#### 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<sub>1</sub>, X<sub>2</sub>, X<sub>3</sub>, ..., X<sub>p</sub> 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<sub>1</sub>, X<sub>2</sub>, X<sub>3</sub>, ..., X<sub>p</sub>

- **Climatic data**
- Spectroscopic image
- Genetic data
- Company data
- Economic indicators
- Internet SN data

- Y
- -> weather forecast
- -> content of ingredient
- -> disease risk
- -> portfolio allocation-> GDP
- -> 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

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

#### 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)

#### 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<sup>1,†</sup>, Ge Zhang<sup>2,†</sup>, Guangyun Sun<sup>1</sup>, Subba Rao Indugula<sup>1</sup>, Hong Cheng<sup>1</sup>, Dubravka Havas-Augustin<sup>3</sup>, Natalija Novokmet<sup>3</sup>, Dusko Rudan<sup>3</sup>, Zijad Durakovic<sup>3</sup>, Sasa Missoni<sup>3</sup>, Ranajit Chakraborty<sup>4</sup>, Pavao Rudan<sup>3</sup> and Ranjan Deka<sup>1</sup>\*

<sup>1</sup>Center for Genome Information, Department of Environmental Health, University of Cincinnati, Cincinnati, OH, USA

<sup>2</sup>Human Genetics Division, Cincinnati Children's Hospital, Cincinnati, OH, USA

<sup>3</sup>Institute for Anthropological Research, Zagreb, Croatia

<sup>4</sup>Center for Computational Genomics, University of North Texas Health Science Center, Forth Worth, TX, USA

© 2012 The Authors Annals of Human Genetics © 2012 Blackwell Publishing Ltd/University College London Annals of Human Genetics (2012) 00,1-7

Genome-Wide Association of Uric Aci



Figure 1 Manhattan plot of GWA single-locus *P*-values. The two horizontal dash lines indicate significant thresholds at  $5 \times 10^{-8}$  and  $5 \times 10^{-6}$ . Six regions that reach suggestive genome-wide significance ( $P < 5 \times 10^{-6}$ ) 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,...

# Genomics Meets Glycomics—The First GWAS Study of Human N-Glycome Identifies HNF1 $\alpha$ as a Master Regulator of Plasma Protein Fucosylation

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

> European Journal of Human Genetics (2011) 19, 341–346 © 2011 Macmillan Publishers Limited All rights reserved 1018-4813/11 www.nature.com/eihg

#### 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<sup>1</sup>, Ge Zhang<sup>1,2</sup>, Nina Jeran<sup>3</sup>, Dubravka Havas-Augustin<sup>3</sup>, Sasa Missoni<sup>3</sup>, Wen Niu<sup>1</sup>, Subba Rao Indugula<sup>1</sup>, Guangyun Sun<sup>1</sup>, Zijad Durakovic<sup>3</sup>, Nina Smolej Narancic<sup>3</sup>, Pavao Rudan<sup>3</sup>, Ranajit Chakraborty<sup>4</sup> and Ranjan Deka<sup>\*,1</sup>

#### 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 data points per individual!

#### ~ 3x coverage

<u>File View Tracks</u>	Regions	Help											
Cow (bosTau6)	•	chrl	•	chr1:79,074,222-7	9, 074, 335	Go 🗂	• •	🤣 🖪 🛪					+
								-					
		┛						—— 114 bp -					
	NAME			79,074,240 bp 	I	79,074,260 bp 	I	79,074,280 E	bp I	79,074,300 bp 		79,074,320 bp 	<u> </u>
		р.	10]										
BTANU1.bam Coverage													
				с т (	тет								_
BTAN01.bam			• •										
						A	-	A I	Т	T A	C A	6 6 <b>C</b>	т
Sequence	→	TC	GATGGACATG/	AGTCTCAGCAAG	CTCCGAGAG	TTGGTGAAGG	A C A G G G A A	б <mark>сст</mark> ббтбт(	GCTGCAGTCT	A T G C G G T T G C C A A	6	А САА СТАА БТБА	C T G A A C T G A A
Gene				· · · · ·				LPP	• • • •	· · · · ·	• • • •		· · · · ·

#### ~ 30x coverage

	330	340	350	360	370	380	390	400 4	10 420	430	440	450	460
SNPSTER4:1:29:1058:615876ACCA/2/1-90	CCADACAAD TTTOTT	TADDATATO	CCCTTOAC	TATAATCAATACT	TCADTCAT	TTTAAT							
SNPS7ER4:1.64:167:1824#7GACCA/2/1-90				AC	TCADODAT	TTOOAT-	····	CTTTTTCTCACTCAT	TTTTCAAAACAC	ATAAAAAT	ANDAAAAAAAA	TTOTTTCTC	CTT
SNPSTER4:1:20:375:957#TGACCA/2/1-90				ATACI	TCATOOAT	TTOAAT-	····	CTTTTTCTCACTCAT	TTTTCAAAACAC	ATAAAAAT	T AO CAAAO AO AA	TTOTTTCTC	C
SMPS7ER4 1:76:1671:2016#7GACCA/1/1-90					00AT	TTUAAT-	····	CTTTTTCTCACTCAT	TTTTCAAAACAC	ATAAAAAT	TADDAAADAAAA	TTOTTTOTC	CTTCCANCAC-
SMPSTER4 1 96:1682:1532#TGACCA/2/1-90		TABBATATO	CCCTTOAC	TATAATAATACT	TCABOOAT	TTOAAT-	····	CTTTTTCTCACTCAT	TTTTCAAAACAC	ATAAA			
SMPSTER4 1:78:97:110007GACCA/1/1-90	CCABACAACTTTOTT	TABBATATO	CCCTTOAC	TATAATUAATACT	TCABOOAT	TTUAAT-	····						
SNPSTER4 1 5:1577 563# TGACCA/1/1-90	CCARACAARTTTOTT	TABBATATO	CCCTTOAC	TATAATGAATACT	CADDOAT	TTOAAT.							
SNPSTER4-1:33:712:362076ACC4/1/1-90	CCABACAAB TTTOTT	ATABBATATO	CCCTTOAC	TATAATOAATACT	CADDOAT	TTLAAT	TAATTOCT.						
SNPSTER4 1 93 1681 18928 TGACCA/2/1-90		O TABBATATO	CCCTTOAC	TATAATGAATACT	CADDOAT	TTOAAT.		CTTTTCTCACTCAT	TTTTCAAAACAC				
SNPSTER4 1 39 1248 18728 TGACCA/2/1.90		ATABBATATA	CCCTTOAC	TATAATOAATACT	TAGGOAT	TTOAAT.		CITTTTCCCACTCAT	TTTTCAAAACACAC				
SNPSTER4-1-2-1360-12220 TGACCA/2/1.90		TABBATAT	CCCTTOAC	TATAATOAATACT	CADDOAT	TTOOAT	TAATTOCT	CITTTCCCACTCAT	TTTTCAAAACACOC				
SMPSTER4-1-26-367-176087640C642/1-90			CCCTTCAC	TATAATHAATACT		TTUAAT	TAATTOCT	CITITICICACICAT	TTTTCAAAACACOC	ATAAA			
SNPSTER4-196-347-9468 TGACCA/2/1.90			CCCTTOAC	TATAATOAATACT	CACCA	TTOAAT	TAATTOCT	CITITCICACICAT	TTTTCAAAACACOC				
SNP 572R4.1.00.341.0408708000427150			CCCTTOAC			TTUAAT		CITET CICACICA	TTTTCAAAACACAC				
SNP 57EP4-1-40-504-0000704/000427-50			CCCTTCAC	TATAATUAATACT	CADOOAT	TTOAAT	TAATTOCT	CT TT CT CACT CA	TTTTCAAAACACAC				
SNPSTER4-1-96-4-403#TG8CC8/2/1-90			C C C III MAC		CALL A	TTURAT	TAATTOCT	CTTTTTCTCACTCA	TTTTCAAAACACAC	ATAAAAAT	TABBAAAAAAAAAAAA	TRITTOTO	C
CNDC7E 04.4.40.004.3000 70 400 4/3/4 00								C C C C C C C	TTTTCAARACACOC				
SIDE STERN A 53-004 4000 TO ACC 4/04 00								C CACICA	TTTTCAAAACACAC	ATAAAAA	ADDAAADADAAA		
SNP 57ER4. 7.53.694.4004 754004 277-50						00.		CT CT CACT CA	CARAACAC	ALAAAAA	AUTAAAUAUAA		
SIDPSTER4.109.1367.888764000000						I LAAT-	····	CHITCHCACTCA	CARAACAC	ATARAAAT	TAU PARAPATAR		
SNPSTER4:7:13:1473:90401GACC4/2/1-90			CCCTTOAC	ATAA TAATAL		AA -	A A A A A A A A A A A A A A A A A A A	CICACICA	TTT CARACACAC	ATAAA			
SNPSTER4:1:26:179:1998#TGACCA/2/1-90			CCCHINAC	TATAATAATAC	CADODA	AAT-	· · · · · · · · · · · · · · · · · · ·	CHITCHCACTCA	TITICAAAACACAA	ATAAA			
SMPSTER4:1:32:687:1613#TGACCA/2/1-90				· · · · · · · · · · · · · · · AC	CABBBAT	DAAT-	····	CHITTCHCACTCA	TTTTCAAAACACOC	ALAAAAA	AUUAAAUAUAA		
SNPSTER4:1:39:1321:61007/SACCA/2/1-90					CAUDUA	TTUAAT-	· · · ·	CHITTELCACICAL	TITTCAAAACAC	ATAAAAAT	LAGGAAAGAGAGAA		
SNPSTER4.1.12.676.1877#1GACCAV21-90				· · · · · · · · · · · · AC	L'AD DO A	T PAAL	A TAAT ILL	CHITTCTCACTCA	TITTCAAAACACAC	ATAAAAA	AUUAAAUAUAA		
SNPSTER4:1:39:1609:2023#7GACCA/2/1-90				AC	CADODAT	TIDAAT-	· · · · · · · · · · · · · · · · · · ·	CHITTELCACICA	TITICAAAACACOC	ATAAAAAT	TADDAAAAGADAA		CTT
SNPSTER4:1:67:1681:1316#TGACCA/2/1-90				· · · · · · · · · · · · AC	CADODAT	TTOAAT -	····	CTTTTTCTCACTCAT	TTTTCAAAACACOC	ATAAAAAT	TADUAAADADAA		C11
SNPSTER4:1:64:1489:1623#7GACCA/2/1-90				· · · · · · · · · · · · · · · · AC	CANDUAT	TIDAAT-	····	CHITTELCACTCAT	TTTTCAAAACACAC	ATAAAAAT	TAGGAAAAGADAA		C
SNPSTER4:1:85:1309:1170#TGACCA/2/1-90				• • • • • • • • • • • ACT	CADDDAT	TTOAAT-	····	CTTTTTCTCACTCAT	TTTTCAAAACAC	ATAAAAAT	TAD BAAABABAAA		CTT
SNPSTER4:1:88:1491:1823#TGACCA/2/1-90				AC	CASSAT	TIDAAT-	····	CTITITCTCACTCAT	TITTCAAAACACHC	ATAAAAAT	TADDAAADADAA	THEFT	CTT
SNPSTER4:1:37:211:499#TGACCA/1/1-90	· · · · · · · · · · · · · · · · · · ·	O TADEATATO	CCCTTMAC	TATAATOAATAC	CABBBAT	TTDAAT-		CTITITCTCACTCA	TTTTCAAAACACOC	ATAA			********
SNPSTER4:1:48:1773:738#TGACCA/2/1-90				ATAC	CABODAT	TIOAAT-	····	CHITTETCACTCAT	TTTTCAAAACAC	ATAAAAAT	TADDAAADADAA	TIGITIECTC	C
SNPSTER4:1:78:1676:627#TGACCA/2/1-90		* * * * * * * * * *	· · · · TOAC	TATAATHAATACI	TCADODAT	TTOAAT -	····	ICTITITCTCACTCAN	TTTTCAAAACAC	ATAAAAAT	TADDAAADADA-		* * * * * * * * * * * *
SNPSTER4:1:78:724:1662#TGACCA/2/1-90				TATAATBAATACI	TCABODAT	TTOAAT-	····	ICT TTTTTCTCACTCAN	TTTTCAAAACACOC	ATAAAAAI	TADDAAADADA-	*********	**********
SNPSTER4:1:7:1007:1087#TGACCA/1/1-90		OTADBATATO	CCCTTOAC	TATAATOAATACI	CADODAT	TIDAAT-		CTTTTTCTCACTCAT	TTTCAAAACAC	ATAA			* * * * * * * * * * * *
SNPSTER4:1:86:576:1259#7GACCA/1/1-90	CCARACAAD TTTOT	U TAUDATAT D	CCCTTOAC	TATAATOAATACI	ECABODAT	TTOAAT-		CTTTTTTCTCACTC					*********
SNPSTER4:1:21:1778:1045#TGACCA/1/1-90	••••••••••••	<b>GTADDATATE</b>	CCCTTMAC	TATAATAATAC	TADODAT	TIDAAT-		CITTUTCICACICA		ATAA			
SNPSTER4:1:60:421:632#TGACCA/2/1-90				ACT	TCABOOAT	TTOAAT-	····	ICTTTTTCTCACTCAT	TTTTCAAAACACOC	ATAAAAAT	TTAO DAAADAAAA		CTT
SMPSTER4:1:18:1045:1894#TGACCA/2/1-90				ACT	CADODAT	TTOAAT-		CITTITCICACTCAT		ATAAAAAT	ITAD GAAAQADAT	TINTICIC	<u>C 1 1</u>
SMPSTER4:1:87:1399:729#TGACCA/2/1-90				ACT	TCABOOAT	TTUAAT-	····	CITITICICACICAT	TTTTCAAAACACOC	ATAAAAAT	ITAGO AAADADAA	TINTICIC	<b></b>
SMPSTER4:1:32:461:1967#TGACCA/2/1-90				AC	CADODAT	TTOAAT-	····	CTTTTTCTCACTCAT	TTTTCAAAACAC	ATAAAAAT	I TADUAAADADAA	TTOTTTTTC	CTT
SMPSTER4:1:64:1057:712#TGACCA/2/1-90				ACT	TCABBOAT	TTUAAT-	····	ICTTTTTCTCACTCAT	TTTTCAAAACACOC	ATAAAAAT	ITADO BAADADAA	TTOTTTCTC	CTT
SMPSTER4:1:78:645:1579#TGACCA/2/1-90				AC	TCADODAT	TTOAAT-	····	ICTTTTTCMCACTCAT	TTTTCAAAACACOC	ATAAAAAT	I TAU GAAAGADAA	TT <mark>OTITICIC</mark>	CTT
SNPSTER4:1:82:1356:376#TGACCA/2/1-90				AC	TCADODAT	TTOAAT-	····	CITITCICACICAT	TTTTCAAAACACDC	ALAAAAA	TAD CAAAO ABAA	TTOTTTTNIC	C
SMPSTER4:1:85:88:617#TGACCA/1/1-90	CCADACAABITIDII	CTADEATATE	CNCTTOAC	TATAATAATAATACI	TCADDO AT	TTOAAT-	····						
SMPSTER4:1:96:1740:1829#TGACCA/2/1-90			TOAC	TATAAT AATACT	TCABBBAT	TTOAAT-	···· OTAATTOCT	CTTTTTCTCACTCAT	TTTTCAAAACAC	ATAAAAAT	TADDAAADADA-		
SMPSTER4:1:37:1295:434#TGACCA/2/1-90			TOAC	TATAATOAATACI	T CABOOAT	TTOAAT-	· · · · · · · · · · · · · · · · · · ·	CTTTTTCTCACTCAT	TTTTCAAAACACOC	ATAAAAAT	TTAD BAAADAAA		
SMPSTER4:1:97:604:750#7GACCA/2/1-90			TOAC	TATAATGAATACI	TCAGOGAT	TTOAAT-	· · · • • • • • • • • • • • • • • • • •	CTTTTTCTCACTCAT	TTTTCAAAACAC	ATAAAAAT	TADOAAAGADA-		
SNPSTER4:3:71:1068:1401#TGACCA/1/1-38				AAT AATAC	T CABODAT	TTLAAT -		CTTT					
g/[89161218/w/]NC_000023.9[NC_000023_133460000tb133460600/	1-601CCAUACAAUTTTett	GTADEATATO	CCCTTOAC	TATAATAATAC	CADOOAT	TTOAATO	TAASTAATTOCTI	CTITICCACTCAT	TTTTCAAAACAC	ATAAAAAT	ADDAAADAAAA	TTUTTTCTC	CITCCALCACC
Conse	nsus				100					-			
	CCAGACAAGTTTGTT	GTAGGATATO	CCCTTGAC	TATAATGAATACT	TCAGGGATI	TTGAATG	··· · · · · · · · · · · · · · · · · ·	CTTTTTCTCACTCAT	TTTTCAAAACACGC	ATAAAAATI	TAGGAAAGAGAA	TOTTTTCTC	CTTCCAGCAC -

# 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



#### 23andMe is the best place to take a personalized journey through your DNA.

23andMe offers access to the following, at one price with NO subscriptions:

- Over 200 online health & traits reports
- The largest genealogical DNA database in the world
- Updates on new genetic discoveries that are personalized for your DNA by our experts

\$299 add to cart

Ships in 1-2 business days. Expect results 2-3 weeks after we receive your sample.



#### 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