Cell type deconvolution

Mikhail Dozmorov

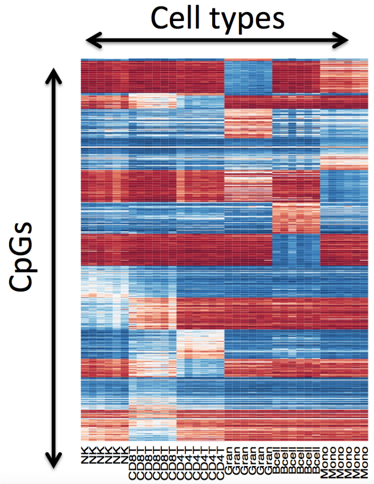
Spring 2018

Characterizing Cell-types

- Goal: Whole transcriptome (and epigenome) profiles of individual cell-types
- Problem: Whole-tissue sample is composed of two or more distinct cell types
- Example: Blood samples comprise a mixture of predominantly 5 cell types
    - Neutrophils
    - Lymphocytes
    - Monocytes
    - Basophils
    - Eosinophils

Blood is a mixture of many cell types

Whole blood cell types: T cells (CD8T, CD4T, Natural Killer), B cells, Granulocytes, Monocytes Bioconductor data package available: library(FlowSorted.Blood.450k)



Characterizing Cell-types

- Technically challenging to measure whole transcriptome expression from single-cells
- Approach: Computational deconvolution of cell mixtures
    - Reference-based
    - Reference-free
    - Semi-reference-free

Reference-based cell-type deconvolution

- Using an existing reference DNA methylation (DNAm) database of cell types that are thought to be present in the tissue of interest
- Estimated fractions are relative (can be absolute if the dominant cell type is known)
- Estimated fractions can then be used as covariates in supervised multivariate regression models to infer differentially methylated cytosines (DMCs) that are independent of changes in cell-type composition.

Advantages

- Absolute or relative cell-type fractions can be estimated in each individual sample
- If required, they can be easily combined with batch-correction methods such as COMBAT
- The model itself is relatively assumption free

Reference-based cell-type deconvolution

- Using an existing reference DNA methylation (DNAm) database of cell types that are thought to be present in the tissue of interest
- Estimated fractions are relative (can be absolute if the dominant cell type is known)
- Estimated fractions can then be used as covariates in multivariate regression models to infer differentially methylated cytosines (DMCs) that are independent of changes in cell-type composition.

Disadvantages

- Require knowledge of the main cell types that are present in the tissue. Reliable reference DNAm profiles must be available for these cell types
- Cannot deal with unknown confounding factors
- Assume that cell–cell interactions in the sample do not affect the DNAm profiles of the individual cell types
- Reference profiles could be confounded by factors such as age or genotype

Reference-free cell-type deconvolution

- Inferring from the full data matrix ‘surrogate variables’, which include sources of data variation that are driven by cell-type composition
- These surrogate variables are inferred from the data without the need for a reference DNAm database and are used as covariates in the final supervised multivariate regression model to infer DMCs that are independent of changes in cell-type composition and other cofounders

Advantages

- There is no requirement to know the main cell types in a tissue or to have reference DNAm profiles; hence, in principle, they are applicable to any tissue type
- De novo (unsupervised) discovery of novel cell subtypes
- Allow for the possibility that cell–cell interactions alter the profiles of individual cell types
- Can adjust simultaneously for other confounding factors, known or unknown

Reference-free cell-type deconvolution

- Inferring from the full data matrix ‘surrogate variables’, which include sources of data variation that are driven by cell-type composition
- These surrogate variables are inferred from the data without the need for a reference DNAm database and are used as covariates in the final supervised multivariate regression model to infer DMCs that are independent of changes in cell-type composition and other cofounders

Disadvantages

- Without further biological input, they cannot provide estimates of cell-type fractions in individual samples
- Performance is strongly dependent on model assumptions, which are often not satisfied

Semi-reference-free cell-type deconvolution

- Inferring surrogate variables representing variation due to cell-type composition but that, unlike a purely 'reference-free' approach, does so by using partial prior biological knowledge of which cytosine–guanine dinucleotides (CpGs) differ between cell types
- Infer the surrogate variables from the reduced data matrix, projected on this set of selected features

Advantages

- Allow for the possibility that cell–cell interactions alter the DNAm profiles of individual cell types
- If required, can be combined with batch-correction methods such as COMBAT
- More robust to incomplete knowledge of underlying cell types in the tissue of interest
- Can provide approximate relative estimates of cell-type fractions in individual samples

Semi-reference-free cell-type deconvolution

- Inferring surrogate variables representing variation due to cell-type composition but that, unlike a purely ‘reference-free’ approach, does so by using partial prior biological knowledge of which cytosine–guanine dinucleotides (CpGs) differ between cell types
- Infer the surrogate variables from the reduced data matrix, projected on this set of selected features

Disadvantages

- Performance is still strongly dependent on model assumptions, which may not be satisfied
- Inference of absolute cell-type fractions in individual samples remains challenging
- The ability to resolve highly similar cell types is limited

Computational Deconvolution of cell mixtures

- Venet, D., F. Pecasse, C. Maenhaut, and H. Bersini. “Separation of Samples into Their Constituents Using Gene Expression Data.” Bioinformatics (Oxford, England) 17 Suppl 1 (2001): S279-287.
    - Proffered a linear relationship: “Any cellular type present in the tissue contributes differently to the measured expression of a given gene”
    - “. . . start directly from the gene expression data obtained on the composite samples to determine mathematically the profile of expression of the cellular types present.”

Computational Deconvolution of cell mixtures

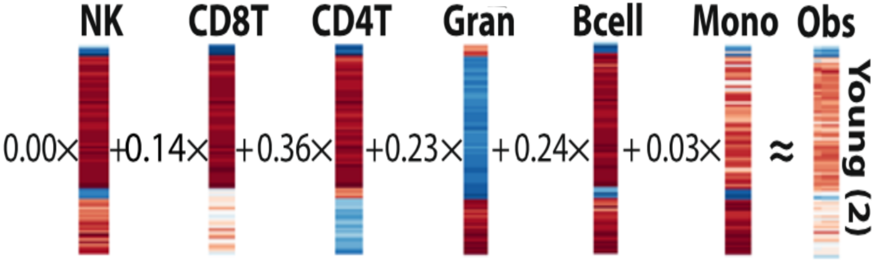
- M - matrix of heterogeneous measures, genes (rows) by samples (columns)
- G - cell signatures, genes (rows) by cell types (columns)
- C - cell proportions, cell types (rows) by samples (columns)

Heterogeneous measures can be expressed as a linear combination of cell signatures and proportions

$$M_{ij} = \sum_{k}^{Nct} G_{ik} C_{ik}$$

- gene i, sample j, number of cell types Nct

Computational Deconvolution of cell mixtures



In matrix notation

$$M = GC$$

Venet, D., F. Pecasse, C. Maenhaut, and H. Bersini. "Separation of Samples into Their Constituents Using Gene Expression Data." Bioinformatics (Oxford, England) 17 Suppl 1 (2001): S279-287.

Estimating cell proportions given signature

- Abbas et al, 2009
- Gong et al, 2011
- Kuhn et al, 2011
- Qiao et al, 2012
- Houseman et al, 2012
- Zhong et al, 2013
- Liebner et al, 2014
- Chikina et al, 2015

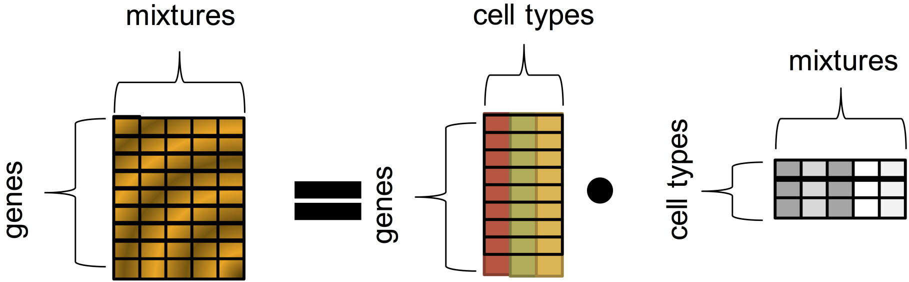
Gene expression deconvolution

- Gene expression deconvolution is an emerging technique analogous to in silico flow cytometry
    - The deconvolution methods aim to computationally resolve a GEP into its component cell types (virtual tissue dissection)
- Expression Deconvolution
    - Requires a signature matrix consisting of marker genes and their expression values
    - Also requires a biological mixture
    - There is also a vector which contains the cell subset of the mixture in the signature matrix

Modeling Cell Mixtures

Mixtures (X) are a linear combination of signature matrix (S) and concentration matrix (C)

$$X_{m\times n} = S_{m\times k} \cdot C_{k\times n}$$



Previous Work

- Coupled Deconvolution: Given: X, infer: S, C
  - NMF. Repsilber, BMC Bioinformatics, 2010
  - Minimum polytope. Schwartz, BMC Bioinformatics, 2010
- Estimation of Mixing Proportions: Given: X, S, infer: C
  - Quadratic Programming. Gong, PLoS One, 2012
  - LDA. Qiao, PLoS Comp Bio, 2012
- Estimation of Expression Signatures: Given: X, C, infer: S
  - csSAM. Shen-Orr, Nature Brief Com, 2010

Cell type composition

- The epigenome will vary from cell type to cell type. Blood is composed of many cell types.
- Houseman (2012) BMC Bioinformatics showed that this can (will) confound studies of DNA methylation performed on blood samples.
- Reinius (2012) PLoS One has flow-sorted blood data on 450k.
- Obviously, other tissues can be affected. See Guintivano (2013) Epigenetics for brain.

Cell-type deconvolution algorithms

Name	Description	Programming language	Web links
CP/QP	Reference-based method using constrained projection	R	https://github.com/sjczheng/EpiDISH
RPC	Reference-based robust partial correlations	R	https://github.com/sjczheng/EpiDISH
CIBERSORT	Reference-based support vector regressions	R	https://github.com/sjczheng/EpiDISH
SVA	Surrogate variable analysis (reference-free)	R	www.bioconductor.org/SVA package
ISVA	Independent surrogate variable analysis (reference-free)	R	https://cran.r-project.org/package=isva
RefFreeEWAS	Reference-free deconvolution	R	https://cran-r-project.org/package=RefFreeEWAS
RefFreeCellMix	Reference-free or semi-reference-free NMF using recursive QP	R	https://cran-r-project.org/package=RefFreeEWAS
MeDeCom	Reference-free or semi-reference-free constrained and regularized NMF	R	http://github.com/lutsik/MeDeCom
EDec	Like RefFreeCellMix but applied to breast cancer or tissue	R	https://github.com/BRL-BCM/EDec
RUV/RUVm	Removing unwanted variation	R	http://www.bioconductor.org/missMethyl package
CancerLocator	Inference of tumour burden and tissue of origin from plasma cfDNA	Java	https://github.com/jasminezhoulab
MethylPurify	Tumour purity estimation from WGBS or RRBS data	Python	https://pypi.python.org/pypi/MethylPurify
InfiniumPurify	Tumour purity estimation from Illumina Infinium data	Python	https://bitbucket.org/zhengxiaoqi/