Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

Lecture 1 – course introduction

Paul Pavlidis

January 6 2014

Today's topics

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

Your instructors

- Dr. Jenny Bryan – Associate Professor of Statistics/MSL
  - jenny@stat.ubc.ca
- Dr. Gaby Cohen-Freue – Assistant Professor of Statistics
  - gcohen@stat.ubc.ca
- Dr. Paul Pavlidis – Associate Professor of Psychiatry/CHiBi
  - paul@chibi.ubc.ca
- TAs: Luolan (Gloria) Li (lli@bcgsc.ca) and Shaun Jackman (sjackman@gmail.com)

Office hours: please contact us by email to set up a meeting

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 (new this year)

Principal Component Analysis

Clustering

Classification

Resampling and bootstrap

Model selection and regularization

Gene sets and gene networks

Course mechanics

Course web site

http://www.ugrad.stat.ubc.ca/~stat540/

- Lecture notes
- Lab notes
- Assignments

Discussion group: Google groups

http://groups.google.com/group/stat540_2014

  - TAs will add you
  - Privacy issues: to be addressed Wednesday

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: R help
  - 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)
- Make sure you can access journals online (e.g. via the UBC VPN)
  - http://it.ubc.ca/services/email-voice-internet/myvpn/setup-documents
  - Some resources are via SpringerLink, which requires use of the UBC network for access.

Evaluation

- **Homeworks**
  - Two assignments worth 25 points each
- **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

New for 2014: we may suggest/require that incremental progress be submitted along the way

Group projects

- Starts today – start thinking about it
- A few minutes for group project pitches on Jan 20 and Jan 22.
- Form groups by Fri Jan 24 (3-4 people)
- Friday Jan 31: initial project proposals
- Feb 28: final project proposal
- 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, where the data comes from published sources.
- 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

Each value =
Measurement of one gene one sample

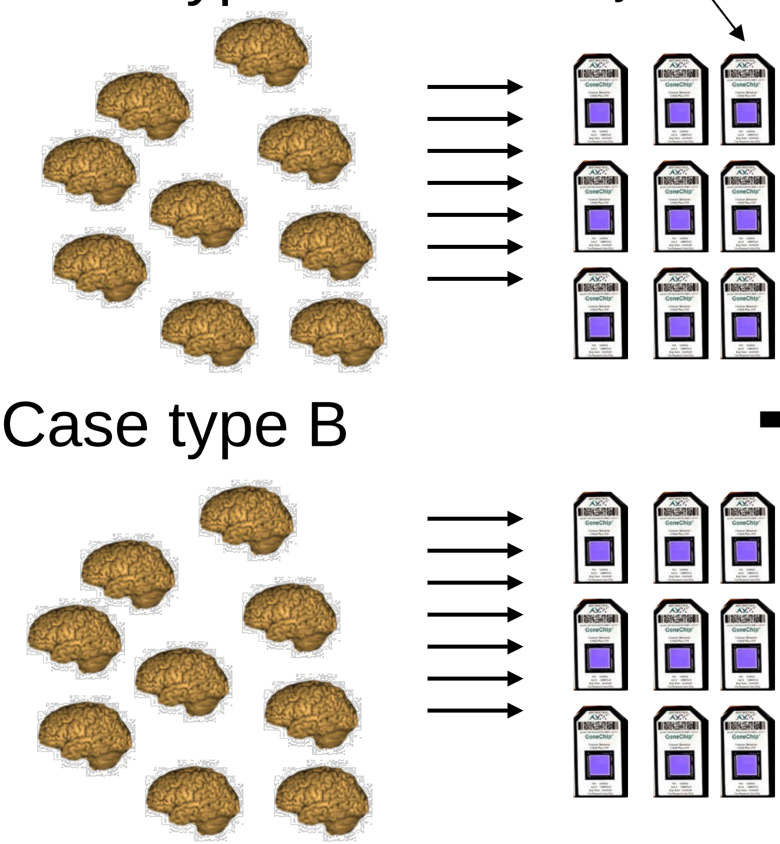
One of these gives us measurements for every gene, in a sample

Samples

Case type A

Assays

Case type B



One row per gene (20000)

One column per sample

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	5066
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

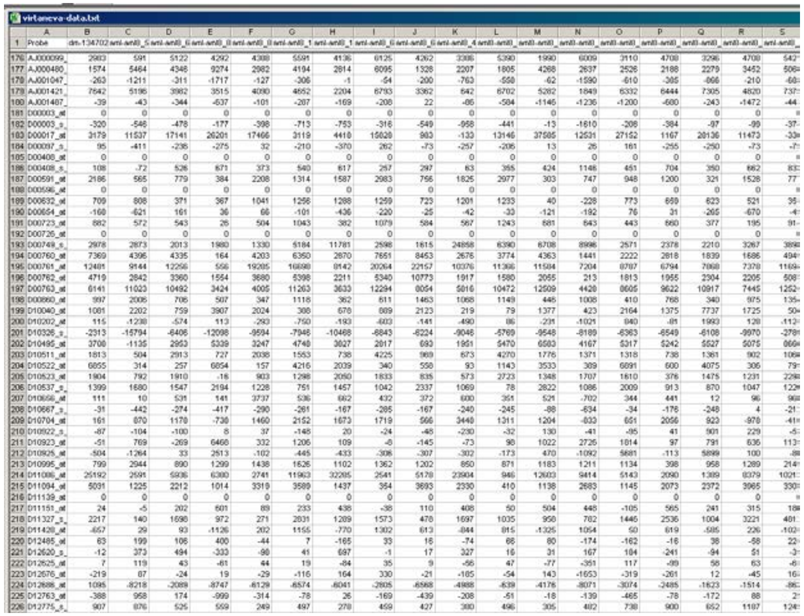
...

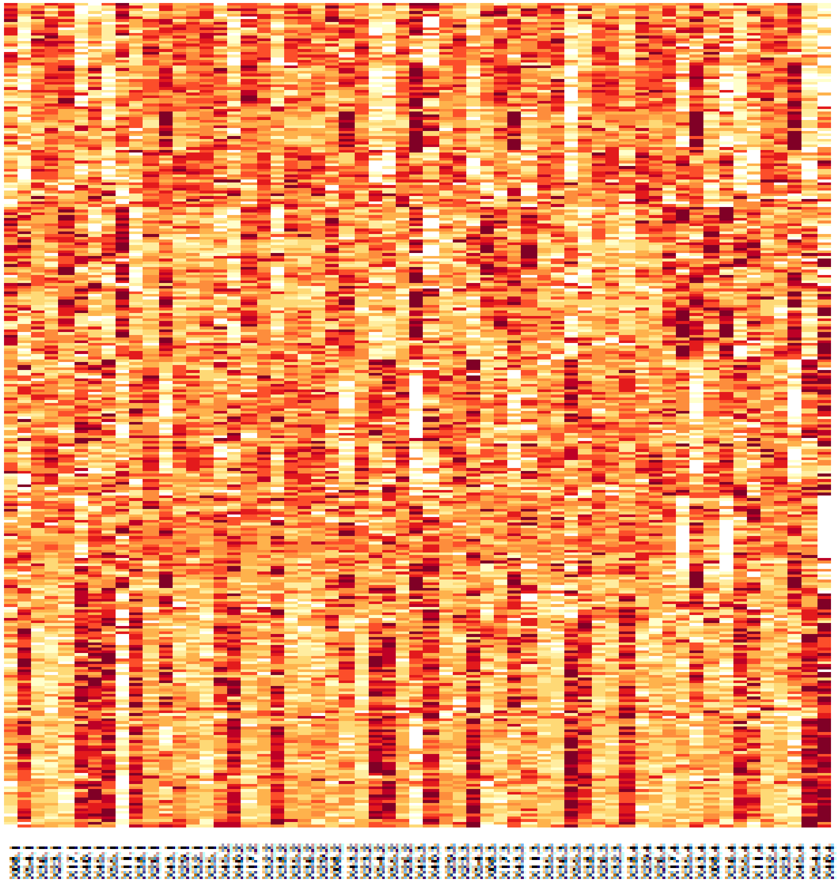
217	D11151_at	24	-5	202	601	89	233	438	-38	110	408	50	504	448	-105	565	241	315	180
218	D11327_s_	2217	140	1698	972	271	2831	1289	1573	478	1697	1035	958	782	1446	2536	1004	3221	4812
219	D11428_at	-657	29	93	-1126	202	1155	-770	1302	613	-844	815	-1325	1054	50	619	-585	226	-1024
220	D12485_at	63	199	106	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

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)

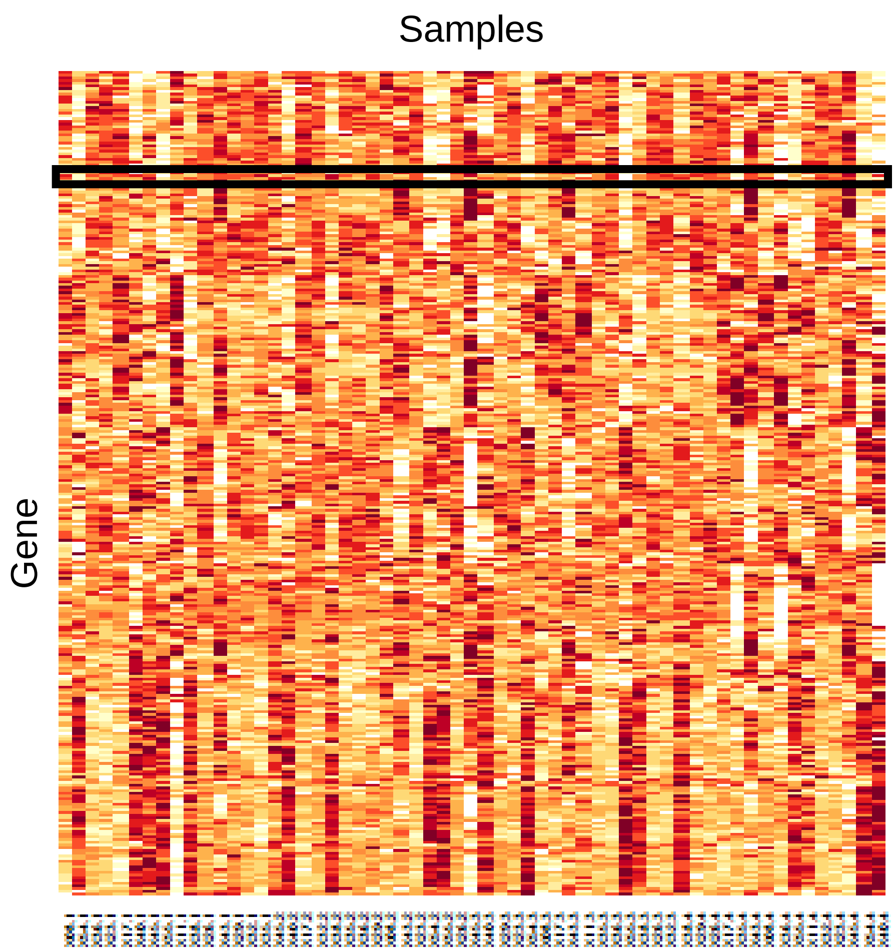
Alternative representation

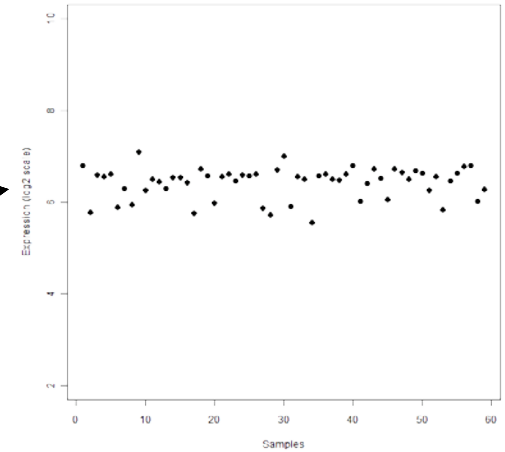




Lighter colours mean higher levels of gene expression ("activity")
Only show part of the data!

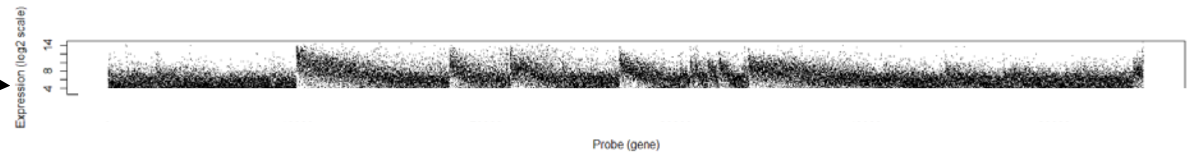
Profile for a gene

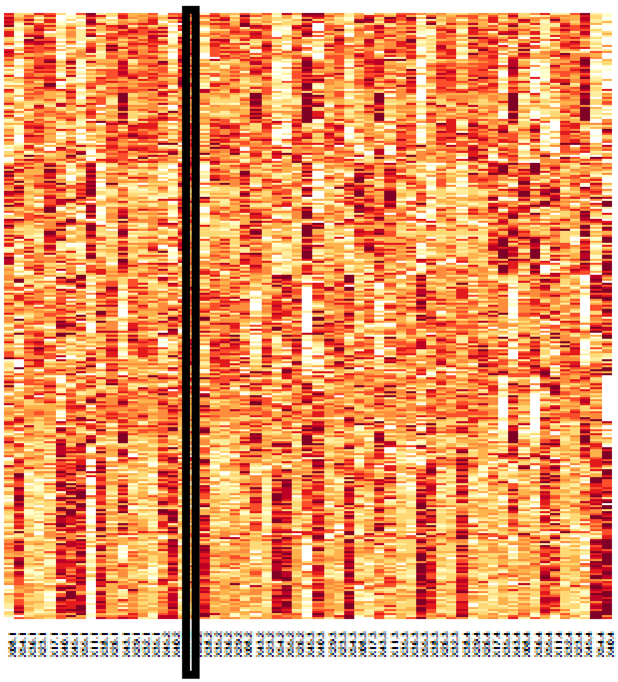




Profile for a gene. This is a 59-element vector

Profile for a sample





Each column is a 20000 element vector

(for many, many more rows)

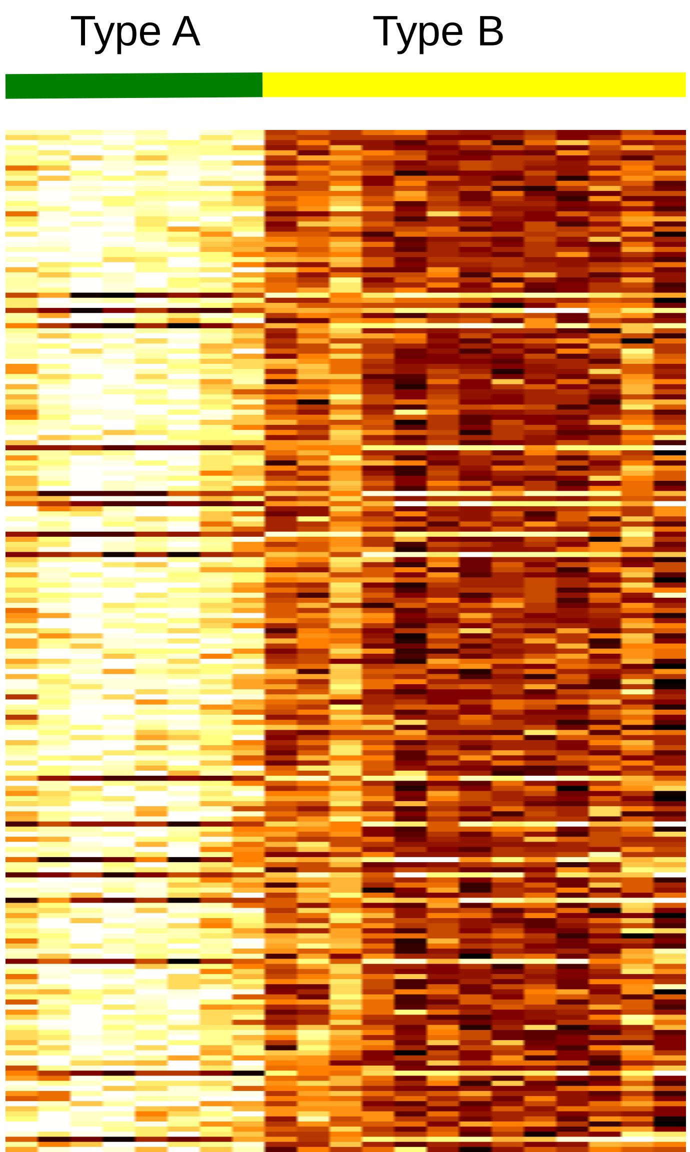
*This is a schematic. The graph and color map don't match

One type of analysis

- I've ranked the genes by how different they are between types A and B (t-statistic)
- 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?



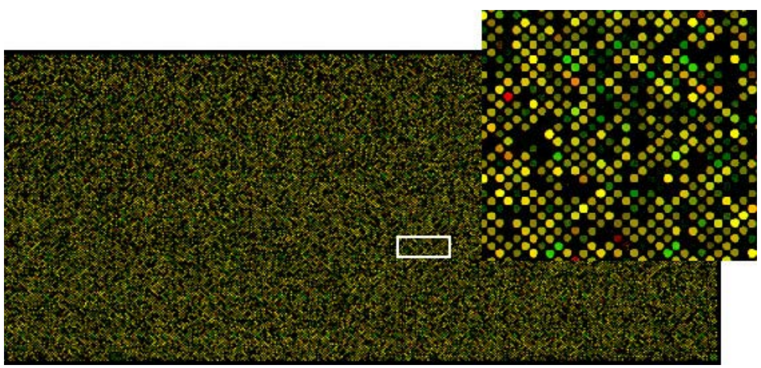
Pitfalls and challenges

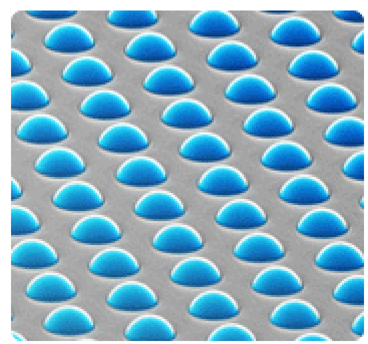
- Signals can be small and buried in lots of non-signals; False positives are a danger.
- Need to detect outliers, batch effects and other confounds
- Can we make better use of the fact that we're testing 20,000 genes than just doing a t-test on each one?
- Data sets (and questions) are often much more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

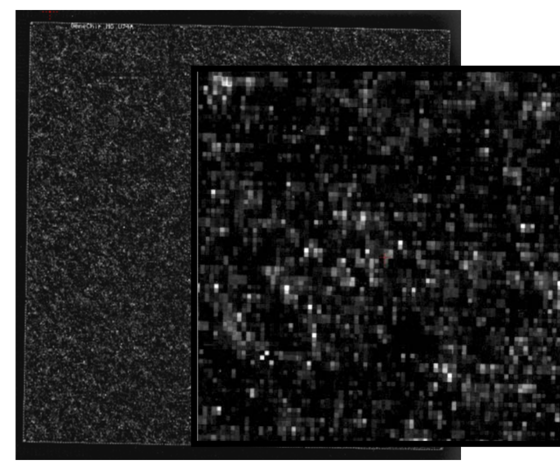
High-dimensional technologies

- 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







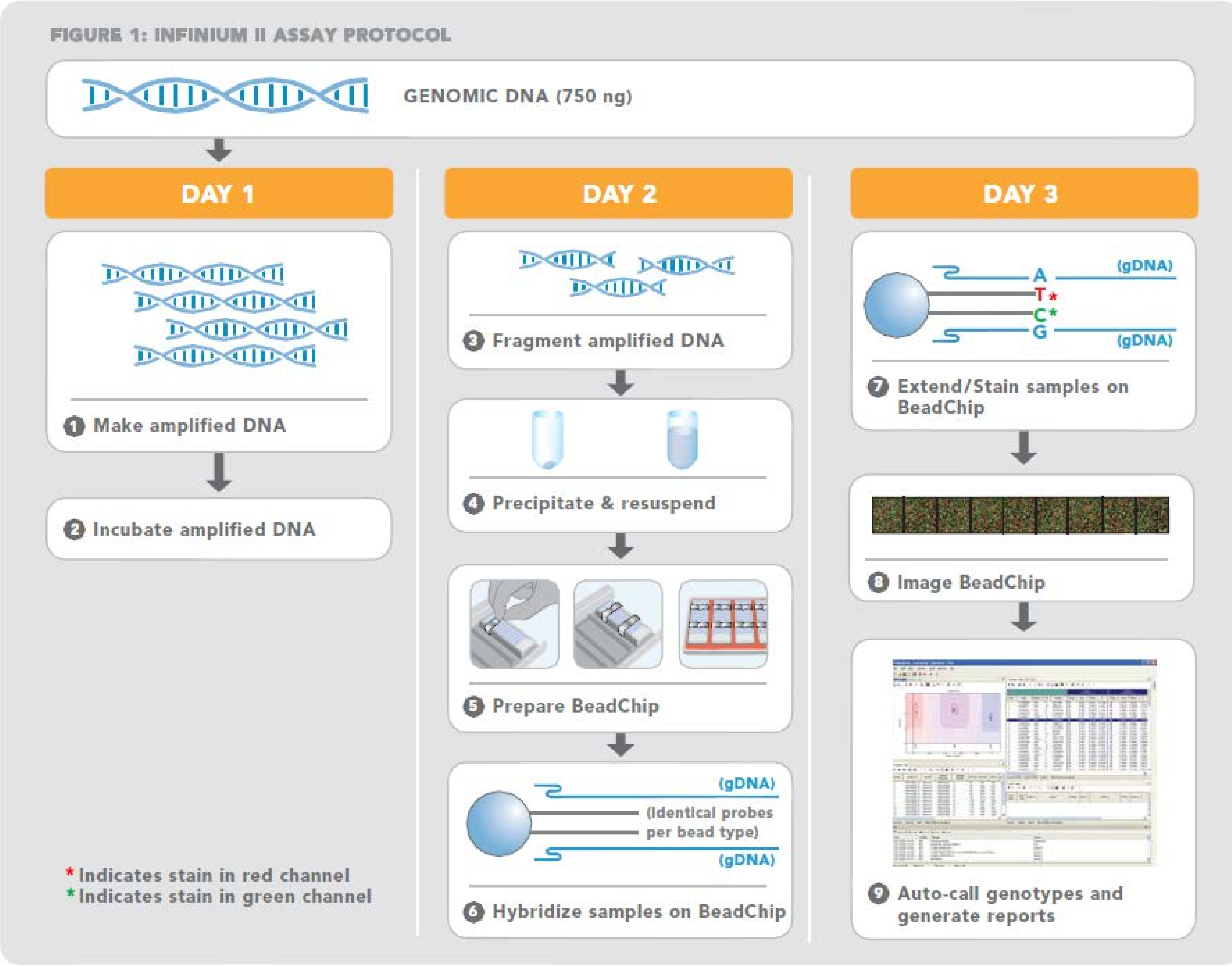
Agilent SurePrint

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

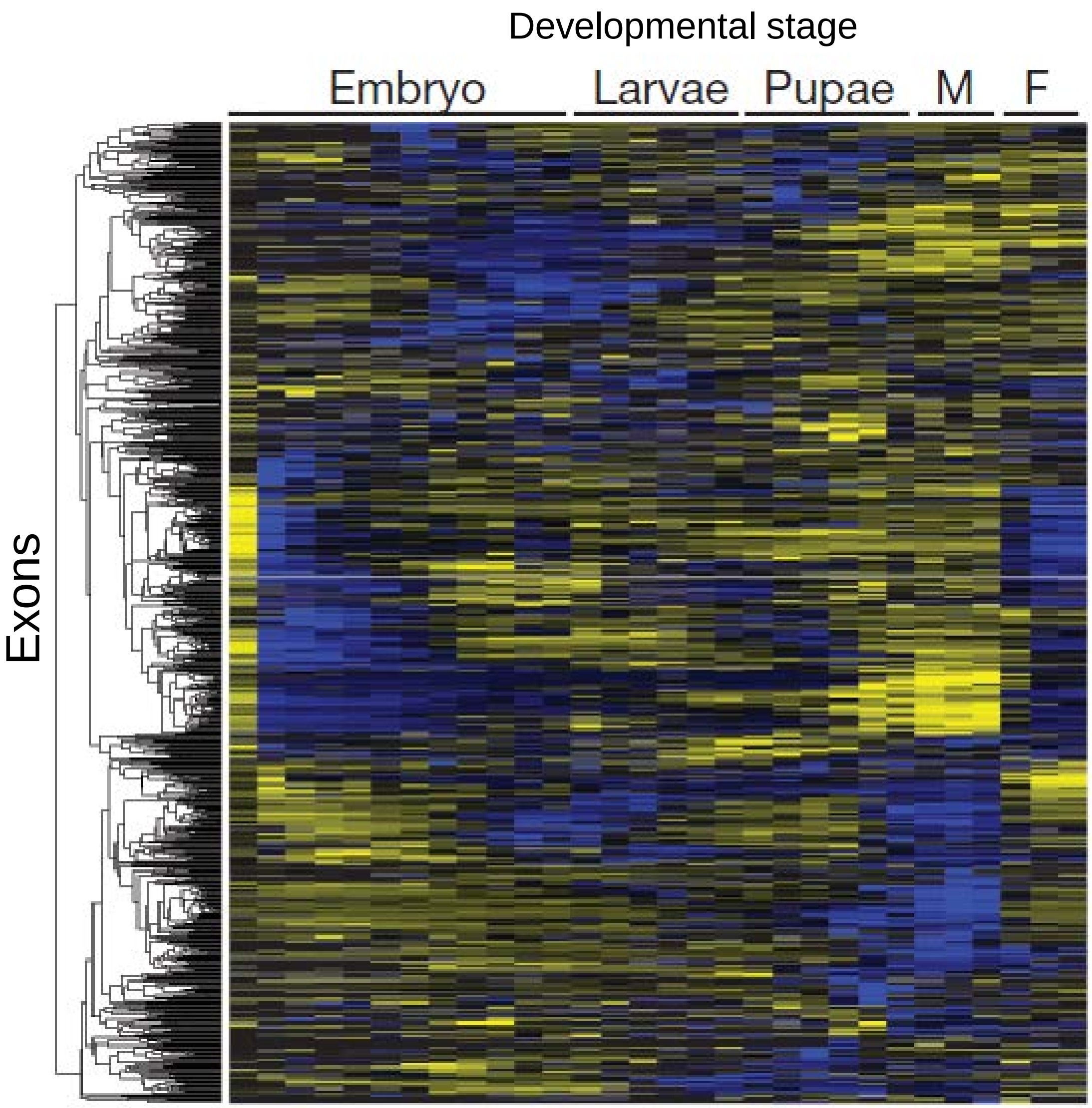
Other methods

- 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 $\rightarrow$ 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



Colour indicates use pattern of the exon

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

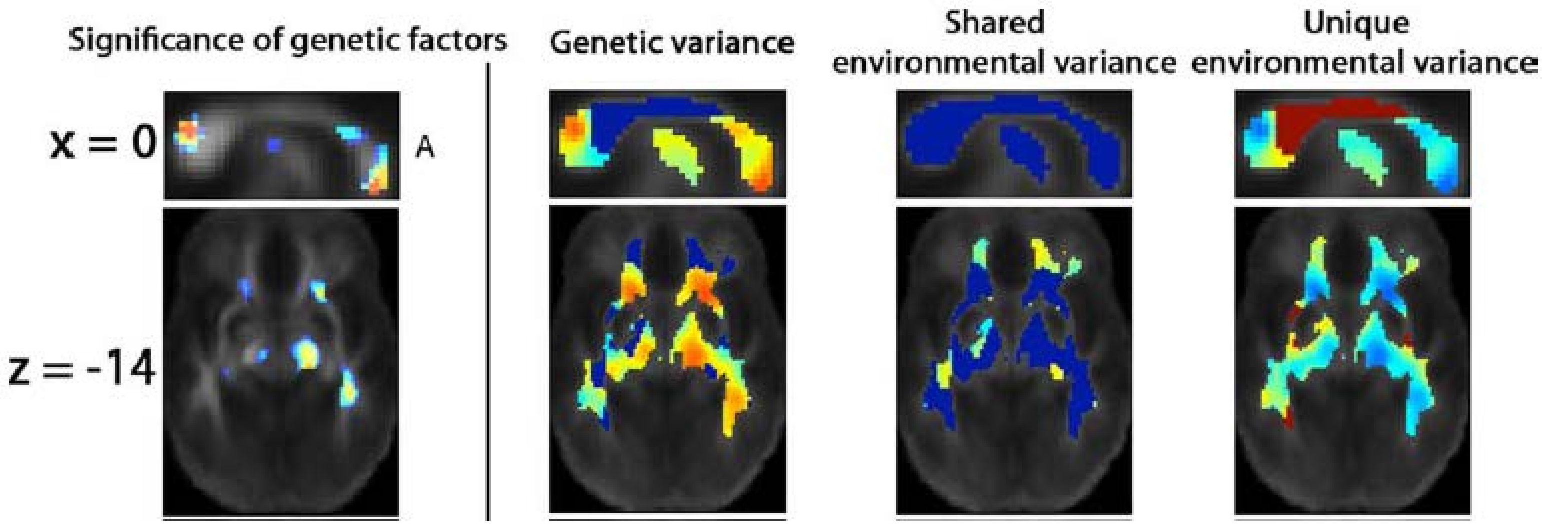
The developmental transcriptome of *Drosophila melanogaster*

Brenton R. Graveley[1*], Angela N. Brooks[2*], Joseph W. Carlson[3*], Michael O. Duff[1*], Jane M. Landolin[3*], Li Yang[1*], Carlo G. Artieri[4], Marijke J. van Baren[5], Nathan Boley[6], Benjamin W. Booth[3], James B. Brown[6], Lucy Cherbas[7], Carrie A. Davis[8], Alex Dobin[8], Renhua Li[4], Wei Lin[8], John H. Malone[4], Nicolas R. Mattiuzzo[4], David Miller[9], David Sturgill[4], Brian B. Tuch[10,11], Chris Zaleski[8], Dayu Zhang[7], Marco Blanchette[12,13], Sandrine Dudoit[14], Brian Eads[9], Richard E. Green[15], Ann Hammonds[3], Lichun Jiang[4], Phil Kapranov[8], Laura Langton[5], Norbert Perrimon[16], Jeremy E. Sandler[3], Kenneth H. Wan[3], Aarron Willingham[17], Yu Zhang[4], Yi Zou[7], Justen Andrews[9], Peter J. Bickel[6], Steven E. Brenner[2,17], Michael R. Brent[5], Peter Cherbas[7,9], Thomas R. Gingeras[8,18], Roger A. Hoskins[3], Thomas C. Kaufman[9], Brian Oliver[4] & Susan E. Celniker[3]

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

2212 • The Journal of Neuroscience, February 18, 2009 • 29(7):2212–2224

Behavioral/Systems/Cognitive

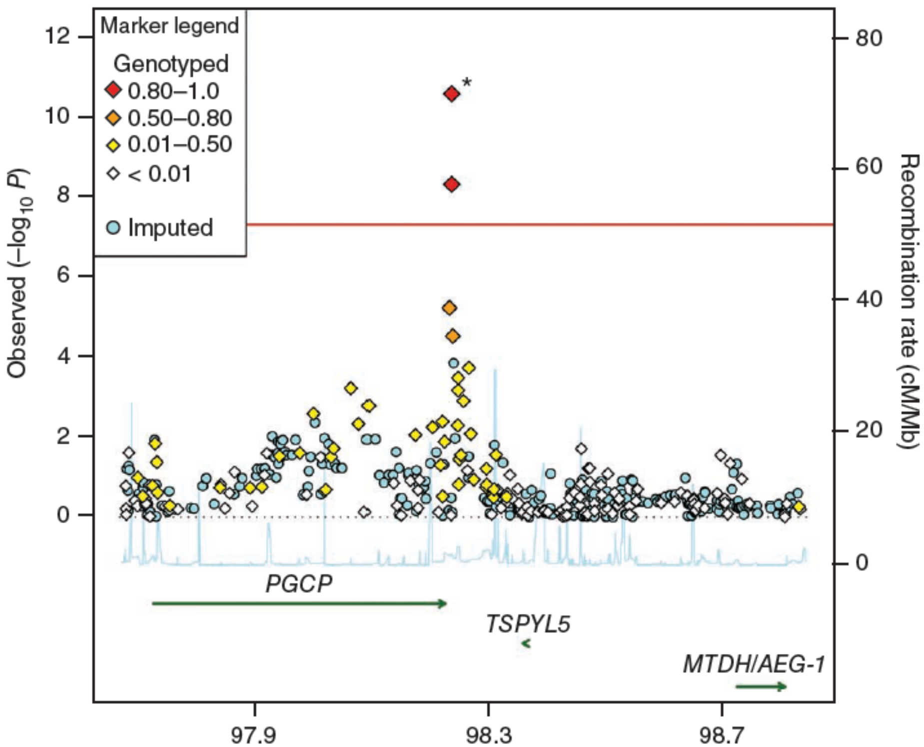
Genetics of Brain Fiber Architecture and Intellectual Performance

Ming-Chang Chiang,[1] Marina Barysheva,[1] David W. Shattuck,[1] Agatha D. Lee,[1] Sarah K. Madsen,[1] Christina Avedissian,[1] Andrea D. Klunder,[1] Arthur W. Toga,[1] Katie L. McMahon,[2] Greig I. de Zubicaray,[2] Margaret J. Wright,[3] Anuj Srivastava,[4] Nikolay Balov,[4] and Paul M. Thompson[1]

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



Chromosome 8 coordinate

- 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[15], Anine H Stam[7], Claudia M Weller[15], Axel Heinze[17], Katja Heinze-Kuhn[17], Ingrid Goebel[5,6], Guntram Borck[5,6], 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 Lucae[14], Federica Tozzi[39], Pierandrea Muglia[39,40], Jeffrey Barrett[1], Jaakko Kaprio[2,24,41], Markus Färkkilä[4], Leena Peltonen[1,2,42,48], 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[7,15], 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