Progression and Gene Expression in Cervical Cancer

N. Fillmore[1], P.F. Lambert[2], P. Ahlquist[3], and M.A. Newton[4]

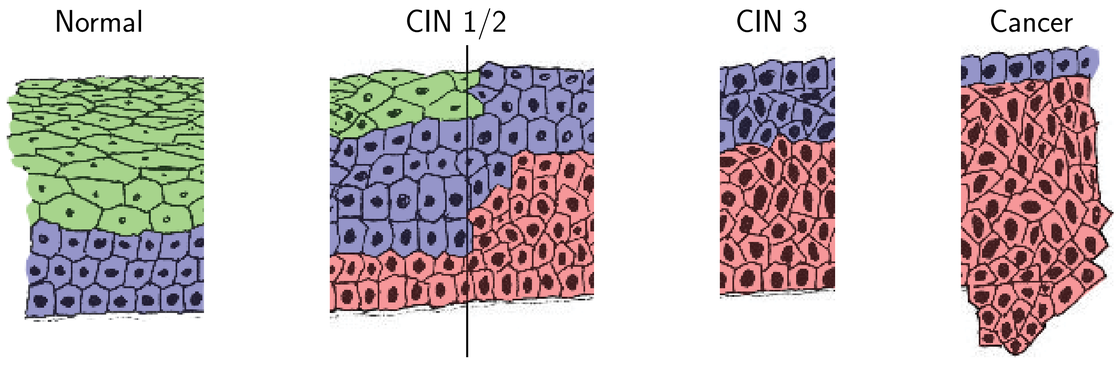
[1]Computer Sciences, [2]McArdle Lab. for Cancer Research, [3]Inst. for Molecular Virology and Howard Hughes Medical Inst., [4]Statistics and BMI

Goal

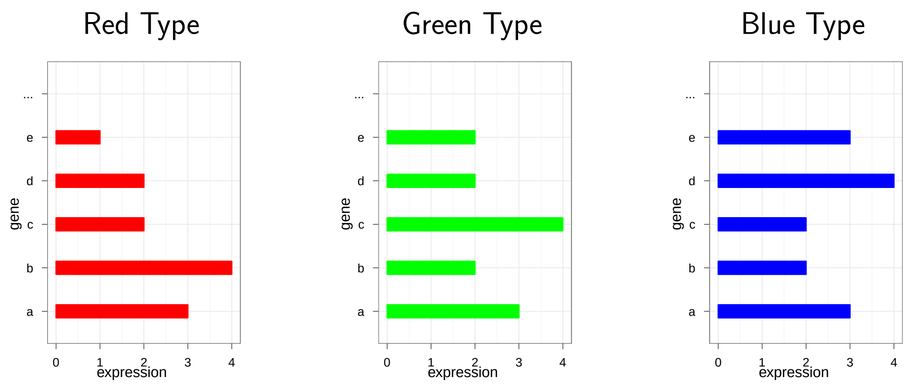
Develop a statistical model of changes in gene expression through four stages in the development of cervical cancer, and use this model to understand aspects of cervical cancer progression.

Model - Overview

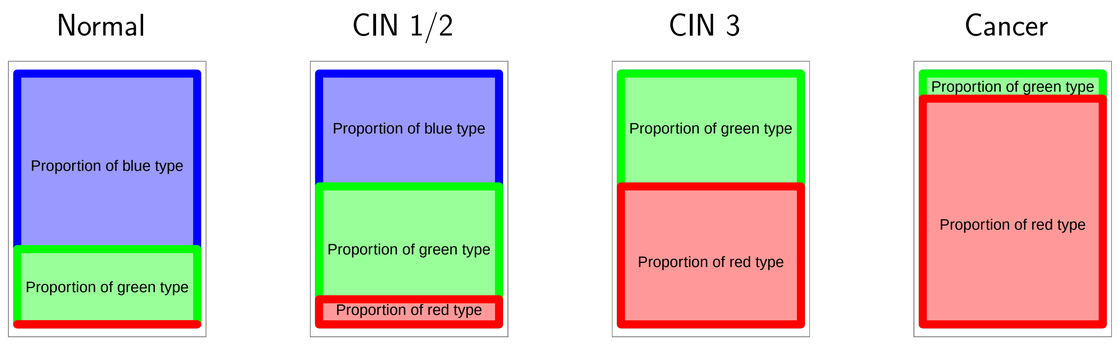
- Tissue at each stage of the progression leading to cervical cancer is composed of cells of several different types, mixed together; different stages are associated with different relative proportions of each type:[a]



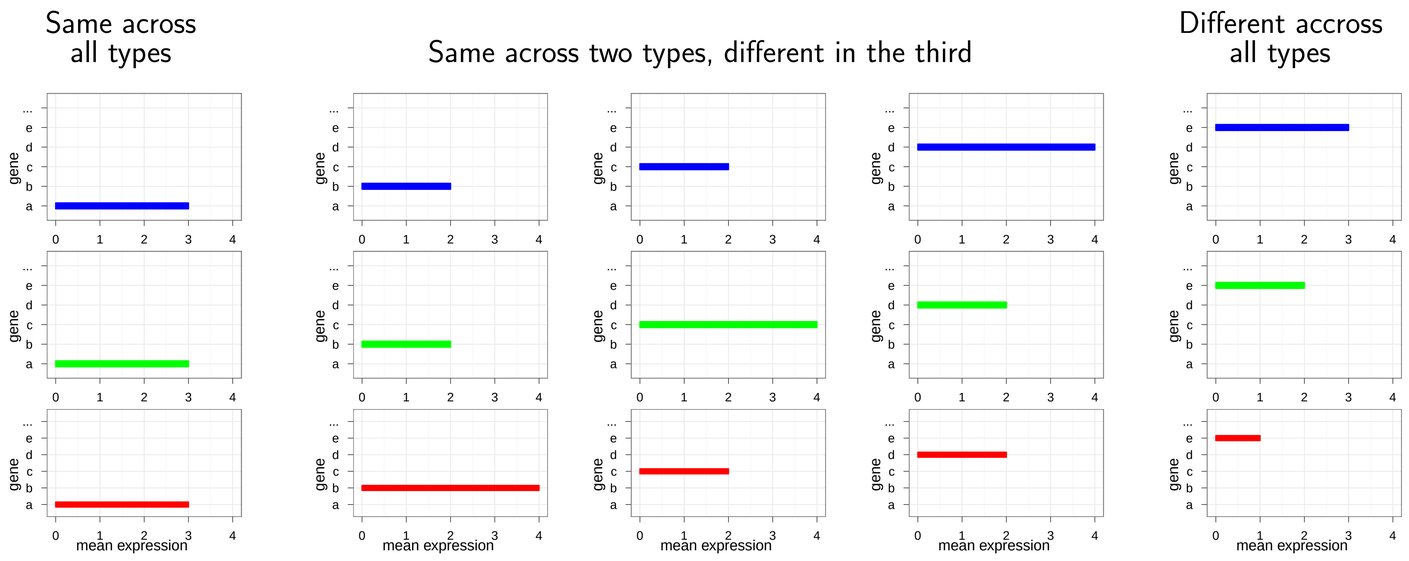
- Each type of cell in a tissue sample has a separate "pure" gene-expression profile:



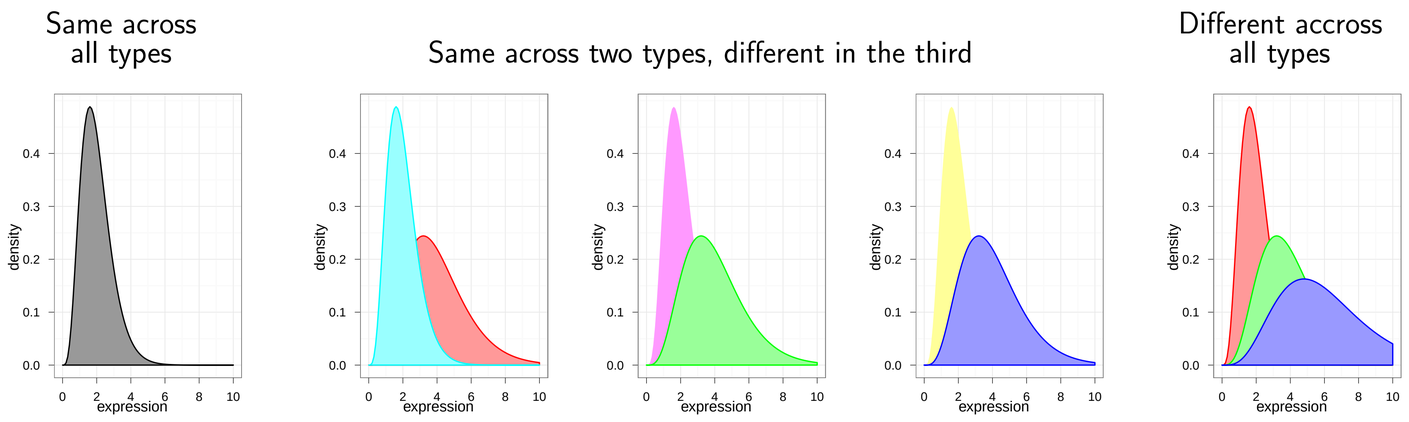
- Since the cells in each tissue sample are all mixed together, the observed gene-expression profile is a weighted average of the pure type-specific profiles; the weights are the proportions of cells of each type at each stage of the progression:



- 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 expression level shared across patients, genes following the pattern, and types contained in the subgroup; each specific expression measurement is assumed to follow a gamma distribution around the mean, with a shape parameter shared by all genes, patients, and types.



The mean expression levels are also assumed to follow a gamma distribution, with a single grand mean and shape.[b]

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

[b] Gamma-gamma model from Kendziorski et al. (2003).

Model - Details

Fixed quantities:

- $n = 128$ tissue samples, indexed by i.
- σ_i - the stage of each tissue sample.
- $G \approx 54,000$ genes, indexed by g.
- T types, indexed by t.
- J patterns of differential expression, indexed by j; J is a function of T.

Parameters of interest:

- $\pi_1, \ldots, \pi_J$ - coefficients of mixture over patterns of differential expression.
- $p_{\sigma,t}$ - proportion of cells of type t in tissue at stage σ.
- a - shape parameter for distribution around each subgroup's mean.
- a_0 - shape parameter for distribution of subgroup means around the grand mean.
- ν - scale parameter for distribution of subgroup means around the grand mean.

Random variables:

- Z_g - gene g's expression pattern; follows Categorical(π).
- $\Lambda_{j,\mathcal{T}}$ - $a/\Lambda_{j,\mathcal{T}}$ is the mean expression level within subgroup $\mathcal{T}$ of expression pattern j; $\Lambda_{j,\mathcal{T}}$ follows Gamma(a_0, ν).
- $X_{i,g,t}$ - expression level of gene g within cells of type t in tissue sample i; follows Gamma($a, \lambda_{z_g,\mathcal{T}}$), where $\mathcal{T}$ is the subgroup of expression pattern z_g that contains type t.
- $S_{i,g}$ - overall expression level of gene g in tissue sample i; $S_{i,g} = \sum_{t=1}^{T} p_{\sigma_i,t} X_{i,g,t}$.

$S_{i,g}$ is observed; all other variables are latent.

Data

- 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.
- Data is from the Study to Understand Cervical Cancer Early Endpoints and Determinants.
- A previous analysis (M.A. Newton) identified genes showing various patterns of differential expression among the four stages.

Estimation - In Progress

- Markov chain Monte Carlo simulation of the parameters and the latent variables, given the observed expression levels.
- Posterior mean estimate of each parameter.

Other Work

- With Colin Dewey and Bo Li, I am working on principled evaluation of de novo transcriptome assemblies from RNA-seq data.

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