SOC-B CENTRE FOR DOCTORAL TRAINING IN BIOSOCIAL RESEARCH

‘OMICS & SYSTEMS BIOLOGY

Adam Stevens
Lecturer in Child Health and Growth
Division of Developmental Biology and Medicine
INTRODUCTIONS
Overview

- What is biological information?
- Types of ‘omic data
- Basic analysis and tools
- Network analysis of ‘omic data
WHAT IS BIOLOGICAL INFORMATION?
Genotype vs. Phenotype

• **Genotype**
  - The genetic makeup of an organism
  - An organism’s complete set of genes
  - Instructions for building and maintaining
    - Formation of proteins, regulation of metabolism
  - Genetic traits
  - Internally coded - not observed
  - Copied during cell division & reproduction

• **Phenotype**
  - Observable physical properties of an organism
  - Appearance, development & behavior

*Genes + Environment = Phenotype*
Central Paradigm of Molecular Biology

Genes to messenger to proteins

DNA → mRNA → Protein

Biological information flow

- Replication
- Transcription
- Translation
- Protein

- DNA Information
- RNA Flow/Process
- Protein Function
Central dogma of molecular biology

- **DNA Replication**
  - DNA
  - RNA
  - Protein
  - Metabolite

- **Transcription**
  - DNA Replication
  - RNA

- **Translation**
  - RNA
  - Protein

- **Genomics**
  - DNA

- **Transcriptomics**
  - RNA

- **Proteomics**
  - Protein

- **Metabolomics**
  - Metabolite
DNA & Genomics

- Deoxyribonucleic acid (DNA)
  - Polymer of nucleotides
  - Sequence of nucleotides is responsible for carrying and retaining hereditary information in a cell – Base Sequence

- Double helix of complementary base pairs
- Nitrogenous base
  - Adenine and Thymine
  - Cytosine and Guanine
- Phosphate group
- Deoxyribose

https://courses.lumenlearning.com/microbiology/chapter/structure-and-function-of-dna/
**RNA & Transcripomics**

- Ribonucleic acid (mRNA)
  - Result of TRANSCRIPTION
    - Information encoded with the DNA sequence of one or more genes is TRANSCRIBED into a strand of RNA – RNA transcript
      - Single stranded
      - A, G, C, U (T)

Translation: DNA to mRNA to Protein

By: Suzanne Clancy, Ph.D. & William Brown, Ph.D. (Write Science Right) © 2008 Nature Education
Proteins & Proteomics

A GUIDE TO THE TWENTY COMMON AMINO ACIDS

Amino acids are the building blocks of proteins in living organisms. There are over 500 amino acids found in nature, however, the human genetic code only directly encodes 20. ‘Essential’ amino acids must be obtained from the diet, whilst non-essential amino acids can be synthesised in the body.

Chart Key:
- ALIPHATIC
- AROMATIC
- ACIDIC
- BASIC
- HYDROXYLIC
- SULFUR-CONTAINING
- AMIDIC
- NON-ESSENTIAL
- ESSENTIAL

<table>
<thead>
<tr>
<th>Name</th>
<th>Three Letter Code</th>
<th>Chemical Structure</th>
<th>DNA Codons</th>
<th>IUPAC</th>
<th>Common Name</th>
<th>Common Name</th>
<th>Essential</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine (A)</td>
<td>Ala</td>
<td>( \begin{array}{c} \text{H} \ \text{N} \ \text{CO} \end{array} )</td>
<td>GCT, GCC, GCA, GGG</td>
<td>( \text{H} )</td>
<td>Alanine</td>
<td>Alanine</td>
<td>Yes</td>
</tr>
<tr>
<td>Glycine (G)</td>
<td>Gly</td>
<td>( \begin{array}{c} \text{H} \ \text{N} \ \text{CO} \end{array} )</td>
<td>GGT, GGC, GGA, GGG</td>
<td>( \text{H} )</td>
<td>Glycine</td>
<td>Glycine</td>
<td>No</td>
</tr>
<tr>
<td>Isoleucine (I)</td>
<td>Ile</td>
<td>( \begin{array}{c} \text{H} \ \text{N} \ \text{CO} \end{array} )</td>
<td>ATT, ATC, ATA</td>
<td>( \text{H} )</td>
<td>Isoleucine</td>
<td>Isoleucine</td>
<td>Yes</td>
</tr>
<tr>
<td>Leucine (L)</td>
<td>Leu</td>
<td>( \begin{array}{c} \text{H} \ \text{N} \ \text{CO} \end{array} )</td>
<td>CTT, CTC, CTA, CTG, TTA, TGG</td>
<td>( \text{H} )</td>
<td>Leucine</td>
<td>Leucine</td>
<td>Yes</td>
</tr>
<tr>
<td>Proline (P)</td>
<td>Pro</td>
<td>( \begin{array}{c} \text{H} \ \text{N} \ \text{CO} \end{array} )</td>
<td>CCT, CCC, CGA, CGG</td>
<td>( \text{H} )</td>
<td>Proline</td>
<td>Proline</td>
<td>Yes</td>
</tr>
<tr>
<td>Valine (V)</td>
<td>Val</td>
<td>( \begin{array}{c} \text{H} \ \text{N} \ \text{CO} \end{array} )</td>
<td>GTT, GTC, GTA, GTG</td>
<td>( \text{H} )</td>
<td>Valine</td>
<td>Valine</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: This chart only shows those amino acids for which the human genetic code directly codes for. Selenocysteine is often referred to as the 21st amino acid, but is encoded in a special manner. In some cases, distinguishing between asparagine/aspartic acid and glutamine/glutamic acid is difficult. In these cases, the codes asx (B) and glx (D) are respectively used.
Metabolites & Metabolomics

- Low molecular weight products/components of metabolism
- Metabolism is a complex interplay of chemical reactions that occur within cells
- Diverse roles
  - Energy metabolism
  - Amino acid metabolism
  - Lipid metabolism
- Primary metabolism – directly involved in growth, development
- Secondary metabolism – end products of metabolism
  - Antibiotics, steroids etc
Central dogma of molecular biology

- **DNA** → Genomics
- **RNA** → Transcriptomics
- **Protein** → Proteomics
- **Metabolite** → Metabolomics

Genotype → Phenotype
What is ‘Omic Data

Omic data sets include:

- Genetics (SNPs, CNVs)
- Transcriptomics (Affymetrix, RNAseq)
- Epigenomics (DNA methylation, histone mods)
- ChIPseq
- Metabolomics
- Proteomics
- Phosphoproteomics

All have specific quality control (QC) issues and difficulties in analysis
All rely on the use of a false discovery rate correction (FDR) for analysis

Multi-omic analysis is the integration of different omic data sets
TYPES OF 'OMICS DATA?
Why ‘omics’?

- Adopts an holistic view of all ‘molecules’ that make up a cell, tissue or organism
- Universal approach/Hypothesis-generating
- No analysis bias
- Many applications
  - ‘BIOMARKER’ discovery
  - Early detection/population screening
  - Increasing understanding of disease aetiology
  - Drug discovery & toxicity and efficacy screens
Biomarkers of disease

In medicine, a **biomarker** is a measurable indicator of the severity or presence of some disease state. More generally a **biomarker** is anything that can be used as an indicator of a particular disease state or some other physiological state of an organism.
Experimental Design

- Case vs. Control (an appropriate one!)
- Samples numbers?
  - How many is sufficient?
  - Test cohort + validation cohort?
- Quality of samples
  - Collection/storage/processing
- Confounding factors?
  - Age, gender, environment
Is this good experimental design?

- Metabolomics investigation of liver failure from plasma samples

**Table 1. Demographic Information of the Healthy Group and Liver Failure Patient Group Investigated**

<table>
<thead>
<tr>
<th></th>
<th>healthy group</th>
<th>patient group</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n = 23)</td>
<td></td>
<td>(n = 24)</td>
</tr>
<tr>
<td>Gender (male/female)</td>
<td>15/8</td>
<td>21/3</td>
</tr>
<tr>
<td>HBsAg</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>Age (year)</td>
<td>27.39 ± 9.24</td>
<td>46.77 ± 13.35</td>
</tr>
<tr>
<td>ALT (U/L)</td>
<td>&lt;40</td>
<td>172.63 ± 147.49</td>
</tr>
<tr>
<td>TB (μmol/L)</td>
<td>&lt;12</td>
<td>457.33 ± 135.48</td>
</tr>
<tr>
<td>PT (s)</td>
<td>&lt;14</td>
<td>26.06 ± 15.14</td>
</tr>
<tr>
<td>MELD score</td>
<td>/</td>
<td>24.68 ± 8.38</td>
</tr>
</tbody>
</table>

*Abbreviations: ALT, alanine aminotransferase; TB, total bilirubin; PT, prothrombin time; MELD, model for end-stage liver disease. The value is represented as the form of mean ± SD.*
Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring


See all authors and affiliations

Science 15 Oct 1999:
Vol. 286, Issue 5439, pp. 531-537
DOI: 10.1126/science.286.5439.531

Generic approach to cancer classification based on gene expression monitoring by DNA microarrays applied to human acute leukemias

38 Affymetrix microarrays with 6,817 probes
27 from childhood acute lymphoblastic leukemia
11 from adult acute myeloid leukemia
Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning

Louise C. Kenny\textsuperscript{a,*}, Warwick B. Dunn\textsuperscript{b}, David I. Ellis\textsuperscript{b}, Jenny Myers\textsuperscript{a}, Philip N. Baker\textsuperscript{a} and the GOPEC Consortium, and Douglas B. Kell\textsuperscript{b,*}

- Pre-eclampsia - Pregnancy-induced hypertension which may affect mother and foetus

<table>
<thead>
<tr>
<th></th>
<th>Normal outcome</th>
<th>Preeclampsia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n=87$</td>
<td>$n=87$</td>
</tr>
<tr>
<td>Age</td>
<td>30 (19–43)</td>
<td>31 (19–41)</td>
</tr>
<tr>
<td>Parity</td>
<td>0 (0–2)</td>
<td>0 (0–2)</td>
</tr>
<tr>
<td>BMI (weight/height\textsuperscript{2})</td>
<td>25 (19–46)</td>
<td>26 (18–46)</td>
</tr>
<tr>
<td>Max (S) BP (mm Hg)</td>
<td>122 (96–147)</td>
<td>162 (138–220)*</td>
</tr>
<tr>
<td>Max (D) BP (mm Hg)</td>
<td>80 (60–93)</td>
<td>110 (90–140)*</td>
</tr>
<tr>
<td>Delivery gestation</td>
<td>40 + 4</td>
<td>37 + 0*</td>
</tr>
<tr>
<td>(weeks + days)</td>
<td>(34 + 3 to 42 + 0)</td>
<td>(26 + 3 to 41 + 1)</td>
</tr>
<tr>
<td>Birth weight (g)</td>
<td>3420 (2380–4420)</td>
<td>2410 (590–4300)*</td>
</tr>
<tr>
<td>IBR (centile)</td>
<td>34 (10–99)</td>
<td>8 (0–99)*</td>
</tr>
</tbody>
</table>

Median (range).
Pre-eclampsia vs normal outcome.
\(p < 0.0001\).
Genomics

- Genome = total DNA of a cell or an organism
- Human genome = 3.2 billion bases and estimated > 30,000 protein coding genes

- Seeking mutations or alterations that may contribute towards a certain disease!
  - *i.e.* the genes BRCA1 and BRCA2 cause 60% of all cases of hereditary breast and ovarian cancers
  - BUT not a single mutation - there are >800 different mutations in BRCA1 alone

TOOLS: Whole Genome Sequencing

Sequencing

The Whole Genome Sequencing (WGS) Process
WGS is a laboratory procedure that determines the order of bases in the genome of an organism in one process. WGS provides a very precise DNA fingerprint that can help link cases to one another allowing an outbreak to be detected and solved sooner.

1. DNA Extraction
   Scientists take bacterial cells from an agar plate and treat them with chemicals that break them open, releasing the DNA. The DNA is then purified.

2. DNA Shearing
   DNA is cut into short fragments of known length, either by using enzymes "molecular scissors" or mechanical disruption.

3. DNA Library Preparation
   Scientists make many copies of each DNA fragment using a process called polymerase chain reaction (PCR). The pool of fragments generated in a PCR machine is called a "DNA library."

4. DNA Library Sequencing
   The DNA library is loaded onto a sequencer. The combination of nucleotides (A, T, C, and G) making up each individual fragment of DNA is determined, and each result is called a "DNA read."

5. DNA Sequence Analysis
   The sequencer produces millions of DNA reads and specialized computer programs are used to put them together in the correct order like pieces of a jigsaw puzzle. When completed, the genome sequence containing millions of nucleotides (in one or a few large pieces) is ready for further analysis.
Cost of Sequencing

PacBio Sequencer
£300K
Single-Molecule, Real-Time (SMRT) technology
www.pacb.com
Transcriptomics

- Measuring the transcriptome
  - All the mRNA that is transcribed at a given point in a given cell or organism
- Provides direct knowledge of gene regulation and protein content information
  - Which genes are actively expressed
- Began in 1990s and is now a widespread discipline
  - Many technological advances
  - Now a routine, ‘simple’ process

- TOOLS: Microarrays & RNA-Seq
DNA Microarrays

A microarray is an orderly arrangement of rows and columns on a surface like a glass slide. Each of the spots on an array contains single-stranded DNA molecules that correspond to a single gene. An array can contain a few, or thousands, of genes.
DNA Microarrays

Converted to numerical values for interpretation

Limitation: Results limited to what probes you have on your chip
RNA Seq

- High throughput sequencing with computational methods
- No reference sequence
- Large dynamic range
- Sequences every RNA molecule and profiles the expression of a particular gene by counting the number of times its transcript has been sequenced
- Expression levels!

SAMPLE RNA → Fragmentation → RNA FRAGMENTS

RNA FRAGMENTS → Reverse transcription & amplification → cDNA FRAGMENTS

Sequencing Machine → READS
Proteomics

- Investigating the entire complement of proteins within a cell, tissue or organism.
- >100,000 proteins
- Large dynamic range

Some questions
- When & where proteins are expressed
- The involvement of proteins with particular phenotypes
- How proteins are modified or how they interact with each other

- Blood, urine, tissues
Proteomic Tools

- Mass spectrometry

Cell Culture or Tissue
Proteins for top-down analyses come from a variety of sources.

Extraction of Proteins
Proteins are extracted and denatured.

Separation of Proteins
Proteins are separated, most often by molecular weight, to reduce sample complexity and ensure maximal identification of intact proteins.

Automated Data Analysis
Intact proteins are identified in an automated fashion using ProSightPC software, including characterization of post-translational modifications, sequence polymorphisms, and cleavage sites.

Analysis by LC-MS/MS
Intact proteins are analyzed by LC-MS/MS on Orbitrap-based mass spectrometers.

http://planetorbitrap.com/data/fe/image/Workflows_TopDownProt(1).png
Proteomic Tools

https://www.creative-proteomics.com/services/2d-electrophoresis.htm
### Proteomics Data

#### Sample Names – often 10's with different classes

<table>
<thead>
<tr>
<th>id</th>
<th>Protein ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Abundance of protein? in each sample**

Do statistical analysis to decide which features change between classes.
Metabolomics

• Investigating the entire complement of low molecular weight molecules in an sample of interest.

• ~ 5000 metabolites in a biological sample from a human
  • GREATLY impacted by environment!
  • GREATLY impacted by time!

• Experimental design & sample collection is VERY important!
# Biological Matrices

**Primary Sources**
- Serum
- Plasma
- CSF
- BAL
- Saliva
- Semen
- Urine
- Faeces
- Sweat
- Breath

**Secondary Sources**
- Tissue Biopsies
  - Brain
  - Nerve
  - Lung
  - Pancreas
  - Liver
  - Heart
  - Gut
  - Skin

**Additional Sources**
- Animal Models
- Mammalian Cell Culture
- IVF culture medium
Chromatography linked to Mass Spectrometry

Higher affinity to stationary phase

Lower affinity to stationary phase

Serum: GC-MS or LC-MS

Urine: GC-MS or LC-MS
Chromatography linked to Mass Spectrometry

Hippuric acid

$m/z=179.17$

UHPLC-MS (electrospray (ESI))
Soft ionisation technique, intact parent mass ion detected, but many adducts can be produced

GC-MS (electron impact)
Much greater degree of fragmentation as higher energy ionisation process

Matching of the chromatographic retention time and fragmentation mass spectra between a sample analyte and a reference standard is required for definitive id.
We have *ca.* 1600 analytes in our GC-MS library

Metabolomics Data

Sample Names – often 100’s with different classes

<table>
<thead>
<tr>
<th>id</th>
<th>RT (min)</th>
<th>m/z</th>
<th>ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Use RT & m/z to identify the metabolite – **MAJOR CHALLENGE**!
Do statistical analysis to decide which features change between classes
"Data does not equal information; information does not equal knowledge; and, most importantly of all, knowledge does not equal wisdom. We have oceans of data, rivers of information, small puddles of knowledge, and the odd drop of wisdom."

Nix, H., *A national geographic information system - an achievable objective?*, in Keynote address. 1990: AURISA.
What are the similarities and differences between data, information, knowledge, and wisdom?
The challenge of biomarkers?


- Multi-Centre, large cohort studies
Biomarker Discover: From lab to bedside

Conception (objectives, collaborations, design of experiment)

To the clinic

Representative Cohort study \((n=1000s)\) with analysis of candidates (LC-MS and/or assay)

Set of candidate metabolites

Hypothesis validation (independent sample set 2)

Hypothesis generation study 1 (independent sample set 1)
SOME EXAMPLES......
Personalised/Precision Medicine

• Is this the future of healthcare?
  • 2015 State of the Union address President Obama announced that he was launching the Precision Medicine Initiative

• Assess the genotype (SNPs) and phenotype (metabolome) of a patient before they undergo any treatment
  • Population monitoring & data collation
  • Seeking cures & preventative screening

• Offering a well-designed screening program at a reasonable cost may not always be possible due to the numerous associated challenges
  • monetary limitations (labour and consumable costs) as well as ethical, legal and social considerations for an opt-in test
(Metabol)’omics for the masses?

• Wearable technologies – smartphones, smart-watches, health bands, necklaces, glucose monitoring contact lenses
  • Innovations for collecting personal information

• mPower - mobile Parkinson’s Disease study that attempts to research the occurrence, presentation and management of PD symptoms via survey telemetry data using a smartphone app

• Smart-phone app to monitor the association between pain and the weather for people suffering from rheumatoid arthritis
Patients with chronic pain commonly believe their pain is related to the weather. Scientific evidence to support their beliefs is inconclusive, in part due to difficulties in getting a large dataset of patients frequently recording their pain symptoms during a variety of weather conditions. Smartphones allow the opportunity to collect data to overcome these difficulties. Our study *Cloudy with a Chance of Pain* analysed daily data from 2658 patients collected over a 15-month period. The analysis demonstrated significant yet modest relationships between pain and relative humidity, pressure and wind speed, with correlations remaining even when accounting for mood and physical activity. This research highlights how citizen-science experiments can collect large datasets on real-world populations to address long-standing health questions. These results will act as a starting point for a future system for patients to better manage their health through pain forecasts.
The future cycle of metabolomics precision medicine-based research and healthcare where academia, industrial partners, corporate data analytics work with patients’ wearable data collection devices to provide health monitoring solutions.
Twin research for a healthy future

Researching the link between our genes, the environment, and common diseases

Looking to collaborate?

We aim to facilitate and encourage the sharing of TwinsUK data and samples with the world’s scientific community to promote and contribute to scientific research and generate new knowledge. Find out more by visiting our data access pages below.

Collaborate
TWINS UK

- Comprehensive study
- Unique design with internal controls
- Well documented & controlled
- Wealth of scientific publications
The super smelller

The woman who can smell Parkinson's disease

By Elizabeth Quigley
BBC Scotland news

22 October 2015
Biomarker detection

Gaining distance on biomarker discovery......
Data Repositories

MetaboLights
MetaboLights is a database for Metabolomics experiments and derived information. The database is cross-species, cross-technique and covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments. MetaboLights is the recommended Metabolomics repository for a number of leading journals.

More about us

Quick tour
**MTBLS1: A metabolomic study of urinary changes in type 2 diabetes in human compared to the control group**

**Abstract**

Type 2 diabetes mellitus is the result of a combination of impaired insulin secretion with reduced insulin sensitivity of target tissues. There are an estimated 150 million affected individuals worldwide, of whom a large proportion remains undiagnosed because of a lack of specific symptoms early in this disorder and inadequate diagnostics. In this study, NMR-based metabolomic analysis in conjunction with uni- and multivariate statistics was applied to examine the urinary metabolic changes in human type 2 diabetes mellitus patients compared to the control group. The human population were unmedicated diabetic patients who have good daily dietary control over their blood glucose concentrations by following the guidelines on diet issued by the American Diabetes Association. Note: This is part of a larger study, please refer to the original paper below.

**Authors:** Reza Salek, Jules Griffin

---

**Organism(s)**

Homo sapiens

**Study Design**

EFO:diabetes mellitus  
EFO:metabolic syndrome  
Urine global profiling  
CHMN:nuclear magnetic resonance spectroscopy  
NCIT:Human Study Subject  
unrelated metabolites
The PRIDE PRoteomics IDENTifications (PRIDE) database is a centralized, standards compliant, public data repository for proteomics data, including protein and peptide identifications, post-translational modifications and supporting spectral evidence. PRIDE is a core member in the ProteomeXchange (PX) consortium, which provides a single point for submitting mass spectrometry based proteomics data to public-domain repositories. Datasets are submitted to PRIDE via ProteomeXchange and are handled by expert biocurators.
PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments. More...
The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.
ArrayExpress – functional genomics data

ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.

Browse ArrayExpress

Data Content

Updated today at 03:00

- 70691 experiments
- 2243389 assays
- 46.13 TB of archived data
Human Metabolome Database
### Search Parameters

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Age</th>
<th>Effect</th>
<th>Biofluid</th>
<th>P-Value</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dehydroepiandrosterone</td>
<td>Adult (&gt;18 years old)</td>
<td>Decrease</td>
<td>Blood</td>
<td></td>
<td>The metabolomics ...</td>
</tr>
<tr>
<td>Geno Metabolomics</td>
<td>Adult (&gt;18 years old)</td>
<td>Increase</td>
<td>Blood</td>
<td></td>
<td>The metabolomics ...</td>
</tr>
<tr>
<td>L-Carnitine (HMDB0000064)</td>
<td>Adult (&gt;18 years old)</td>
<td>Increase</td>
<td>Blood</td>
<td></td>
<td>The metabolomics ...</td>
</tr>
<tr>
<td>Creatine (HMDB0000064)</td>
<td>Adult (&gt;18 years old)</td>
<td>Increase</td>
<td>Blood</td>
<td></td>
<td>The metabolomics ...</td>
</tr>
<tr>
<td>Cholesterol (HMDB0000067)</td>
<td>Adult (&gt;18 years old)</td>
<td>Increase</td>
<td>Blood</td>
<td></td>
<td>The metabolomics ...</td>
</tr>
<tr>
<td>Glycocholic acid (HMDB00000138)</td>
<td>Adult (&gt;18 years old)</td>
<td>Increase</td>
<td>Blood</td>
<td></td>
<td>The metabolomics ...</td>
</tr>
<tr>
<td>Ornithine (HMDB00000214)</td>
<td>Adult (&gt;18 years old)</td>
<td>Increase</td>
<td>Blood</td>
<td></td>
<td>The metabolomics ...</td>
</tr>
<tr>
<td>Pyroglutamic acid (HMDB00000267)</td>
<td>Adult (&gt;18 years old)</td>
<td>Increase</td>
<td>Blood</td>
<td></td>
<td>The metabolomics ...</td>
</tr>
<tr>
<td>Urea (HMDB00000294)</td>
<td>Adult (&gt;18 years old)</td>
<td>Increase</td>
<td>Blood</td>
<td></td>
<td>The metabolomics ...</td>
</tr>
<tr>
<td>3-Hydroxybutyric acid (HMDB00000357)</td>
<td>Adult (&gt;18 years old)</td>
<td>Increase</td>
<td>Blood</td>
<td></td>
<td>The metabolomics ...</td>
</tr>
</tbody>
</table>
Standardisation

- MSI formed in 2005 to unify and to engage with the growing metabolomics community so that experiments can be reproduced by others and are based on solid sample collection, analysis and data processing. – Complete Transparency!
- Working group now working on how to perform experimental design better.
- Pre-requisite for publication in Metabolomics

Metabolite Identification
MSI Level 1 – Definitive
MSI Level 2 - Putative
NETWORK ANALYSIS

T-distributed stochastic neighbourhood embedding (tSNE)
Developmental trajectories

**Minimal Spanning Tree (MST)**


**Principal Curve (PC)**

Furlan, A. et al. (2017). Multipotent peripheral glial cells generate neuroendocrine cells of the adrenal medulla. Science 357, eaal3753
RNA Velocity of Early Human Embryo

- Reanalysis of >1500 cells from Petropoulos et al. (2016) represented as points, coloured based on age
- UMAP visualisation – dimensionality reduced – cells positioned based on transcriptome similarity
- Velocity streams overlaid on top – captures developmental direction

RNA Cartography

Individual Omic Data Set Analysis

What does a p-value threshold mean?
What does a Fold change cut off mean?

<table>
<thead>
<tr>
<th>varID</th>
<th>Gene Symbol</th>
<th>q-Value</th>
<th>Fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td>203820_s_at</td>
<td>IGF2BP3</td>
<td>1.19574e-16</td>
<td>3.03628</td>
</tr>
<tr>
<td>203819_s_at</td>
<td>IGF2BP3</td>
<td>9.37733e-11</td>
<td>2.62628</td>
</tr>
<tr>
<td>240143_at</td>
<td>---</td>
<td>2.07997e-08</td>
<td>1.76953</td>
</tr>
<tr>
<td>206569_at</td>
<td>---</td>
<td>3.21346e-07</td>
<td>1.32887</td>
</tr>
<tr>
<td>228988_at</td>
<td>ZNF711</td>
<td>1.25012e-05</td>
<td>1.56114</td>
</tr>
<tr>
<td>201417_at</td>
<td>SOX4</td>
<td>1.33199e-05</td>
<td>1.71911</td>
</tr>
<tr>
<td>207996_s_at</td>
<td>C18orf1</td>
<td>1.33199e-05</td>
<td>1.38872</td>
</tr>
<tr>
<td>203038_at</td>
<td>PTPRK</td>
<td>1.33317e-05</td>
<td>1.65887</td>
</tr>
<tr>
<td>201310_s_at</td>
<td>C5orf13</td>
<td>1.38711e-05</td>
<td>1.42943</td>
</tr>
<tr>
<td>222344_at</td>
<td>---</td>
<td>1.67269e-05</td>
<td>1.40778</td>
</tr>
<tr>
<td>213808_at</td>
<td>ADAM23</td>
<td>1.88826e-05</td>
<td>1.61604</td>
</tr>
<tr>
<td>214744_s_at</td>
<td>RPL23</td>
<td>2.49258e-05</td>
<td>1.29967</td>
</tr>
</tbody>
</table>

Age group 2 Vs Rest, ANOVA, q<0.1 = 1524 probe-sets

Lists and Cut-offs!!
Fisher’s Exact Test / Hypergeometric Test

Pick 5 marbles

P(2 are red)

Gene Ontology
Enrichment of ‘causal transitive triangles’
What is Network Biology

Biological networks are:

• “Scale free”
• Resistant to random error
• Exhibit “small world” properties

Collective dynamics of 'small-world' networks

Duncan J. Watts & Steven H. Strogatz

\[ L \propto \log N \]

small-world network - most nodes are not neighbours of one another, but most nodes can be reached from every other by a small number of steps.

A Proteome-Scale Map of the Human Interactome Network


19 Co-first author
20 Co-senior author

DOI: http://dx.doi.org/10.1015/j.cell.2014.10.050
Human Interactome.

**Biogrid 3.1.88**
- Proteins = 14,334
- Interactions = 65,710

**Biogrid 3.2.105**
- Proteins = 18,107
- Interactions = 217,927
What is a network model

FIFA World Cup 2014

http://blog.physicsworld.com/2014/06/19/a-network-analysis-of-the-fifa-world-cup/
Network Topology Is Associated with Essential Function

Sun et al (2010), BMC Genomics 11 S5

Red = Cancer Genes, Black = Essential Genes, Grey = Control Genes

A comparative study of cancer proteins in the human protein-protein interaction network

Jingchun Sun1,2, Zhongming Zhao1,2,3*

Network properties of human disease genes with pleiotropic effects

Sreenivas Chavali*, Fredrik Barrenas1, Kartik Kanduri and Mikael Benson
Network Clusters are associated with hierarchy of biological function.

Hierarchy of Functions and Pathways ranked by centrality of cluster

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Core nodes of cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUMO2</td>
<td>SUMO2, UBC, MDM2, TPS3, EEF1A1, PRKDC, RPS4X, SRRM2, RPS13, RPS20</td>
</tr>
<tr>
<td>SKP1</td>
<td>SKP1, BTRC, CUL1, GSK3B, CTNNB1, SKP2, NFKBIA, CLSPN, FBXW11, FBXO6</td>
</tr>
<tr>
<td>GSK3B</td>
<td>GSK3B, AXIN1, APP, AKT1, MAPT, CTNNB1, ELAVL1, KIF5B, AXIN2, HIPK2</td>
</tr>
<tr>
<td>NCOA1</td>
<td>NCOA1, NCOA6, ESR1, PPARγ, MLL3, RXRA, ESR2, ESR4, NCOA6, VDR</td>
</tr>
<tr>
<td>EZH2</td>
<td>EZH2, EED, SUZ12, EZH1, JARID2, SON, SRSF7, NRF1, FBL, HDAC1</td>
</tr>
<tr>
<td>TCEB2</td>
<td>TCEB2, VHL, CUL5, TCEB1, CUL2, TCEB3, ASB9, STK16, NEDD8, COP56</td>
</tr>
<tr>
<td>NCOB2</td>
<td>NCOB2, NCOA6, HDAC3, BCL6, RARA, AR, ANKK1, THR8, HDAC1, KDM5B</td>
</tr>
<tr>
<td>DIABLO</td>
<td>DIABLO, XIAP, BIRC2, BIRC6, UBE2D4, BIRC3, BIRC7, TRAF2, BIRC5, ELAVL1</td>
</tr>
<tr>
<td>NBN</td>
<td>NBN, MRE11A, MDC1, RAD50, BRCA1, H2AFX, FANCQ1, ATM, TP53P1, ATX</td>
</tr>
<tr>
<td>TSC1</td>
<td>TSC1, RHEB, TSC2, YWHAE, MTor, YWHAB, RAF1, RPTOR, BECN1, MAST3</td>
</tr>
</tbody>
</table>

Cellular responses to stress, Apoptosis, Circadian clock, DNA repair, Metabolism of Proteins (q<1x10^{-3}).

G1/S DNA damage checkpoint (q<6.2x10^{-4}).

Cell Cycle, Circadian Clock (q<1x10^{-5}).

Wnt & Prolactin signalling (q<1.0x10^{-5}).

Apoptosis, Signal transduction (q<0.01).

Wnt, P13/AKT signalling (q<0.01).

Transcriptional regulation of white adipocyte differentiation, Regulation of lipid metabolism by PPARα, Circadian Clock, Mitochondrial biogenesis (q<0.01).

Cellular Senescence, Epigenetic regulation of gene expression (q<2x10^{-4}).

Cellular response to hypoxia (q<1x10^{-5}).

TGF-β Signalling (q<0.05).

Transcriptional regulation of white adipocyte differentiation, Regulation of lipid metabolism by PPARα, Circadian Clock, Mitochondrial biogenesis (q<0.01).

Immune system, Apoptosis (q<0.05)

Toll-Like Receptors Cascades (q<1.0x10^{-4}).

DNA Repair, Cell Cycle (q<1.0x10^{-4})

G2/M DNA damage checkpoint (q<6.7x10^{-4}).

Cellular response to stress (q<0.01), Insulin, IGF1, mTOR, P13K signalling (q<1.0x10^{-5}).
Constraint Based Modelling

**a. Topological enrichment**

- High-throughput data integration
- Enriched regions of change

**b. Constraining the solution space**

- For context-specific flux distributions
- High-throughput data integration
- Upregulated
  - Downregulated

**c. Comparison**

- Simulated fluxes
- High-throughput data
- High flux
  - Low flux
  - Comparison
  - Upregulated
    - Downregulated
  - Comparing objectives to match $^{13}$C fluxomic data
  - Sum of fluxes
  - Biomass yield
  - ATP yield
**Why are methods of prioritising Omic data important?**

**Answer:** Lists don’t necessarily deliver the solution!

Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase


**Affiliations** | **Contributions** | **Corresponding author**


Received 13 July 2010 | Accepted 11 July 2011 | Published online 17 August 2011

Gene set analysis of Transcriptomic data

**Systems Analysis by constraint based modelling predicts correctly**
Many structural elements can be used to model networks

- Tie
- Reciprocity
- Activity
- Popularity
- Triads
- Brokerage

No single rule explains why ALL network ties occur
Integration of Multi-Oomic Data

Boosting Signal-to-Noise in Complex Biology: Prior Knowledge Is Power

Trey Ideker, Janusz Dutkowski, and Leroy Hood

“Network Biology is a primary tool”

The statistical power of omics experiments can be enhanced through bioinformatics methods that decrease noise through the use of (1) complementary datasets, and (2) incorporation of prior knowledge about the system (e.g., aggregating measurements from entities that belong to the same pathway). This results in an effective decrease in the False Discovery Rate (FDR), at a given t statistic cutoff, or in the ability to relax such cutoff while maintaining the same FDR.
CITE-Seq (Cellular Indexing of Transcriptomes and Epitopes by Sequencing)

Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage.

Integration of Networks to Enhance Clinical Prediction

Similarity network fusion for aggregating data types on a genomic scale

Bo Wang1,2, Aziz M Mezlini1,2, Feyyaz Demir1,2, Marc Finne1, Zhuowen Tu3, Michael Bendno1,2, Benjamin Haibe-Kains4,5 & Anna Goldenberg1,2

![Diagram of networks](image)

Figure 1 | Illustrative example of SNF steps. (a) Example representation of mRNA expression and DNA methylation data sets for the same cohort of patients. (b) Patient-by-patient similarity matrices for each data type. (c) Patient-by-patient similarity networks, equivalent to the patient-by-patient data. Patients are represented by nodes and patients’ pairwise similarities are represented by edges. (d) Network fusion by SNF iteratively updates each of the networks with information from the other networks, making them more similar with each step. (e) The iterative network fusion results in convergence to the final fused network. Edge color indicates which data type has contributed to the given similarity.
Mapping the Human ‘Diseasome’

Researchers created a map linking different diseases, represented by circles, to the genes they have in common, represented by squares.

Related Article: Redefining Disease, Genes and All
• A central goal of genetics is to understand the links between genetic variation and disease.

• Intuitively, one might expect disease-causing variants to cluster into key pathways that drive disease etiology.

• But for complex traits, association signals tend to be spread across most of the genome—including near many genes without an obvious connection to disease.

• We propose that gene regulatory networks are sufficiently interconnected such that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes and that most heritability can be explained by effects on genes outside core pathways.

• We refer to this hypothesis as an “omnigenic” model.
An Expanded View of Complex Traits: From Polygenic to Omnigenic

Model: Most genes affect disease risk through highly connected cellular networks

Degrees of separation from core genes

Low 1 2 3 4 5 6 >7 High

Cumulative distribution

Heritability explained

Proportion of genes

Degrees of separation from core genes

Autoimmune GWAS hits affect shared and tissue-specific regulation of immune cells

Autoimmune GWAS SNPs

Null SNPs

eQTLs

Immune gene exp.

CNS gene exp.

Liver gene exp.

Genes with effects through cellular network

Expressed  Not expressed

Immune cellular network

CNS cellular network

Liver cellular network

Cell - Volume 169, Issue 7, p1177–1186, 15 June 2017
Hypernetworks

Hypernetworks

Thanks

Kat Hollywood
Senior Experimental Officer