1 Search and Alignment

Genome Sequencing

Illumina (50-300nt), $> 10^9$ sequences

PacBio ($10^4 - 0.5 \times 10^5$nt), $> 10^5$ sequences

NanoPore ($10^4 - 1.5 \times 10^5$), 10^5 sequences

$q = -10\log_{10}(p)$ p error probability ($\sim 1\%$)

Suffix Tree

The suffix tree for string S of length n: • Has exactly n leaves

- Every **internal** node has at least two children
- Space and Construction: $O(n)$
- Search: $O(p+k)$ or $O((p+k)\log n)$

K-mer

- Search: $O(p)$
- Space: $4n + |\Sigma|^k$

Suffix Array

- Sorted list of all suffixes of a string S
- Can be generated by a depth-first traversal of the suffix tree
- Space: $O(n)$
- Search: $O(p \log n)$
- $L_p = min(k : P \le S_{A[k]}\ or\ k = n+1)$
- $R_p = max(k : S_{A[k]} < P\#\ or\ k = 0)$ with # > any symbol

BWT with FM index

- Space: $O(n)$ Search: $O(p)$
- The $k-th$ occurrence of the character c in L corresponds to the $k-th$ occurrence of character c in F
- C[c]: total number of occurrences of characters $< c$ in L
- Occ(c,k): number of times c occurs in $L[1,k]$
- $LF(i) = C[L(i)] + Occ(L[i], i)$

Needleman-Wunsch Global Alignment

$$min\begin{cases} d_{i-1,j-1} + c(a_i, b_j) \\ d_{i-1,j} + c(a_i, -) \\ d_{i,j-1} + c(-, b_j) \end{cases}$$

Complexity: $\Theta(mn)$

Hirschberg algorithm

Space: $O(max(m,n))$ Time: $O(mn)$

Banded Alignment

- d is an upper bound for the distance
- Space,Time: $O(d \times max(m,n))$
- Δ is the cost for indel
- $Z = \left(-\left\lceil \frac{t}{2\Delta} - \frac{n-m}{2} \right\rceil, \left\lceil \frac{t}{2\Delta} + \frac{n-m}{2} \right\rceil\right)$

Approximate Matching

Initializing the first row in the dynamic programming matrix to 0 (first-row-to-zero-trick) allows for multiple starting positions in S.

Smith-Waterman Local Alignment

$$max\begin{cases} 0 \\ d_{i-1,j-1} + s(a_i, b_j) \\ d_{i-1,j} - \delta(a_i, -) \\ d_{i,j-1} - \delta(-, b_j) \end{cases}$$

Substitution Scores

$$S(A,B) = \sum_i \log \frac{\overbrace{p_{a_i b_i}}^{model}}{\underbrace{q_{a_i} q_{b_i}}_{random}}$$

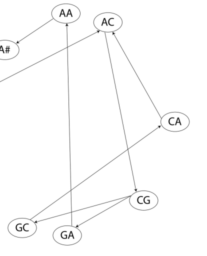
BWT on De Bruijn Graphs

Index construction is similar to BWT on trees:

- sort nodes lexicographically by labels
- assign incoming labels of incoming edges to nodes
- mark first position of every interval with identical node labels
- augment alphabet in `labels` to mark last outgoing edges

Representation only stores topology of the graph, not the frequency of occurrence in T.

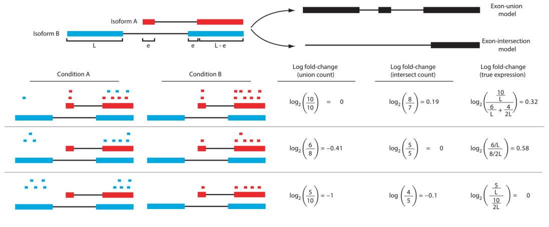
first	labels	nodes
1	A	#A
1	A	A#
1	G	AA
1	#	AC
0	C	AC
1	G	CA
1	A	CG
1	C	GA
1	C'	GC



2 RNA-Sequencing & Gene Expression

Gene Expression Estimation

RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes. Sensible to genomic variation. Alternative transcripts/RNA-processing may lead to differential read counts



Isoform A
Isoform B
L
e
e
L - e
Exon-union model
Exon-intersection model
Condition A
Condition B
Log fold-change (union count)
Log fold-change (intersect count)
Log fold-change (true expression)
$\log_2\left(\frac{10}{10}\right) = 0$
$\log_2\left(\frac{8}{7}\right) = 0.19$
$\log_2\left(\frac{\frac{10}{L}}{\frac{6}{L}+\frac{4}{2L}}\right) = 0.32$
$\log_2\left(\frac{6}{8}\right) = -0.41$
$\log_2\left(\frac{5}{5}\right) = 0$
$\log_2\left(\frac{6/L}{8/2L}\right) = 0.58$
$\log_2\left(\frac{5}{10}\right) = -1$
$\log\left(\frac{4}{5}\right) = -0.1$
$\log_2\left(\frac{\frac{5}{L}}{\frac{10}{2L}}\right) = 0$

- the same gene can contain multiple, partially overlapping transcripts
- ignoring the transcript structure can lead to estimation biases (depending on the gene model used for counting)

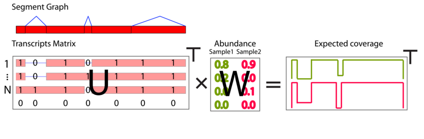
rQuant

- P set of genomic positions; R_p number of reads covering position p; $D_{t,p}$ expected read coverage for transcript t at position p. Repeat until convergence:
- Optimize transcript weights w_t: $\min_w \sum_p \mathcal{L}\left(\sum_t w_t D_{t,p}, R_p\right)$
- Optimize profile weights $D_{t,p}$: $\min_p \sum_p \mathcal{L}\left(\sum_t w_t D_{t,p}, R_p\right)$

Problems with Transcript Quantification

- Abundances cannot be unambiguously determined with single-end reads; (use paired-end reads)
- Solution may be unstable: a small change in reads can cause large changes in estimated abundances
- Read coverage is not uniform over the transcript

Transcript Reconstruction



$$\min_{U,W} L(U^T \times W, C) + \gamma \times N$$

Simple Linear Model

- Assumptions: normality and independence of residuals, homoscedasticity, linearity, additivity
- Modeling count data, gene abundance is the number of successes in a fixed amount of time $\Longrightarrow$ Poisson
- Problem: Poisson can't model overdispersion (caused by excess zeros,correlation/groupings in samples,unobserved variables) $\Longrightarrow Var(X) > E(X)$
- Solutions: variance stabilizing transform or $X \sim NB(p,r) \Longrightarrow Var(X) = E(X) + E^2(X)/r$. With negative binomial we can fit the variance.

Generalized Linear Model

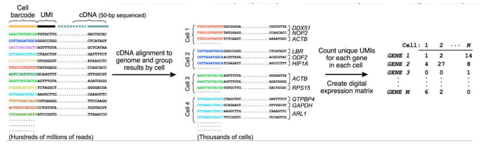
$p(x|\eta) = h(x)\exp\{\eta^\intercal t(x) - a(\eta))\}$

$a(\eta) = log\int h(x)\exp\{\eta^\intercal t(x)\}dx$

$L(\theta) = \log(\prod_n^N h(y_n)\exp\{\eta_n y_n - a(\eta_n))\}$

Poisson: $p(k) = e^{-\lambda}\lambda^k/k!$ Mean = Variance = λ

3 Single-Cell Expression



Peculiarities of SC Data

- Zero-inflated
- Increased variance
- Reveals rare cell population and distinct cell types/states

scNorm

- The global correction factor we used to normalize bulk RNA-Seq does not work well for single-cell data
- scNorm uses quantile regression to estimate the dependence of transcript expression on sequencing depth for every gene.
- Genes with similar dependence are then grouped, and a second quantile regression is used to estimate scale factors within each group
- Within group adjustment for sequencing depth is then performed using the estimated scale factors to provide normalized estimates of expression.

PCA

Orthogonal linear transformation; first components explains the largest variance; doesn't work well with non linear data

tSNE

- Nonlinear dimensionality reduction technique: converts similarities between data points to joint probabilities and tries to minimize the KL-Divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data
- Cluster sizes and distance between clusters (only local distance is preserved) mean nothing. Sometimes one can see shapes in random noise.
- perplexity parameter equivalent to variance σ^2 (range $[5, 50]$, default 30)

UMAP

- Any distance can be plugged into UMAP, not only euclidean distances
- The distributions are not normalized

$\Longrightarrow$ UMAP much faster than tSNE

- Uses binary cross-entropy as a cost function instead of the KL-divergence. Nearest Neighbors instead of perplexity
- Better preserves global structure; Not limited to the first 2-3 dimensions
- min_dist ($[0.001, 0.5]$ 0.1): Larger values ensure embedded points are more evenly distributed, while smaller values $\Longrightarrow$ more accurate local structure
- n_neighbors ($[2, 100]$ 15): Determines the number of neighboring points used in local approximations of manifold structure. Larger values will result in more global structure being preserved at the loss of detailed local structure.

4 Variant Calling

MAQ Algorithm

- Given read z coming from position u on reference sequence x
- Assume that error are independent at site of the read: $p(z|x,u) = \prod_i p(z_i|x, u_i)$
- Assume that $p(u|x)$ is uniformly distributed
- Model $p(z|x,u)$ as: $p(z|x,u) = \prod_i 10^{-\frac{Q_i}{10}}$
- The posterior will be: $p(u|x,z) = \dfrac{p(z|x,u)p(u|x)}{\sum_{v=1}^{L-I+1} p(z|x,v)p(v|x)}$
- In practice, summing over reference sequence omitted for some well chosen constants

MAQ Genotype Calling

- Assume we observe k nucleotides b and $n-k$ nucleotides with allele a
- Assume that our prior is: $P(<a,b>) = \begin{cases} (1-r)/2 & a! = b \\ r & a = b \end{cases}$
- We can model the likelihood (assuming independence): $P(D|<a,b>) = \binom{n}{k}(0.5)^k(1-0.5)^{n-k}$, $P(D|<b,b>) = \binom{n}{n-k}(1-\epsilon)^k(\epsilon)^{n-k}$, $P(D|<a,a>) = \binom{n}{k}(\epsilon)^k(1-\epsilon)^{n-k}$
- We get the posterior with bayes and call the genotype: $\hat{g} = arg\max p(g|D)$ • (Problem) Linkage Blocks: local SNPs are highly correlated, probability is not trivial unless we assume independence

Haplotype Caller

- **Identify active regions**; sliding window along the reference, count mismatches/indels, trigger active region to be processed over a threshold
- **Assemble plausible haplotypes**: assemble a k-mer graph with the reads; weight each path based on the read count evidence; prune unlikely paths (bubbles)
- **Determine per read likelihoods**: PairHMM to determine the likelihood of haplotypes given a read; Determine most likely allele for each read; Take the highest probability of haplotypes (among those that contain the allele) given reads

Reference:	ATCGATCATAGCTAGCTGCG
Haplotype 1:	ATCGA-CATAGCTAGCTGCG
Haplotype 2:	ATGGATCATAGCTTGCTGCG
Haplotype 3:	ATCGA-CATAGCTTGCTGCG

Haplotypes

*	R	1	2	3
1	0.01	0.02	0.03	0.04
2	0.09	0.06	0.07	0.08
3	0.10	0.11	0.01	0.02

Reads

Alleles

-	T	
0.04	0.03	1
0.08	0.09	2
0.11	0.10	3

Reads

Bayesian model

Prior of the genotype — Likelihood of the genotype

$$\Pr\{G|D\} = \frac{\Pr\{G\}\Pr\{D|G\}}{\Sigma_i \Pr\{G_i\}\Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$

$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2}\right) \text{ where } \overbrace{G = H_1 H_2}^{\text{Diploid assumption}}$$

$\Pr\{D|H\}$ is the haploid likelihood function

4.1 Reference Free with De Bruijn

- Assemble input sequencing data into (colored) de Bruijn graph
- Identify local variants as bubbles in the graph
- Compute path quantification for bubbles on the read data
- Derive ranking or likelihood score to prioritize variants
- Calling is more difficult if variants have a distance of less than k to each other or long insertion are handled

Somatic Variant Calling

- **Main challenges**:
- Purity: contamination of normal cells with cancer cells • Tumor purity $= \dfrac{\text{tumor cells}}{\text{normal + tumor cells}}$ • The higher the purity, the easier the task
- Tumor heterogeneity • More complex mutations:not only SNPs and indels • No reliance on diploid assumption • Somatic mutations are not randomly distributed (driver genes)

5 GWAS

Advantages

- No family tree needed, but just bulk genotyping data
- Translatable to clinic quickly • Highly reproducible
- GWAS can detect variants located in poorly understood regions of the genome

Disadvantages

- Limited to large effects and common variants • Linkage Disequilibrium will make it difficult to identify specific causal variant • Typically population can stem from different geographic regions. • Missing heritability: height is roughly 80% heritable but GWAS can only explain 45%

Testing for association

Allele	Cases (with AMD)	Controls (without AMD)	Total Alleles
C	a	b	a+b
T	c	d	c+d
Total Alleles	a+c	b+d	a+b+c+d

- **Fisher Exact Test**: $p = \dfrac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$
- χ^2 **test**: $\chi^2 = \sum_{i=1}^n \dfrac{(O_i - E_i)^2}{E_i}$ $E_1 = \dfrac{(a+b)(a+c)}{(a+b+c+d)}$ Df = (r-1)x(c-1)

Linear Regression

- $Y = \beta_0 + X_1\beta_1 + \epsilon$
- $Y \in \{0,1\}$ $\mathcal{R}$ (phenotype)
- $X_1 \in \{0,1,2\}$ (AA,AB,BB)
- $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \ne 0$
- $t_{n-2} = \dfrac{\hat{\beta}_1}{s_{\beta_1}}$ • $s_{\beta_1} = \sqrt{\dfrac{1}{n-2}\dfrac{\sum_i(y_i - \hat{y}_i)^2}{\sum_i(x_i - \overline{x})^2}}$
- p-values are uniformly distributed under the hypothesis H_0

Multiple testing correction

- P(reject at least once) = 1-P(do not reject)= $1 - (1-0.05)^N$
- We are testing 3 millions positions with GWAS $\Longrightarrow N = 10^6$
- **Bonferroni approach**: All tests are independent (assumption)
- Given $p_1, \dots, p_m$ p-values, then we reject the Null hypothesis for each: $p_i \le \alpha/m$

Population Structure

- Let X be a genotype matrix (#patients)x(#SNPs)
- Do the PCA on $K = XX^T$
- Use PCs as covariates in the association analysis

Linear Mixed Models

- Accounting for structure between individuals (not just population dependency!)
- $y = X\beta + u + \epsilon$
- $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$
- $u \sim \mathcal{N}(0, \sigma_K^2 K)$
- u is a vector of polygene background effects
- K is the kinship relatedness matrix

Meta-analysis

- Combining p-values for a given SNP from k studies: $\chi_{2k}^2 = -2\sum_{i=1}^k \log p_i$
- A log of a uniform follows an exponential distribution. Factor 2 yields chi-squared

SIFT

- Identify protein which overlaps mutational position of interest
- Homology search (Find all similar protein sequences) using PSI-BLAST (position weight based)
- Multiple sequence alignment from PSI-BLAST • Calculate probabilities
- if the probability of amminoacid appearing in that poisiton is < 0.05 then mark as deleterious

6 Ontologies

Basic Formal Ontology (BFO)

- Define **universals** (classes) and **particulars** (instances)
- **continuant** are persistent objects that preserve their identity over time (cellular components)
- **occurrent** is an entity that happens / develops through time and describes an event that continuants participate in (biological process)

Gene Ontology

- Three sub-ontologies:
- **Molecular function**: describes the biochemical activity of a product (enzymatic reaction)
- **Biological process** : describes a biological objective (change of cell state,regulation)
- **Cellular component**: describes location inside the cell where the product is active • Relational links between the GO concepts form a graph structure that can be used for annotation propagation or inference:
- is_a, part_of, instance_of,regulates

Term for Term Testing

- $m_t \subset M$ subset of M with annotation t
- $n_t \subset N$ subset of N with annotation t
- We use the hypergeometric test to compute whether our observation represents a significant enrichment
- $P(X_t = k) = \dfrac{\binom{m_t}{k}\binom{m-m_t}{n-k}}{\binom{m}{n}}$
- H_0: no positive association of term t and study set n
- H_1 : there is an overrepresentation of t in the study set
- $P(X_t \ge n_t | H_0)$
- Use corrective measures on the resulting p-values (Bonferroni)

Gene Set Enrichment Analysis

- Given a list L of n items pre-ranked by a feature of interest (e.g., genes by differential expression between two samples), assess whether distribution of terms annotating a subset S of L is associated with the given ranking
- Compare fractions of items in S vs. fraction of items not in S relative to their ranks r_j up to a given position i in the ranked list L
- $ES = \max_i |P_{hit} - P_{miss}|$
- $P_{hit}(S,i) = \sum_{g_j \in S\ \forall j \le i} \dfrac{|r_j|^p}{n_r}$, with $n_r = \sum_{g_j \in S} |r_j|^p$
- $P_{miss}(S,i) = \sum_{g_j \notin S\ \forall j \le i} \dfrac{1}{n - n_s}$, with $n_s = |S|$ • **Significance Assessment**: • generate k random gene sets M_i (with k typically >1000) • compute empirical distribution of $ES(M_i)$from the random set • asses significance of $ES(S)$ relative to the empirical distribution

Human Phenotype Ontology(HPO)

- We can define the similarity of two terms t_1, t_2 sharing ancestors $A(t_1, t_2)$
- sim(t_1, t_2) = $\max_{a \in A(t_1, t_2)} - \log p(a)$
- $p(a)$ is the probability of term a measured as its frequency of annotation over all diseases in the database
- We can define the similarity of two diseases d_1, d_2:
- $sim(d_1 -> d_2) = avg\left[\sum_{s \in d_1} \max_{t \in d_2} sim(s,t)\right]$
- To break the asymmetry of the distance we have:
- $sim(d_1, d_2) = \dfrac{sim(d_1 -> d_2) + sim(d_2 -> d_1)}{2}$