Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

Lecture 1 – course introduction

Paul Pavlidis

January 5 2015

Today's topics

- What the course is about
- Course mechanics
- Introduction to high-dimensional biology

Your instructors

- Dr. Gaby Cohen-Freue – Assistant Professor of Statistics
  - gcohen@stat.ubc.ca
- Dr. Paul Pavlidis – Professor of Psychiatry/CHiBi
  - paul@chibi.ubc.ca
- Dr. Sara Mostafavi – Assistant Professor of Statistics / Medical Genetics – saram@stat.ubc.ca
- TAs: Evan Durno (wdurno@gmail.com), Alice Zhu (jingyunalice@gmail.com)

Course audience

- Researchers who want to know how to analyze large data sets from biological studies
- Genomics-focused, but information is broadly applicable

- Statistics students might find the math parts easy
- Biology students might find the biology easy
- We are counting on you to help make it work: help your peers!

Prerequisites

Officially, none. But:

- **Statistics** – You should have already taken university level "Statistics 101". You'll get a refresher, but you should be prepared to get comfortable thinking about things like "probabilities" and "specificity".
- **Biology** – **No requirements**, but you are expected to learn things like the difference between a DNA and RNA and a gene and a genome. We assume you are here because you are interested in biology and will pick it up.
- No **R** experience required but you must be prepared to do a lot of self-guided learning.
- You'll use your own computer to run R. If you can't install R on your computer, ask us for options.

What you can expect to learn

- Conceptual and practical knowledge you need to handle large biological data sets
  - Less about specific types of data, more about generally applicable approaches and principles
- You will be able to critically evaluate analyses in the literature
- Implementation of analyses using the R/Bioconductor computing environment

Not about:

- Formal mathematical theory underpinning the approaches
- Gory details of how to analyze any particular type of data at a low level

Topics covered

Probability foundations

Exploratory data analysis

Data QC and preprocessing

Basic statistical inference ("one gene at a time")

Large-scale inference ("genome-wide")

Count-based data (e.g. RNA-seq) analysis

DNA methylation analysis

Principal Component Analysis

Clustering

Classification

Resampling and bootstrap

Model selection and regularization

Gene sets and gene networks

Course mechanics

Course web site

http://stat540-ubc.github.io/

- Lecture notes
- Lab notes
- Assignments

Much interaction via Github (discussions, submission)

https://github.com/STAT540-UBC

Lectures

- ESB 4192
- Lectures shared among three professors
- Notes provided on web before class

Sections/Labs

- Wednesdays in room ESB 1042
- Officially from 12-1, but we will start at 11
  - 11-12: Lab available
  - 12-1: TA Office hour (this week: Mol. Bio. Primer)
- Self-guided exercises to help you learn to use R for analysis.
- Using your own computer (other options possible)
- Exercise material will be made available ahead of time
- Towards end of course, more time devoted to working on group projects.

Readings

- No textbook, but we can give suggestions
- Lectures often come with suggested background papers (reviews or primary literature)
- Helpful to access journals online (e.g. via the UBC VPN)
  - http://it.ubc.ca/services/email-voice-internet/myvpn/setup-documents

Evaluation

- **Homeworks**
  - Two assignments worth 20 points each
  - +5 points each for peer evaluation
- **Group project**
  - Planning + project + poster session – 40 points
- 10 Points for "other"
  - e.g. Preparedness, participation.

Homework assignment

- One for February, one for March.
- Involve detailed analysis of real data
- Deliverables include a short report and R code
- Two weeks from assignment to due date
- Lateness penalties

Group projects

- Starts today – start thinking about it
- A few minutes for group project pitches on Jan 19 and Jan 21.
- Form groups by Fri Jan 23 (3-4 people)
- Friday Jan 30: initial project proposals
- Feedback to groups Feb 13 – Proposals finalized by Feb ~15
- Work on projects over rest of term
- Final session of the course is the poster session

Group projects: where do they come from?

- Historically, almost all projects have been based on a data set provided by a student (i.e., collected in their lab).
- Occasionally, instead based on an idea from a student, using published data.
- If you need help thinking up an idea for a project let us know. But this has never been needed before (beyond refinement). If you are unsure of where you are going to get a project from, wait until you hear the project pitches.

Examples of past group projects

- Genomic copy number alterations for prognosis of prostate cancer
- Learning about proteins from other proteins: Protein Database Prediction
- Conditional epistasis profiling in yeast
- Epigenetic biomarkers for cancer diagnosis
- Comparative metagenomics : metabolic potential
- Epigenome and transcriptome in rice strains
- Analysis of HPV E2 protein on host gene expression
- Effects of Mutations in Histone Modifying Enzymes on Gene Expression Profiles
- Methodological considerations in analysis of Illumina Infinium methylation data
- Gene expression in invasive ragweeds
- Modeling time-course expression of SET domain-containing genes in mouse embryos
- Gene expression in blood of humans with asthma challenged with allergen

2011 and 2012 project titles, paraphrased

High-dimensional biology

1. What is it
2. What kinds of methods are used to analyze it
3. Some examples

Collecting data the low-dimensional way

- Pick one variable (e.g. "activity of a protein") and study it under various conditions.
- Repeat this for another variable
- Usually "hypothesis-driven"
- Powerful, but knowledge accumulates slowly and synthesis is difficult

Biology is complicated

- Thousands of "parts"
- Limitations of the "one thing at a time approach" – how do the parts work together?
- Technology enabling increasingly detailed analyses – measure many things in parallel

- Drawback: Fishing expeditions?

Defining "high dimensional"

- Large number of features measured in each sample/subject/individual ("high content")
  - Genes, proteins, DNA sites, brain regions, etc.
- Not *usually* talking about huge numbers of samples (e.g. individuals studied) –
  - often10s, but can be 1000s (some genetics studies)
- Studies can sometimes be "non-hypothesis driven"

Example of a question answered with a high-dimensional approach

- Tumor type A is deadly and type B is more easily treatable (but still bad)
- Telling A from B is difficult
  - Cells look the same, etc. – we only find out by seeing what happens to the patients.
- We know that cancer is a "gene" disease

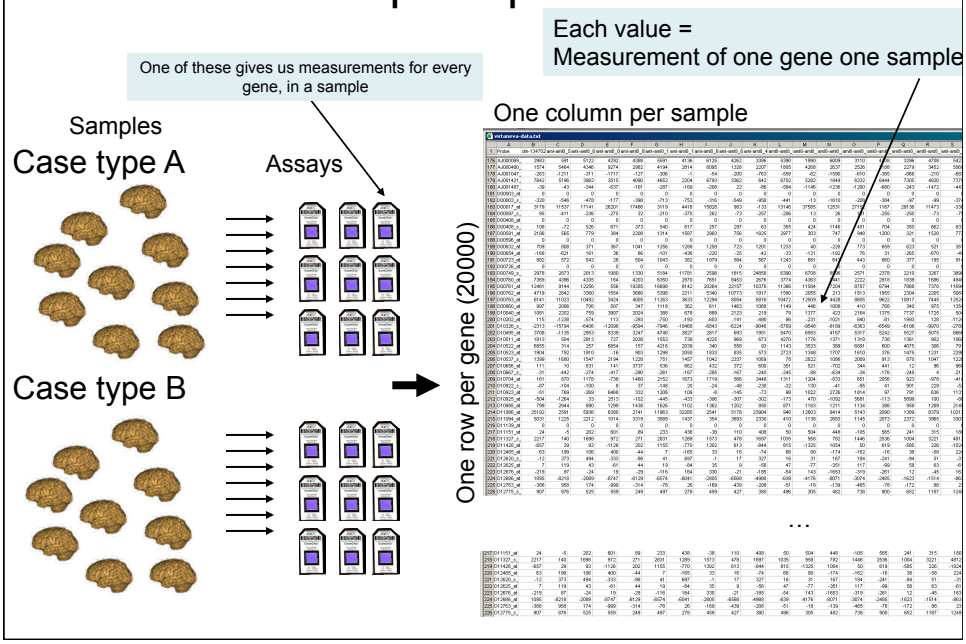
Questions:

- Where is the difference?
- Can we find new targets for drugs or for diagnosis?
  - (Drug targets are usually proteins, encoded by genes)

Looking for insight from genomics

- Since cancer is a disease of genes, let's look at the genes - not just one, but all of them
- We are hypothesizing that there is *some* difference in genes between the two types, if only we could find it
- But we're not starting with a *specific* hypothesis. We're going to test thousands of hypotheses
- In this example, we're going to look at "gene expression levels" – a measure of "how active" is each gene.

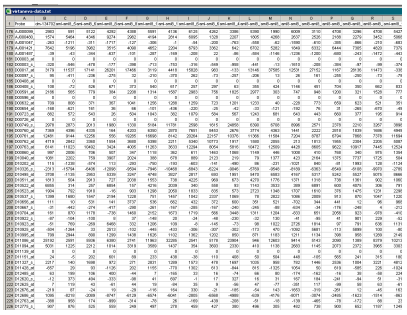
Example experiment

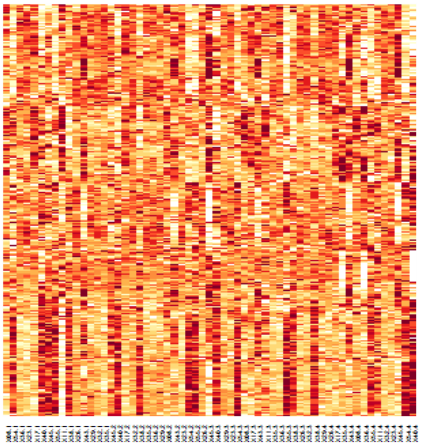


A partial list of things to assay

- DNA/Chromatin
  - Genotypes, copy numbers ("mutations" and variants)
  - DNA methylation
  - Chromatin state (histone marks, transcription factors …)
- RNA
  - Quantification of transcripts (protein coding, non-coding)
  - Transcript variants (splicing, editing)
- Proteins
  - Detection, Quantification
  - Binding and complexes
- Metabolites and other small molecules
- Phenotypic screens
  - RNAi (etc.)
  - Genetic interactions
- Cellular composition of a sample (cell types)

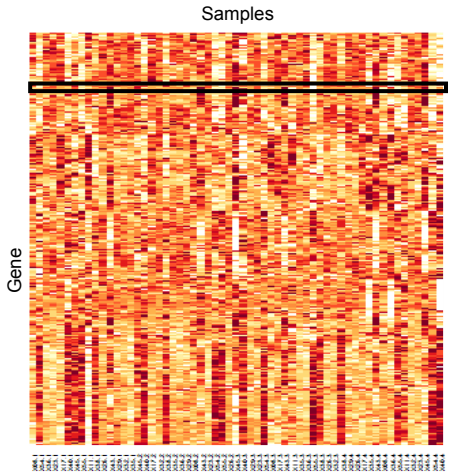
Alternative representation

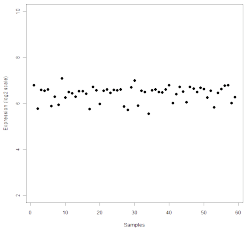




Lighter colours mean higher levels of gene expression (“activity”)
Only show part of the data!

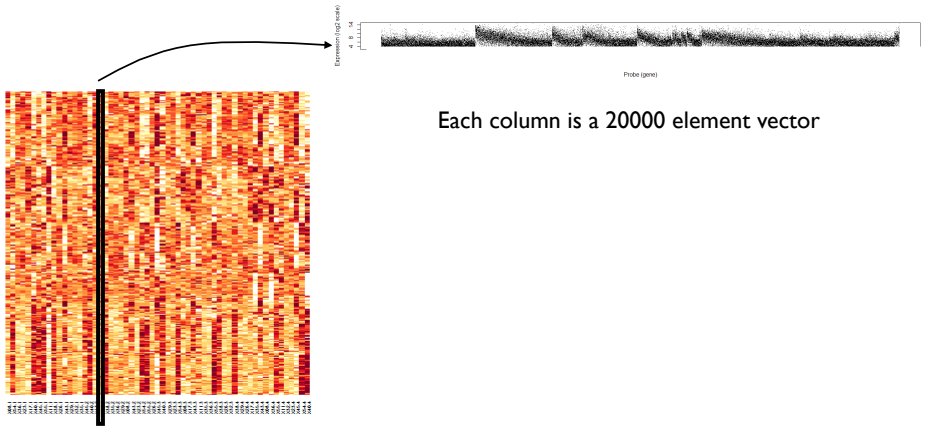
Profile for a gene





Profile for a gene. This is a 59-element vector

Profile for a sample



(for many, many more rows)

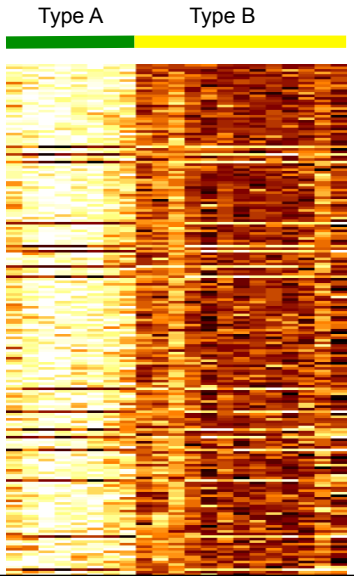
Each column is a 20000 element vector

*This is a schematic. The graph and color map don't match

One type of analysis

- I've ranked the genes by how different they are between types A and B (t-statistic)
- Only the first few genes are shown
- Though it can be a lot more complicated, most "high-dimensional" studies boil down to something like this, at least in part

What's the big deal?



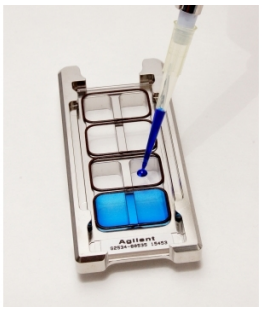
Pitfalls and challenges

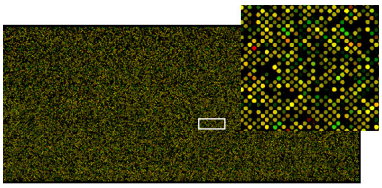
- Signals can be small and buried in lots of non-signals; False positives are a danger.
- Need to detect outliers, batch effects and other confounds
- Can we make better use of the fact that we're testing 20,000 genes than just doing a t-test on each one?
- Data sets (and questions) are often much more complex
- Getting just a list of "hits" isn't enough – can we understand something more about the "system"

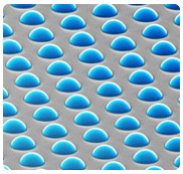
High-dimensional technologies

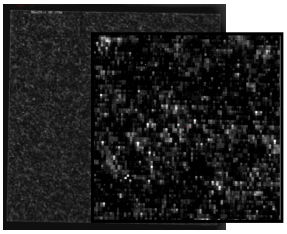
- DNA & RNA sequencing
  - Transcriptomes, exomes, full genomes
- Complex gene library construction
  - Expression vectors, protein tags, knockdowns
- Microarrays and other robotic/parallel tech.
  - Screens, high-content assays …
- Mass spectroscopy
- Flow cytometry
- Imaging

Microarrays









Agilent SurePrint

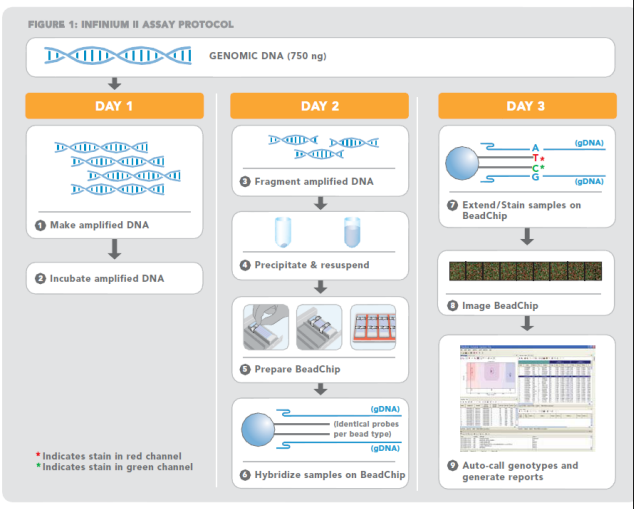
Illumina Beadarray

Affymetrix Genechip

SNP arrays

- Similar idea to the RNA arrays, but hybridize genomic DNA, and probe is designed to be sensitive to the allele
- Intensities are converted into a "call", with a quality score.
- Low-quality calls are usually simply treated as missing data.
- DNA methylation arrays involves specialized versions of these (+bisulphile convesrion)

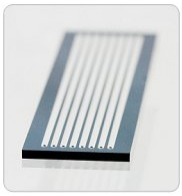
Illumina



Sequencing-based assays

- Instead of using hybridization to a designed probe, determine the DNA sequence of the sample
- Several competing platforms

- Genotyping: Compare to a reference
- RNA: quantify how many times you see a sequence



Up to eight samples can be loaded onto the flow cell for simultaneous analysis on the Illumina Genome Analyzer.

Illumina HiSeq

Analysis modes

- What is the general toolkit available for the analysis of data?
- How are these specialized for high-dimensional data?

Exploratory analysis

- The first thing you do with your data
- Graphs and other visualizations, often combined with data reduction
- Use to spot problems, formulate hypotheses
- Often rely on power of human brain
- Data reduction essential to make exploration tractable for large data sets, even then it can be a challenge
- Follow up with more formal analysis

Model fitting and hypothesis testing

- Formally test a specific question about the data
- Is what I see "statistically significant"?
- False positives are a major risk in large data sets
- Can exploit repeating structure of the data to improve ability to find true positives

Unsupervised learning

- "Learn" undiscovered groupings in the data
- Clustering -- how do my samples or features group together?
- Useful as an exploratory technique as well as "data mining" when backed with quantitative analyses
- Example: Finding previously unknown groups of subjects based on a gene profile

Supervised learning

- Can I predict an unmeasured feature of a sample from a measured one?
- Less common than unsupervised learning, most commonly used in clinically-oriented settings – development of biomarkers
- Example: predicting tumour drug response based on gene profiles

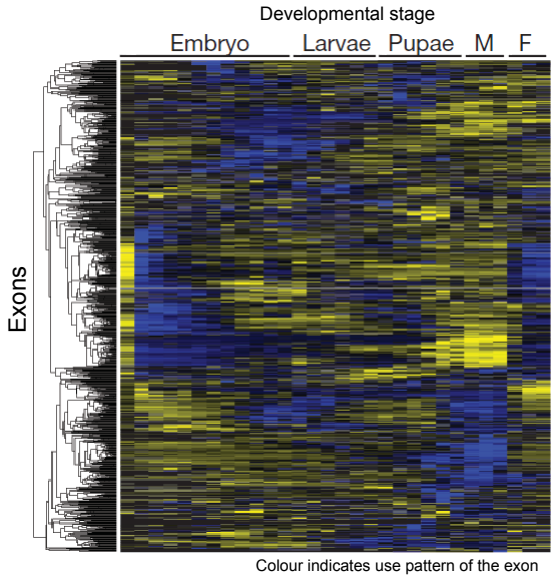
Other methods

- Many analyses just give a list of genes
- "Downstream" analysis needed to make sense of it - "biological interpretation"
  - Overlay/combine/compare with other data
  - Transform one data set into another type of data at a different granularity
    - Genes → pathways
- Usually these end up returning to exploratory etc. modes

More examples

- Illustrate some real-life cases of high-dimensional data
- We hope to teach you enough in the course to do at least primitive versions of these analyses
- … or at least be able to read the papers
- … even if it's a type of experiment we don't teach in detail.

Example 1: Analysis of RNA in fly lifespan with RNA-seq



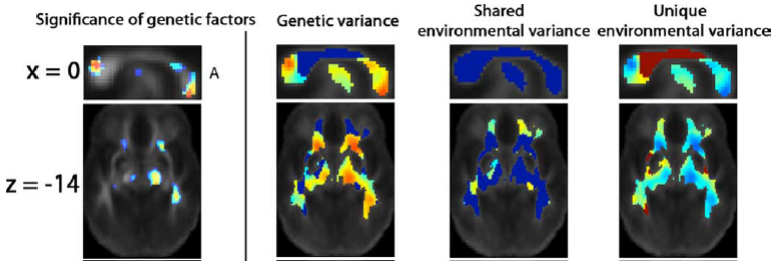
- 30 developmental stages
- Analysis at the exon level
- Heat maps
- Clustering
- GO enrichment
- Generates many new hypotheses

The developmental transcriptome of *Drosophila melanogaster*

http://www.nature.com/nature/journal/vaop/ncurrent/full/nature09715.html

Example 2: How much of brain structure* differences are accounted for by:

- Relatedness (twins)
- IQ



* Fractional anisotropy / White matter integrity

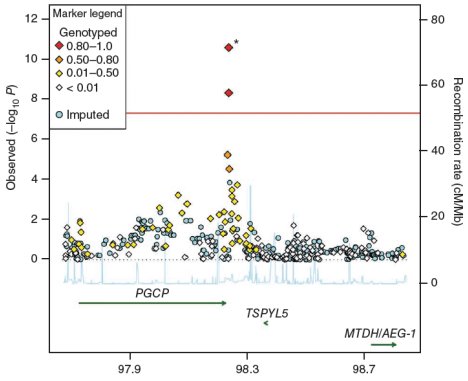
Behavioral/Systems/Cognitive

Genetics of Brain Fiber Architecture and Intellectual Performance

http://www.ncbi.nlm.nih.gov/pubmed/19228974

- 92 identical or fraternal twins
- 1.5 million voxels per subject
- Linear models, factor analysis
- Multiple test correction
- Heat maps
- Genetics explains 80% of the variance
- Brain structures correlated with IQ ~0.3

Example 3: Genetics of migraine



- 13,500 individuals
- 429,912 DNA sites tested
- Analysis* with multiple test correction to identify markers associated with migraine
- One site is "A" in 0.267 of the migraine-affecteds but only 0.216 of the controls
- 40% higher risk of migraine if you have "A"

Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1

Verneri Anttila[1,2,*], Hreinn Stefansson[3], Mikko Kallela[4], Unda Todt[5,6], Gisela M Terwindt[7], M Stella Calafato[1,8], Dale R Nyholt[9], Antigone S Dimas[1,10,11], Tobias Freilinger[12,13], Bertram Müller-Myhsok[14], Ville Artto[4], Michael Inouye[1,15], Kirsi Alakurtti[1,2], Mari A Kaunisto[2,16], Eija Hämäläinen[1,2], Boukje de Vries[17], Anine H Stam[7], Claudia M Welter[13], Axel Heinze[17], Katja Heinze-Kuhn[17], Ingrid Goebel[17,4], Guntram Borck[18,4], Hartmut Göbel[17], Stacy Steinberg[3], Christiane Wolf[14], Asgeir Björnsson[3], Gretar Gudmundsson[18], Malene Kirchmann[19], Anne Hauge[19], Thomas Werge[20], Jean Schoenen[21], Johan G Eriksson[16,22–24], Knut Hagen[25], Lars Stovner[25], H-Erich Wichmann[26–28], Thomas Meitinger[29,30], Michael Alexander[31,32], Susanne Moebus[33], Stefan Schreiber[34,35], Yurii S Aulchenko[36], Monique M B Breteler[36], Andre G Uitterlinden[37], Albert Hofman[36], Cornelia M van Duijn[36], Päivi Tikka-Kleemola[38], Salli Vepsäläinen[4], Susanne Lucas[4,5], Federica Tozzi[39], Pierandrea Muglia[39,40], Jeffrey Barrett[1], Jaakko Kaprio[2,24,41], Markus Färkkilä[4], Leena Peltonen[1,2,42,43], Kari Stefansson[3], John-Anker Zwart[25,43], Michel D Ferrari[7], Jes Olesen[19], Mark Daly[42], Maija Wessman[2,16], Arn M J M van den Maagdenberg[7,45], Martin Dichgans[12,13], Christian Kubisch[5,6,44,45], Emmanouil T Dermitzakis[11], Rune R Frants[15] & Aarno Palotie[1,2,42,46,47] for the International Headache Genetics Consortium

VOLUME 42 | NUMBER 10 | OCTOBER 2010 NATURE GENETICS

http://www.nature.com/ng/journal/v42/n10/abs/ng.652.html

* Cochran-Mantel-Haenszel, like Fisher's exact test