Big data in genetics

Slides: `bit.ly/ICQG2016`

These are slides for a 10-min introduction on Big data in genetics, given at the International Conference on Quantitative Genetics on 16 June 2016.

Source: `https://github.com/kbroman/Talk_ICQG2016`
Slides: `http://bit.ly/ICQG2016_nonotes`
With notes: `http://bit.ly/ICQG2016`

- More samples
- More phenotypes
- More genomic-y things
- More time points, tissues, environments, treatments
- More artifacts

But more is not necessarily better.

What is big data? It's more data.

But more is not necessarily better. The effort to measure more stuff on more things can lead us towards cheap measures that give crap data.

On the other hand, there are huge opportunities here. I'm particularly excited about intermediate, refined phenotypes that may have simpler genetic architecture and get us closer to mechanisms.

Moving data around can be a feat

Analysis results often considerably bigger than data.

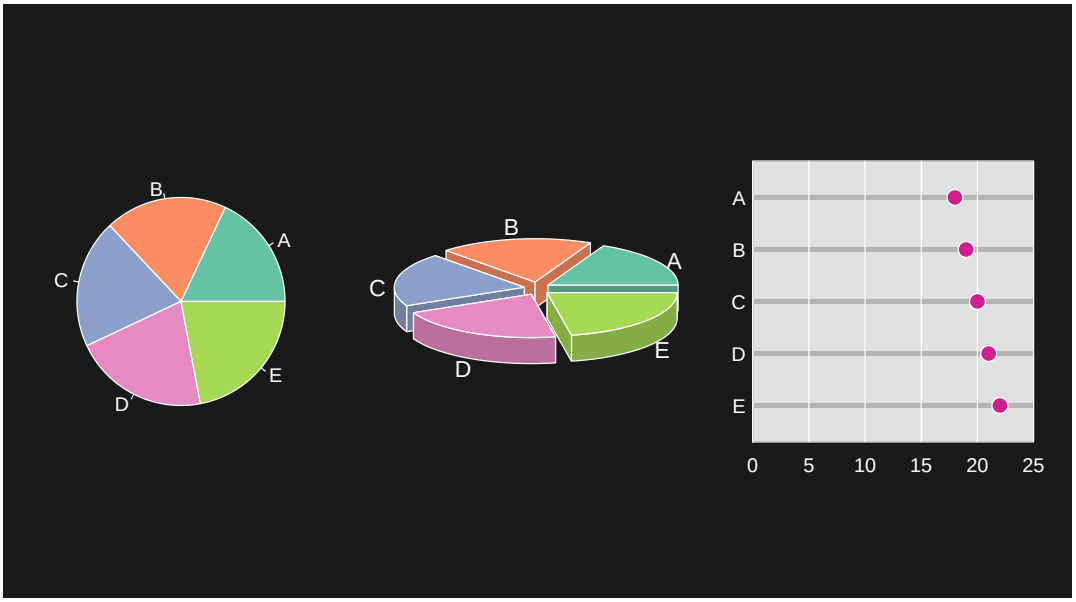
Saving results to disk can take as long as the calculations themselves.

I recently said to a collaborator, “I’ll just copy this to a USB stick.” But it was going to take an hour. Needed to use an ethernet cable instead.

Data visualizations are critical

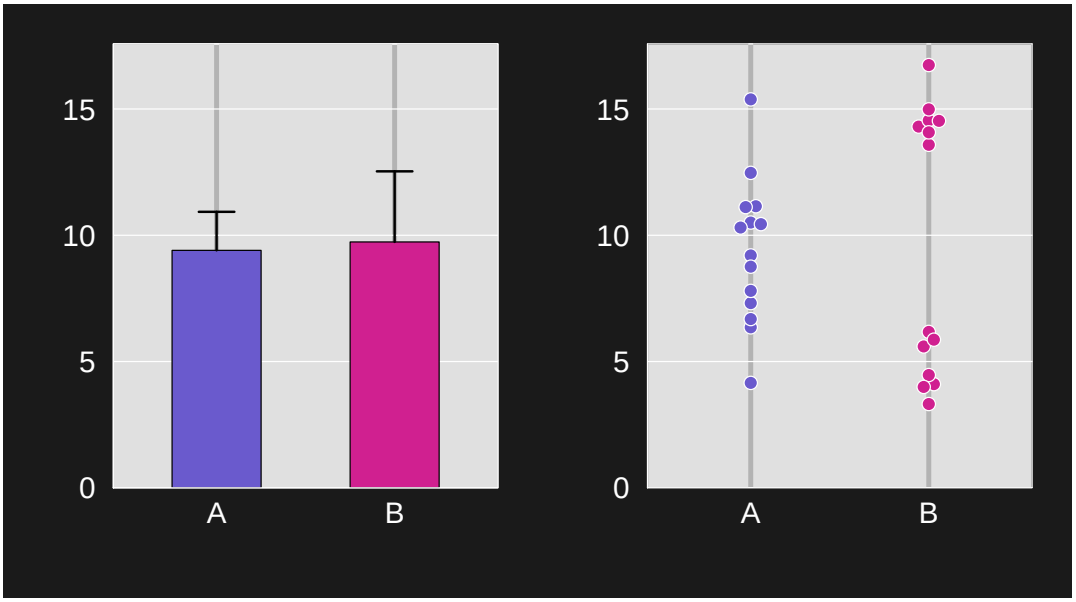
With big data (and big results), we rely more on data visualizations, and we need to do a better job at this.

Pie charts suck



Pie charts are ineffective because humans are terrible at quantifying areas.

Bar charts suck



Reducing results to two numbers is almost always bad.

Tables suck

rs ID	All	HDL_CRP	HDL_LDL	HDL_TG	LDL_CRP	TG_CRP	TG_LDL
rs3764261	3.115800e-01	2.610400e-31	6.167100e-31	7.179700e-33	3.857500e-01	2.301000e-01	2.475900e-01
rs1532624	5.934000e-01	7.134300e-24	1.286400e-23	1.477200e-24	4.096300e-01	1.467800e-01	1.844000e-01
rs2794520	2.936500e-13	4.404900e-22	3.241900e-01	2.812900e-01	2.030400e-22	5.021100e-23	8.300500e-01
rs7499892	3.460300e-02	3.303200e-16	1.553200e-16	1.935800e-20	6.642000e-01	3.166200e-01	6.211500e-01
rs2592887	3.707500e-10	6.883400e-17	9.482500e-02	1.104900e-01	5.918400e-17	8.024200e-17	7.710100e-01
rs646776	5.625600e-02	6.116800e-02	2.055800e-14	2.914200e-01	2.389500e-15	2.819600e-01	1.587700e-15
rs1532085	9.795800e-01	2.046400e-11	2.626800e-11	2.063700e-15	8.229500e-01	2.304300e-01	2.445000e-01
rs1811472	6.488000e-10	1.028000e-14	1.075200e-01	1.334100e-01	7.085600e-15	8.336700e-15	7.862900e-01
rs12093699	3.487700e-08	2.803100e-14	8.704700e-01	9.775600e-01	1.324500e-13	5.136100e-14	8.555500e-01
rs2650000	2.150800e-06	3.117800e-11	4.840500e-01	5.395800e-01	1.412700e-11	1.175100e-11	7.312200e-01
rs6728178	1.348000e-02	3.726700e-06	3.718900e-11	8.567700e-09	1.192200e-07	1.192900e-06	5.170100e-10
rs6754295	1.244400e-02	4.028300e-06	3.838300e-11	1.715000e-08	9.608700e-08	2.323300e-06	8.179200e-10
rs693	5.549800e-02	3.253100e-02	4.795900e-11	3.963400e-03	1.410000e-10	1.015900e-02	1.324700e-10
rs7953249	6.533200e-06	2.201500e-10	4.063400e-01	4.969100e-01	1.016400e-10	8.025500e-11	7.893600e-01
rs1169300	1.736100e-05	9.062700e-10	5.325500e-01	7.450100e-01	3.118100e-10	1.515200e-10	6.399000e-01
rs2464196	1.619700e-05	9.053200e-10	6.220200e-01	7.569200e-01	4.075100e-10	1.744700e-10	7.528400e-01
rs673548	4.967500e-02	4.309800e-06	2.014500e-10	3.655500e-09	1.150800e-06	5.222900e-07	1.055400e-09
rs415799	2.000800e-01	2.320400e-07	1.493100e-07	2.216300e-10	7.457500e-01	2.088500e-01	3.640400e-01
rs676210	5.072700e-02	5.251900e-06	2.883700e-10	5.535600e-09	1.364200e-06	7.256100e-07	1.583900e-09
rs174546	1.379500e-01	1.590200e-01	8.556400e-07	9.688800e-03	1.623200e-05	1.427300e-02	4.819700e-10
rs102275	1.678600e-01	1.112900e-01	5.655100e-07	9.205100e-03	1.723700e-05	1.722200e-02	7.111600e-10
rs1260326	4.928500e-01	1.036300e-01	2.940800e-01	7.494900e-10	8.537200e-02	1.110700e-09	1.140400e-09
rs261336	6.096900e-01	1.066600e-04	1.045000e-03	9.195300e-10	7.777300e-02	2.454000e-04	5.129400e-04
rs174537	1.410500e-01	1.549200e-01	1.474200e-06	1.113400e-02	3.039100e-05	1.637500e-02	1.306000e-09
rs1535	1.565800e-01	1.949600e-01	1.776900e-06	1.539500e-02	2.517300e-05	2.155300e-02	1.698300e-09
rs174556	6.854800e-02	3.880900e-01	1.614800e-06	5.138000e-02	5.632100e-06	5.600000e-02	1.846800e-09
rs10096633	7.343600e-01	1.076800e-05	1.327200e-05	2.542900e-09	6.869700e-01	7.121200e-08	2.132500e-08
rs735396	4.019400e-04	3.576200e-08	4.346600e-01	4.457100e-01	1.036900e-08	2.650500e-09	4.197800e-01
rs3923037	1.198500e-03	2.762800e-03	3.162700e-08	1.440400e-06	9.422300e-07	6.535200e-06	4.137700e-09
rs2126259	1.323700e-02	4.359000e-09	9.557500e-08	1.423400e-04	5.889200e-06	2.068100e-04	6.961300e-04
rs9989419	9.182800e-01	1.922700e-08	1.983700e-08	4.919800e-09	9.527800e-01	8.393700e-01	8.597400e-01
rs780094	6.157600e-01	2.610900e-01	6.568900e-01	7.159300e-09	2.491000e-01	2.042200e-08	1.189700e-08
rs11668477	4.263400e-01	4.826000e-02	8.350000e-09	1.810800e-02	2.422300e-08	4.300100e-02	3.161500e-08
rs11265260	7.116800e-06	4.158100e-08	2.523900e-02	2.956100e-02	1.047600e-08	9.316100e-09	3.308700e-01
rs1800961	6.390900e-01	1.094000e-08	1.636900e-07	1.074400e-07	1.981400e-01	1.465700e-03	1.085900e-02
rs754524	4.032900e-02	1.337000e-01	1.529600e-08	1.141300e-01	3.017000e-08	3.648500e-01	2.833300e-08
rs2075650	7.106400e-03	3.324200e-04	5.500700e-04	3.998700e-04	2.922700e-07	2.092800e-08	1.028800e-05
rs255049	8.151200e-01	2.079900e-07	2.294600e-08	1.371900e-07	3.933200e-01	4.514200e-01	8.102700e-02
rs166358	5.721200e-01	7.791800e-07	8.254800e-07	3.621500e-08	4.471400e-02	5.396600e-01	1.480400e-02

We mostly care about qualitative differences; tables are terrible for that.

Tools matter

- Need better than toy implementations.
- No more “The attached is similar to the code we used.”
- Work together on common tools.

Too often, we focus on the methods and write toy software implementation that are sufficient for the methods paper but not for anything else.

Often we don’t provide any software for our new methods. And we may not take sufficient care in ensuring the computational reproducibility of our work.

And main academic incentives are to make a new tool rather than to contribute to others’ tools. Working together on common tools would be more useful for the community.

Recognize tool makers

- Novelty of methods isn't everything
- Need a home for tool makers in academics

Academics (tenure, grants, awards) rewards novel methods far more than useful software.

Folks interested in tool development need to devote considerable effort to things they don't care about, or they'll leave for a lucrative data science industry job, where their talents are more strongly rewarded.

We need to fix this. It's a cultural problem.

Training

- Statistical/computational methods
  - Summer Institute in Statistical Genetics
  - Short Course on Systems Genetics
- Data manipulation and management
  - `datacarpentry.org`
- Software engineering
  - `software-carpentry.org`

Academics (tenure, grants, awards) rewards novel methods far more than useful software.

Folks interested in tool development need to devote considerable effort to things they don't care about, or they'll leave for a lucrative data science industry job, where their talents are more strongly rewarded.

We need to fix this. It's a cultural problem.

Be open

- Open data
- Open software
- Open manuscripts

Openness of data, software, and manuscripts is better for the community.