Probability Models for RNA Assembly and Analysis

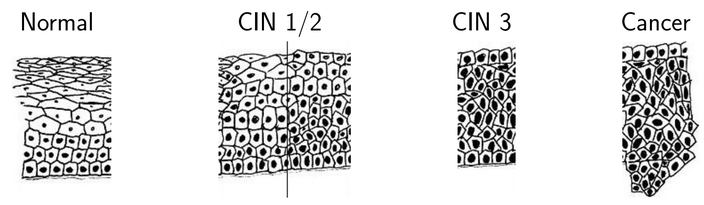
Nathanael Fillmore and others (vide infra)
University of Wisconsin, Madison, Computer Sciences

Progression and Gene Expression in Cervical Cancer with P.F. Lambert, P. Ahlquist, and M.A. Newton

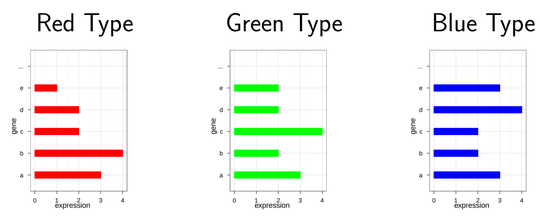
Goal: Develop a statistical model of changes in gene expression through four stages in the development of cervical cancer, and use this model to understand aspects of cervical cancer progression.

Model

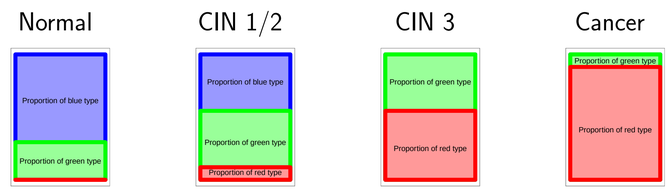
- Tissue at each stage of the progression leading to cervical cancer is composed of cells of several different types, mixed together; different stages are associated with different relative proportions of each type:[a]



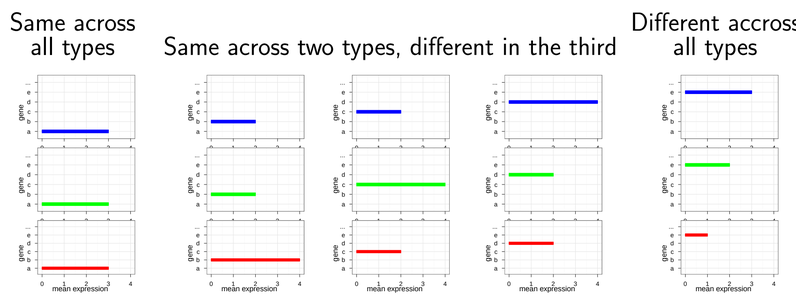
- Each type of cell in a tissue sample has a separate "pure" gene-expression profile:



- Since the cells in each tissue sample are all mixed together, the observed gene-expression profile is a weighted average of the pure type-specific profiles; the weights are the proportions of cells of each type at each stage of the progression:



- Each gene follows a particular pattern of differential expression across the cell types:



- Each subgroup of types within each differential expression pattern is associated with a common mean log expression level shared across patients, genes following the pattern, and types contained in the subgroup; each specific log expression measurement is assumed to follow a normal distribution around the mean, with common variance shared by all genes, patients, and types.

The mean log expression levels are also assumed to follow a normal distribution, with a single grand mean and variance.

[a] Figure from `http://staffwww.dcs.shef.ac.uk/people/D.Walker/research/probe_cin.jpg`.

Computational challenges solved

- The likelihood has a closed form, even though the overall (heterogeneous) expression's distribution is only specified indirectly.
- We evaluate the likelihood efficiently using a block Cholesky decomposition.
- We find the maximum likelihood estimate of all parameters using a Newton-style interior point method.

Validation

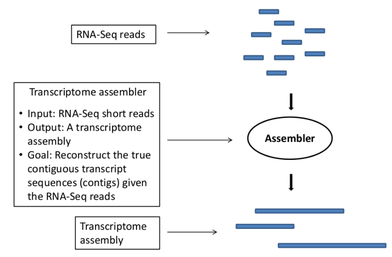
- Experiments with simulated data show that the true parameters can be recovered effectively, if the model assumptions hold.

Preliminary results

- Each of 128 cervical tissue samples (24 normal, 36 CIN 1/2, 40 CIN 3, 28 cancerous) was measured by an Affymetrix whole genome microarray, which contains about 54,000 probe sets. (From the SUCCEED project.)
- We have used our procedure to get an estimate for the parameters.
- This tells us also the distribution (i) of differential-expression patterns and (ii) of the type-specific mean log expression levels.
- Much information is available and under analysis.

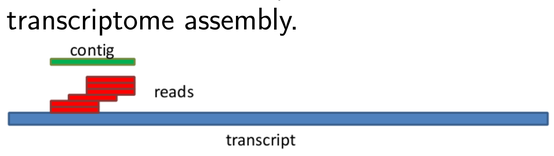
RSEM-Eval: A Probabilistic Transcriptome Assembly Evaluator with B. Li and C.N. Dewey

Goal: Determine which transcriptome assembly, out of several candidate assemblies, is best based only on the RNA-Seq data from which the assemblies were constructed, without a reference. To do so, try to model as accurately as possible the process of RNA-Seq read generation and the process of ideal transcriptome assembly.



Model

- The full joint distribution:
  $$\boldsymbol{P}(\text{contigs}, \text{coverages}, \text{reads}) = \underbrace{\boldsymbol{P}(\text{contigs}, \text{coverages})}_{\text{prior}} \underbrace{\boldsymbol{P}(\text{reads}|\text{contigs}, \text{coverages})}_{\text{likelihood}}.$$
- Prior: Based on ideal transcriptome assembly.



- Likelihood: Based on RSEM, with a modification due to the fact that we are conditioning on an assembly, not a full transcriptome.
- Our score:
  $\boldsymbol{P}(\text{contigs}|\text{reads}) \propto \boldsymbol{P}(\text{contigs}, \text{reads}) = \int \boldsymbol{P}(\text{contigs}, \text{coverages}, \text{reads})\, \boldsymbol{d}\text{coverages}$.
- Approximate by BIC, i.e., "width times height".

Alignment-based meta-evaluation criteria

- Oracleset: ideal assembly - the best we can achieve.
- Precision:
  - For each contig $\boldsymbol{a}$ in the assembly, let $\boldsymbol{b}(\boldsymbol{a})$ be the best-aligned oracleset element. If the best alignment $(\boldsymbol{a}, \boldsymbol{b}(\boldsymbol{a}))$ is still bad, we ignore it.
  - Nucleotide: Precision is the fraction of bases in the assembly that exactly match their best-aligned counterparts in the oracleset.
  - Pair: Precision is the fraction of pairs of bases, $\boldsymbol{k}$ positions apart, in which both ends of the pair exactly match their best-aligned counterparts in the oracleset.
  - Transcript: Precision is the fraction of transcripts in the assembly that have sufficiently high identity with their best-aligned counterparts in the oracleset.
- Recall is the reverse, interchanging the assembly and the oracleset.
- F1 is the harmonic mean of precision and recall.

$k$mer-based meta-evaluation criteria

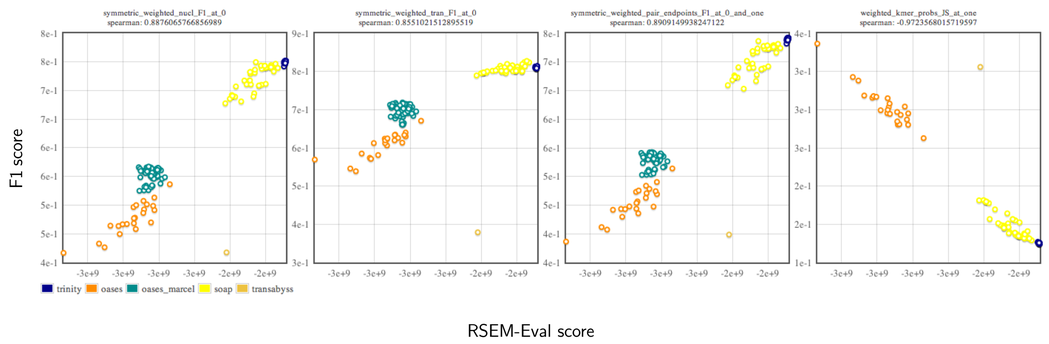
- Oracleset: ideal assembly - the best we can achieve.
- Let $\boldsymbol{K} = \{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}^k$ be the set of all possible $\boldsymbol{k}$mers.
- Each assembly induces a probability distribution $\boldsymbol{\mu}$ over $\boldsymbol{K}$ by counting how many times each $\boldsymbol{k}$mer occurs in the assembly and normalizing.
- The oracleset also induces such a probability distribution $\boldsymbol{\nu}$.
- The Jensen-Shannon divergence between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ is
  $$\frac{1}{2}\boldsymbol{KL}\left(\boldsymbol{\mu} \,\middle\|\, \frac{1}{2}(\boldsymbol{\mu}+\boldsymbol{\nu})\right) + \frac{1}{2}\boldsymbol{KL}\left(\boldsymbol{\nu} \,\middle\|\, \frac{1}{2}(\boldsymbol{\mu}+\boldsymbol{\nu})\right)$$
  where $\boldsymbol{KL}$ is the KL divergence.

Data

- Data: 50 M single-end 76bp reads from a mouse.[b]
- Assemblies: We generated about 150 assemblies by running different assemblers with different parameter settings.
- Reference (for meta-evaluation): RefSeq.
- We have also run this experiment on a simulated mouse dataset, a different (real) mouse dataset, and a yeast dataset.

[b] Grabherr, M. G. et al. (2011). Nature Biotech. 29, 644-652.

Results



N. Fillmore was supported by an NLM training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM 5T15LM007359).