Big Data Seminar Series: OMICS Data

March 15, 2023
Katerina Kechris
Departments of Biostatistics and Informatics
Colorado School of Public Health
Outline

1. Current omics technologies
2. Examples of analyses
3. Common themes in omics data analysis
4. Discussion on how to plan your omics study
5. Questions
Part 1: Technologies
Technologies

1. Microarrays (RNA/DNA)

2. Sequencing (RNA/DNA)

3. Mass-spectrometry (proteins/metabolites)

DNA

• **Genome** (whole genome sequencing, WGS)
  • Within and across population
  • Across species

• **Exome**

• **Single nucleotide polymorphisms (SNPs)**

• **Chromosome conformations** (3C/Hi-C)

[https://www.creativebiomart.net/](https://www.creativebiomart.net/)
DNA Modifications & Interactions

• DNA methylation (epigenome) (methyl-Seq)
• Histone modifications (epigenome) (ChIP-Seq)
• DNA binding proteins (e.g., transcription factor) (ChIP-Seq)
• Chromatin accessibility (ATAC-Seq)

https://www.creativebiomart.net/
RNA

- mRNA (transcriptome) (RNA-Seq)
- RNA binding proteins (e.g., splicing factors) (CLIP-Seq)
- Methylation RNA (epitranscriptome) (MeRIP-Seq)
- Other species
  - miRNA, lncRNA, etc
  - 16s rRNA (microbiome)
RNA

Diamantopolous et al., 2018  ATM
Single-Cell vs Bulk Cell

Spatial Transcriptomics

Heatmap of top 5 marker genes in each layers

https://blog.bioturing.com
Proteins

• Abundance
• Structure
• Protein-protein interactions
• Post-translational modifications (e.g., phosphoproteomics, glycoproteomics)

Metabolites

- Types of small molecules
  - Lipids – lipidomics
  - Exogenous factors– exposomics
  - Diet - nutrigenomics
- Toxicology (changes due to chemicals/drugs)
- Metabolic reactions (e.g., fluxomics)

https://thorax.bmj.com/content/69/9/876
Multi-Omics

http://melgen.org/multi-omics-approach/ Vilne & Shunkert 2018
Multi-Omics

From same cell, simultaneous detection of mRNA & chromatin accessibility (e.g., Multiome 10X Genomics)
Large-scale Projects & Databases
Large-scale Projects & Databases

hmdb
The Human Metabolome Database

HUPO
HUMAN PROTEOME ORGANIZATION
translating the code of life
Multiple-Cohorts & Populations
Resources @ AMC

Genomics Shared Resource Home Page

The Genomics and Microarray Shared Resource at University Of Colorado Denver Cancer Center is an advanced, state-of-the-art DNA and Protein microarray and Next Generation (NextGen) DNA sequencing technology center providing crucial research support for investigators interested in using:

- Next Generation Sequencing:
  - Illumina HiSeq 2500/4000 sequencing
  - Illumina MiSeq sequencing
  - LifeTech IonPGM sequencing

- DNA Microarray:
  - Illumina BeadArrays
  - Agilent Microarrays

University of Colorado School of Medicine Biological Mass Spectrometry Facility
Discover the possibilities of personalized medicine
Part 2: Examples
Study 1: Epigenetics & Type 1 Diabetes (T1D)
with Jill Norris (Epidemiology, ColoradoSPH) & Randi Johnson (Biomedical Informatics, SOM)

- DNA methylation link between genetic susceptibility & environmental exposure in T1D
- Most studies on individuals already diagnosed with T1D
- Goal: Study pre-disease DNA methylation changes associated with later development of T1D

**Study Design:** DNA methylation measured in cord blood at birth and blood during childhood prior to onset of clinical T1D from Diabetes Autoimmunity Study in the Young (DAISY) cohort (n=174)

**Platform:** Illumina BeadChip Array

**Analysis:** longitudinal mixed model, meta-analysis, region-based analysis

Johnson et al., (2020) Longitudinal DNA methylation differences precede type 1 diabetes *Scientific Reports*
Study 2: Protein-Metabolite Networks in Chronic Obstructive Pulmonary Disease (COPD)

- Most biomarker studies focus on single molecules, but panels have shown to improve prediction
- Examine proteins & metabolites to find phenotype specific networks as candidate biomarkers

**Study Design:** proteins and metabolites measured in blood on COPDGene cohort subjects (n=1008)

**Platform:** Metabolon, SOMAScan

**Analysis:** sparse canonical correlation analysis, adjusting cell counts

Mastej et al., (2020) Identifying Protein-metabolite Networks Associated with COPD Phenotypes. *Metabolites*
Study 3: Role of miRNA in Alcohol Related Behaviors

- Increasing role of miRNA in alcohol related behaviors
- Role of miRNAs as mediators of the genetic effect on behaviors

**Study Design:** miRNA expression measured in brain of recombinant inbred panels in mice; genotypes, gene expression in brain available and behavioral phenotypes
**Platform:** small RNA sequencing
**Methods:** Bayesian Network Analysis

With Laura Saba, Boris Tabakoff (SSPPS), Paula Hoffman (SOM)

BMC Genomics
Part 3:
Common Themes
Common Themes Among Omics Projects

1. Study Design and Planning
2. Data Storage
3. Processing Data
4. Multiple Testing Comparisons
5. Enrichment Analysis
6. Validation
1. Study Design

• Simple 2-group comparisons (e.g. differential expression/abundance)
  • Many investigators can do by themselves

• More complex models
  • More processing time

• Data driven network analysis
  • Need a higher sample size (dozens)

• Machine learning
  • Needs the highest sample size (hundreds)

• Talk to CIDA for designs outside of a simple 2-group comparison

Data Collection Questions: RNA-Seq example

• Communicate with the core/company collecting data is key to figure out best technology for your needs
• Do you want bulk or single cell-specific level?
• What type(s) of RNA do you want to look at?
  • mRNA only (polyA selection or possibly Tag-Seq)
  • Long non-coding and other longer types (total RNA)
  • miRNA and other smaller RNAs
  • Rare RNA types like fusion genes? (longer paired-end reads)
• What level are you looking on quantitating your data on?
  • Gene level only
  • Isoform specific level
  • Reconstruct your own transcriptome (need deep sequencing)
2. Data Storage

• Depends on core/company generating the data
• Raw data backup
• Software can now perform on a compressed file (e.g. fastq.tar.gz)
• Allow 3-4x the amount of the raw data as empty space computing
• Plan for where analysis will be conducted:
  • Local Server
  • Cloud computing
  • Galaxy
• Long term storage

**RNA-Seq .fastq**
Size = \# reads * (100 + 2*readLength)
Example: 100 million reads with a read length of 150 = 40G

**Methylation Array .idat**
450K ~ 7MB
EPIC ~ 11MB
2 files per sample
3. Processing Data

- Much more processing time than traditional data
- Raw data is provided as 1 (or 2) files/sample
- Example of RNA-Seq pre-processing steps:
Normalization

Process of removing (or minimizing) non-biological variation

• RNA-Seq
  • Reads/Fragments Per Kilobase per Million (RPKM/FPKM)
  • Transcripts per Million (TPM)
  • Quantile
  • Weighted Trimmed Mean of Log Expression Ratios (M values) (TMM)
  • DESeq Median of Ratios (geometric mean & scaling factor)
  • Removal of Unwanted Variation (RUV)
  • Surrogate Variable Analysis (SVA)

• Metabolomics (MS):
  • Locally estimated scatterplot smoothing (LOESS)
  • Systematic Error Removal using Random Forest (SERRF)
  • Median
  • Quantile
  • Cross-Contribition Compensating Multiple Standard Normalization (CRMN)
  • SVA
  • RUV
  • R/MSprep evaluates best method for metabolomics MS data

• Methylation Arrays:
  • Subset-quantile within array normalization (SWAN)
  • Normal-exponential using out-of-band probes (Noob)
  • Single-sample Noob (ssNoob)
  • Functional normalization (Funnorm)

• Microarrays:
  • Robust Multichip Average (RMA)
  • Guide to Probe Logarithmic Intensity Error (PLIER)

R/Normalyzer:
A Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets

No Standard Method!
QC Visualization: Evaluating Normalization Density Plots – Methylation Array Example

RLE Plots: Relative Log Expression

RAW Counts

Normalized Counts

Data Reduction for QC Purposes

- Principal Component Analysis (PCA)
- Factor Analysis
- Singular Value Decomposition (SVD)
- Independent Component Analysis (ICA)
- t-Distributed Stochastic Neighbor Embedding (tSNE)
- Uniform Manifold Approximation and Projection (UMAP)

https://dataconomy.com/
QC PCA Plots

Clustering by biological or technical factors
Depending on study design
Sample Level QC: Dendrograms
Feature Level QC

- Detection above background threshold
- Coefficient of variation (CV) threshold
- No set feature QC for any technology
4. Multiple Testing

• Same statistical test for every feature
  • Example: 20,000 genes, then you have 20,000 tests
  • If you leave alpha = 0.05 you would expect 1,000 false positive results (Yikes!)

• Perform correction for multiple testing

• Bonferroni
  • Multiple the p-value by the # of tests performed
  • Most conservative and considered too harsh

https://xkcd.com/882/
False Discovery Rate (FDR)

- Adjusts each p-value differently depending on rank

**FALSE POSITIVES**

Same Distribution

**FALSE NEGATIVES**

Different Distributions
False Discovery Rate (FDR)

- Tries to estimate your distribution of non-significant p-values
5. Enrichment & Over-representation Analysis

- Big picture at system level
- Static (Over-representation)
- Fluid (Enrichment)
  - Gene Set Enrichment Analysis (GSEA)

<table>
<thead>
<tr>
<th></th>
<th>Candidates</th>
<th>Genome (background)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Pathway</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not in Pathway</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Background Set is Important

- What is present in study sample type
  - Example: if looking at lung tissue you would not expect all genes to be expressed in the lung regardless of study design

- Arrays certain genes are over-represented
  - Various number of probes/gene
  - Example: Illumina’s EPIC array (DNA Methylation) there is a range of 1 to 1,487 probes/gene, with a median of 20 probes per gene
    - R/missMethyl takes into account how many probes are designed on array

<table>
<thead>
<tr>
<th></th>
<th>Candidates</th>
<th>Genome (background)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Pathway</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not in Pathway</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6. Validation

• Journals wanting more validation

• Reproduce quantitation:
  • High-throughput methods are not the gold standard in quantitation
  • Gene expression: qRT-PCR
  • Methylation: Pyrosequencing
  • Metabolomics: Targeted or internal standard

• Functional validation:
  • Gene knock-down or knock-out methods
  • Use different dataset (publically available) to show this effect

• Multi-omics integration:
  • Gene candidate in both RNA-Seq and ChIP-Seq
  • Correlation among DNA methylation and gene expression
Part 4: 
Planning Your Study
Planning your study

1. Talk to core to plan experiment & discuss
   • Technology
   • Protocol options
   • Timeline
   • Sample handling and prep

2. Plan for computing needs (software, hardware) & data storage
Planning your study

3. Work with CIDA
   • More complex study design (e.g., multiple time points, biological/treatment groups)
   • More complex analyses (e.g., alternative splicing, transcriptome reconstruction, gene fusion)
   • CIDA provides not only analysis support, but also grant support
   • Include CIDA in planning with core

4. Budget time and effort for data analysis (biggest bottleneck)
Questions?