A Bayesian model for gene expression on MERFISH data

Johannes Köster[1,2,3], Xiaole Shirley Liu[1,3]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health
[2]Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School
[3]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute

Introduction

Multiplexed error-robust fluorescence in-situ hybridization (MERFISH, Chen et al. (2015)) is a new approach to measure RNA molecules of hundreds of different genes in single cells in parallel, while retaining spatial information.
MERFISH uses the following strategy to label hundreds of RNA species in parallel with a limited set of colors (see Fig. 1):

- RNA molecules are labeled with multiple probes
- N hybridization rounds are performed
- in each round, different probes are marked fluorescently, generating a binary pattern for each RNA molecule
- RNA species are identified by comparing the pattern against the designed probsets
- a modified 4-bit distance hamming code is used for robustness

We present a Bayesian model to reliably estimate gene expression and differential gene expression on MERFISH data for any number of cells. The presented approach will be available soon as a set of fast, parallelized command line utilities, implemented on top of Rust-Bio (Köster 2015).

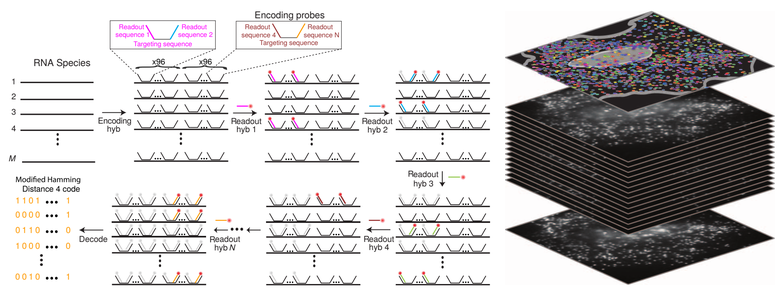

Fig. 1: MERFISH workflow (Chen et al. 2015).

Estimating gene expression

The obtained binary words for each RNA molecule are assigned to RNA species (i.e., genes). Thereby, single bit errors can be corrected, obtaining an exact and a corrected count for each gene. Chen et al. (2015) report prior probabilities for making $1 \to 0$ or $0 \to 1$ errors, which can be used to derive the probability of an exact or corrected call or miscall and the probability to miss a molecule.
Let D be the given data, i.e., the binary words obtained assigned to a particular RNA species. We obtain the likelihood of an expression x by summing over all possibilities to have x_c correct calls, x_m miscalls, and $x - x_c$ misses as

$$\Pr(D|X = x) = p_{\text{miss}}^{x-x_c} \cdot \sum_{i=0}^{\min\{|D_e|,x\}} \binom{|D_e|}{i} \binom{|D| - |D_e|}{x_c - i} \cdot p_{\text{exact call}}^{i} \cdot p_{\text{corrected call}}^{x_c-i} \cdot p_{\text{exact miscall}}^{|D_e|-i} \cdot p_{\text{corrected miscall}}^{x_m-|D_e|+i}.$$

Using Bayes theorem, we can calculate the posterior probability $\Pr(X = x|D)$ for an expression x (see Fig. 2 and 3). For a set of n cells with data $\mathcal{D}$, the posterior probability for a mean expression $\overline{X} = x$ in all cells is

$$\Pr(\overline{X} = x|\mathcal{D}) = \sum_{\mathbf{x}:n^{-1}\sum_{i=1}^{n}\mathbf{x}_i=x} \prod_{i=1}^{n} \Pr(X = \mathbf{x}_i|D_i).$$

The probability mass function can be calculated in pseudo-polynomial time with dynamic programming.

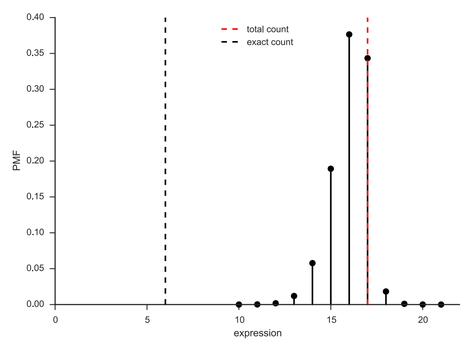

Fig. 2: Probability mass function for a typical gene with low counts.

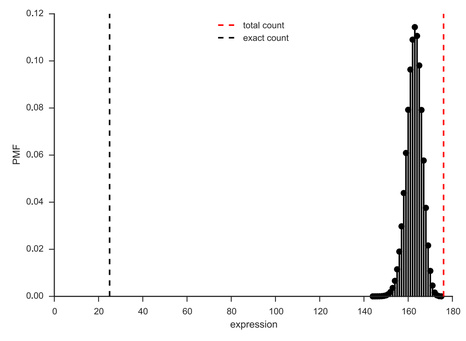

Fig. 3: Probability mass function for a typical gene with high counts. Our Bayesian model reveals that the raw counts are too optimistic here.

Estimating differential gene expression

The posterior probability for a given fold change f between the sets of cells $\mathcal{D}$ and $\mathcal{D}'$ can then be calculated as

$$\Pr(F = f|\mathcal{D}, \mathcal{D}') = \sum_{x} \Pr(\overline{X} = x|\mathcal{D}) \Pr(\overline{X} = fx|\mathcal{D}').$$

On top of this probability, we can calculate the conditional expectation for the fold change and the corresponding standard deviation (see Fig. 4). We can further obtain the posterior error probability (PEP) for differential expression as $\Pr(|F| \leq f_{\text{min}}|\mathcal{D}, \mathcal{D}')$ and use it to control the false discovery rate (FDR) in a Bayesian way as described by Muller, Parmigiani, and Rice (2006).

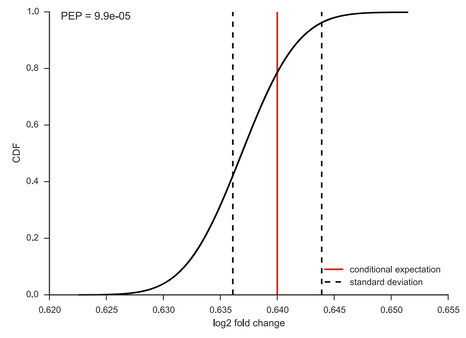

Fig. 4: Cumulative distribution function of fold change for an example gene between two sets of cells.

References

Chen, K. H., a. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang (2015). "Spatially resolved, highly multiplexed RNA profiling in single cells". In: *Science*.
Köster, J. (2015). "Rust-Bio: a fast and safe bioinformatics library". In: *Bioinformatics*.
Muller, P., G. Parmigiani, and K. Rice (2006). "FDR and Bayesian multiple comparisons rules". In: *Bayesian Statistics 8* 1995, pp. 349–370.

LaTeX TikZposter