Very Many Variables and Limited Numbers of Observations

The p>>n Problem in Current Statistical Applications

Prof. Johann SÖLKNER

BOKU Vienna, Austria

Contents

- The p>>n problem
- Satistics vs data mining
- Variable selection procedures
- Prune, bundle, shrink and learn
- Real life examples
  - Genetics

The p>>n problem

- Consider Y the variable of interest and a "small" set of observations
- Given a very large set of potentially explanatory variables, X_1, X_2, X_3, …, X_p where p is large, then there is a high probability of finding an X that correlates with Y even if no real correlation exists in the domain

The p>>n problem

- Different X are often highly correlated
- Signals can drown in noise
- Computational challenges -> curse of dimensionality

Example problems

$X_1, X_2, X_3, \ldots, X_p$	->	Y
Climatic data	->	weather forecast
Spectroscopic image	->	content of ingredient
Genetic data	->	disease risk
Company data	->	portfolio allocation
Economic indicators	->	GDP
Internet SN data	->	influential users

- Different X are often highly correlated
- Signals can drown in noise

Statistics

- It is pointless attempting a definition of a discipline as broad as statistics (Hand, 1999)
- Emphasis on inference
- Probability and hypothesis testing
- Model building, understanding causative relations

Data Mining

- Results are models or patterns
  - Predictive power is the driver
- Makes also use of ideas, tools, and methods from other areas
  - database technology
  - machine learning
  - …

Statistics vs Data Mining

- Common aim: dicover structure in the data
- Some (statisticians) claim:
  Statistics is rigorous, data mining is ad hoc

- Huge amounts of data (often routinely collected) in many fields render statistics and data mining closer than the disciplines used to be

Variable selection

- Find influential variables
- Discard useless ones

- We will discuss variable selection in the multiple linear regression setting

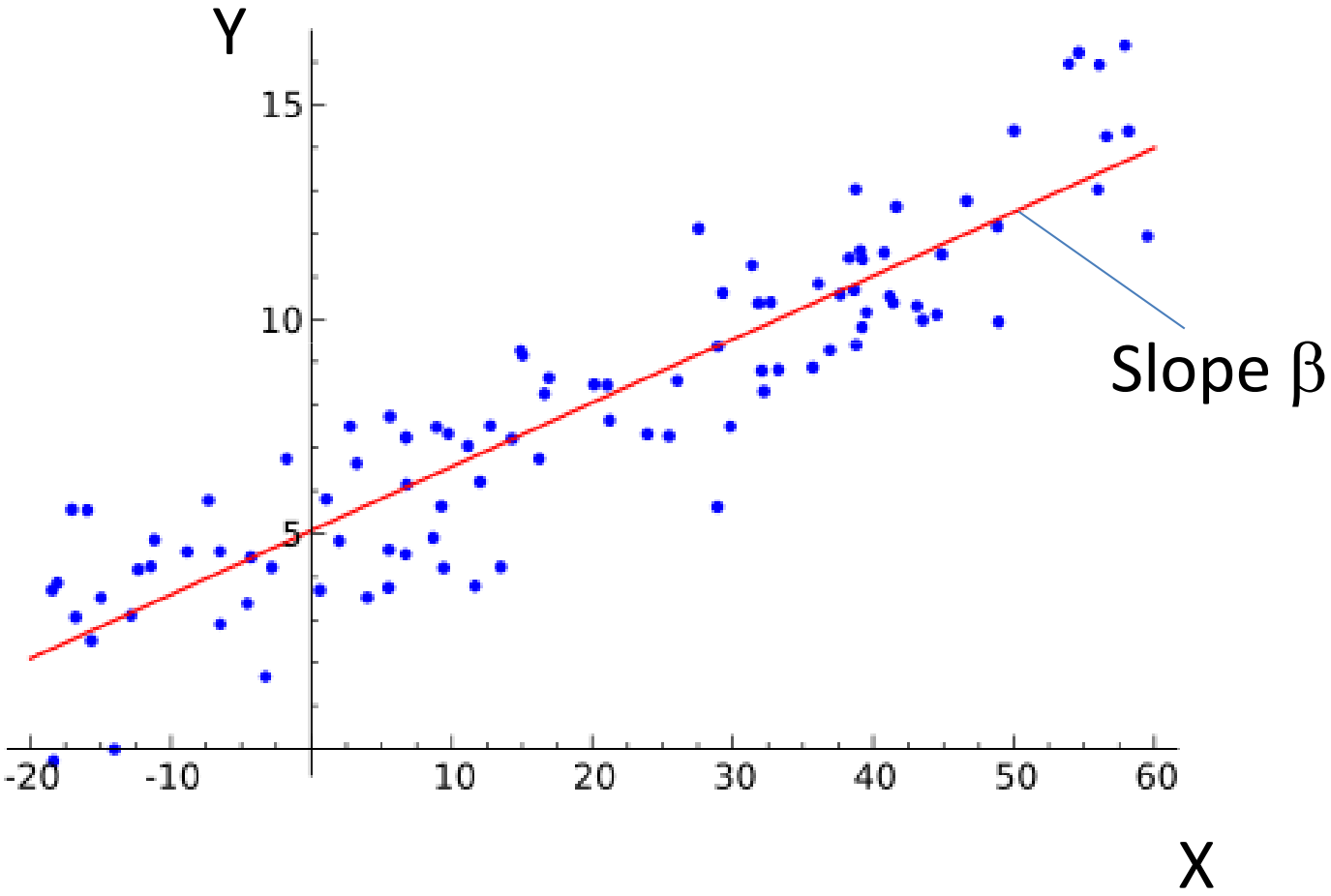
$$Y = X_1\beta_1 + X_2\beta_3 + X_3\beta_3 + \ldots + X_p\beta_p + \varepsilon$$

Variable selection

Regression has (at least) three major purposes:

1. Estimate coefficients in a pre-specified model
2. Discover an appropriate model
3. Predict values for new observations

Linear regression



Multiple linear regression

- Two or more explanatory variables
- Describes the state of nature of the dependent variable better when extra explanatory variables add information
- Number of explanatory variables must be smaller than number of observations (technical limitation)

Multiple linear regression

- As the number of explantory variables approaches the number of observations, noise is generated
  - Overfitting
- When some of the explanatory variables are highly correlated, it gets impossible to disentangle their effects
  - Multicollinearity problem

Variable selection methods

- Forward selection
  - Try simple linear regression with alternatively all explanatory variables
  - Keep the one that explains the data best and add (alternatively) all others
  - Continue until no statistcally obvious improvement is made
- Extremely prone to overfitting
  - Stopping rules

Variable selection methods

- Least angle regression (LAR), LASSO
  - Incredibly efficient set of methods of exhaustive search for best model up to a particular dimension (number of explanatory variables)
- Prone to overfitting
- Extensively researched to avoid pitfalls

Variable selection methods

- The simplest method:

Run simple linear regression for each variable separately, plot p-values (Manhattan plots)

Adjust for multiple testing

- Works well in Genome Wide Association Studies (GWAS)

Annals of human genetics

doi: 10.1111/j.1469-1809.2011.00698.x

Genome-Wide Association of Serum Uric Acid Concentration: Replication of Sequence Variants in an Island Population of the Adriatic Coast of Croatia

Rebekah Karns[1,†], Ge Zhang[2,†], Guangyun Sun[1], Subba Rao Indugula[1], Hong Cheng[1], Dubravka Havas-Augustin[3], Natalija Novokmet[3], Dusko Rudan[3], Zijad Durakovic[3], Sasa Missoni[3], Ranajit Chakraborty[4], Pavao Rudan[3] and Ranjan Deka[1*]

[1] *Center for Genome Information, Department of Environmental Health, University of Cincinnati, Cincinnati, OH, USA*
[2] *Human Genetics Division, Cincinnati Children's Hospital, Cincinnati, OH, USA*
[3] *Institute for Anthropological Research, Zagreb, Croatia*
[4] *Center for Computational Genomics, University of North Texas Health Science Center, Forth Worth, TX, USA*

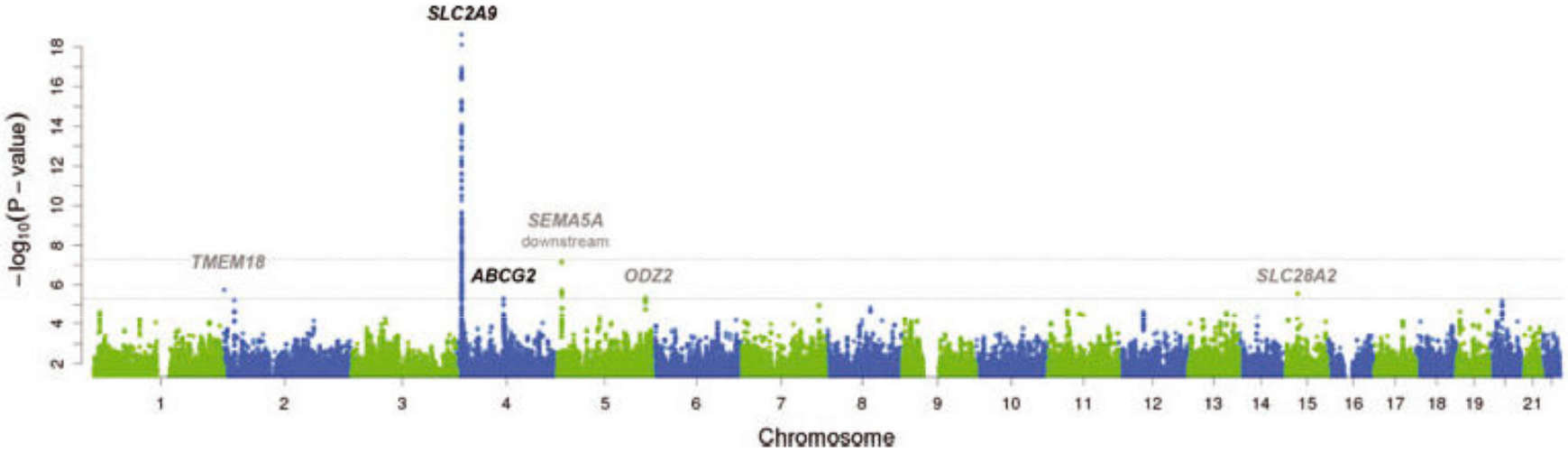


Figure 1 Manhattan plot of GWA single-locus *P*-values. The two horizontal dash lines indicate significant thresholds at 5×10^{-8} and 5×10^{-6}. Six regions that reach suggestive genome-wide significance ($P < 5\times10^{-6}$) are highlighted with names of nearby genes. Gene names in black are previously reported uric acid associated genes.

Prune, Bundle, Shrink and Learn

- Prune
  - Variable selection methods
- Bundle
  - Principal components, partial least squares
- Shrink
  - Ridge regression, BLUP
- Learn
  - Machine learning and other data mining methods

Bundle

- Instead of selecting variables, produce a small set of new variables that are functions of the original variables
- **Principal components regression (PCR)** is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components and applies regression using these new variables

Bundle

- **Partial least squares regression (PLS-R)** is similar to PCR but finds the new variables considering the co-variation of independent variables and the target variable
- Used extensively in various fields, including econometrics
- PCR and PLS-R also suffer from overfitting
- PLS components are (maybe too much) driven by the covariation of independent variables, often no big difference between the two methods

Shrink

- Ridge Regression (Hoerl and Kennard, 1970)
- Makes multiple regression with p>n possible with a technical trick
  - Matrix algebra
  - Change the diagonal values of the coefficient matrix X‘X
- Regresses estimates towards he mean (shrink)
- Has desirable statistical properties

Purge and Shrink

- Elastic Net

  LASSO – Ridge Regression hybrid

- Very desirable statistical properties

  – Normally better than LASSO or Ridge Regression

Learn

- Machine learning algorithms
  - Support vector machine (SVM) and derivatives
- In SVM the basic idea is to map the data x into a high-dimensional feature space F via a nonlinear mapping function, and to do linear regression in this space (cf. Boser et al., 1992; Vapnik, 1995)
- Just a different set of techniques of pooling information

Prune, Bundle, Shrink and Learn

- Extreme rush in development
- Combining approaches
- Frequentist, Bayesian, Data Mining

- Prediction of future records is the dominant driver

- Very similar prediction with conceptually very different methods…

Note

- More about LASSO, ridge regression, elastic net and other methods by Patrik Waldmann

- School of Biometrics, Thursday

Applications - Genetics

- The SNP era
  (soon to be whole genome sequence era)
- Human Genome consists of ~ 3,000,000,000 nucleotides (A,C,G,T) arranged in 23 pairs of chromosomes
- 99% identical for all humans, 1% (30,000,000) variable
- SNP: Single Nucleotide Polymorphism
  - Variable locus (one of the 30,000,000)

Applications - Genetics

- SNP chips with 500,000 – 2,500,000 SNPs available at a low price of € 100-250
- Extremely much information
- Genome Wide Association Studies
  - Genotype cases and controls (often panels of controls are available free of cost)
- Many other interesting questions
  - Inbreeding, Admixture,…

OPEN ACCESS Freely available online PLoS GENETICS

Genomics Meets Glycomics—The First GWAS Study of Human N-Glycome Identifies HNF1α as a Master Regulator of Plasma Protein Fucosylation

Gordan Lauc[1,2☯], Abdelkader Essafi[3☯], Jennifer E. Huffman[3☯], Caroline Hayward[3☯], Ana Knežević[2], Jayesh J. Kattla[4,5], Ozren Polašek[6,7], Olga Gornik[2], Veronique Vitart[3], Jodie L. Abrahams[4,5], Maja Pučić[1], Mislav Novokmet[1], Irma Redžić[2], Susan Campbell[3], Sarah H. Wild[8], Fran Borovečki[7], Wei Wang[9,10,11], Ivana Kolčić[7], Lina Zgaga[7], Ulf Gyllensten[12], James F. Wilson[8¶], Alan F. Wright[3¶], Nicholas D. Hastie[3¶], Harry Campbell[8¶], Pauline M. Rudd[4,5¶], Igor Rudan[8,11¶*]

European Journal of Human Genetics (2011) 19, 341–346

www.nature.com/ejhg

ARTICLE

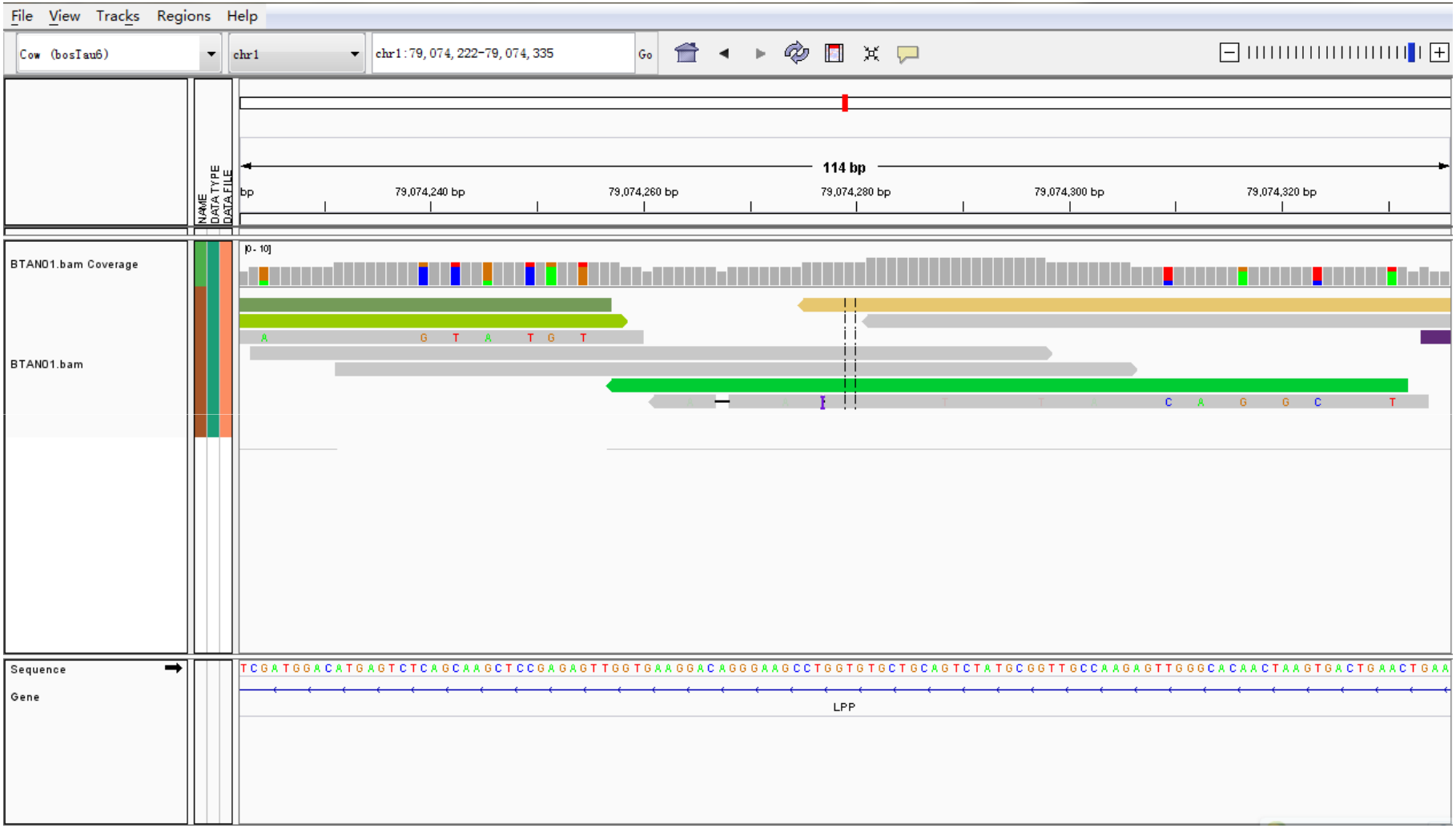
Replication of genetic variants from genome-wide association studies with metabolic traits in an island population of the Adriatic coast of Croatia

Rebekah Karns[1], Ge Zhang[1,2], Nina Jeran[3], Dubravka Havas-Augustin[3], Sasa Missoni[3], Wen Niu[1], Subba Rao Indugula[1], Guangyun Sun[1], Zijad Durakovic[3], Nina Smolej Narancic[3], Pavao Rudan[3], Ranajit Chakraborty[4] and Ranjan Deka*[,1]

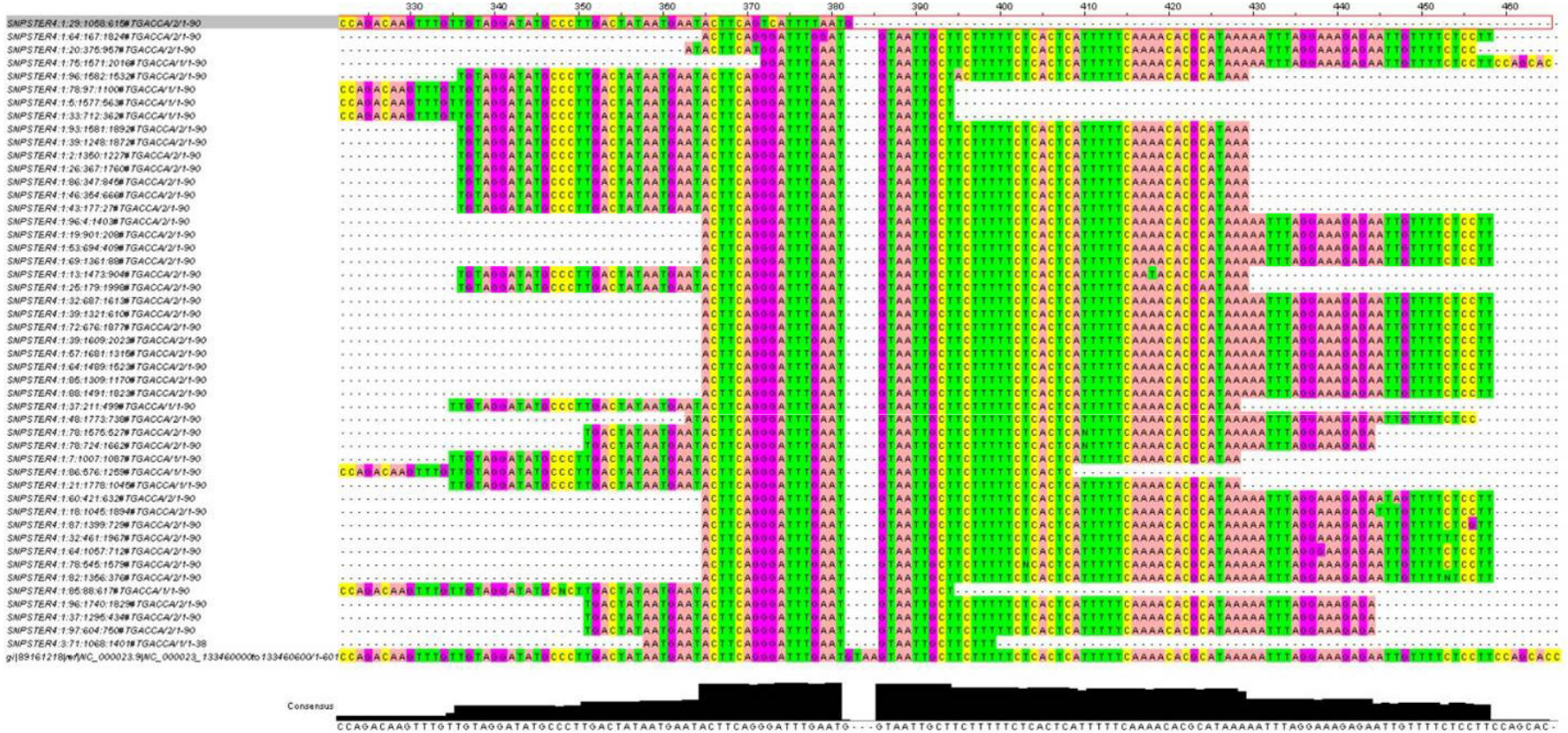
Whole genome sequence

- The 1000 Dollar genome is reality !!
- Reference genome, sequenced at extremely high cost
- Next Generation Sequencing techniques provide extremely many short reads (50-300 bases long), most of which can be quickly and successfully aligned to the reference genome
- 5x, 10x, 30x, 80x, … coverage
- Up to 240,000,000,000 data points per individual!

~ 3x coverage


File View Tracks Regions Help
Cow (bosTau6)
chr1
chr1:79,074,222-79,074,335
Go
114 bp
bp
79,074,240 bp
79,074,260 bp
79,074,280 bp
79,074,300 bp
79,074,320 bp
NAME
DATA TYPE
DATA FILE
BTAN01.bam Coverage
BTAN01.bam
Sequence
Gene
LPP

~ 30x coverage



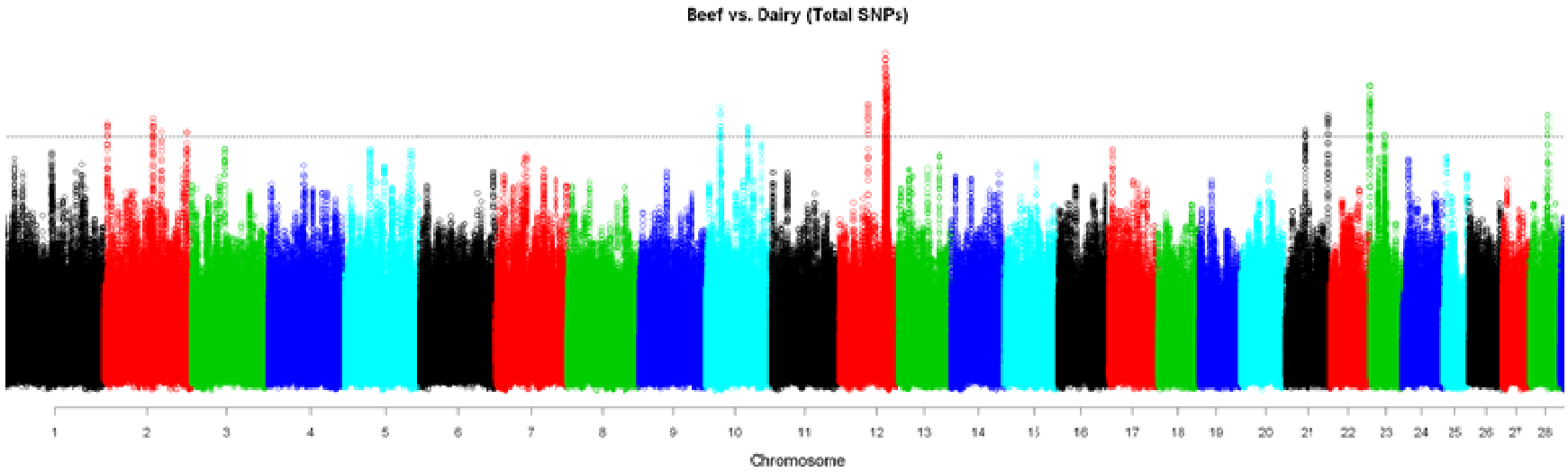
BOKU cattle sequence analysis

- Sequences from the Bovine HapMap consortium
- 70 bulls, 100 more to come
- 3x, 10x, 3 bulls 80x
- Partners of 1000 bull genomes project

- Analysis of data
  - SNP detection
  - Selection signatures
- Much work
  - Bioinformatics, Statistics, Data mining

First selection signature results

- Variation of linkage disequilibrium of dairy vs beef cattle (2 breeds each)



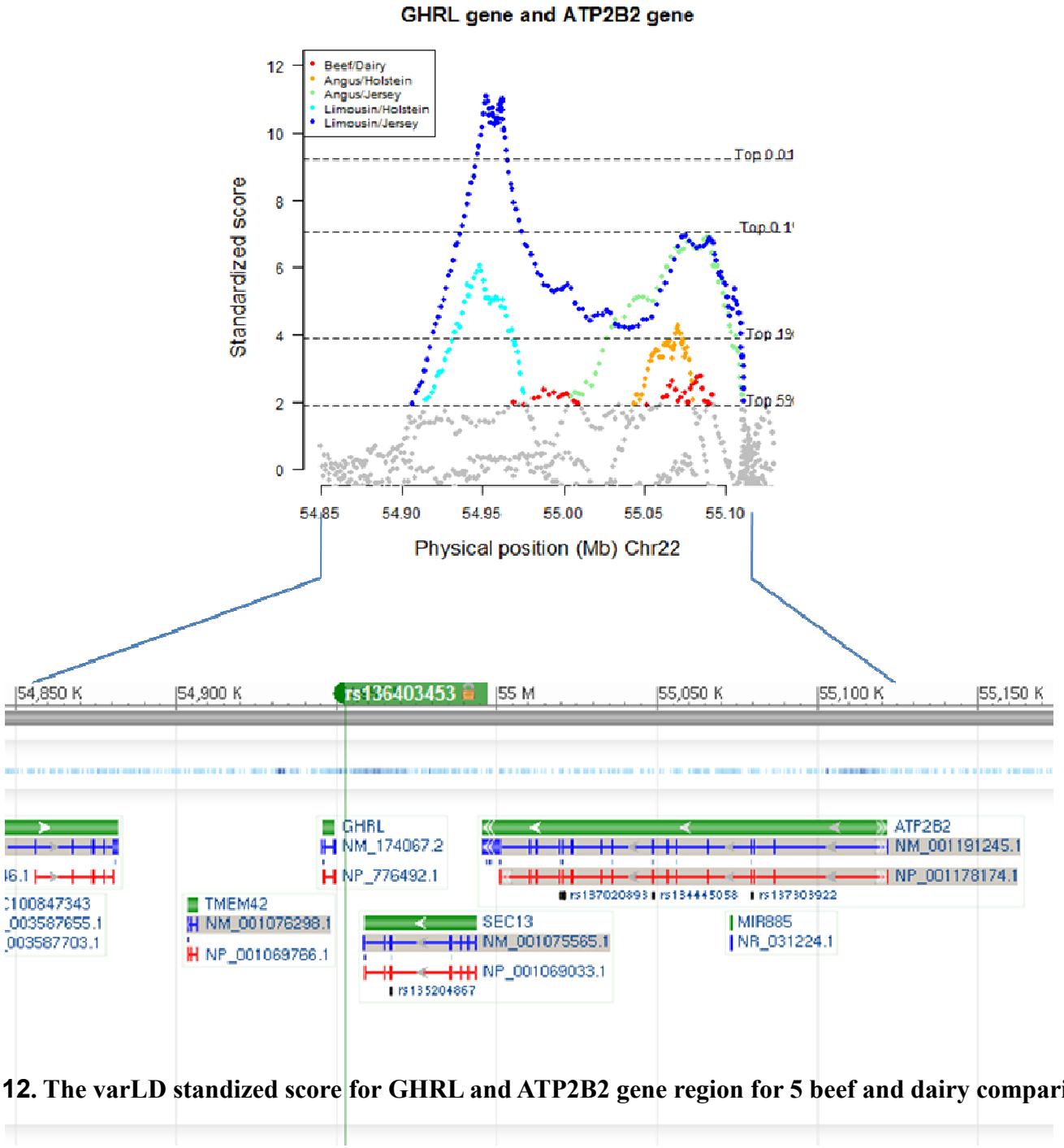


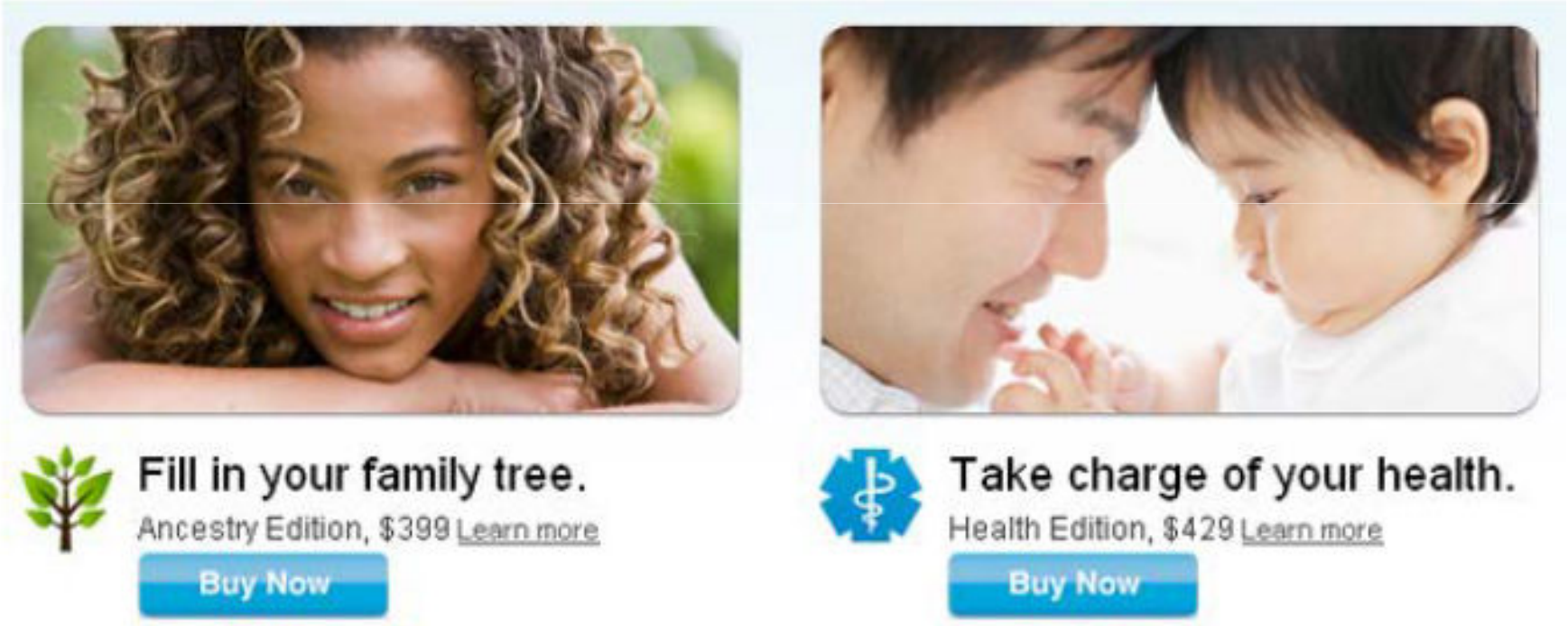
Figure 12. The varLD standized score for GHRL and ATP2B2 gene region for 5 beef and dairy comparisons.

Consumer genomics

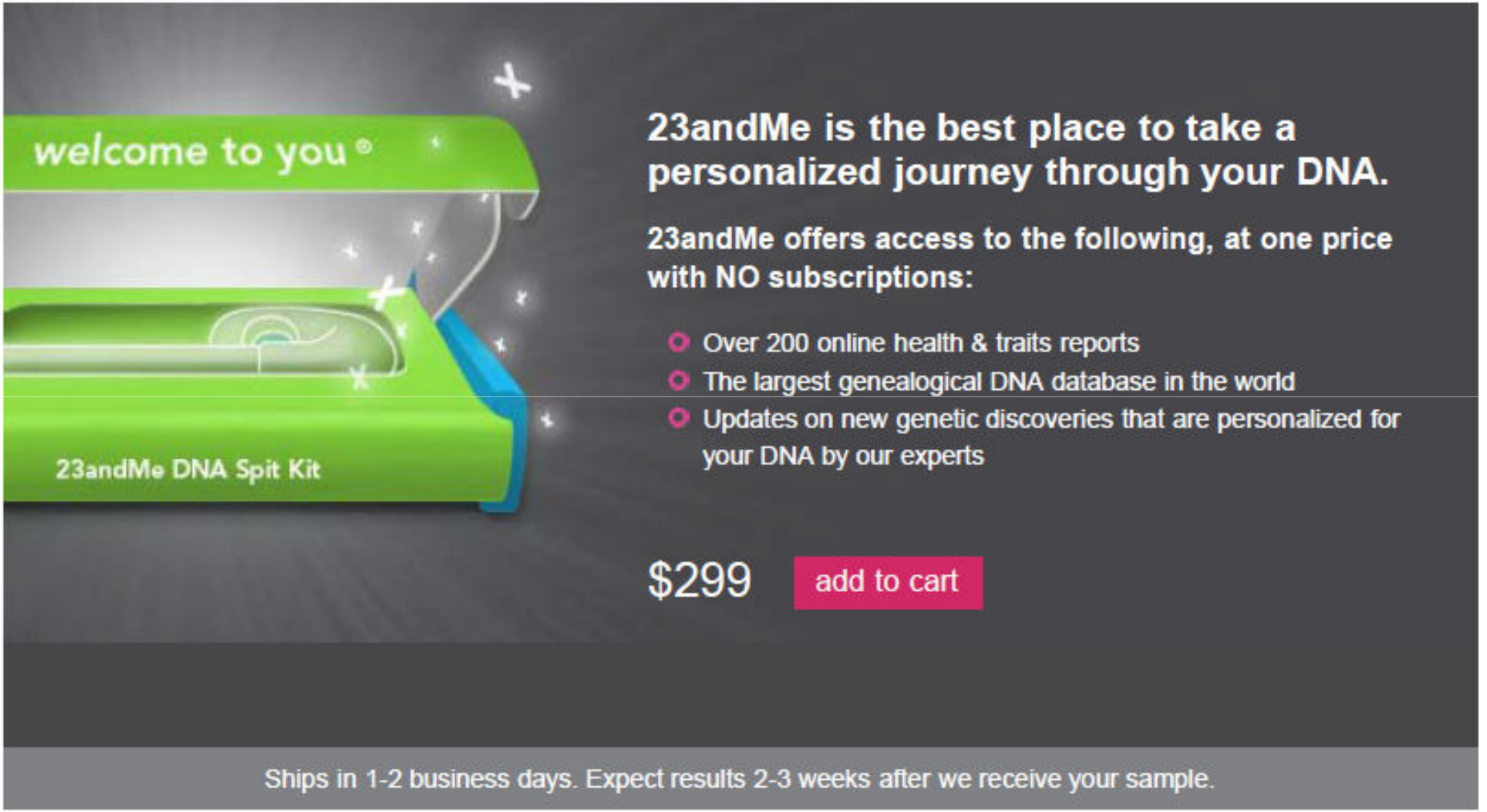
- Companies offering genomic profiling
  - 23andMe
  - DeCODEme
  - Navigenics
  - …
- Genetic risk profiling
  - Risks for up to 200 diseases (e.g., Diabetes Type 2)
  - Relative to population average
  - Regular updates based on newest scientific findings
- Ancestry profiling

Consumer genomics

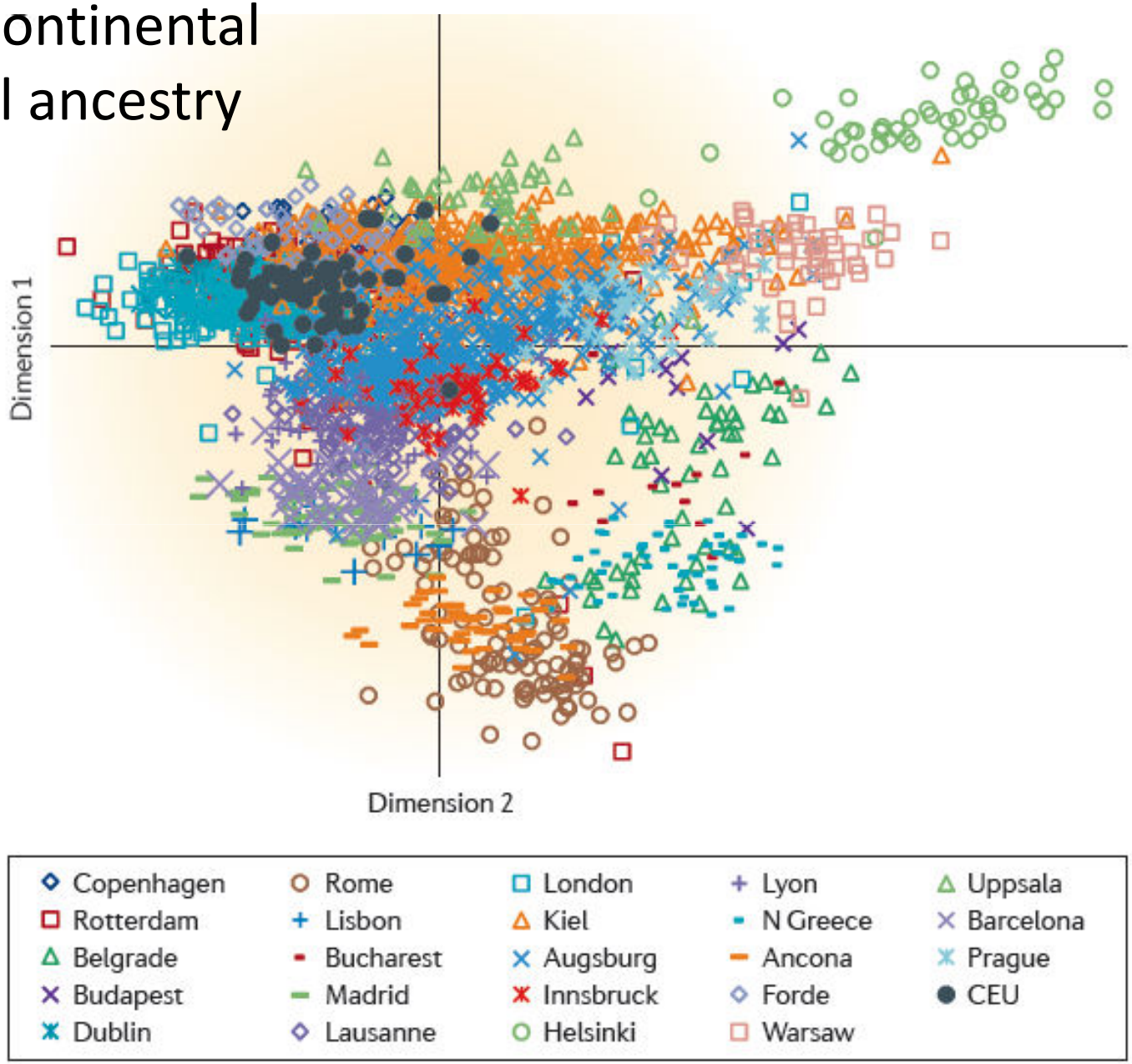
23andMe (2011)



23andMe store 2012



From continental to local ancestry



Note

- Use of SNP data to predict phenotypes, individual levels of admixture and levels of inbreeding

- School of Biometrics, Thursday

Conclusions

- Huge amounts of data everywhere
  - Think only about internet usage data!
- Statisticians need to get familiar with data mining tools, quickly
- Pedictive capacity is often more important than causal analysis
  - Not so in GWAS, though
- Statistics and data mining: Happy marriage !

Announcement

- International postgraduate course

Livestock conservation genomics:
Data, Tools and Trends

- Luna Hotel, Pag, 1.–7. October 2012
- Organizers: Ino Curik (University of Zagreb)
  Johann Sölkner (BOKU Vienna)
- International top scientists presenting