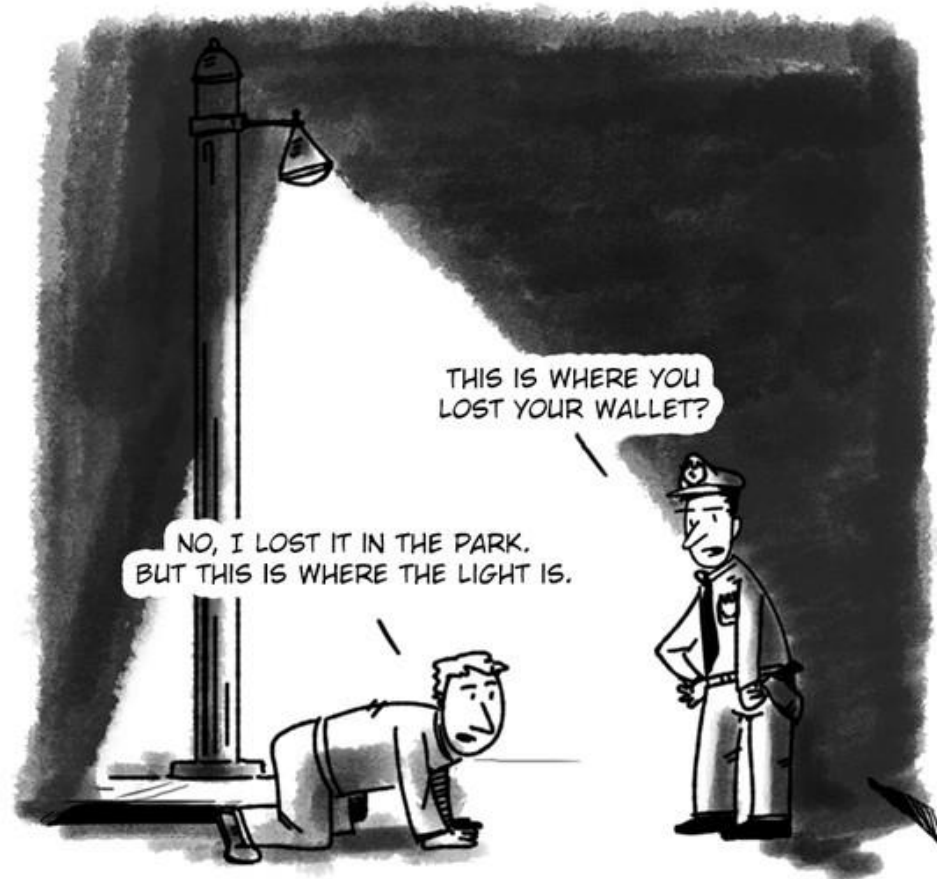


# Analyzing large-scale genomics data

Bas Heijmans  
Molecular Epidemiology  
Leiden University Medical Center  
The Netherlands  
[bas.heijmans@lumc.nl](mailto:bas.heijmans@lumc.nl)

# 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  $\neq$  1 gene  
→ hypotheses (!) and discoveries on the full complexity of biology.
- Exploiting natural variation  
→ The human as model organism

# Learning objectives

1. SPSS 2<sup>nd</sup>
2. '*R*' 1<sup>st</sup>

# 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

T-TEST GROUPS=sex(0 1)  
/MISSING=ANALYSIS  
/VARIABLES=age  
/CRITERIA=CI(.95).

→ **T-Test**

[DataSet2] I:\studies&projecten\LLS BioBank\Phenotypes\LLS\_gen3.sav

**Group Statistics**

	Geslacht	N	Mean	Std. Deviation	Std. Error Mean
leeftijd tijdens afname	female	1921	69,5819	17,36366	,39617
serumscreening	male	1438	67,7512	15,18307	,40039

**Independent Samples Test**

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
leeftijd tijdens afname	Equal variances assumed	91,667	,000	3,189	3357	,001	1,83077	,57417	,70501	2,95652
	Equal variances not assumed			3,250	3277,145					
serumscreening	Equal variances assumed	91,667	,000	3,189	3357	,001	1,83077	,57417	,70501	2,95652
	Equal variances not assumed			3,250	3277,145					

IBM SPSS Statistics Processor is ready | Cases: 1,400 | Unicode:ON



# 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 <sup>6</sup>	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 $7 \times 10^6$	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 <sup>6</sup>	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

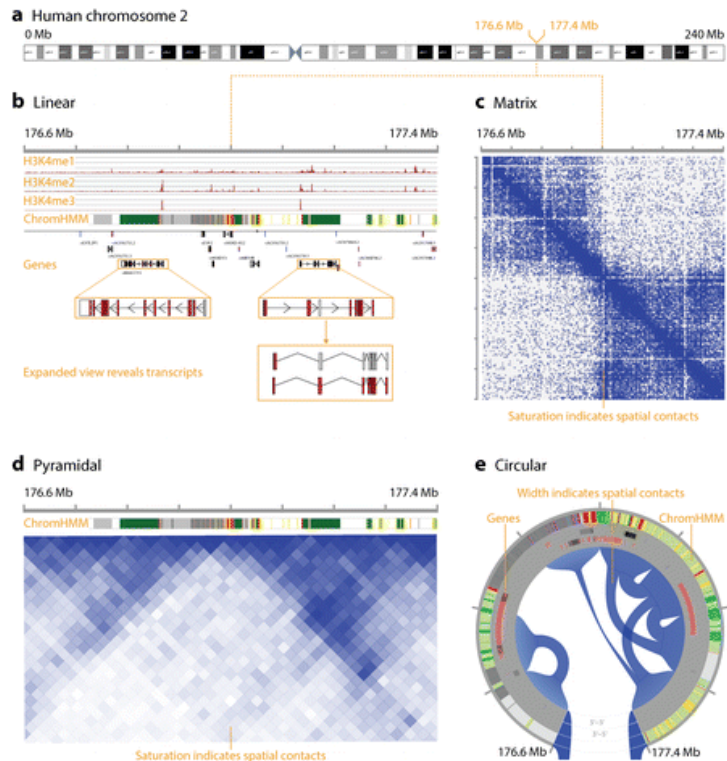
- Click-fest
- Complex output
- Ugly graphs
- Black-box: need to trust developers

# Why?

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

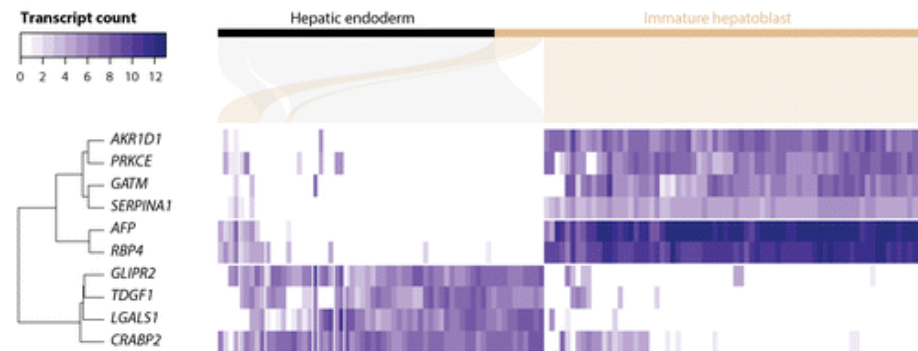
- Many different formats of data files
- Data require preprocessing (quality control, normalization)
- Many tests (thousands, millions, billions)
- Novel methods
- Computational intensive methods
- Smart figures to make sense of data
- Visualizations to make sense of results
- Linking to external knowledge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

# Visualizations

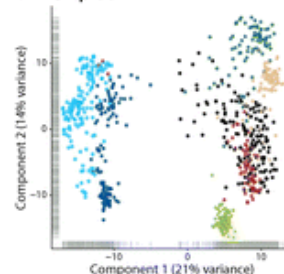


O'Donoghue SI, et al, 2018.  
Annu. Rev. Biomed. Data Sci. 1:275–304

**a Sankey diagram, tree graph, and heat map**

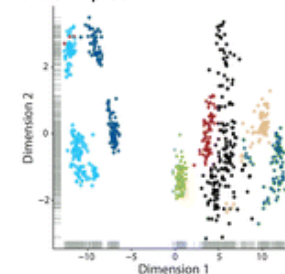


**b PCA plot**

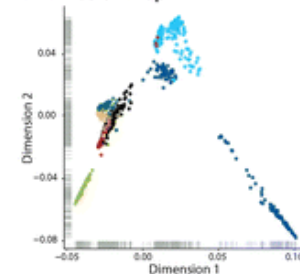


• Definitive endoderm • Endothelial • Hepatic endoderm • Immature hepatoblast • IPSC • Mature hepatocyte • Mesenchymal stem cell

**c t-SNE plot**



**d Diffusion map**



O'Donoghue SI, et al, 2018.  
Annu. Rev. Biomed. Data Sci. 1:275–304

# 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.