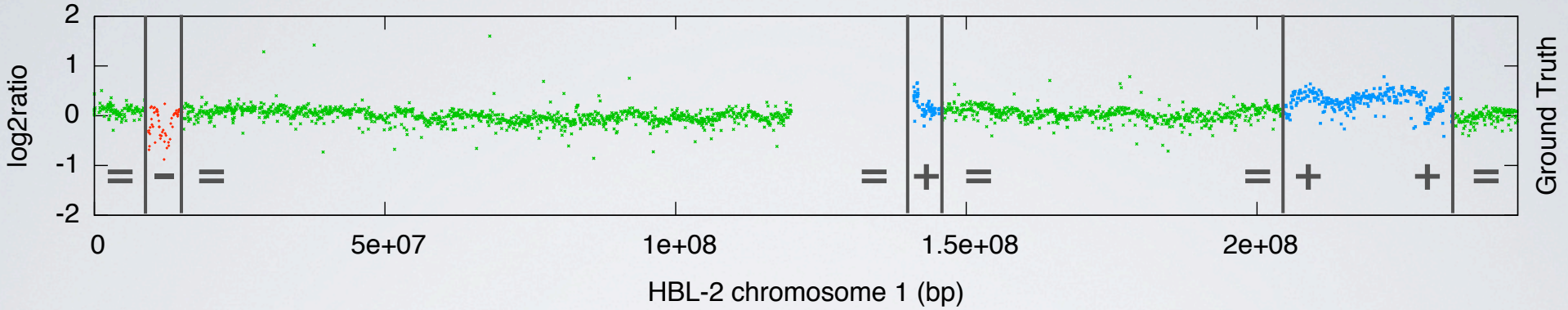
# Segmenting observation sequences

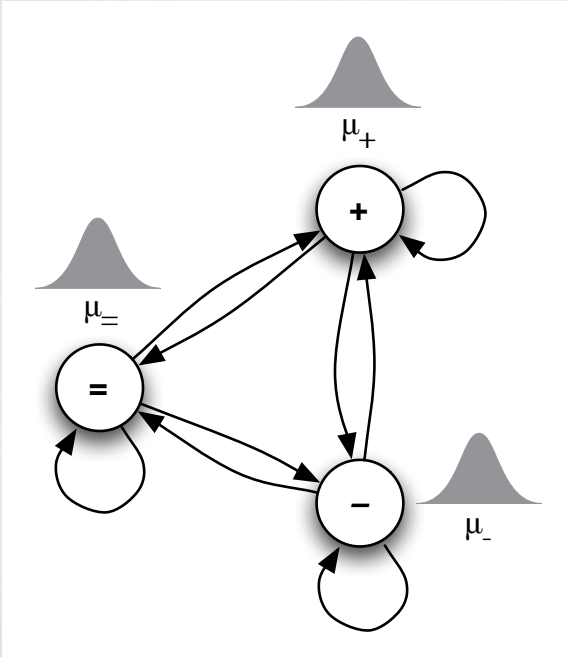



- Hidden Markov Model (HMM) with continuous emissions
- States = copy numbers

# Segmenting observation sequences




- Estimate Maximum-Likelihood HMM parameters using EM (Baum-Welch)
- Segmentation = most likely state path for observations (Viterbi path)

# Maximum Likelihood vs Bayesian

$$\theta_{ML} = \arg\max_{\theta} P(O \mid \theta)$$

$$Q^* = \arg\max_{Q} P(Q \mid O, \theta_{ML})$$

$$P(Q_t^* = '+' \mid O, \theta_{ML})$$

$$P(Q \mid O) = \int_{\theta} P(Q \mid O, \theta) P(\theta \mid O) \, d\theta$$

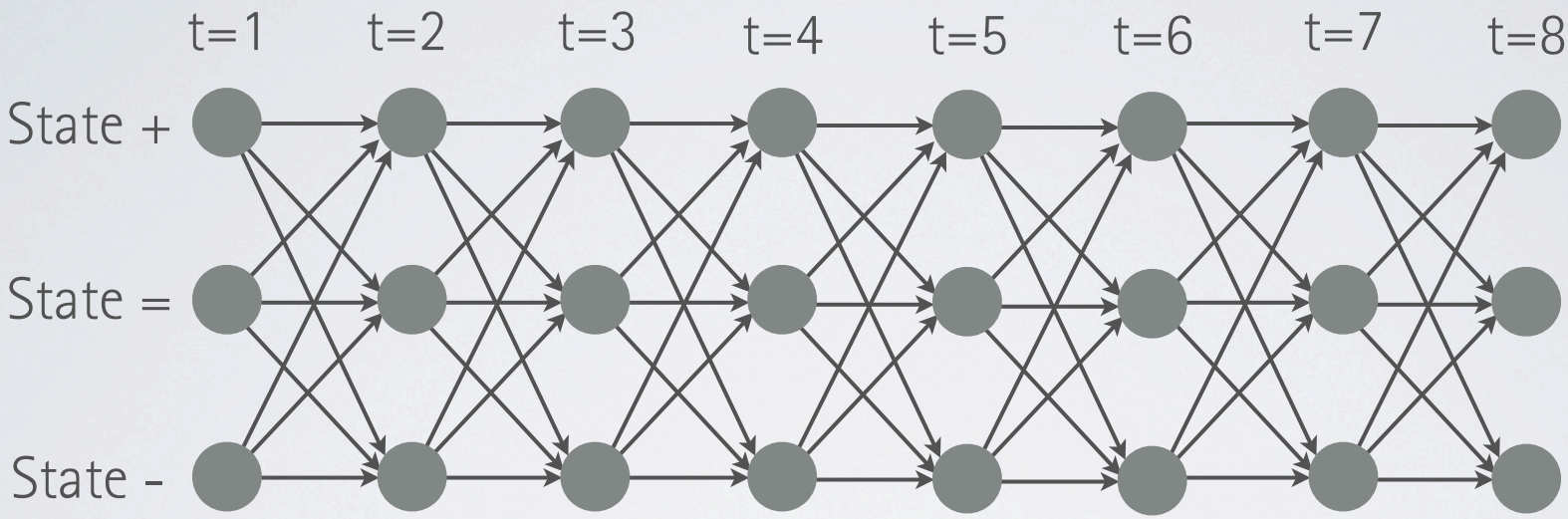
# Markov Chain Monte Carlo in HMM

- Draw $\theta^{(1)},\theta^{(2)},\theta^{(3)},\theta^{(4)}\ldots$ and $Q^{(1)},Q^{(2)},Q^{(3)},Q^{(4)}\ldots$ from $P(Q,\theta \mid O)$

- Estimate marginal distribution

$$P(Q_t = '+' \mid O) = \frac{1}{n}\sum_i I\{Q_t^{(i)} = '+'\}$$

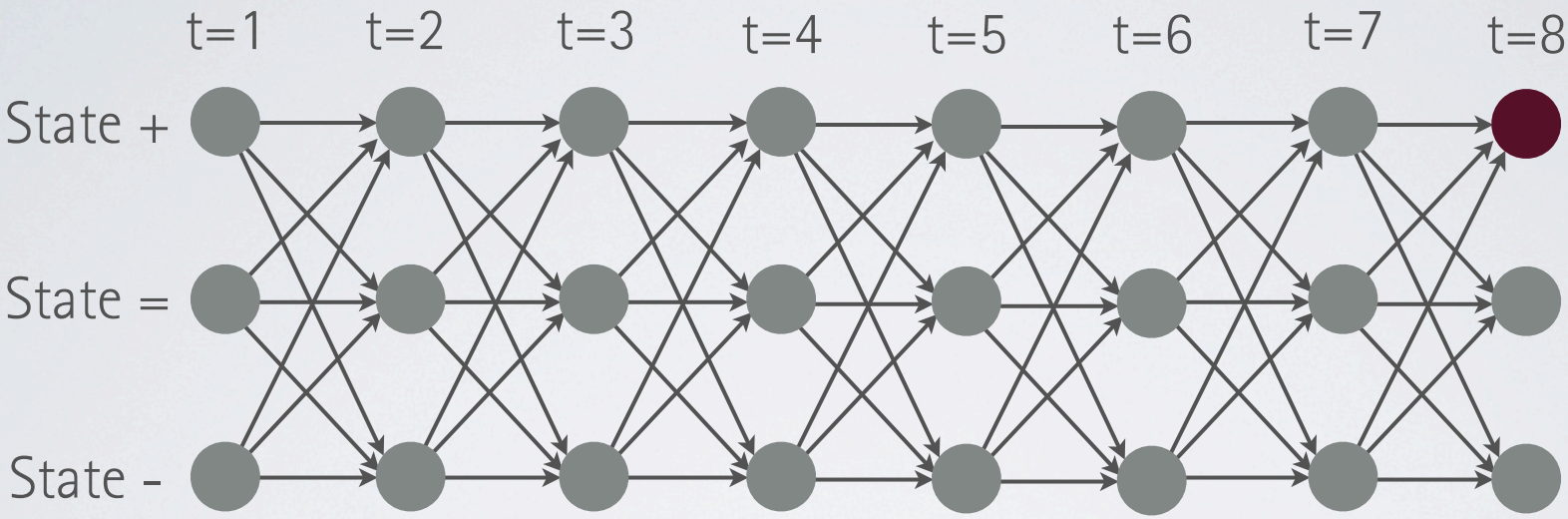
# Forward-Backwards Gibbs (FBG)




- Compute Forward-Variables for i,t $P(O_1...,O_t,Q_t=i|\theta^{(k)})$
- Sample last state $Q_T$ from $P(O_1...,O_T,Q_T|\theta^{(k)})$

Scott, Journal of the American Statistical Association, 2002

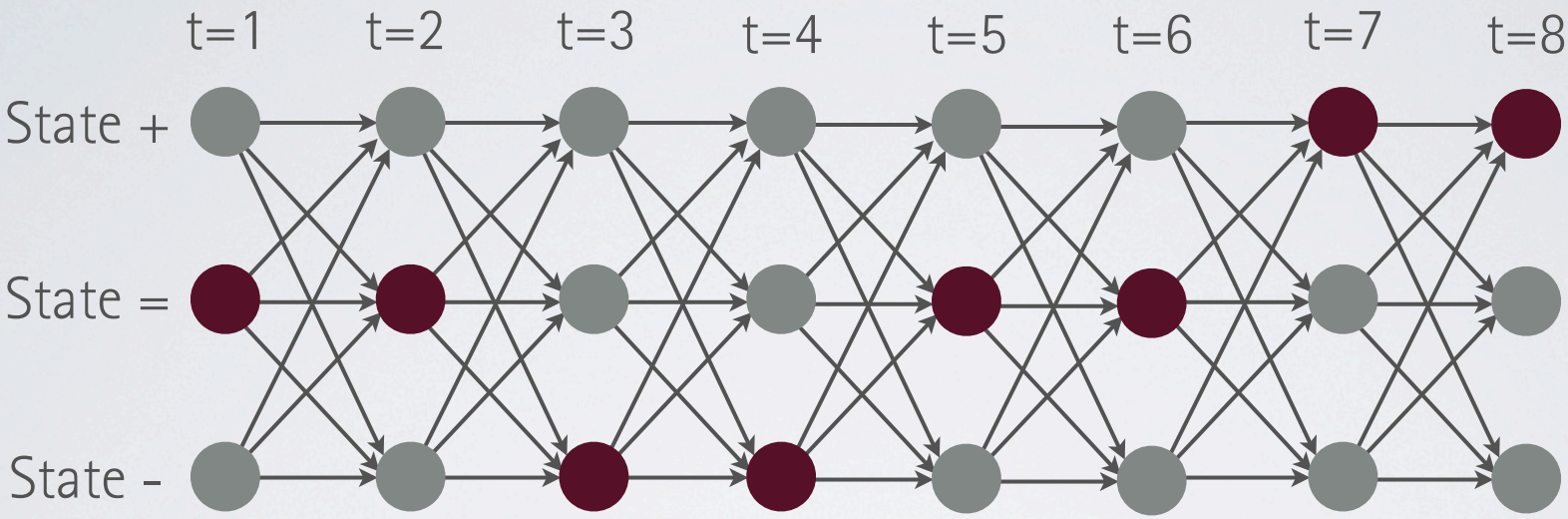
# Forward-Backwards Gibbs (FBG)




- Sample states backwards from $P(Q_t, O_1 \ldots, O_t \mid \theta^{(k)}) P(Q_{t+1} \mid Q_t)$

Scott, Journal of the American Statistical Association, 2002

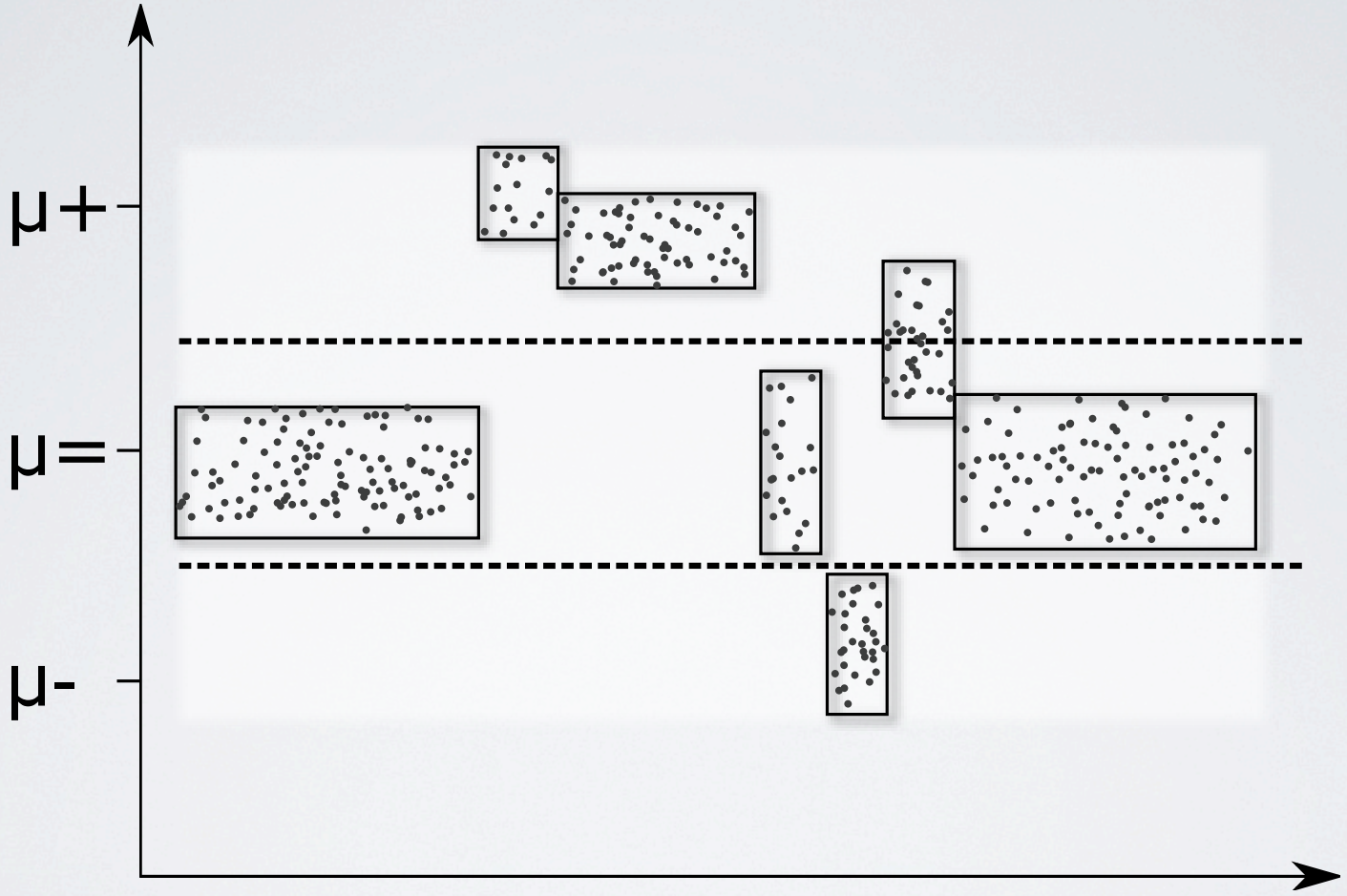
# Forward-Backwards Gibbs (FBG)



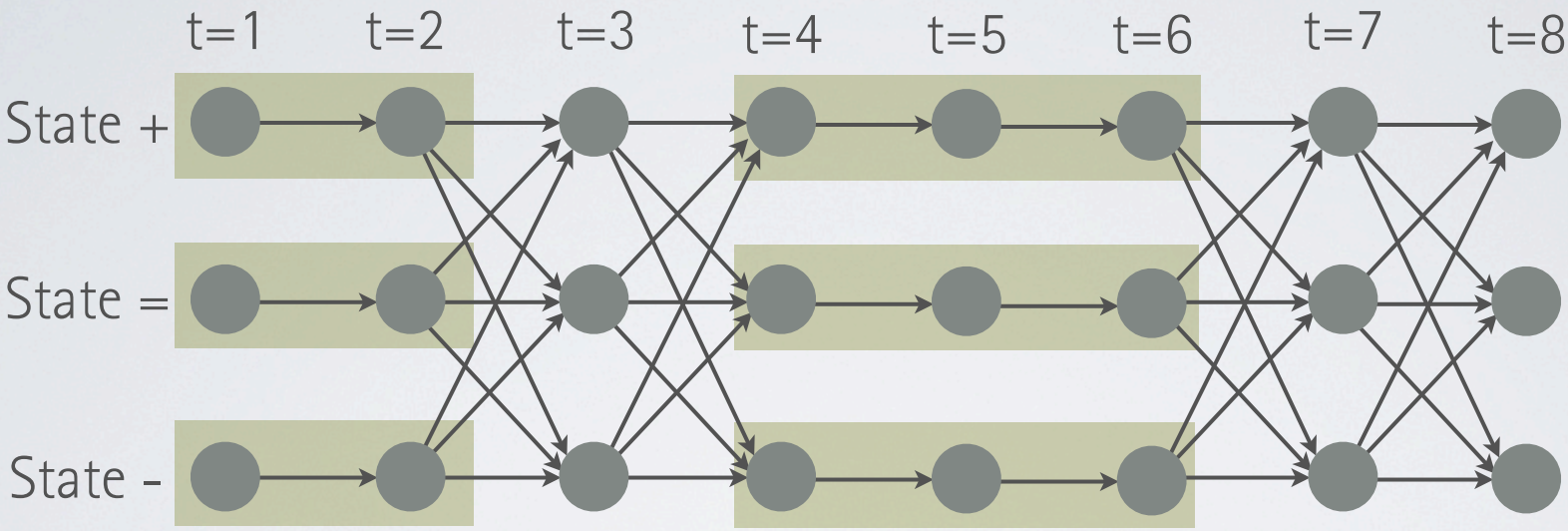

- Sample states backwards from $P(Q_t, O_1...,O_t \mid \theta^{(k)})P(Q_{t+1} \mid Q_t)$
- Sample new parameters $\theta^{(k+1)}$

Scott, Journal of the American Statistical Association, 2002

# Relations for Bayesian HMM



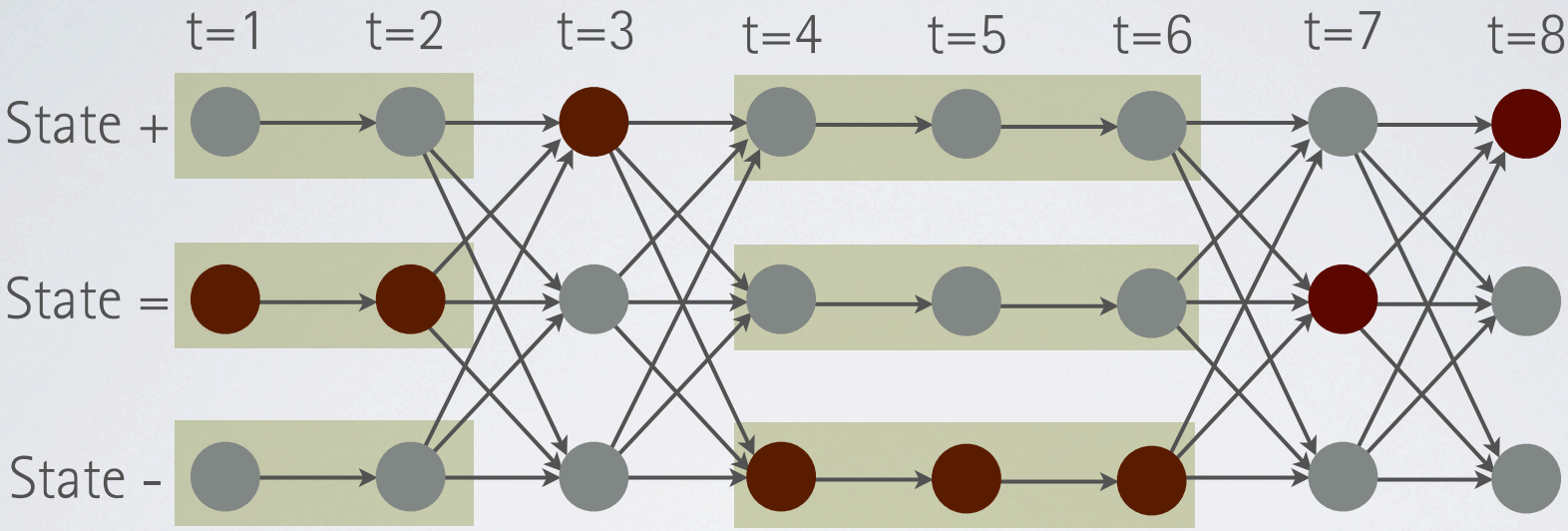
# Approximate sampling




- Assume same state for those piece-wise constant segments

Scott, Journal of the American Statistical Association, 2002

# Approximate sampling




- Sample states backwards from $P(Q_t, O_1...,O_t \mid \theta^{(k)})P(Q_{t+1} \mid Q_t)$
- Once per block!

Scott, Journal of the American Statistical Association, 2002

# Evaluation

- Compare to gold standard evaluation for respective dataset
- Consider as two class problems: normal vs. aberration
- Report
  - Recall = TP / (TP + FN)
  - Precision = TP / (TP + FP)
  - F1-measure = 2 x recall x precision / (recall + precision)

# Why not to use ML ...

| Method | | F1 | Recall | Precision | Likelihood |
|---|---|---|---|---|---|
| FBG | | 0.82±0.00 | 0.84±0.00 | **0.88±0.00** | – |
| EM: initial parameters from prior | ML | 0.65 | **0.90** | 0.50 | 15158 |
| | best | 0.85 | 0.84 | 0.86 | 9616 |
| | avg. | 0.76±0.09 | 0.86±0.03 | 0.68±0.12 | 13744 |
| EM: initial parameters sampled uniformly | ML | 0.64 | **0.90** | 0.50 | 15159 |
| | best | **0.86** | 0.84 | **0.88** | 9136 |
| | avg | 0.54±0.24 | 0.77±0.21 | 0.46±0.27 | 13457 |

All methods run for 810s (1 run FBG, ~1000 repetitions for EM, 30-50 steps each)

# Why to use approximate sampling...

| Method | | F1 | Recall | Precision | Time |
|---|---|---|---|---|---|
| FBG | | 0.82±0.00 | **0.84±0.00** | 0.88±0.00 | 810s |
| Approximate sampling | w=1 | 0.85±0.00 | 0.83±0.00 | 0.88±0.00 | 72s (11x) |
| | w=2 | 0.87±0.00 | 0.83±0.00 | 0.91±0.00 | 21s (40x) |
| | w=3 | **0.89±0.00** | 0.83±0.00 | **0.95±0.00** | 13s (62x) |
| | w=4 | 0.84±0.08 | 0.77±0.11 | **0.95±0.01** | 13s (62x) |
| | w=5 | 0.71±0.17 | 0.60±0.22 | **0.95±0.01** | 13s (62x) |
| | w=6 | 0.79±0.07 | 0.69±0.10 | **0.96±0.01** | 14s (60x) |

# Imbalanced problems

- Consider data as one-dimensional (ignore spatial information)
- Fit mixture: weights $\alpha_i$
- Consider $r=\alpha_{max}/\alpha_{min}$
- Model selection for $r > 10{,}000{,}000$ ?

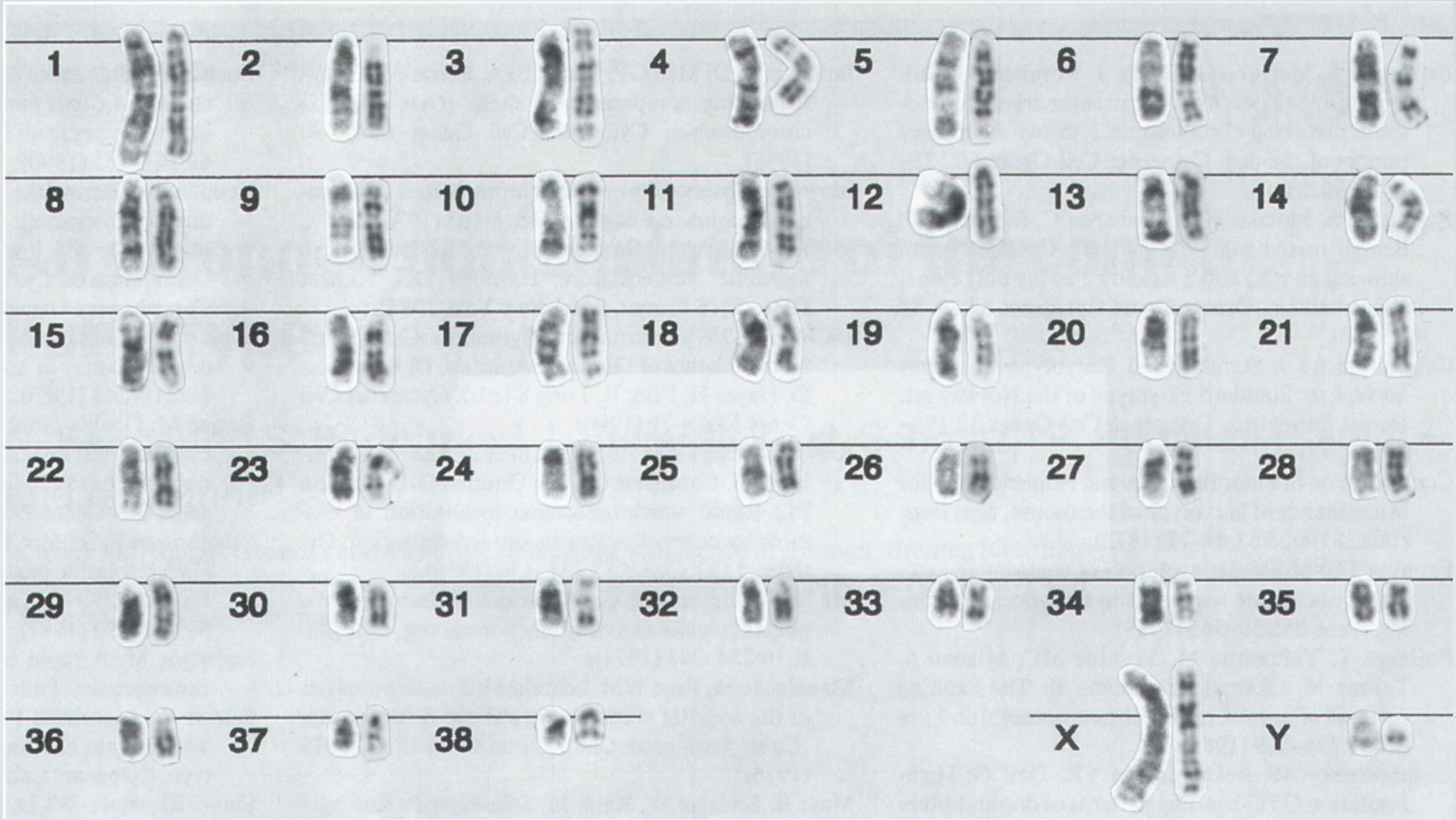
# Numbers

- ~50,000 observations (arrayCGH)
- ~500,000 observations (SNPArrays)
- Tens of Millions (sequencing)
- Thousands of samples now; 100,000s soon

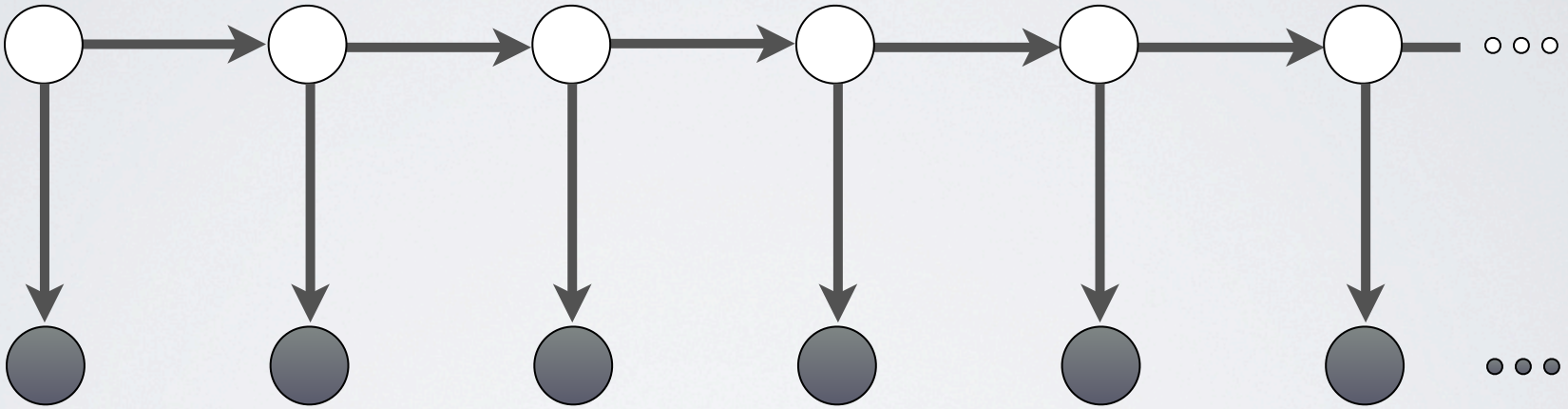
# Regions of equal sequence composition

*Story #3*

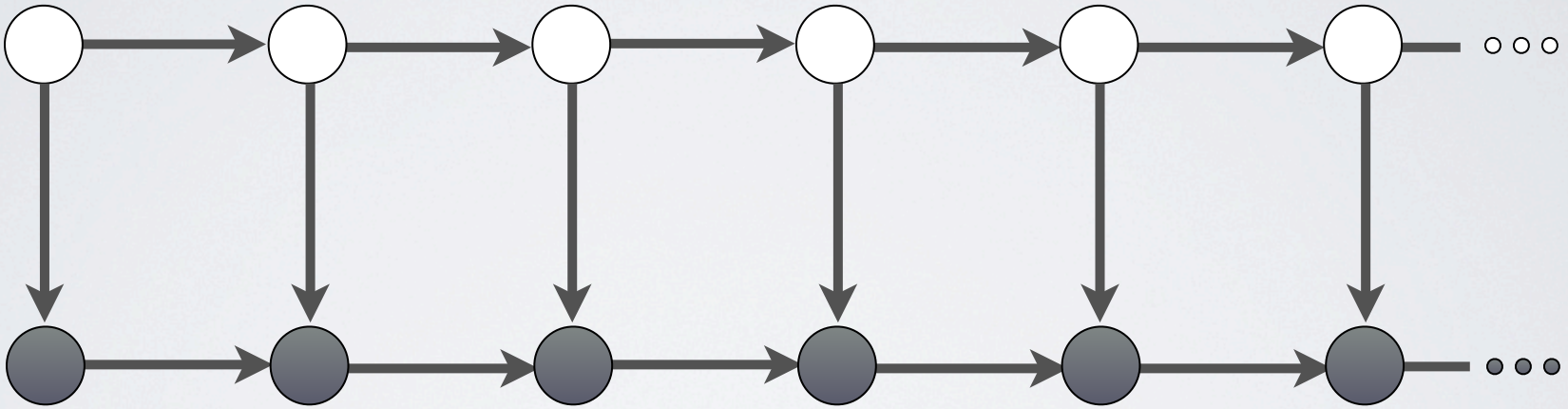
# Chromosome Bands



# Hidden Markov Model



# Hidden Markov Model with higher order emissions



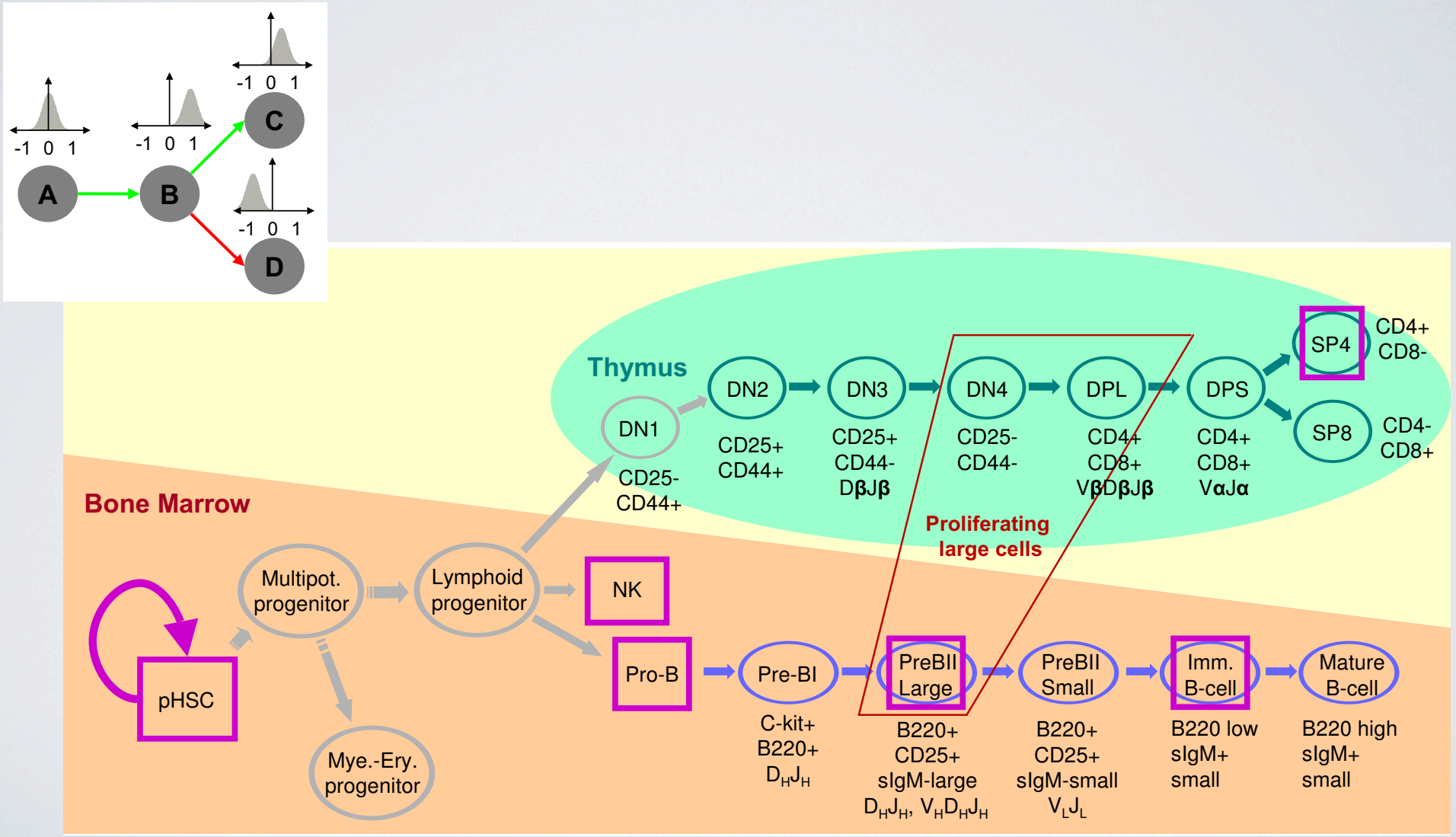
# Numbers

- DNA sequences of 50,000-3,000,000,000
- #hidden states: 4-16
- Many genomes

# Outlook

- Selected models important part of the analysis
- Unsupervised, automated model selection necessary
- Important problems (HMM):
  - {A,C,G,T} instead of binary observations
  - N=3,4,5 hidden states instead of N=2

# Stem cells to specialized cells



# Acknowledgements


- Benjamin Georgi (University of Pennsylvania)
- Md Pavel Mahmud (Rutgers)

# Thanks.