Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

Lecture 1 – course introduction

Paul Pavlidis

January 4 2017

Today's topics

- What the course is about
- Course mechanics – details and up-to-date info will be at https://stat540-ubc.github.io/

(These lecture notes are just an overview)

- A primer on molecular biology
- Introduction to high-dimensional biology

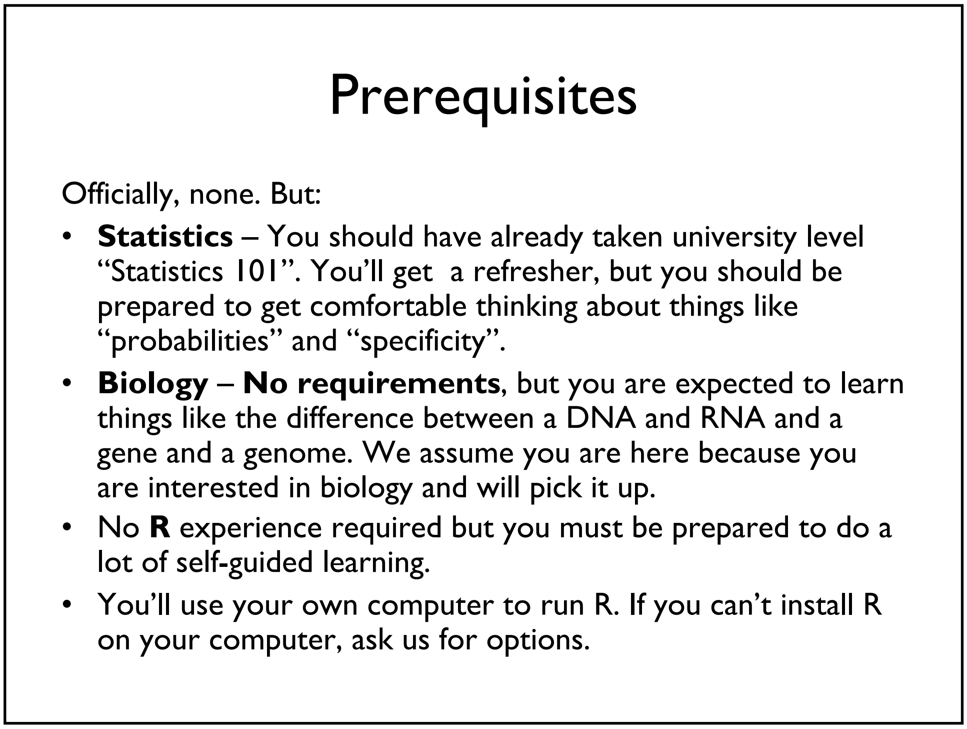
Your instructors

- Dr. Paul Pavlidis – Professor of Psychiatry/Michael Smith Labs
  - paul@msl.ubc.ca
  - Bring administration questions to me
- Dr. Amrit Singh – Dept. Pathology & Lab Medicine /Centre for Heart and Lung Innovation @ St. Paul's
- Dr. Rob Balshaw – Dept. of Statistics and BC CDC
- TAs: Farnush Farhadi (farnush.farhadi@gmail.com), Santina Lin (Part-time; hello@santina.me)

Course audience

- Researchers who want to know how to analyze large data sets from biological studies
- Genomics-oriented, but focus on broadly applicable statistical approaches

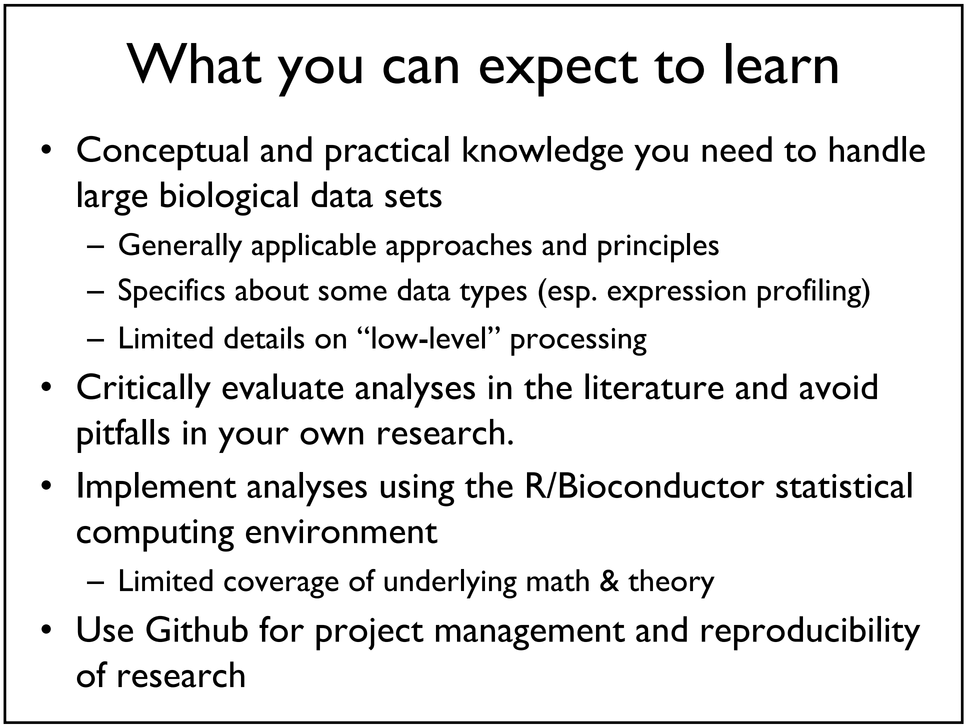
- Statistics students might find the math parts easy
- Biology students might find the biology easy
- We are counting on you to help make it work: help your peers!



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
  - Generally applicable approaches and principles
  - Specifics about some data types (esp. expression profiling)
  - Limited details on “low-level” processing
- Critically evaluate analyses in the literature and avoid pitfalls in your own research.
- Implement analyses using the R/Bioconductor statistical computing environment
  - Limited coverage of underlying math & theory
- Use Github for project management and reproducibility of research

Topics covered in the course

Probability foundations

Exploratory data analysis

Data QC and preprocessing

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

Large-scale inference ("genome-wide") – multiple testing

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

DNA methylation analysis

Principal Component Analysis

Clustering

Classification

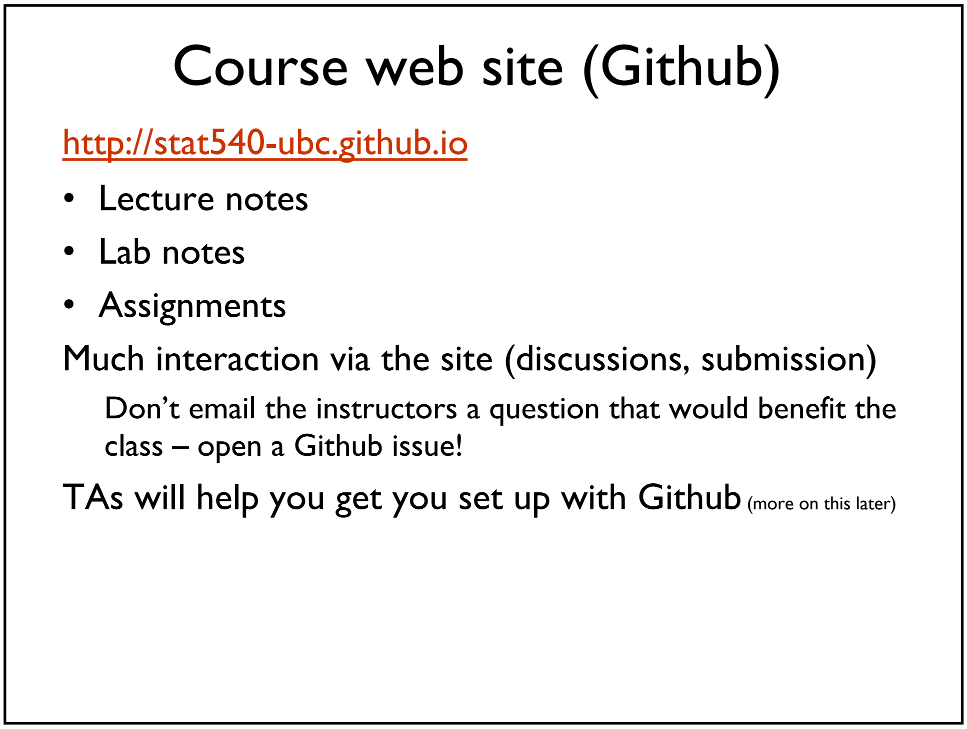
Resampling and bootstrap

Model selection and regularization

Gene sets and gene networks

https://stat540-ubc.github.io/subpages/syllabus.html

Course mechanics



Course web site (Github)

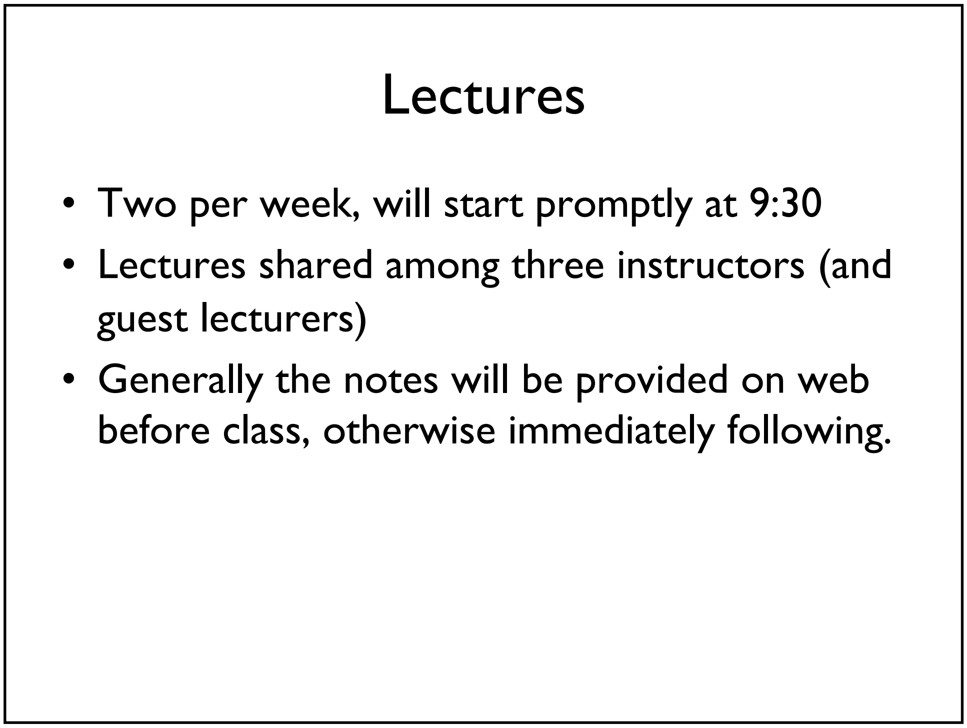
http://stat540-ubc.github.io

- Lecture notes
- Lab notes
- Assignments

Much interaction via the site (discussions, submission)

Don’t email the instructors a question that would benefit the class – open a Github issue!

TAs will help you get you set up with Github (more on this later)



Lectures

- Two per week, will start promptly at 9:30
- Lectures shared among three instructors (and guest lecturers)
- Generally the notes will be provided on web before class, otherwise immediately following.

Seminars (“Labs”)

- Wednesdays in room ESB 1042
- 11AM-1PM – please plan on coming for one hour. For first session (TODAY):
  - 11-12:30 R and github setup help
  - 12:30-1:00 Mol. Bio. Primer

What’s in the labs?

- First half of course: Self-guided exercises using R
  - Using your own computer (other options possible)
  - Exercise material will be made available ahead of time
- Towards end of course, 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)
  - https://it.ubc.ca/services/email-voice-internet/myvpn

Evaluation

- **Homework**
  - Assignment worth 30 points – in two parts (plus one ‘setup’ assignment worth 5 points)
- **Group project**
  - Planning + project + poster session + report = 60 points
- **Short writing assignment**
  - 5 points

(Note changes from last year!)

Homework assignment

- Involve detailed analysis of real data
- Deliverables include a short report and R code
- Will have two weeks from assignment to due date
- Lateness penalties

Short writing assignment

- Revised for this term
- Meant to develop your skills relating course material to the literature
- Select, read, summarize and critique a recent paper from the ‘omics literature
- Details to follow

Group projects

- Starts **today** – start thinking about it
- A few minutes for group project pitches later this month during lecture time, also proposed via github
- Form groups by Fri Jan 25 (4-5 people) and provide an initial rough project proposal
- You’ll get feedback from instructors
- Proposals finalized by Feb15
- Work on projects over rest of term
- Final session of the course is the poster session

Dates may change – refer to course site

Group projects: where do they come from?

- Nearly all projects have been based on a data set provided by a student (i.e., collected in their lab).
- Occasionally 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

Your first assignment

- Due in two weeks (Jan 18) – 5 points
- See https://github.com/STAT540-UBC/STAT540-UBC.github.io/blob/master/homework/practice_assignment/practice_assignment.md
- You will get to:
  - Set up your github environment
  - Get R installed and cover some basics
  - Make your first github commits.

First step ASAP: complete survey at https://goo.gl/forms/P8pRUppM2XLIJNQL2

Setting up for STAT540, 2017

Hi, let me help you set up a GitHub repo in the STAT540 GitHub organization !

* Required

Molecular biology in 5 slides

Ignoring many exceptions and complications!

- **DNA**: linear arrangement of nucleotides ('bases'), contains information to construct the organism; provides mechanism of heritability
  - Every time a cell divides, the DNA is copied **(replicated)**
- **Gene**: a stretch of DNA that is transcribed into a functional RNA
  - Simplest genomes contain ~1000 genes (e.g.: E. coli has ~4000, yeast has ~6000)
  - Most multicellular organisms have ~15-25k genes (e.g. worms, flies, vertebrates)
- **Genome**: the full complement of DNA in one cell
  - The sequence of the DNA is the **genotype**; can refer to a specific place ("locus") or overall.
  - The properties of the organism produced is the **phenotype**
- **RNA**: Immediate read-out of a gene, also made of nucleotides ("transcript")
  - Complication: Splicing. Primary transcript is of exon and intron regions, latter are removed; "**exome**" is the set of all exons. A single processed transcript is a messenger or mRNA.
- **Protein**: Major working parts of cells, encoded by genes (via RNA) and made ("translated") by the ribosome (a big molecular machine)
  - Proteins are strings of amino acids ("polypeptides"); 3 nucleotides code one AA
- DNA, RNA and protein (plus many other types of molecules used by cells) are produced using chemicals (from air and food) and energy from sunlight (directly or indirectly via food) = "**metabolism**", a process which involves the function of (at least) hundreds of genes.

The human genome

(a typical mammalian genome)

- ~6 billion bases of DNA over 23 pairs of chromosomes (3 billion bases in "haploid")
- Of the total 46 chromosomes, you got 23 from mom and 23 from dad
- ~10-15% of the human genome is functional – the rest is "junk"
  - 10-15% includes exons of genes and regulatory elements
  - Balance includes bulk of introns and intergenic regions
- About 20,000 protein-coding genes +some RNA-only genes

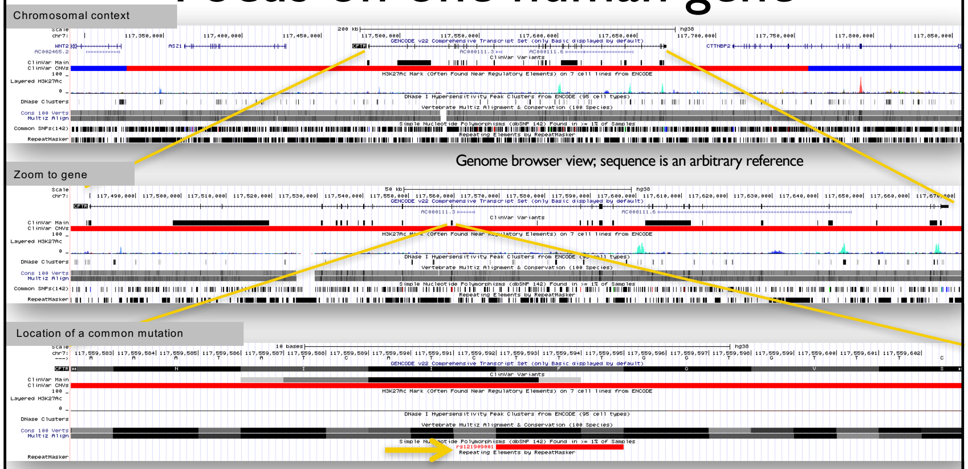
Molecular regulation

- Not all genes are active in any given cell
  - Many organisms are made of different types of cells – the differences are established by changing which genes are active
  - All organisms regulate which genes are active depending on the environment
- Regulation happens at multiple levels (transcription, translation, post-translation)
- System of signals, receptors, switches = complex "wiring" of genes with each other and with the environment – goal of "systems biology" is to understand this in full detail.

Genetic variation

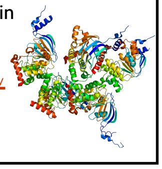
- Differences in DNA sequence between two individuals of the same species (different genotypes) – humans: millions of differences
- Most are inherited from mom and dad, but also acquired during your life.
  - Depending on when acquired, will affect all or part of your body.
- Most variation has no effect on phenotype ("neutral" or nearly)
  - Occurs in a junk region, is silent (degeneracy of amino acid code), or affects an "unimportant" gene
- Some variation is "deleterious"
  - Slightly deleterious: increases your risk of disease; can be hard to detect
  - Highly deleterious: mutations "cause" disease; e.g. Cystic fibrosis
- Even more rarely variation can be "beneficial"
- In population genetics "deleterious" and "beneficial" are defined by effects on **reproductive success** (how many descendants you have), but we can also talk about it in terms of the effect on the gene's biochemical function.

Focus on one human gene



Genome browser view; sequence is an arbitrary reference

- CFTR: Gene “for” Cystic fibrosis – encodes a protein involved in regulating cell fluid levels
- Most common mutation is deletion of 3 bases (“delF508”) causing an unstable protein
  - One copy: probably beneficial (in population genetics terms; retained in the population)
  - Two copies (no “good” copy) – associated with CF risk
- http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&position=chr7%3A117475813-117672093&hgsid=466552281_9bE3VrJY4rVhIAbe4NUU7CvQSmxV
- http://www.cfmedicine.com/htmldocs/cftext/genetics.htm
- https://en.wikipedia.org/wiki/Cystic_fibrosis_transmembrane_conductance_regulator
- http://www.genet.sickkids.on.ca/CftrDomainPage.html?domainName=NBD1
- http://www.rcsb.org/pdb/explore/explore.do?structureId=1XMI

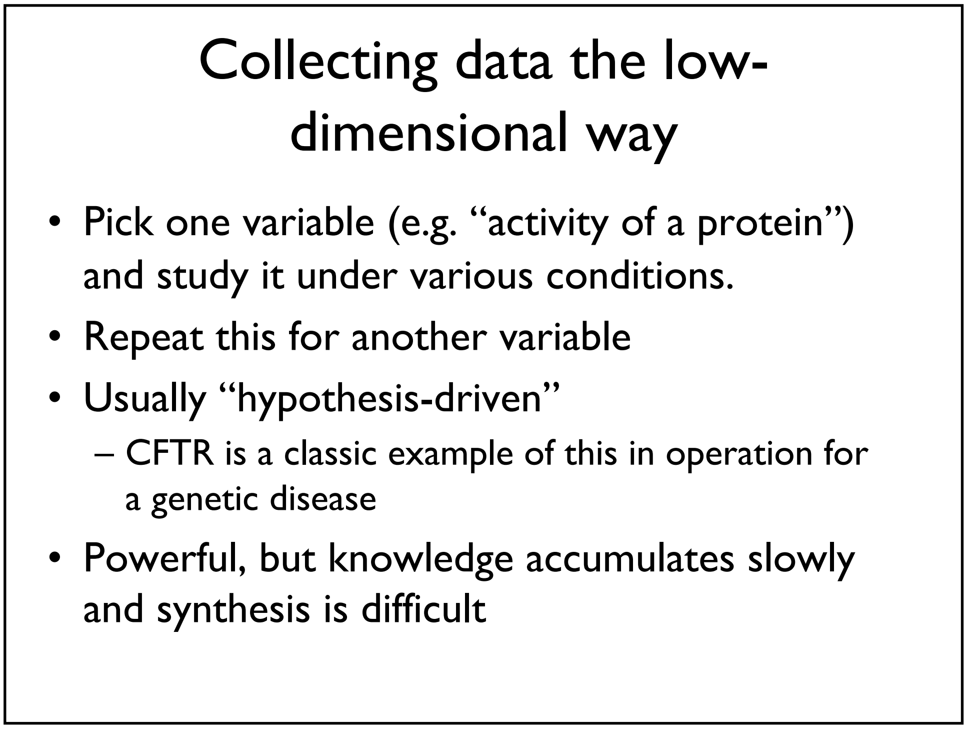


Key points…

- Thousands of “moving parts” (e.g. genes)
- Hugely complex interactions and regulation – poorly understood
- Many genes have poorly understood function
- Genetic variation and environment interact in complex ways
- Many diseases have a genetic component still to be understood.
  - Same applies to non-humans (plant disease resistance, etc.)
- Reductionist paradigm: conceptually works backwards from the phenotype to the genotype, attempting to resolve the steps in between

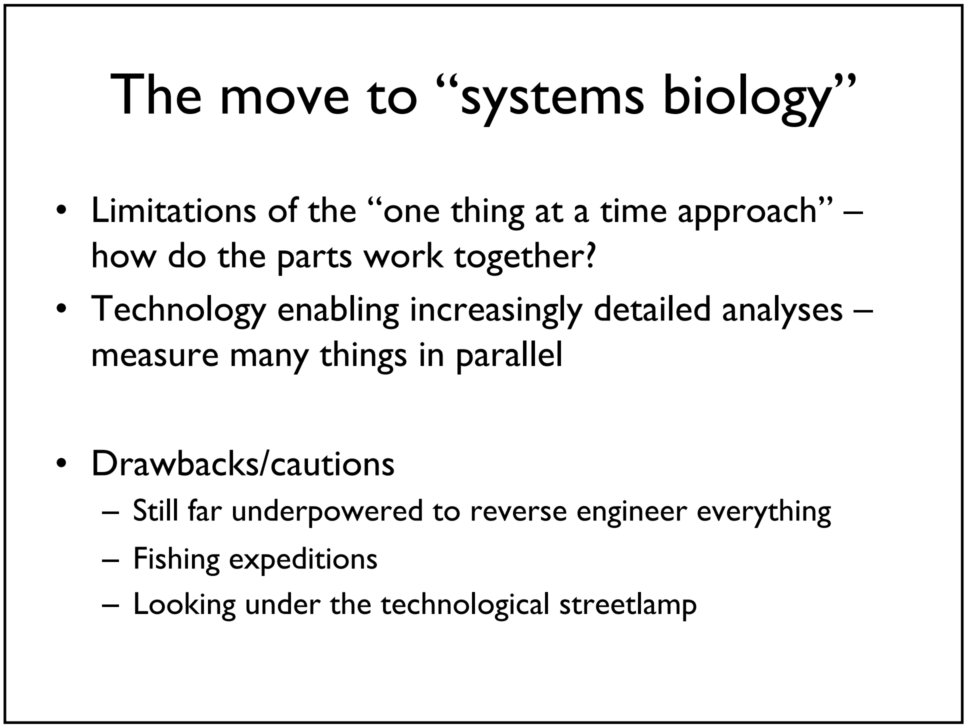
High-dimensional biology

1. What, why and how
2. Overview methods are used to analyze it



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”
  - CFTR is a classic example of this in operation for a genetic disease
- Powerful, but knowledge accumulates slowly and synthesis is difficult



The move to “systems biology”

- 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

- Drawbacks/cautions
  - Still far underpowered to reverse engineer everything
  - Fishing expeditions
  - Looking under the technological streetlamp

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)

Many things to assay…

In this course, we are largely concerned with quantitative measurements of genomic features specifically:

- Gene transcript levels (RNA)
- DNA methylation (Lecture 15 and Seminar 8)

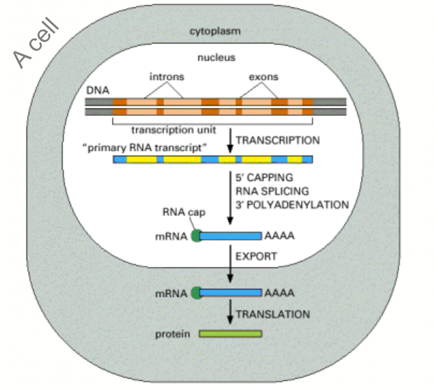
Contrast with analysis of DNA sequence

- Closest analogue in course scope is Genome Wide Association Analysis (GWAS)
- Characterization of rare variants in individual genomes is another major area of study (Lecture 23)

However, many of the statistical approaches and issues are relevant to many data types

- Protein levels (quantitative proteomics)
- Metabolomes (quantitative small molecule analysis)
- Microbiomes (quantitative analysis of microbe species distributions)
- Functional brain imaging (units of analysis are brain regions)

Gene expression



(Diagram shows just one gene)

At RNA level, regulatory points include:

- Rate of transcription
- Rate of modification (splicing etc.) and export to the cytoplasm
- Rate of degradation

Alberts, Molecular Biology of the Cell

Why study gene expression?

- The **"readout" of genetic variation** is partly in gene expression – it can help us understand the link between genetics and phenotypes
- **Genes expression is regulated** and changes in response to environment, disease, age, etc.
- Genes are expressed at an appropriate time and place: we can learn (or guess) about gene function by comparing expression patterns (**guilt-by-association**)
- The pattern of gene expression can be used as a '**fingerprint' of the state the sample** was in at the time of measurement – a biomarker
- Examining the details about which genes are relevant to the fingerprint ("differentially expressed") give **insight into the process/disease/condition** of interest.

High-dimensional technologies

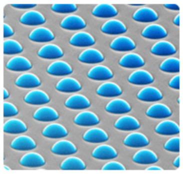
In this course 99% of the experiments we discuss involve one of two basic technologies:

Microarrays

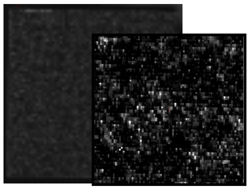
Massively parallel nucleotide sequencing

Microarrays

Illumina Beadarray



Affymetrix Genechip

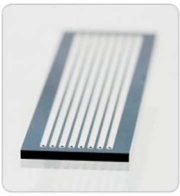


- Substrate glass slide contains specific short DNA probes for genes of of interest.
- Each spot is one probe sequence; possibly multiple spots for a given gene.
- **Hybridize** a labeled mixture of RNA from sample.
- Readout: expression level for many genes for one sample

- Variants of same idea can be used for DNA site-specific sequence and methylation analysis

Sequencing-based assays

- Instead of using hybridization to a designed probe, determine many (millions) of short randomly-selected sequences from the sample.
- RNA: quantify how many times you see a sequence (~mRNA molecule)
- Same technology can also be used for DNA Methylation and genome sequencing.



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

Illumina

A typical genomics study

Motivation:

- Tumor type A is deadly while type B is more treatable
- Telling A from B is difficult using "conventional" means
  - Cells look the same, etc. – we only find out by seeing what happens to the patients.

Questions:

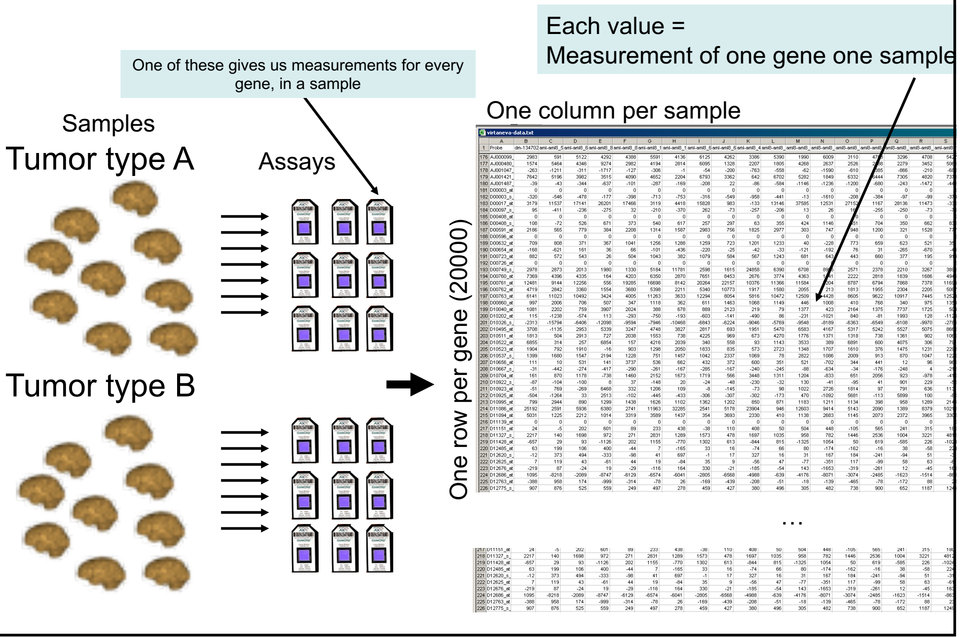
- Can we characterize the differences better?
- Can we find new targets for drugs or for diagnosis?
  - Drug targets are usually proteins, encoded by genes
- Are there other subtypes that can be described?

Looking for insight from genomics

- 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.

This example is only partly realistic: these days, DNA sequencing would probably be done as well. Looking at RNA is still a useful adjunct to examining the DNA.

Example experiment



Each value = Measurement of one gene one sample

One of these gives us measurements for every gene, in a sample

One column per sample

Samples

Tumor type A

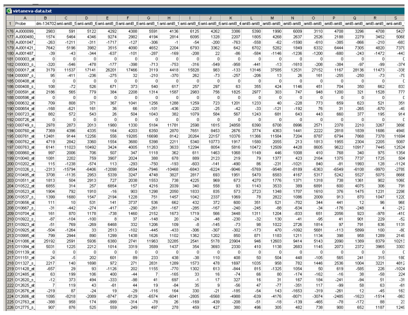
Assays

One row per gene (20000)

Tumor type B

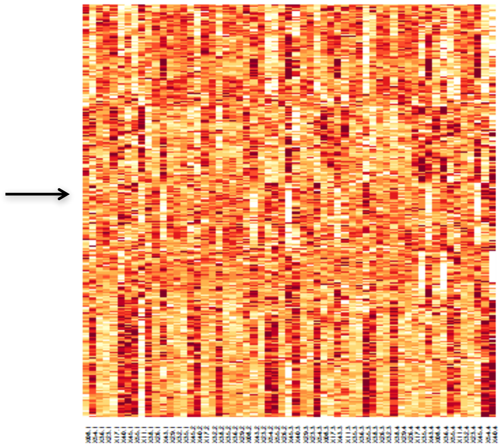
…

Alternative representation



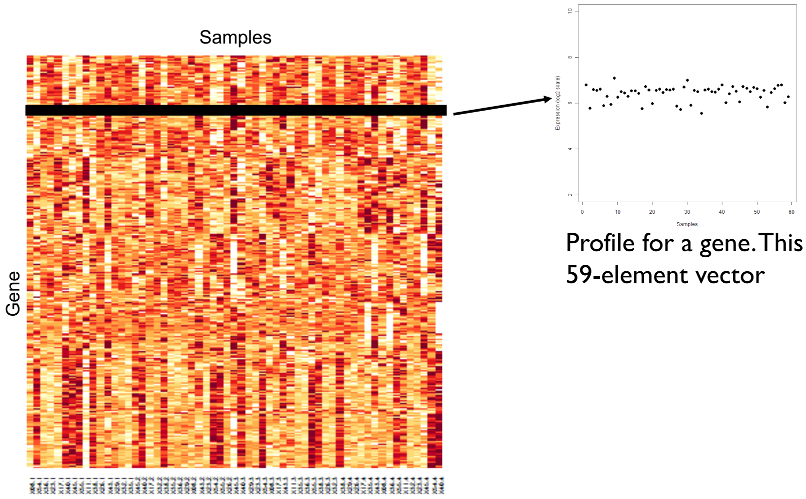
virtaneva-data.txt

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Probe	dm-134702	aml-aml8_5	aml-aml8_6	aml-aml8_8	aml-aml8_8	aml-aml8_1	aml-aml8_1	aml-aml8_6	aml-aml8_6	aml-aml8_4	aml8-aml8_	aml8-aml8_	aml8-aml8_	aml8-aml8_	aml8-aml8_	aml8-aml8_	aml8-aml8_	aml-aml8_
176	AJ000099_	2983	591	5122	4292	4388	5591	4136	6125	4262	3386	5390	1990	6009	3110	4708	3296	4708	5427
177	AJ000480_	1574	5464	4346	9274	2982	4194	2814	6095	1328	2207	1805	4268	2637	2526	2188	2279	3452	5068
178	AJ001047_	-263	-1211	-311	-1717	-127	-306	-1	-54	-200	-763	-558	-62	-1590	-610	-385	-866	-210	-683
179	AJ001421_	7642	5196	3982	3515	4090	4652	2204	6793	3362	642	6702	5282	1849	6332	6444	7305	4820	7379
180	AJ001487_	-39	-43	-344	-637	-101	-287	-169	-208	22	-86	-584	-1146	-1236	-1200	-680	-243	-1472	-443
181	D00003_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
182	D00003_s_	-320	-546	-478	-177	-398	-713	-753	-316	-549	-958	-441	-13	-1610	-208	-384	-97	-99	-374
183	D00017_at	3179	11537	17141	26201	17466	3119	4418	15828	983	-133	13146	37585	12531	27152	1167	28136	11473	-338
184	D00097_s_	95	-411	-236	-275	32	-210	-370	262	-73	-257	-206	13	26	161	-255	-250	-73	-79
185	D00408_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
186	D00408_s_	108	-72	526	671	373	540	617	257	297	63	355	424	1146	451	704	350	662	833
187	D00591_at	2186	565	779	384	2208	1314	1587	2983	756	1825	2977	303	747	948	1200	321	1528	777
188	D00596_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
189	D00632_at	709	808	371	367	1041	1256	1288	1259	723	1201	1233	40	-228	773	659	623	521	351
190	D00654_at	-168	-621	161	36	66	-101	-436	-220	-25	-42	-33	-121	-192	76	31	-265	-670	-49
191	D00723_at	882	572	543	26	504	1043	382	1079	584	567	1243	681	643	443	660	377	195	914
192	D00726_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
193	D00749_s_	2978	2873	2013	1980	1330	5184	11781	2598	1615	24858	6390	6708	8996	2571	2378	2210	3267	3898
194	D00760_at	7369	4396	4335	164	4203	6350	2870	7651	8453	2676	3774	4363	1441	2222	2818	1839	1686	4949
195	D00761_at	12481	9144	12256	556	19285	16698	8142	20264	22157	10376	11366	11584	7204	8787	6794	7868	7378	11694
196	D00762_at	4719	2842	3360	1554	3680	5398	2211	5340	10773	1917	1580	2055	213	1813	1955	2304	2205	5087
197	D00763_at	6141	11023	10492	3424	4005	11263	3633	12294	8054	5816	10472	12509	4428	8605	9622	10917	7445	12524
198	D00860_at	997	2006	706	507	347	1118	362	611	1463	1068	1149	446	1008	410	768	340	975	1354
199	D10040_at	1081	2202	759	3907	2024	388	678	889	2123	219	79	1377	423	2164	1375	7737	1725	504
200	D10202_at	115	-1238	-574	113	-293	-750	-193	-603	-141	-490	86	-231	-1021	840	-81	1993	128	-1124
201	D10326_s_	-2313	-15794	-6406	-12098	-9594	-7946	-10468	-6843	-6224	-9046	-5769	-9548	-8189	-6363	-6549	-6108	-9970	-2789
202	D10495_at	3708	-1135	2953	5339	3247	4748	3827	2817	693	1951	5470	6583	4167	5317	5242	5527	5075	8666
203	D10511_at	1813	504	2913	727	2038	1553	738	4225	969	673	4270	1776	1371	1318	738	1361	902	1060
204	D10522_at	6855	314	257	6854	157	4216	2039	340	558	93	1143	3533	389	6891	600	4075	306	791
205	D10523_at	1904	792	1910	-16	903	1298	2050	1833	835	573	2723	1348	1707	1610	376	1475	1231	2298
206	D10537_s_	1399	1680	1547	2194	1228	751	1457	1042	2337	1069	78	2822	1086	2009	913	870	1047	1220
207	D10656_at	111	10	531	141	3737	536	662	432	372	600	351	521	-702	344	441	12	96	968
208	D10667_s_	-31	-442	-274	-417	-290	-261	-167	-285	-167	-240	-245	-88	-634	-34	-176	-248	4	-212
209	D10704_at	161	870	1178	-738	1460	2152	1673	1719	566	3448	1311	1204	-833	651	2056	923	-978	-410
210	D10922_s_	-87	-104	-100	8	37	-148	20	-24	-48	-230	-32	130	-41	-95	41	901	229	-52
211	D10923_at	-51	769	-269	6468	332	1206	109	-8	-145	-73	98	1022	2726	1814	97	791	636	1131
212	D10925_at	-504	-1264	33	2513	-102	-445	-433	-306	-307	-302	-173	470	-1092	5681	-113	5899	100	-80
213	D10995_at	799	2944	890	1299	1438	1626	1102	1362	1202	850	871	1183	1211	1134	398	958	1289	2149
214	D11086_at	25192	2591	5936	6380	2741	11963	32285	2541	5178	23904	946	12603	9414	5143	2090	1389	8379	10213
215	D11094_at	5031	1225	2212	1014	3319	3589	1437	354	3693	2330	410	1138	2683	1145	2073	2372	3965	3303
216	D11139_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
217	D11151_at	24	-5	202	601	89	233	438	-38	110	408	50	504	448	-105	565	241	315	180
218	D11327_s_	2217	140	1698	972	271	2831	1289	1573	478	1697	1035	958	782	1446	2536	1004	3221	4812
219	D11428_at	-657	29	93	-1126	202	1155	-770	1302	613	-844	815	-1325	1054	50	619	-585	226	-1024
220	D12485_at	63	199	106	400	-44	7	-165	33	16	-74	66	80	-174	-162	-16	38	-58	224
221	D12620_s_	-12	373	494	-333	-98	41	697	-1	17	327	16	31	167	184	-241	-94	51	-31
222	D12625_at	7	119	43	-61	44	19	-84	35	9	-56	47	-77	-351	117	-99	58	63	-61
223	D12676_at	-219	87	-24	19	-29	-116	164	330	-21	-185	-54	143	-1653	-319	-261	12	-45	163
224	D12686_at	1095	-8218	-2089	-8747	-6129	-6574	-6041	-2805	-6568	-4988	-639	-4176	-8071	-3074	-2485	-1623	-1514	-863
225	D12763_at	-388	958	174	-999	-314	-78	26	-169	-439	-208	-51	-18	-139	-465	-78	-172	88	23
226	D12775_s_	907	876	525	559	249	497	278	459	427	380	496	305	482	738	900	652	1187	1249

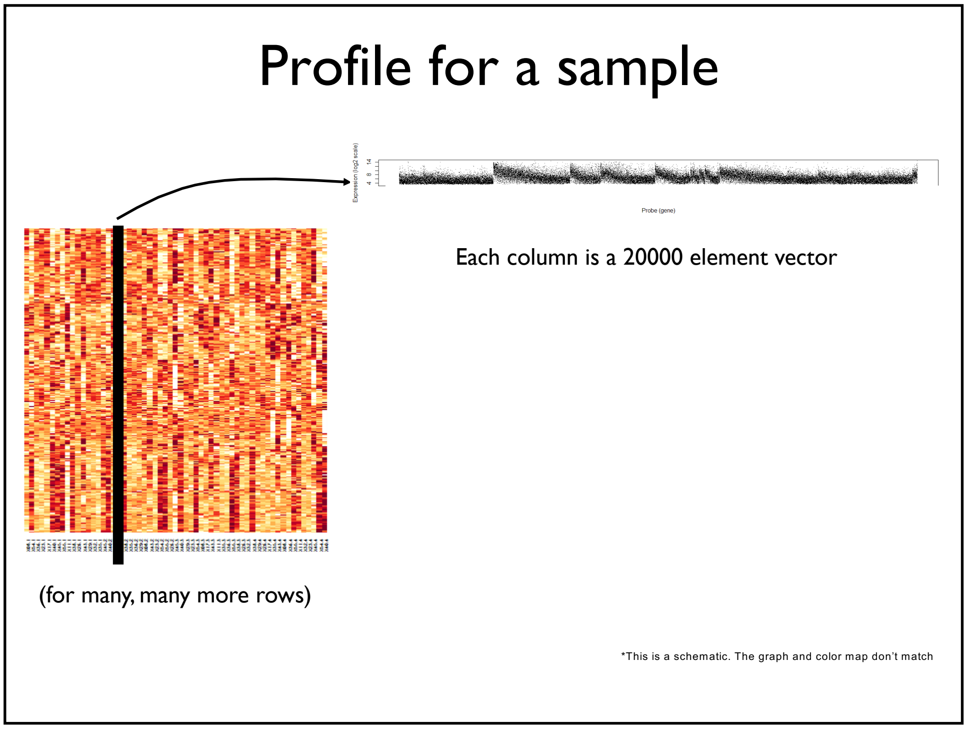


Lighter colours mean higher levels of gene expression (“activity”)
Only show part of the data!

Profile for a gene



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

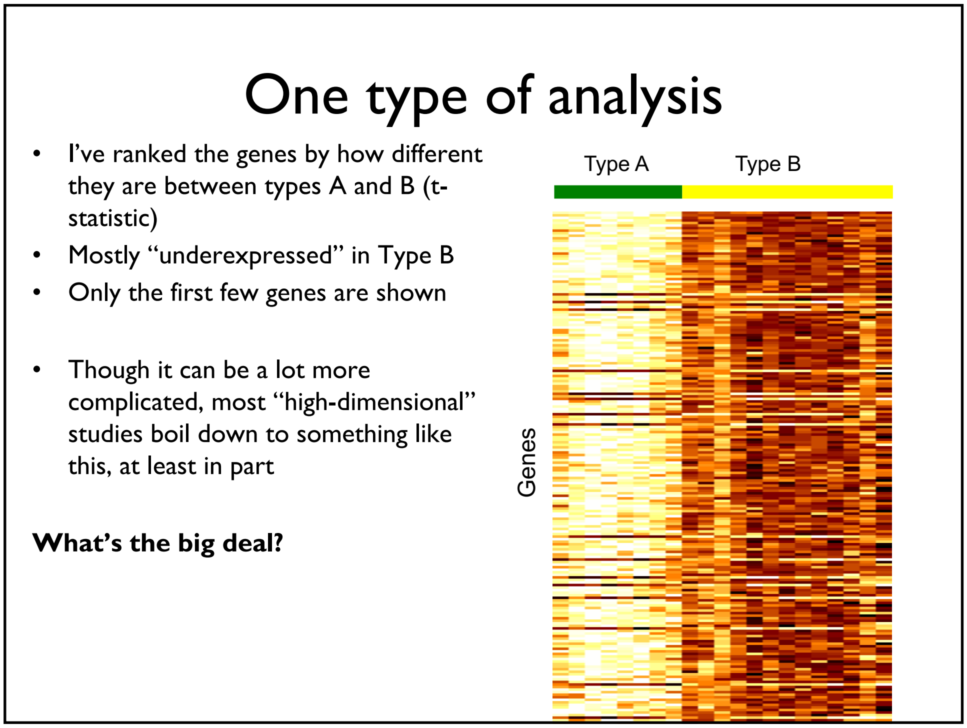


Profile for a sample

Each column is a 20000 element vector

(for many, many more rows)

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

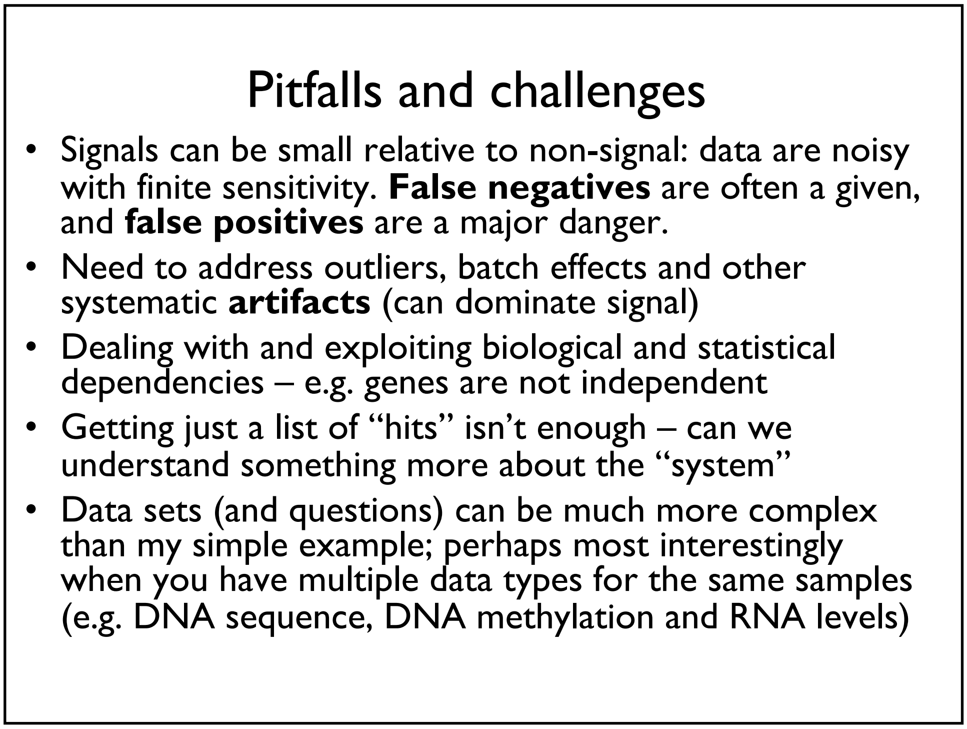


One type of analysis

- I've ranked the genes by how different they are between types A and B (t-statistic)
- Mostly "underexpressed" in Type B
- Only the first few genes are shown

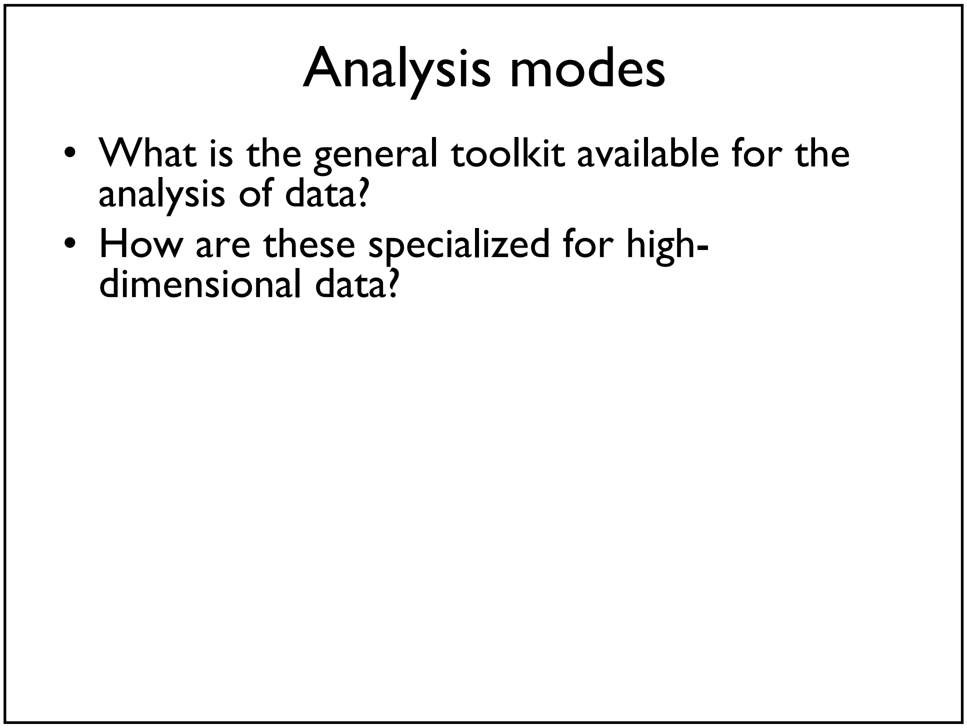
- Though it can be a lot more complicated, most "high-dimensional" studies boil down to something like this, at least in part

What's the big deal?



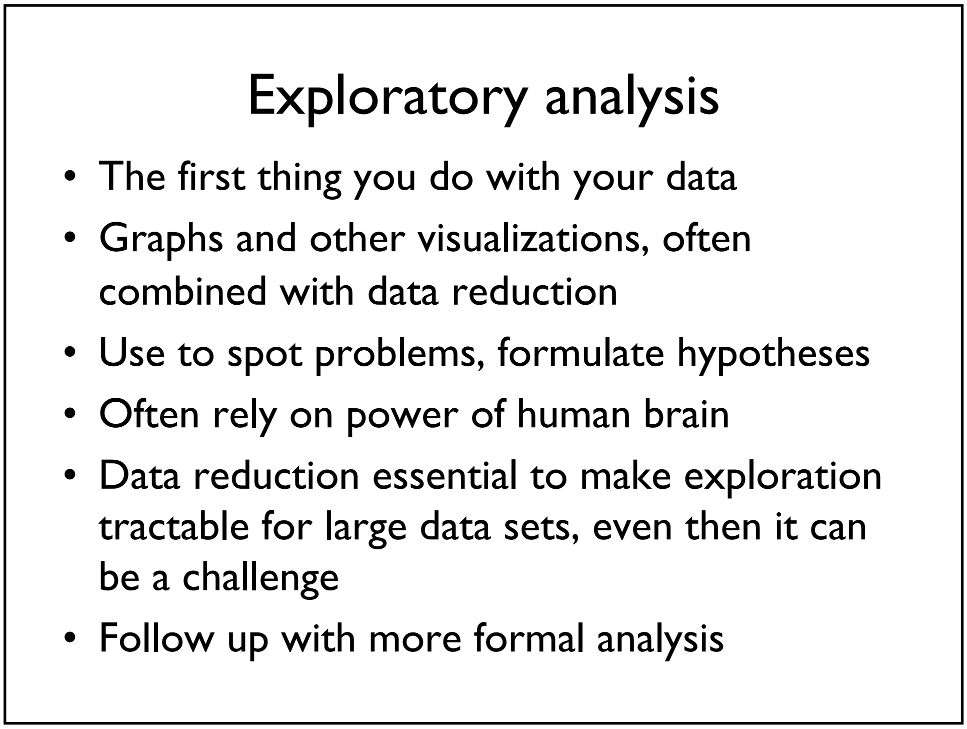
Pitfalls and challenges

- Signals can be small relative to non-signal: data are noisy with finite sensitivity. **False negatives** are often a given, and **false positives** are a major danger.
- Need to address outliers, batch effects and other systematic **artifacts** (can dominate signal)
- Dealing with and exploiting biological and statistical dependencies – e.g. genes are not independent
- Getting just a list of “hits” isn’t enough – can we understand something more about the “system”
- Data sets (and questions) can be much more complex than my simple example; perhaps most interestingly when you have multiple data types for the same samples (e.g. DNA sequence, DNA methylation and RNA levels)



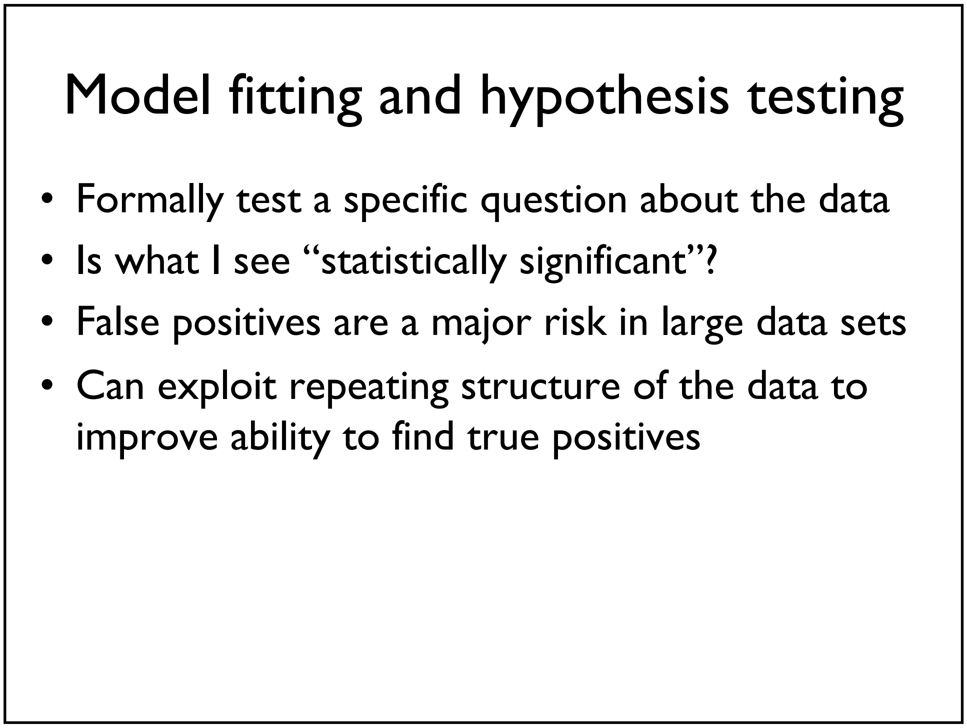
Analysis modes

- 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

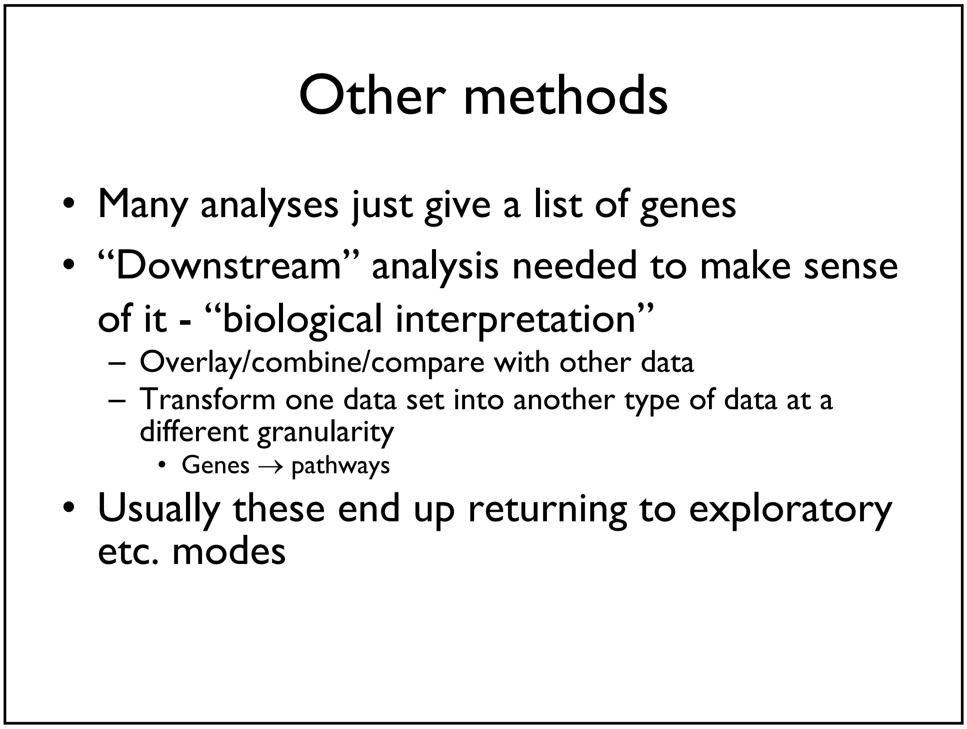
- Formally test a specific question about the data
- Is what I see "statistically significant"?
- False positives are a major risk in large data sets
- Can exploit repeating structure of the data to improve ability to find true positives

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

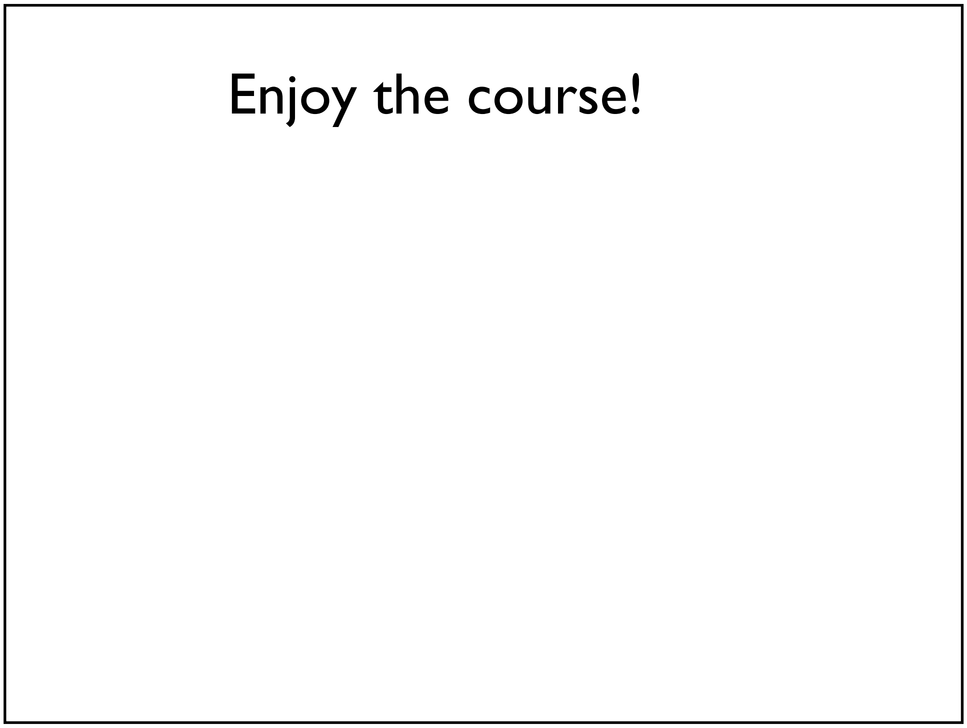
Supervised learning

- Can I predict an unmeasured feature of a sample from a measured one?
- Less common than unsupervised learning, most 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



Enjoy the course!