

1 Normalization

RPKM is a very important normalization tool.
not too sure about tpm, need to double check on that

2 Single Cell

Goals:

1. cells differentiates into sub-celltypes
2. unknown celltype discovery

2.1 Dimensionality reduction

Motivation:

1. high dim data often has lower dim representation w/o much reconstruction error
2. lower dim representation can often represent info about high dim pairwise dist.

Types of dim-red:

- Global methods
 1. all pairwise dist equally impt
 2. lower dim pairwise dist fit high-dim ones
 3. often use magnitude or rank order
- Local methods
 1. only local dist reliable in high dim
 2. more weight on modelling local dist correctly

Methods:

- PCA
 - finds directions with largest variance
 - minimize squared reconstruction error
 - equiv to liner autoencoders
 - Steps of PCA
 1. \bar{X} : mean of all samples(usually rows), adjust $X \rightarrow X' = X - \bar{X}$
 2. covar matrix $C = X'^T X'$
 3. find eigenvectors and eigenvalues of C , ie. all pair of \vec{v}, λ st. $C\vec{v} = \lambda\vec{v}$
 4. eigenvalues can be used to calculate percentage of total variance for each component

$$v_j = 100 \frac{\lambda_j}{\text{total eigenvalue}}$$

This is non-parametric method, do not insist on a parametric encoding function

- Multi-Dimensional Scaling
 - arrange low dim points to minimize diff between pairwise distances in the high and low D space
 - a possible approach: start w a random vector, perform gradient decent
 - is there something to do with PCA? then we don't need iterative method

- Sammon (non-linear autoencoder)
 - with extra layers, much more powerful than PCA, but can be slow to converge, and can get stuck on local optima
 - Multi-Dimensional Scaling(MDS) can be made non-linear by giving higher weights to smaller distances, a popular formula is

$$cost = \sum_{ij} \left(\frac{|x_i - x_j| - |y_i - y_j|}{|x_i - x_j|} \right)^2$$

where x is high-dim dist, and y is low-dim dist

- still slow and get stuck on local optima

2.2 Graph-based method

- address uniform circularity
- **Isomap** is a dim-red technique based on graphs
 - each datapoint is connected to k nearest neighbor in high-dim
 - edge weights = euclidean dist
 - approx of distance = shortest path in contracted graph
- **Probabilistic local MDS**
 - local distances are more imp't than non-local ones
 - in this way all local distances are given equal importance
- **stochastic neighbor embedding(SNE)** has a probabilistic way to decide if a distance is local
 - convert global distances into probability of one datapoint picking another datapoint as its neighbor (what defines a neighbor tho) - still about isomaps?
 - each point in high-dim has a conditional probability of picking any other point as its neighbor
 - distribution (some sort of Gaussian) is over high-dim distances (if high-dim coords unavailable, a similarity / dissimilarity matrix may be used)
 $p_{j|i}$ is the prob. of picking j given starting at i in high-dim.

$$P_{j|i} = \frac{e^{-2d_{ij}^2/2\sigma_i^2}}{\sum_k e^{-d_{ik}^2/2\sigma_i^2}}$$

- having the probabilities potentially allow us to throw away the raw high-dimensional data
- evaluation done using pairwise distance in low dimensional map (shows how well the lower dim representation models high-dim data ig)
 $q_{j|i}$ is the prob of picking j given starting at i in low-dim
- compute the **Kullback-Leibler divergence** between prob. in the high-dim and low-dim spaces (why not just use dist in high dim?) - more space / time efficient
- nearby pts in high-dim should be close in low-dim
- picking σ used to compute p - the radius of the gaussian
 - different radius is needed in different parts of the space to keep the no. of neighbors constant
 - big radius \rightarrow high entropy for distribution over i 's neighbors
 - small radius \rightarrow low entropy
- **Symmetric SNE**

- simpler than stochastic
- works best if different procedures are used for computing p 's and q 's.
- compromise: no longer guarantees that if using same dimension will produce optimal solution
- turn conditional prob into symmetric pairwise probabilities
- Optimization methods for SNE
 - simulated annealing could lead to better global optimization
 - add Gaussian noise to y location in each update
 - spend longer time at noise level where global structure start to form
 - **t-SNE** - use Gaussian at many (infinite) spatial scales, cheaper as we don't have to exponentiate anymore ([why????](#))

2.3 Supervised dim-red: Neural networks

- last few layers have much fewer values than inputs
- use intermediate layers as lower-dim representations
- can easily add prior biological knowledge, such as protein interactions or transcription factors
- essentially, some nodes in the hidden layer are same as before, others are based on biological info

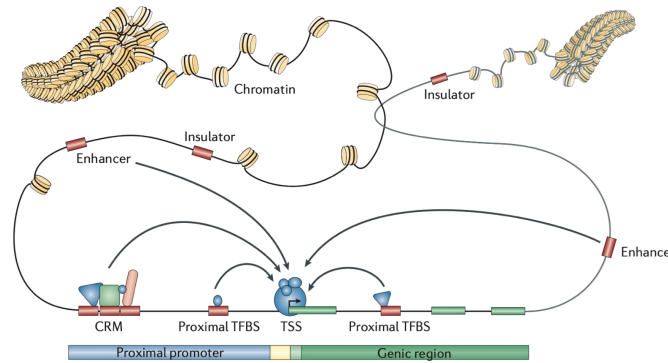
2.3.1 Additional NN architecture: Siamese

- supervised, but not trying to maximize training accuracy
- input: whether each pair is similar
- output: binary label of similar / not similar
- thus directly optimize dim-red layer for KNN

3 ChIP-Seq

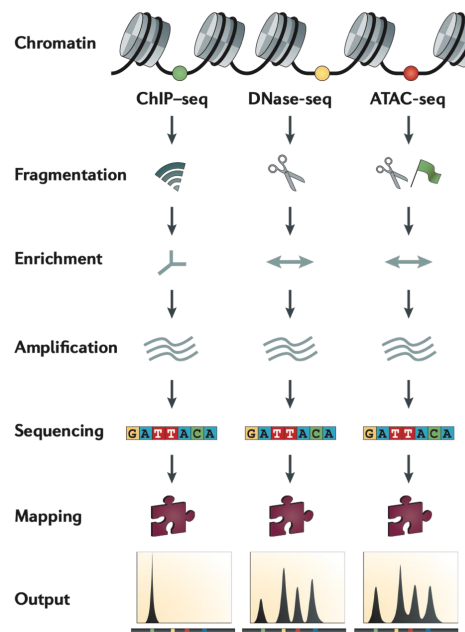
3.1 Background

- transcription factors bind to specific locations on the chromosome
- the binding is highly specific
- this has impact on gene regulation



- Transcription start site (TSS)
- Transcription factor binding sites (TFBS)
- Cis-regulatory module (CRM)
- Proximal promoter and distal enhancer

- The core promoter regions has about 300 TF,
(general transcription machinery, required for the transcription of most things),
another 1500 TFs for others (proximal enhancer/promotor/silencer, only affect some genes)
- comparisons of different whole genome enrichment technologies



3.2 Questions related to TF binding

- Where do they bind
 - some almost always bind to proximal promoters
 - others could bind to many regions
- How does specific binding work
 - exists a consensus motif (most common sequence)
 - looks like this



- shows which nucleotide is most abundant at each position, represented as **Position Weight Matrix (PWM)**
- **Motif**: a **recurring** sequence with biological importance
 - * can be constant or have a few variable elements
 - * often serves as binding site for TFs / proteins
 - * often very short (5-25)
 - * often distant from gene
 - * often has inexact repeating pattern (challenge for identification)
- sometimes also an effect of protein-protein interactions, ie. protein conformation change upon interactions etc?
- to determine binding site, often uses **Protein Weight Matrix**, but it is over-simplifying
- How to identify where they bind
 - genes regulated by the same TF are likely to have the same motif in regulatory region
 - to find genes regulated by the same TF → **comparative genomics** could be very useful
 - eg. knock-off of specific TFs → lower expression of a set of genes
 - group of genes that are co-expressed across many set of experiments are also likely to be regulated by same TF
 - challenges
 - * we don't know the exact motif seq
 - * we don't know how far it will be
 - * motifs can differ slightly across
 - * how to distinguish it from just “random” sequences that happen to repeat?
 - Multiple sequence alignment (MSA)
 - motif-finding based on EM algorithm (MEME-suite - Multiple EM for Motif Elucidation)
- How is it involved in gene regulation
- Is it useful for gene regulatory network?

3.3 ChIP-seq

- technology:
 1. chromatin immunoprecipitation + high throughput sequencing
 2. detect genome-wide location of TF and other binding proteins in labs
 3. find all DNA seq bound by TF-X
 4. try to learn the regulatory mechanisms of a TF or DNA-binding protein
- General strategies to call ChIP-seq peaks
 - ChIP-seq yields distributions for tags from forward and reverse strands
 - overlap of the two can be observed
 - actual binding site of TF should be between the 2 distributions
 - from the difference between the 2 peaks → formulate a single **peak summit**
- MACS: model-based analysis for ChIP-seq

4 cis-regulatory motif

5 Haplotype inference

5.1 Phase

- h_1, h_2, h_3 : unobserved ancestral haplotypes
- h_{4A}, h_{4B} : unobserved haplotypes for individuals
- Circles: alleles, mutations

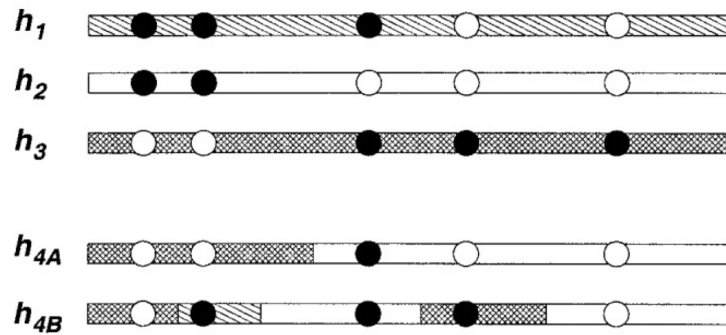
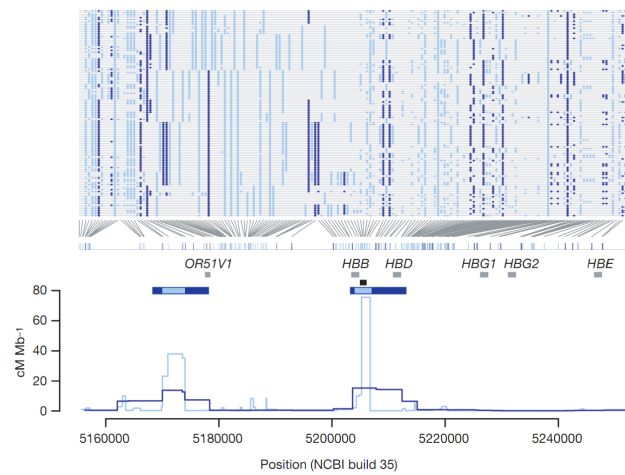


Figure 1

Haplotype Structure and Recombination Rate Estimates: HapMap I vs. HapMap II

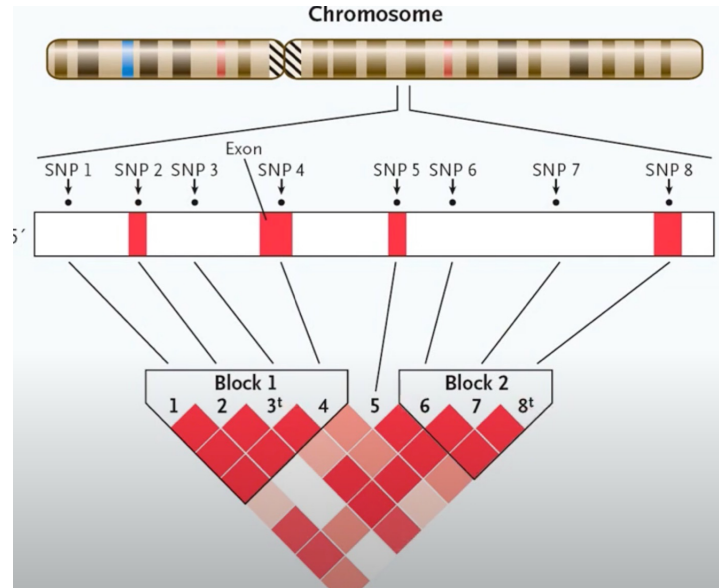


- Genotype imputation
 - Given: SNP array genotype data
 - some has missing data / untyped SNPs
 - much cheaper
 - use HMM
 - more data, more accurate for imputation
- Beagle / SHAPE-IT are most commonly used software

5.2 Linkage disequilibrium (LD)

- reflect rs between alleles at diff loci

- linkage equilibrium: no linkage, not coupled
- disequilibrium \rightarrow linkage
- **LD**: allelic association measure
- calculating LD:
 - assume independence, calculate expected frequency
 - $D_{AB} = P_{AB} - P_A P_B$
- calculate for all loci (SNPs)
- spot-light figures



- neighboring loci shows more linkage
there could be LP blocks on the chromosome, where there's stronger linkage within block and weaker across block
each block is like a voting district / **Marker regions**
- They are **Tag SNPs**
- pretty standard practice
- dense genotyping more expensive
 - get the LP
 - use inference on sparser genotyping
 - multi-phase procedure

6 Population structure

6.1 Background

- def: set of individuals with distinct genetic variations
- eg. ancestral history, lactose intolerance
- Hardy-Weinburg equilibrium

- Under random mating, both allele and genotype frequency remain const
- Current gen:
 - * $D + H + R = 1$
 - * $p = \frac{2D+H}{2}$
 - * $q = \frac{2R+H}{2}$
- Next gen:
 - * $D' = p^2$
 - * $H' = 2pq$
 - * $R' = q^2$
 - * $p' = \frac{2p^2+2pq}{2} = p^2 + pq = p$
 - * $q' = \frac{2p^2+2pq}{2} = q^2 + pq = q$
- given population data, can test if it holds (often chi-square test)
 - * testing is recommended
 - * if fail, means something is going on
 - * or genotyping error
 - * want HWE to hold for control group
 1. compute allele freq from observed data
 2. compute expected genotype freq
 3. compute test statistic (deg of freedom 1)
- Due to **Genetic drift**, even when assumptions of HWE hold, HWE may not hold
- genetic drift
 - change in allele freq due to random sampling
 - all mutations eventually drift to 0 or 1 eventually
 - is neutral
- Wright's F_{ST} (Wright-Fisher model)

$$\binom{2N}{k} p^k q^{2N-k}$$

- Ways how populations evolve
 - population divergence
 - separated into subpopulations with independent selection and drift
 - admixture
 - mixing of population

6.2 Inferring pop structure from genotype data (nowadays mostly just PCA)

Mixture model

- cluster individual into K populations
- does not model admixture
- cluster individual into populations
- probability model for mixture of C Gaussians

$$p(x) = \sum_{i=1}^k p(x|c=i) \cdot p(c=i)$$

- c : labels, $p(c)$ label freq, multinuilli
- x : genotype / alleles, $p(x|c) = \prod_{i=1}^J p(x_i|c)$, assume independence
- can then learn the model with EM
- Inference: $p(c|x)$ to infer cluster label

Admixture model

–