Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics

Charles Gawad, Winston Koh and Stephen R. Quake

Pandu Raharja

Technische Universität München
Ludwig-Maximillians-Universität München

June 20, 2017

- Gawad, Charles, Winston Koh, and Stephen R. Quake. "Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics." Proceedings of the National Academy of Sciences 111.50 (2014): 17947-17952.

Outline

- Motivation
- Methods (+ Results)
- Code session
- Discussion

Tumor Heterogeneity[1]

- **Cancer** – a generic term for cells with uncontrolled growth:
  - Carcinoma: epithetlial origin
  - Leukemia: bone marrow origin
  - Lymphoma: lymph node origin
  - Sarcoma: connective tissue origin
  - *and others*
- Each class is further divided into multiude of categories
- Each category has its own morphological, metabolic, genetic and growth characteristics

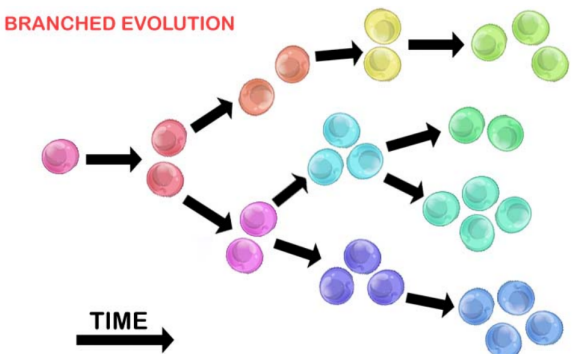
[1] or *why is it almost impossible to "cure" cancer?*

Tumor Heterogeneity

- Concerted efforts have been done to map mutational landscape of major cancer classes
- (TCGA consortium) Kandoth, Ding *et al* , Nature, 2013:
  " *The Cancer Genome Atlas (TCGA) has used the latest sequencing and analysis methods to identify somatic variants across thousands of tumours...Here we present data...3,281 tumours across 12 tumour types...we identified 127 significantly mutated genes...and emerging...cellular processes in cancer...[A]verage number of mutations...varies across tumour types...Mutations in transcriptional factors/regulators show tissue specificity, whereas histone modifiers are often mutated across several cancer types*"
- **TL;DR:** cancer is a damn complex class of diseases

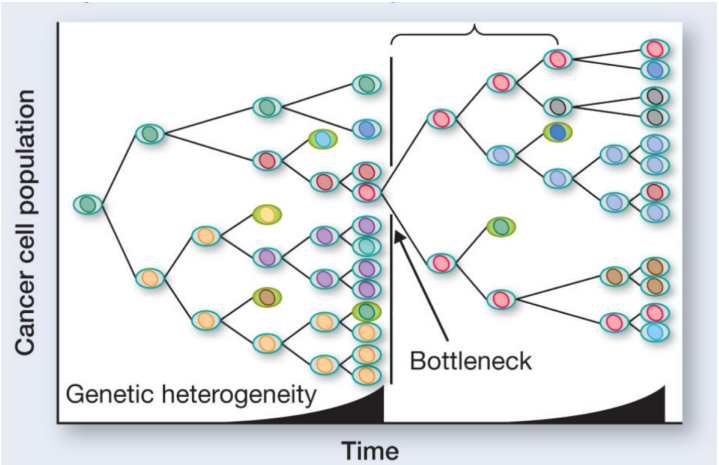
Intratumor Heterogeinity

- Cancer is mostly born out of series of mutation and epigenetic changes
- These changes are inherited along the lineage
- New mutation happening in subpopulation create a new clone



wikipedia

Intratumor Heterogeinity and Cancer Treatment



Swanton, Nature, 2012

This paper's motivation

- Study the dynamics of intratumor heterogeinity:
  - Determination of segretation pattern of tumor clones
  - Reconstruction of clonal tree
  - Identification of common features across clones
  - Discovery of proliferative oncogenic point mutations
  - *etc*
- Provide high resolution dynamics of cancer development (in this case, *acute lymphoblastic leukimia (ALL))*

Methods

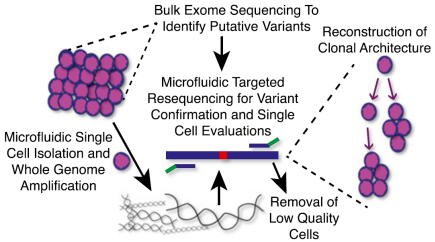
- Single-cell cancer genome sequencing protocol
- Computational methods to segregate leukimia into distinct clonal populations using combined:
  - **Expectation-Maximization (EM)** on **multivariate Bernoulli model** with **AIC** regularization followed by **Multiple Correspondence Analysis**, AND
  - **Clustering** of mutations using Jaccard distance with clone number estimation using **sum of square error**

  followed by clone tree reconstruction using **minimum spanning tree reconstruction** on consensus data

Single-cell cancer genome sequencing

Main goal: identify single-cell variants to be used in further analysis

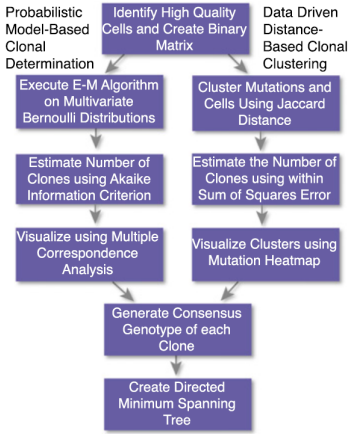
- Bulk Exome Sequencing was conducted to establish reference sequence
- Resequencing of each cell (microfluidic cell isolation and sequencing) to identify cell-specific variants
- Removal of low quality cells



Gawad et al, PNAS, 2014

Computational Methods

Combined **probabilistic** and **clustering** approach



Gawad et al, PNAS, 2014

Probabilistic Approach

Define:

- Vector representation of an individual's single-cell mutational profile $\mathbf{x} := (x_1, x_2, \cdots x_d)$ with $x_i \in \{0, 1\}$ denoting presence of mutation at i-th base.
- Probability of observing mutation $\theta_i = P(x_i = 1)$

We can model mutation as Bernoulli process and hence arrive at following probability of single-cell mutational profile x:

$$P(\mathbf{x}|\Theta) = P(x_1, \cdots, x_d|\Theta) = \prod_{i=1}^{d} \theta_i^{x_i}(1 - \theta_i)^{1-x_i}$$

(This is the model for only one clone, i.e. single lineage)

For different clones $c_1, \cdots c_J$ with π_j the proportion of j-th clone and $\sum_{j=1}^{J} \pi_j = 1$ we can represent the probability as finite mixture of multivariate Bernoulli distribution **for each cell**:

$$P(\mathbf{x}|\Theta) = \sum_{j=1}^{J} \pi_j P(\mathbf{x}|\Theta_j) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i} (1-\theta_{ji})^{1-x_i}$$

The probability for N cells are then:

$$P(\mathbf{x^{(1)}}, \cdots, \mathbf{x^{(N)}}|\Theta) = \prod_{n=1}^{N} (\sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} {\theta_i}^{x_{ni}} (1-\theta_{ji})^{1-x_{ni}})$$

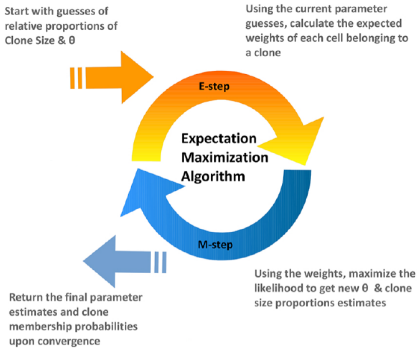
The log-likelihod of parameters $\{\pi_j, \Theta_j\}_{j=1}^{J}$ can then be written as:

$$l(\theta, \pi) = \sum_{n=1}^{N} log(\sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} {\theta_i}^{x_{ni}} (1-\theta_{ji})^{1-x_{ni}})$$

Probabilistic Approach – Expectation Maximization

Estimation of clone membership of each cell using **Expectation Maximization (EM)**:

1. **Estimation step**: assign the clone category weights for each cell using the likelihood $l(\theta, \pi)$, i.e. $E[\mathbf{c}|l]$
2. **Maximization step**: maximize the likelihood to get new θ & clone proportion estimate based on estimated weights, i.e. $argmax_{\{\theta,\pi_1 \cdots \pi_N\}}\{l(\theta, \pi)\}$
3. Repeat steps 1 & 2 until the parameters converge



Gawad et al, PNAS, 2014

Probabilistic Approach – BIC

- During EM step, models with larger number of clones are preferred:

$$l(\theta, \pi) = \sum_{n=1}^{N} log(\sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_i^{x_{ni}} (1 - \theta_{ji})^{1 - x_{ni}})$$

- Overfitting is done by introducing BIC (Bayesian Information Criterion), which would penalize models with more free parameters:

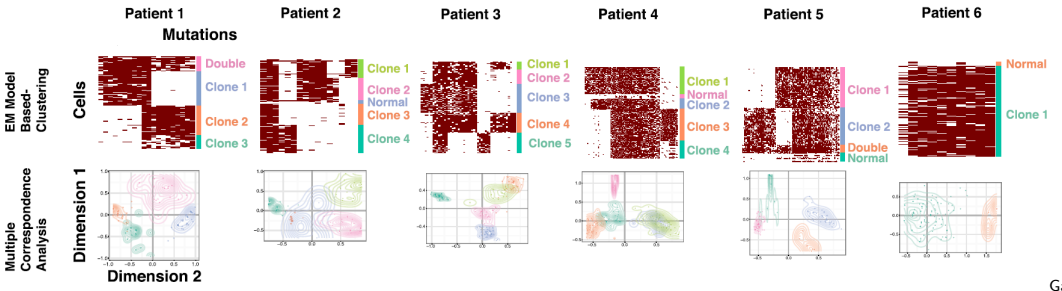
$$BIC = ln(N)J - 2ln(l)$$

Upon this, each cell will be assigned to respective clonal population given the model and observed data.

Probabilistic Approach – Multiple Correspondence Analysis (MCA)

MCA $\sim$ Principal Component Analysis (PCA) on categorical data:

- Representation of the cells in two-dimensional space allows the validation of clone assignment

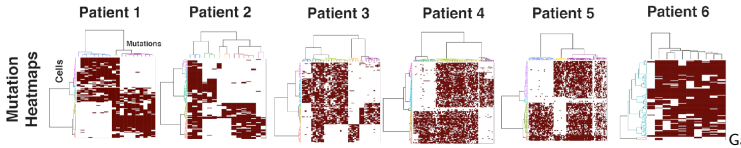


Gawad et al, PNAS, 2014

Clustering Approach – Clustering on Jaccard Distance

$$J(\mathbf{x_1}, \mathbf{x_2}) = \frac{|\{x_i | x_{1i} = 1 \wedge x_{2i} = 1\}|}{|x_1|}$$

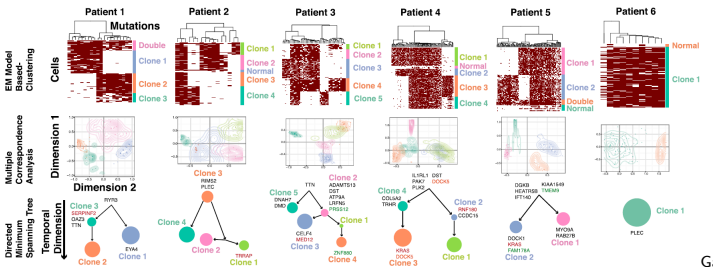
- Hierarirchal clustering over the distances reveals tree-structured relations among cells.
- Number of clones is estimated using sum of square error of elements in a cluster (Ward's method)



Gawad et al, PNAS, 2014

Combined Approach

- Clonal assignment results from both approaches are combined to show validity.
- Clones genealogy inferrence is done by introducing temporal ordering of the cells and applying Maximum Parsimony method (Farris, 1970 & Fitch, 1971).



Gawad et al, PNAS, 2014

Code Session

Data and code repo: github.com/lianchye/Clonal_Analysis.git

Paper review

- **Readability:** 4 – paper is overal readable
- **Reproducibility:** 5 – github repo!
- **Novelty:** 3 – simple and yet powerful method to identify intratumor heterogeneity
- **Impact:** 4 – usable for other cancer classes
- **Aesthetics:** 3 – aesthetic okay, some typos in appendix
- **Structure:** 2 – personally more interested in methods and less in tiny detailed results
- **Overall:** 4 – good paper

References

- Gawad, Charles, Winston Koh, and Stephen R. Quake. "Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics." Proceedings of the National Academy of Sciences 111.50 (2014): 17947-17952.
- Kandoth, Cyriac, et al. "Mutational landscape and significance across 12 major cancer types." Nature 502.7471 (2013): 333-339.
- Swanton, Charles. "Intratumor heterogeneity: evolution through space and time." Cancer research 72.19 (2012): 4875-4882.
- Ward Jr, Joe H. "Hierarchical grouping to optimize an objective function." Journal of the American statistical association 58.301 (1963): 236-244.
- Farris, James S. "Methods for computing Wagner trees." Systematic Biology 19.1 (1970): 83-92.