Analyzing large-scale genomics data

Bas Heijmans
Molecular Epidemiology
Leiden University Medical Center
The Netherlands
bas.heijmans@lumc.nl

FOS course Molecular Data Science – 22 October 2018.

The drunkard's search effect

From 1 to all

- All genetic variants, genes, metabolites
  → comprehensive & representative
  (instead of generalizing a single bit of knowledge)
- Disease ≠ 1 gene
  → hypotheses (!) and discoveries on the full complexity of biology.
- Exploiting natural variation
  → The human as model organism

Learning objectives

1. SPSS 2nd
2. '*R*' 1st

Why?

From traditional data to large-scale (high-dimensional) data:

- Many different formats of data files
- Data require preprocessing (quality control, normalization)
- Many tests (thousands, millions, billions)
- Novel methods
- Computational intensive methods
- Smart figures to make sense of data
- Visualizations to make sense of results
- Linking to external knowledge for interpretation

How to in SPSS

GeneExpression.cel
(Affymetrix)

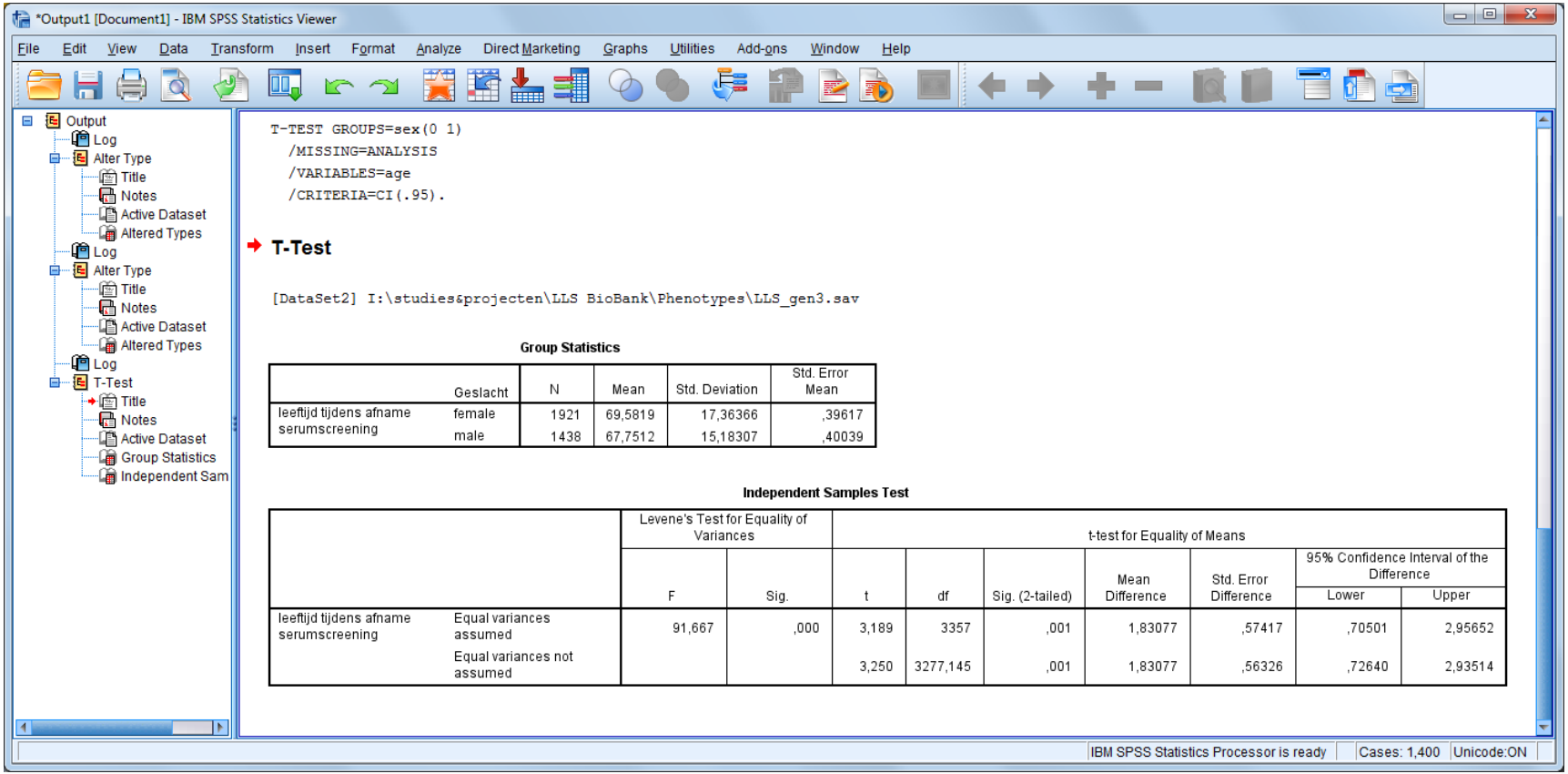
GeneExpression.idat
(Illumina)

- Many different formats of data files.
- Data require preprocessing (quality control, normalization) prior to analysis.

How to in SPSS

Person-id	Expression gene 1	Outcome
1	10	2.3
2	6	0.9
…	…	….
1000	15	1.5

How to in SPSS



How to in SPSS

Person-id	Gene 1	Gene 2	…	Gene 22,703	Outcome
1	10	1		90	2.3
2	6	0		54	0.9
…	…	…	…	…	….
1000	15	3		39	1.5

- 22,703 tests
- Repeat same analysis many times and store results in one data object.

How to in SPSS

Person-id	Variant 1	Variant 2	…	Variant $7x10^6$	Gene 1	Gene 2	…	Gene 22,703	Outcome
1	AG	TT		AT	10	1		90	2.3
2	GG	TC		AA	6	0		54	0.9
…	…	…	…	…	…	…	…	…	…
1000	GG	TC		AT	15	3		39	1.5

- 7M x 22,703 tests
- Distribute computations across processors (parallelization)
- Novel methods

How to in SPSS

Person-id	Variant 1	Variant 2	…	Variant $7x10^6$	Gene 1	Gene 2	…	Gene 22,703	Outcome
1	AG	TT		AT	10	1		90	2.3
2	GG	TC		AA	6	0		54	0.9
…	…	…	…	…	…	…	…	…	…
1000	GG	TC		AT	15	3		39	1.5

- 7M x 22,703 tests
- Smart figures to make sense of data
- Visualizations to make sense of results
  → 0.1 trillion (= 10^{11}) p-values

How to in SPSS

Person-id	Variant 1	Variant 2	…	Variant $7x10^6$	Gene 1	Gene 2	…	Gene 22,703	Outcome
1	AG	TT		AT	10	1		90	2.3
2	GG	TC		AA	6	0		54	0.9
…	…	…	…	…	…	…	…	…	…
1000	GG	TC		AT	15	3		39	1.5

- 7M x 22,703 tests
- Linking to external knowledge for interpretation (e.g. location variant, function of gene)

How to in SPSS

Person-id	Expression gene 1	Outcome
1	10	2.3
2	6	0.9
…	…	….
1000	15	1.5

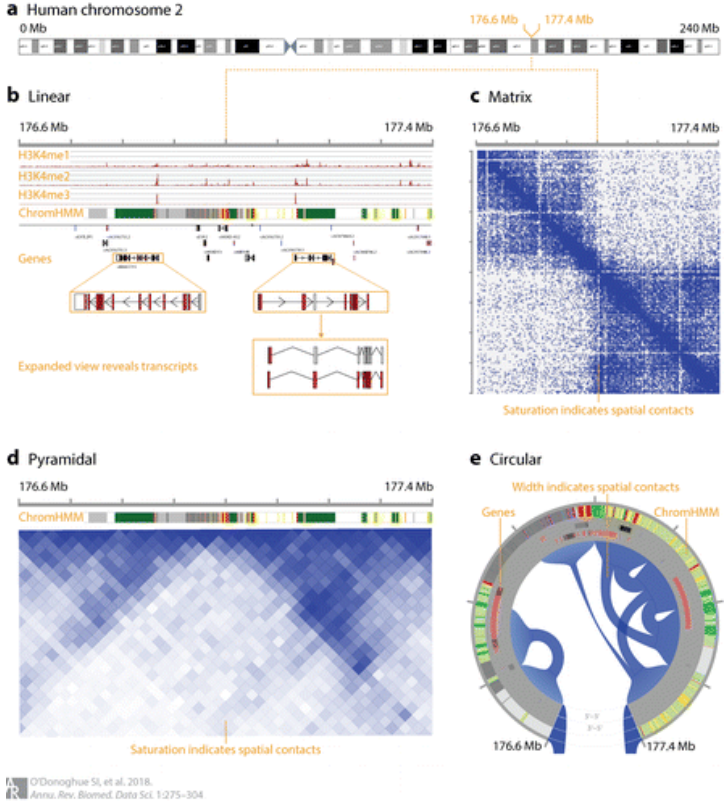
- Click-fest
- Complex output
- Ugly graphs
- Black-box: need to trust developers

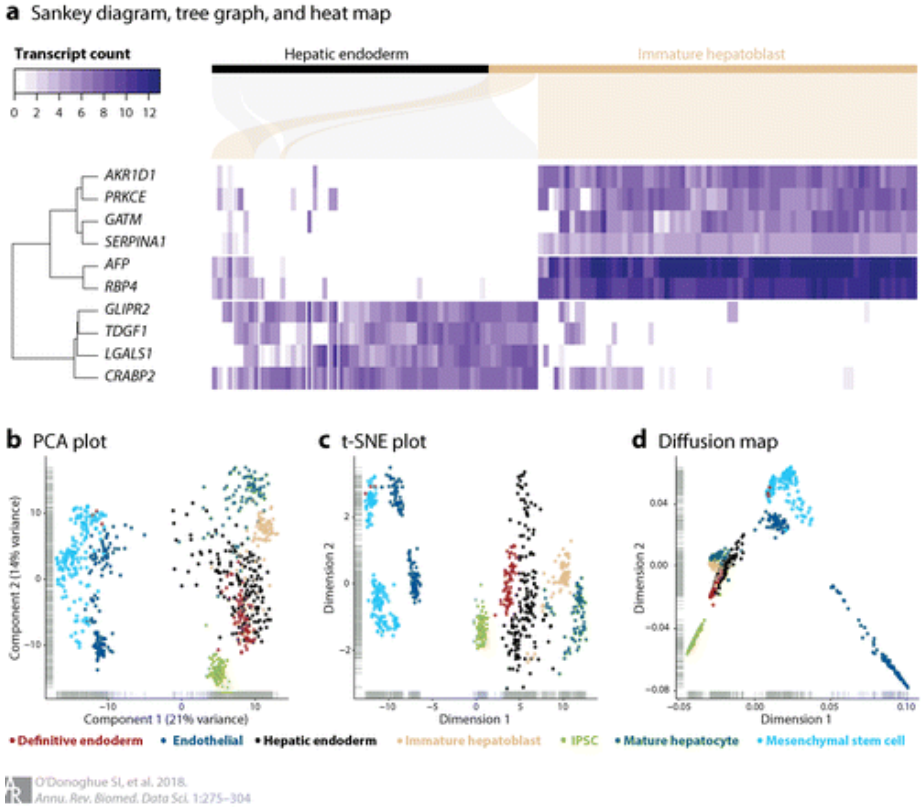
Why?

From traditional data to large-scale (high-dimensional) data

- Many different formats of data files
- Data require preprocessing (quality control, normalization)
- Many tests (thousands, millions, billions)
- Novel methods
- Computational intensive methods
- Smart figures to make sense of data
- Visualizations to make sense of results
- Linking to external knowledge for interpretation

Visualizations





O'Donoghue et al. Annu Rev Biomed Data Sci 2018

R first

- Do not fear the blinking cursor!
- You will find that R is not more complicated than SPSS if scripts are available.
- But: some analyses you will do are!
- Curriculum in transition: this is not an R course (a flavour of R & not all is in R).
- Also: R is not the answer to all issues in bioinformatics.