

# **Statistical Methods for High Dimensional Biology**

**STAT/BIOF/GSAT 540**

Lecture I – course introduction

Paul Pavlidis

January 4 2016

# Today's topics

- What the course is about
- Course mechanics
- A primer on molecular biology (TAs will provide more during lab)
- Introduction to high-dimensional biology

# Your instructors

- **Dr. Sara Mostafavi** – Assistant Professor of Statistics / Medical Genetics – [saram@stat.ubc.ca](mailto:saram@stat.ubc.ca) - Course leader
  - Bring administration questions to her
- **Dr. Paul Pavlidis** – Professor of Psychiatry/Michael Smith Labs
  - [paul@bioinformatics.ubc.ca](mailto:paul@bioinformatics.ubc.ca)
- **TAs: Marjan Farahbod** ([marjan.farahbod@gmail.com](mailto:marjan.farahbod@gmail.com)),  
**Santina Lin** ([hello@santina.me](mailto:hello@santina.me))

# 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
  - 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
- 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

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

# Course mechanics



# Course web site

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

- Lecture notes
- Lab notes
- Assignments

Much interaction via **Github** (discussions, submission)

e.g. 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

- <https://github.com/STAT540-UBC>

# Lectures

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

# 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)
  - <https://it.ubc.ca/services/email-voice-internet/myvpn>

# Evaluation

- **Homework**
  - One assignment worth 30 points
- **Group project**
  - Planning + project + poster session – 50 points
  - Peer evaluations – 10 points
- **Paper critiques**
  - Summarize and critique 2 papers – 10 points total

(Note changes from last year!)

# Homework assignment

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

# Paper reports

- New for this term
- Meant to develop your skills relating course material to the literature
- Two assignments will be given, 5 pts each
- Select, read, summarize and critique a recent paper from the 'omics literature – write one page max, relating to the course content.
- 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 22 (4-5 people)
- Friday Jan 29: initial project proposals
- Feedback to groups Feb 12 – Proposals finalized by Feb ~15
- Work on projects over rest of term
- Final session of the course is the poster session

*Dates are provisional*



# 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

# 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 functional stretch of DNA
  - 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)

- ~3 billion bases of DNA over 23 pairs of chromosomes
- 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

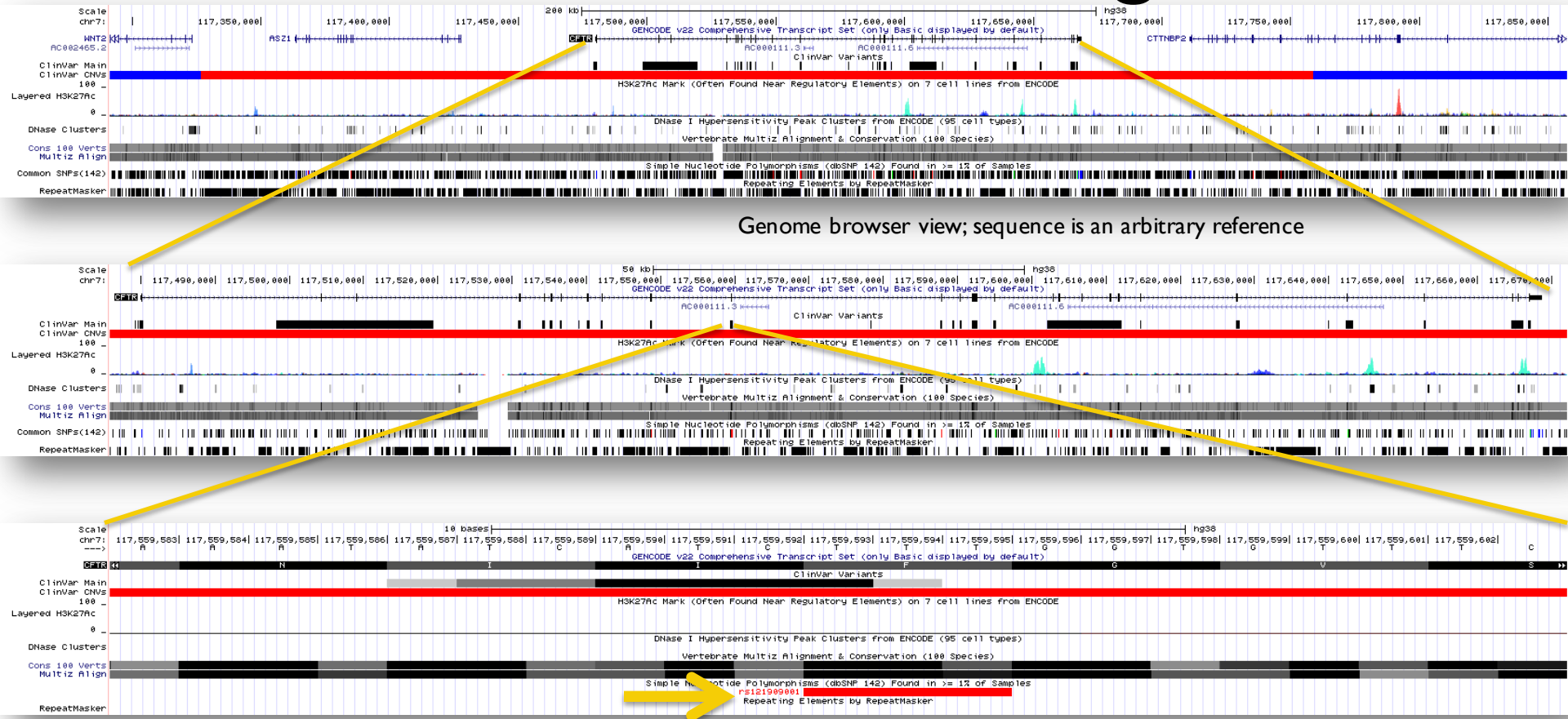
# Molecular regulation

- 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

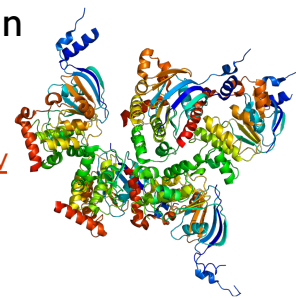
- Differences in DNA sequence between two individuals of the same species (different genotypes) — humans: millions of differences
- Most are inherited from mom and dad, but also acquired during your life.
  - Depending on when acquired, will affect all or part of your body.
- Most variation has no effect 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. Fragile X
- 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



- 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&hgid=466552281\\_9bE3VrJY4rVhIAbe4NUU7CvQSmxV](http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&position=chr7%3A117475813-117672093&hgid=466552281_9bE3VrJY4rVhIAbe4NUU7CvQSmxV)
- <http://www.cfmedicine.com/html/docs/cfext/genetics.htm>
- [https://en.wikipedia.org/wiki/Cystic\\_fibrosis\\_transmembrane\\_conductance\\_regulator](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) –
  - often 10s, but can be 1000s (some genetics studies)

# 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.
  - This example is only partly realistic: DNA sequencing lets us look at tumor genomes directly. Looking at transcripts (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

## Assays

One row per gene (20000)

■ ■ ■

virtaneva-data.txt		C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Proba.	din-134702	am8-am9	6-am8-am9	5-am8-am9	4-am8-am9	1-am8-am9	5-am8-am9	6-am8-am9	4-am8-am9	am8-am9	am8-am9	am8-am9	am8-am9	am8-am9	am8-am9	am8-am9	am8-am9
176	AJ000099	-383	591	5122	4232	4368	5591	4136	6125	4262	3396	5390	1990	6009	3110	4753	3296	4708
177	AJ000490	1574	5464	4346	9274	2982	4194	2014	6095	1328	2207	1805	4269	2637	2526	7198	2279	3452
178	AJ001407	-263	-1211	-311	-1717	-127	-306	-1	-54	-200	-763	-558	-62	-1590	-610	-385	-866	-210
179	AJ001421	7642	5196	3982	3515	4090	4652	2204	6793	3362	642	6702	5282	1849	6332	6444	7305	4820
180	AJ001487	-39	0	-434	-637	-101	-287	-169	-208	22	-86	-584	-1146	-1236	-1200	-680	-243	-1472
181	000003.txt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
182	000003.txt	-320	-546	-47	-177	-309	-713	-763	-316	-549	-959	-44	-1	-1610	-269	-97	-97	-97
183	000013.txt	3179	11537	17141	26201	17466	3119	4418	15828	983	1313	13146	37585	12531	27151	1167	28136	11473
184	000097_s	95	-411	-236	-275	32	-210	-370	262	-73	-257	-206	13	26	11	-255	-250	-73
185	000408.txt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
186	000408_s	108	-72	526	671	373	540	617	257	297	63	355	424	1146	451	704	350	662
187	2106	565	779	384	2208	1314	1487	2983	756	1625	2977	303	747	946	1200	321	1585	873
188	000596_s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
189	000632_s	709	808	371	367	1041	1256	1288	1259	723	1201	1233	40	-228	773	659	623	521
190	000654_s	-168	-621	161	36	66	-101	-436	-220	-25	-42	-33	-121	-192	76	31	-265	-670
191	000723_s	862	572	543	26	504	1043	382	1079	584	567	1243	681	843	643	660	377	195
192	000726_s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
193	000749_s	2978	2873	2013	1980	1330	5146	11781	2598	1615	24959	6309	6709	8936	2571	2378	2210	3267
194	000760_s	7369	9346	9369	4335	164	4203	6350	2870	7651	8453	2676	374	4363	1641	2222	2818	1839
195	000761.txt	12481	9144	12256	556	19285	16698	8142	20264	22157	10376	11366	11584	204	8787	6794	7688	7378
196	000762_s	4719	10232	3060	1554	3680	5398	2211	5340	10707	1917	1580	2055	213	1813	1955	2304	2205
197	000763.txt	6414	11023	10492	3424	4005	11263	3633	12294	8054	5917	1420	12509	4428	8605	9522	10917	7445
198	010511_s	1913	504	231	727	2039	1553	738	4225	860	873	420	1753	1371	1316	736	1361	340
199	010404	1081	2202	759	3907	2024	388	678	889	2123	219	79	1377	423	2164	1375	1737	1725
200	010202	1115	-1238	-574	113	-293	-750	-193	-603	-411	-490	86	-231	-1021	-840	-81	1993	128
201	010326_s	-2313	-15794	-6406	-12098	-9594	-7946	-14068	-6843	-6224	-9048	-5769	-9548	-8189	-6383	-6549	-6108	-9970
202	010495_s	3708	-11135	2953	5339	3247	4748	3827	2611	6933	1951	5470	6583	4167	5317	5242	6527	5075
203	010511_s	1913	504	231	727	2039	1553	738	4225	860	873	420	1753	1371	1316	736	1361	340
204	010522_s	8955	314	257	6854	157	4216	2039	340	558	93	1143	3333	359	6891	600	4075	306
205	010523_s	1904	792	1910	-16	903	1298	2050	1833	835	573	2723	1348	1707	1610	376	1475	1231
206	010537_s	1399	1680	1547	2194	1228	751	1457	1042	2337	1069	78	2822	1086	2009	913	870	1047
207	010606_s	111	10	531	141	3737	536	862	432	372	600	351	521	-702	344	441	12	96
208	010607_s	-31	-442	-274	-417	-280	-261	-167	-285	-167	240	-245	-88	-634	-34	-176	-248	4
209	010704_s	1161	870	1178	738	1460	2152	1673	1719	568	3448	1311	1204	-833	651	2056	923	-978
210	010922_s	-87	-104	-100	8	37	-148	20	-24	-48	-230	-32	130	-41	-95	41	901	229
211	010923_s	-51	769	-269	6468	332	1206	109	-8	-145	-73	98	1022	2726	1814	97	791	636
212	010925_s	-504	-1264	33	2513	-102	-445	-433	-306	-307	-302	-173	470	-1092	5681	-113	5999	100
213	010995_s	199	2444	890	1299	1438	1626	1102	1362	1020	950	871	1183	1211	1134	398	958	1289
214	011006_s	25192	2591	9398	6390	2741	11963	22295	2541	5178	23904	846	12603	8414	5143	2093	1399	8379
215	011094_s	5031	1225	2212	1014	3319	3599	1437	354	3693	2330	410	1138	2683	1145	2073	2372	3965
216	011139_s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
217	011151_s	24	-5	202	601	89	233	438	-38	110	408	50	504	448	-105	565	241	315
218	011327_s	2217	140	1698	972	271	2631	1299	1573	478	1697	1035	958	782	1446	2536	1004	3221
219	011426_s	-657	29	93	-1126	202	1155	-770	1302	813	-844	815	1324	1064	50	619	-585	226
220	021465_s	83	199	106	400	-44	7	-165	33	16	-74	86	80	-174	-162	16	38	-58
221	021620_s	-12	373	494	-333	-98	41	697	-1	17	327	16	31	167	184	-241	-94	51
222	021625_s	7	119	43	-61	44	19	-84	35	9	-56	47	-77	-351	117	-99	58	63
223	021676_s	-219	87	-24	19	-29	-116	164	330	-21	-185	-54	143	-1653	319	-261	12	45
224	021698_s	1095	-8219	-2089	-8747	-6129	-6574	-6041	-2805	-6568	-4988	-639	-4176	-8398	-3074	-2495	-1823	-1514
225	021720_s	-398	958	174	79	-19	-119	-328	-51	-119	-328	-51	-119	-328	-51	-119	-328	-51
226	021775_s	907	876	525	559	249	497	278	459	427	398	436	305	482	738	900	652	1187
...																		
241	U11131_s	24	-5	202	601	89	233	438	-38	110	408	50	504	448	-105	565	241	315
242	U11327_s	2217	140	1698	972	271	2631	1299	1573	478	1697	1035	958	782	1446	2536	1004	3221
243	U11426_s	-657	29	93	-1126	202	1155	-770	1302	813	-844	815	1324	1064	50	619	-585	226
244	U21465_s	83	199	106	400	-44	7	-165	33	16	-74	86	80	-174	-162	16	38	-58
245	U21620_s	-12	373	494	-333	-98	41	697	-1	17	327	16	31	167	184	-241	-94	51
246	U21625_s	7	119	43	-61	44	19	-84	35	9	-56	47	-77	-351	117	-99	58	63
247	U21676_s	-219	87	-24	19	-29	-116	164	330	-21	-185	-54	143	-1653	319	-261	12	45
248	U21698_s	1095	-8219	-2089	-8747	-6129	-6574	-6041	-2805	-6568	-4988	-639	-4176	-8398	-3074	-2495	-1823	-1514
249	U21720_s	-398	958	174	79	-19	-119	-328	-51	-119	-328	-51	-119	-328	-51	-119	-328	-51
250	U21775_s	907	876	525	559	249	497	278	459	427	398	436	305	482	738	900	652	1187

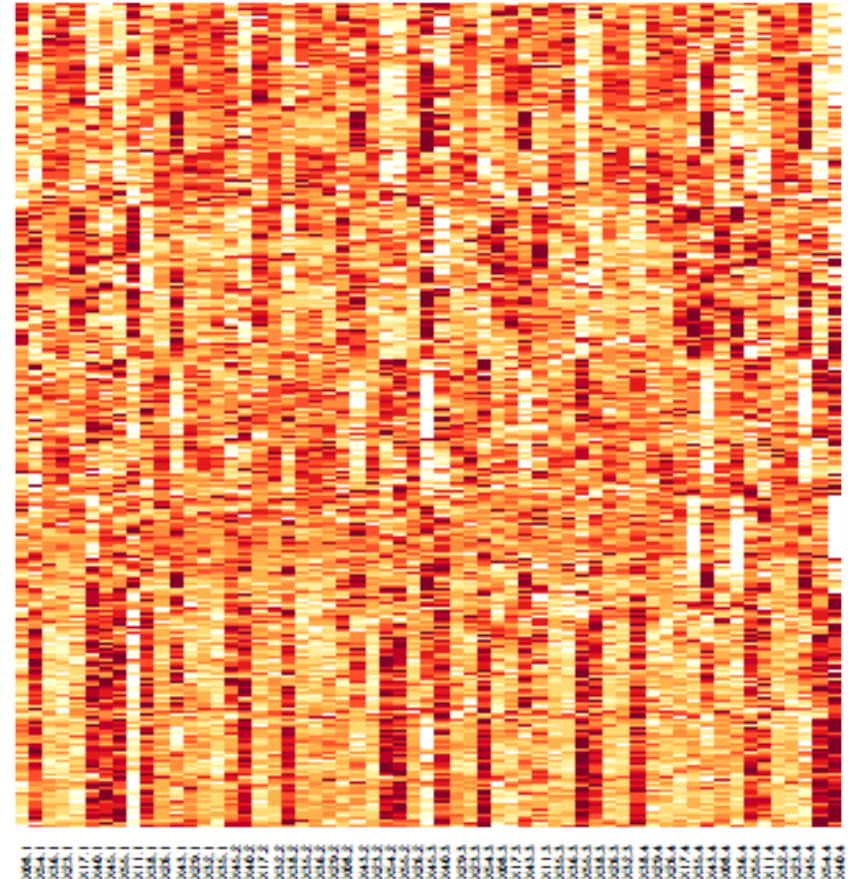
# A partial list of things to assay

- 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)



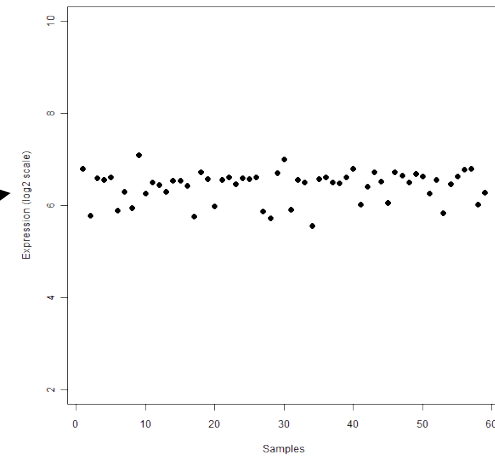
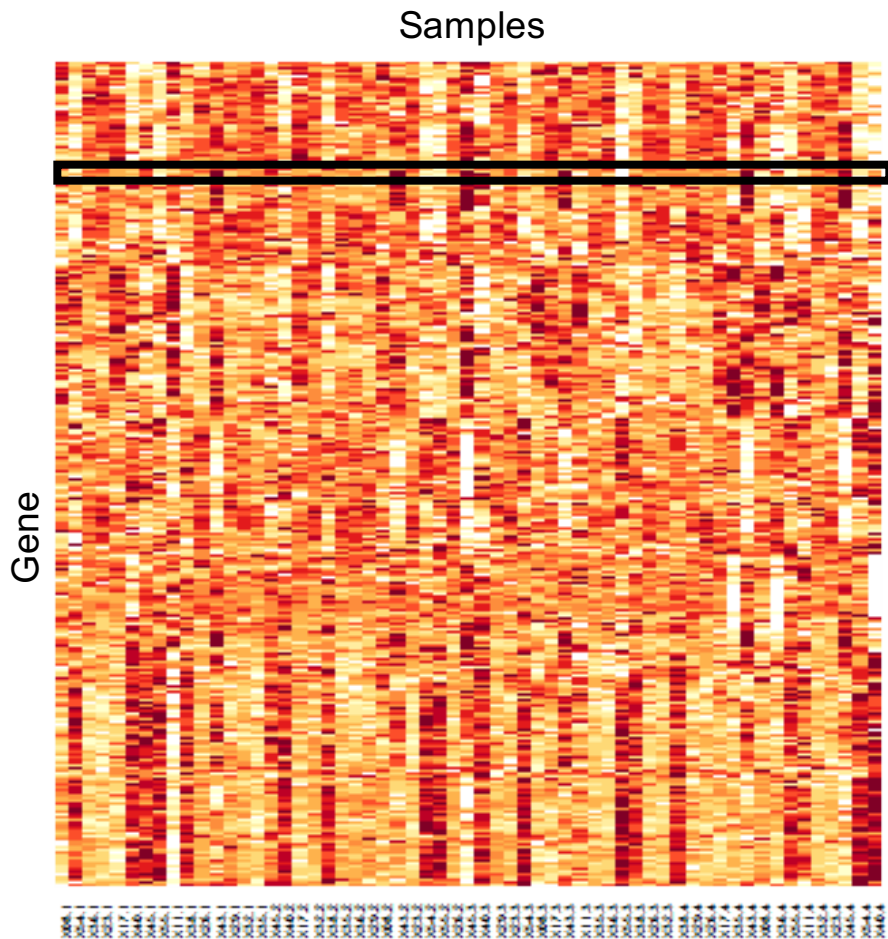
# Alternative representation

vitaceva_data.txt		B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Probe	dm-134702	anti-amiB_5	anti-amiB_6	anti-amiB_7	anti-amiB_8	anti-amiB_9	anti-amiB_10	anti-amiB_11	anti-amiB_12	anti-amiB_13	anti-amiB_14	anti-amiB_15	anti-amiB_16	anti-amiB_17	anti-amiB_18	anti-amiB_19	anti-amiB_20	anti-amiB_21
176	A000099_i	2983	591	5122	4262	4388	5591	4136	6125	4262	3386	5390	1990	6009	3110	4708	3296	4708	5427
177	A000490_i	1574	5484	4346	9274	2982	4194	2814	6095	1328	2207	1805	4269	2637	2526	2188	2279	3452	5068
178	A001047_i	-263	1211	-311	-1717	-127	-306	-1	54	-200	-763	-558	-42	-1580	-610	-385	-896	-210	-658
179	A001421_i	7642	5196	3962	3515	4090	4652	2204	6793	3362	642	6702	5262	1849	6332	6444	7305	4620	7378
180	A001487_i	-39	-43	-344	-637	-101	-287	-169	-208	22	-86	-584	-1146	-1236	-1200	-680	-243	-1472	-443
181	D00003_i	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_i	3179	11537	17141	26201	17466	3119	4416	15628	983	-133	13146	37585	12531	27152	1167	28136	11473	-338
184	D00097_s	85	-411	-236	-275	32	-210	-370	262	-73	-257	-206	13	26	161	-255	-250	-73	-79
185	D00408_i	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	633
187	D00591_i	2186	565	779	384	2206	1314	1587	2953	736	1625	2977	303	747	946	1200	321	1526	777
188	D00596_i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
189	D00632_i	709	808	371	367	1041	1256	1286	1259	723	1201	1233	40	-228	773	669	623	521	351
190	D00654_i	-188	-621	161	36	66	-101	-436	-220	-25	-42	-33	-121	-192	76	31	-265	-670	-46
191	D00703_i	692	572	543	26	504	1043	382	1078	584	587	1243	681	643	443	660	377	195	914
192	D00726_i	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	5164	11781	2598	1615	24558	6390	6708	8996	2571	2378	2210	3267	3698
194	D00760_i	7389	4396	4335	164	4203	6350	2070	7651	3453	2676	3774	4363	1441	2222	2818	1639	1686	4946
195	D00781_i	12481	9144	12256	556	19285	16696	8142	20264	22157	10376	11366	11584	7204	6767	6794	7668	7378	11694
196	D00762_i	4719	2842	3360	1554	3690	5396	2211	5340	10773	1917	1590	2055	213	1813	1955	2304	2205	5007
197	D00763_i	6141	11023	10492	3424	4056	11263	3633	12294	8054	5916	10472	12509	4428	8806	9622	10917	7445	12524
198	D00880_i	997	2086	706	507	347	11118	362	611	1463	1068	1149	446	1008	410	768	340	975	1354
199	D10340_i	1081	2202	759	3907	3024	388	678	889	2123	219	79	1377	423	2164	1375	7737	1725	504
200	D10322_i	115	1236	574	113	-293	-750	-153	-603	-141	-450	86	-231	-1021	940	-81	1993	126	1124
201	D10326_s	-2313	-15794	-6406	-12088	-9594	-7946	-10468	-6843	-6224	-9046	-5769	-9548	-8189	-6363	-6549	-6108	-6970	-2789
202	D10435_i	3708	-1135	2953	5339	3247	4748	3827	2617	693	1951	5470	6693	4167	5317	5242	6527	5075	8666
203	D10511_i	1813	504	2913	727	2038	1553	738	4225	969	673	4270	1176	1371	1318	736	1361	902	1063
204	D10522_i	6895	314	297	6954	157	4216	2039	340	558	93	1143	3533	399	6991	600	4075	306	791
205	D10523_i	1904	792	1910	-16	903	1298	2050	1833	635	673	2723	1348	1707	1610	376	1475	1231	2298
206	D10537_s	1399	1680	1547	2194	1229	751	1497	1042	2337	1069	76	2622	1086	2009	913	870	1047	1226
207	D10656_i	111	10	531	141	3737	536	662	432	372	600	351	521	962	344	441	12	86	968
208	D10687_s	-31	-442	-274	-417	-290	-261	-167	-285	-167	-240	-245	-88	-634	-34	-176	-248	4	-212
209	D10704_i	161	870	1176	-738	1480	2152	1673	1719	566	3448	1311	1204	-633	651	2056	923	-978	-410
210	D10802_s	-97	-104	-100	8	37	-148	20	24	-46	-230	-32	130	-41	-86	41	901	226	-52
211	D10823_i	-51	769	-269	6468	332	1206	109	-8	-145	-73	98	1022	2726	1814	97	791	636	1131
212	D10925_i	-504	-1264	33	2513	-102	-446	-433	-306	-307	-302	-173	470	-1062	5681	-113	9899	100	-80
213	D10995_i	789	2444	980	1299	1438	1626	1102	1362	1202	650	871	1183	1211	1134	398	958	1289	2146
214	D11086_i	25192	2591	5936	6380	2741	11963	32295	2541	5178	23904	946	12603	8414	5143	2090	1389	8379	10213
215	D11094_i	5031	1226	2212	1014	3319	3589	1437	354	3693	2330	410	1136	2683	1145	2073	2372	3965	3303
216	D11139_i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
217	D11151_i	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	1299	1573	478	1697	1035	959	782	1446	2536	1004	3221	4812
219	D11426_i	-657	29	93	-1126	202	1155	-770	1302	613	-844	815	-1325	1054	50	619	-595	226	-1024
220	D12495_i	63	199	106	400	-44	7	-165	33	16	-74	66	80	-174	-162	-16	38	58	224
221	D12630_s	-12	373	494	-333	-98	41	697	-1	17	327	16	31	167	184	-241	-94	51	-31
222	D12632_s	7	119	42	-61	44	19	-84	36	19	-54	36	117	86	581	33	66	83	-61
223	D12676_i	-219	87	-24	19	-29	-116	164	330	-21	-185	-54	-143	-1653	319	-261	12	-45	163
224	D12696_i	1095	6218	-2089	-8747	-6129	-6574	-6041	-2805	-6568	-4988	-639	-417	-8071	-3074	-2495	-1623	-1514	-663
225	D12763_i	-225	915	-174	-999	26	-139	-465	-208	-51	-139	-465	-208	-51	-139	-465	-208	-51	-139
226	D12775_s	907	876	525	559	249	497	278	459	427	390	496	305	482	738	900	652	1197	1249



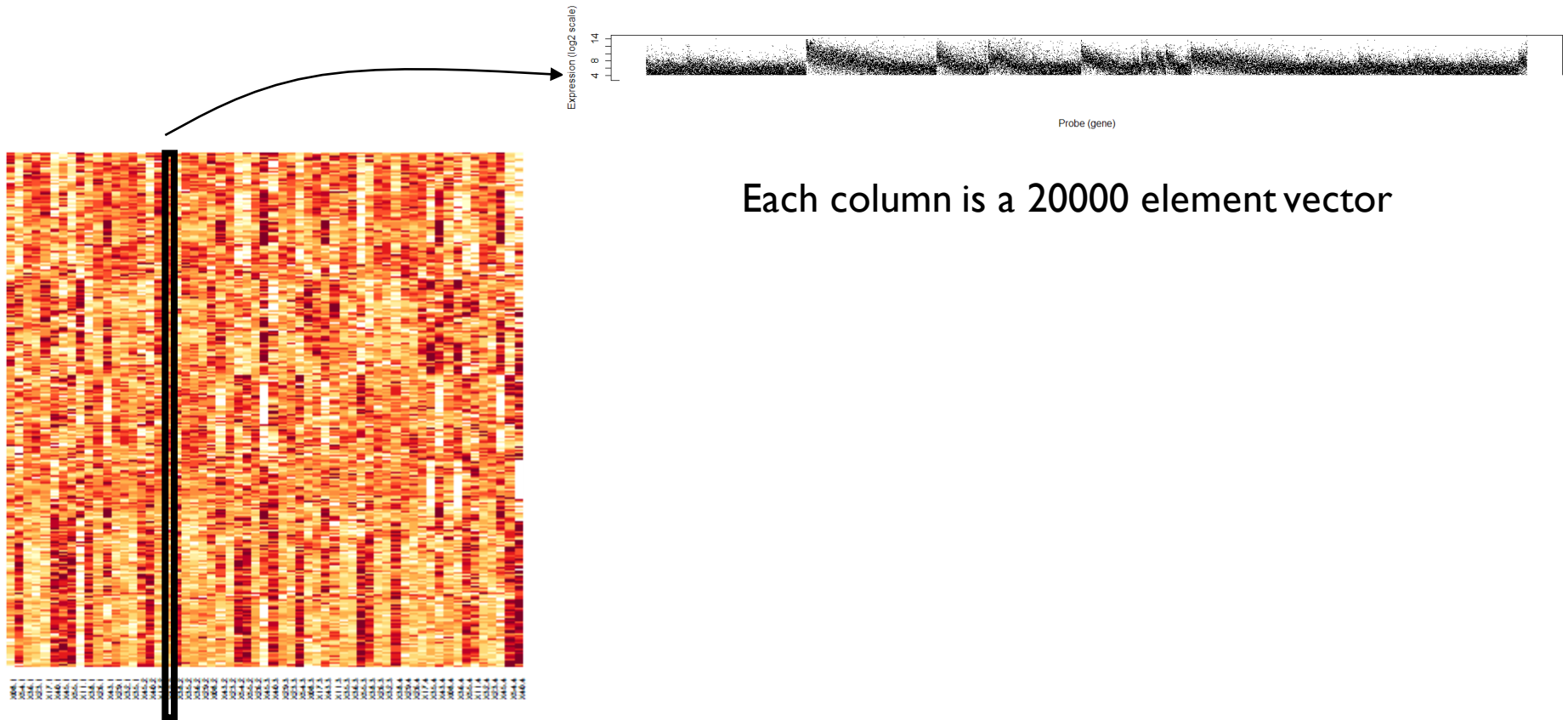
Lighter colours mean higher levels of gene expression (“activity”)  
Only show part of the data!

# 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

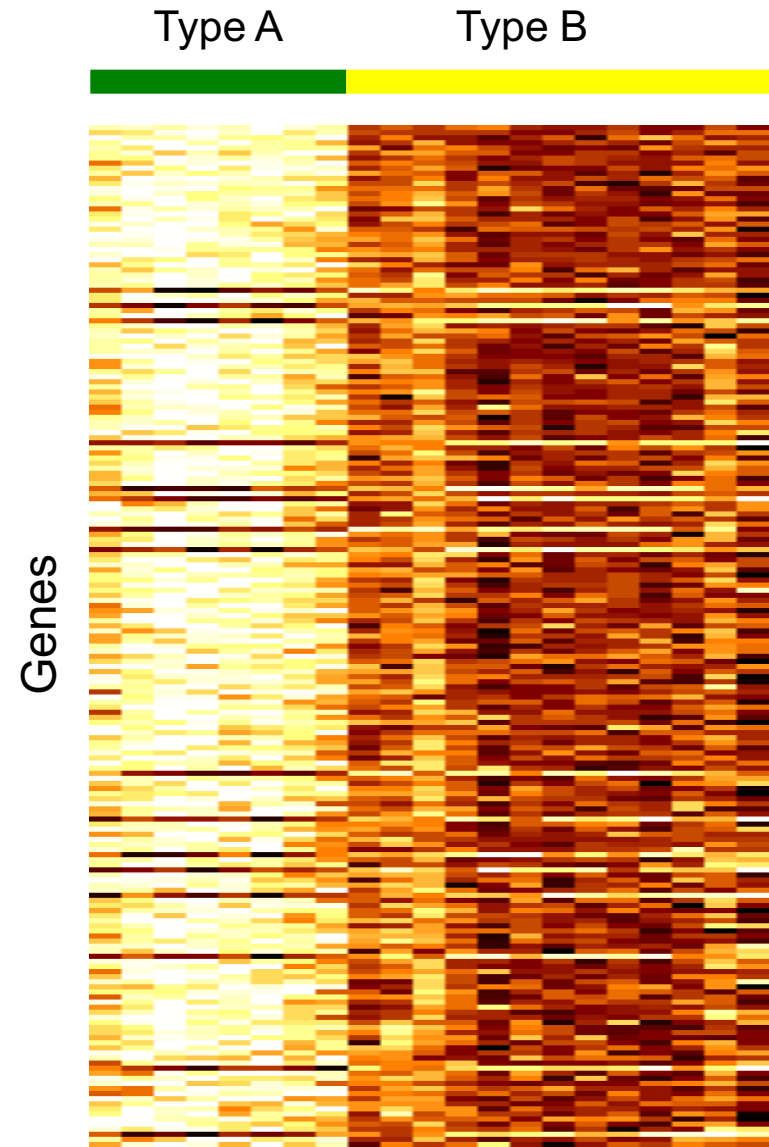
(for many, many more rows)

\*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

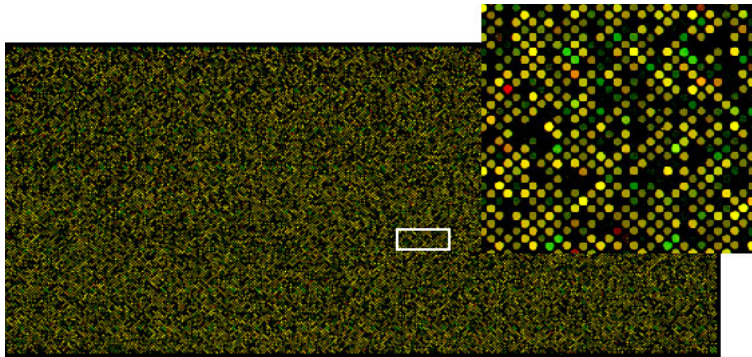
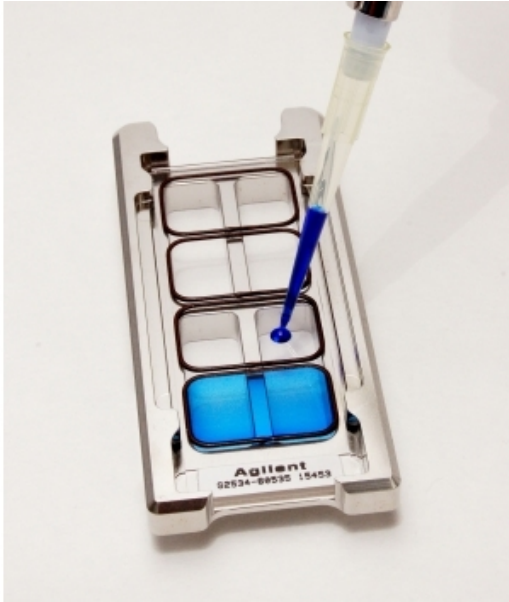
- Signals can be small relative to non-signal: data are noisy with finite sensitivity. **False negatives** are often a given, and **false positives** are a major danger.
- Need to address outliers, batch effects and other confounds
- Dealing with and exploiting biological and statistical dependencies – 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)

# High-dimensional technologies

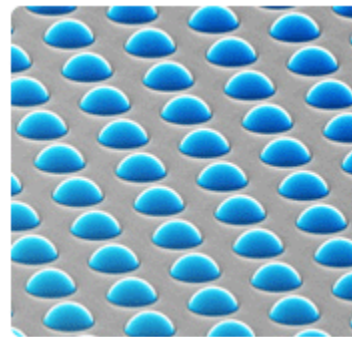
- DNA & RNA sequencing
  - Transcriptomes, exomes, full genomes
- Complex gene library construction
  - Expression vectors, protein tags, knockdowns
- Microarrays and other robotic/parallel tech.
  - Screens, high-content assays ...
- Mass spectroscopy
- Flow cytometry
- ... and many others outside of genomics (e.g. imaging)



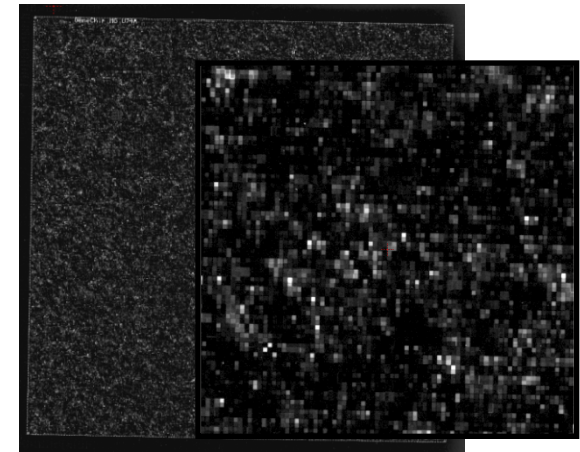
# Microarrays



Agilent SurePrint



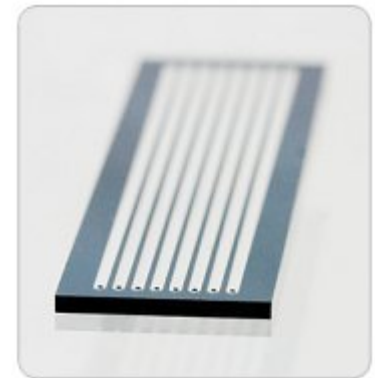
Illumina Beadarray



Affymetrix Genechip

# 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 (~mRNA molecule)



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 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!**