A minicourse on

Genomewide association studies (GWAS)
Part II: Analysis

Matti Pirinen
FIMM, University of Helsinki

December 5th 2012

Contents:

1.Concepts And Rationale

2.Technology and Data

3.Statistics and Analysis

4.Results and Hot topics

- Digestive system disorder
- Cardiovascular disorder
- Metabolic disorder
- Immune system disorder
- Neurological disorder
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurement
- Other measurement
- Chemical compound
- Biological process
- Cancer
- Other disease
- Other trait
- Trait mapping in progress

EMBL-EBI

National Human Genome Research Institute

GWAS Statistics

- Linear model, ML-estimates, Wald's test, p-value, power (R EXAMPLE), odds (WTCCC 2007 BOX on significance)
- Bayesian model, marginal likelihood, Bayes factor (BF) (Wakefield 2009, Vukcevic 2009 Ch. 5&6) (R EXAMPLE)

Effect size

- Quantitaitve: phenotypic scale (e.g., cm, std dev)
- Disease: odds-ratio (~relative risk for rare diseases)
  - Logistic regression
- Meta-analysis: Combining results over studies
  - Inverse variance weighting
  - Forest plots
- Effect size is not (the only) measure of importance of a GWAS finding! (Lander 2011, BOX 1: HMGCR and statins)

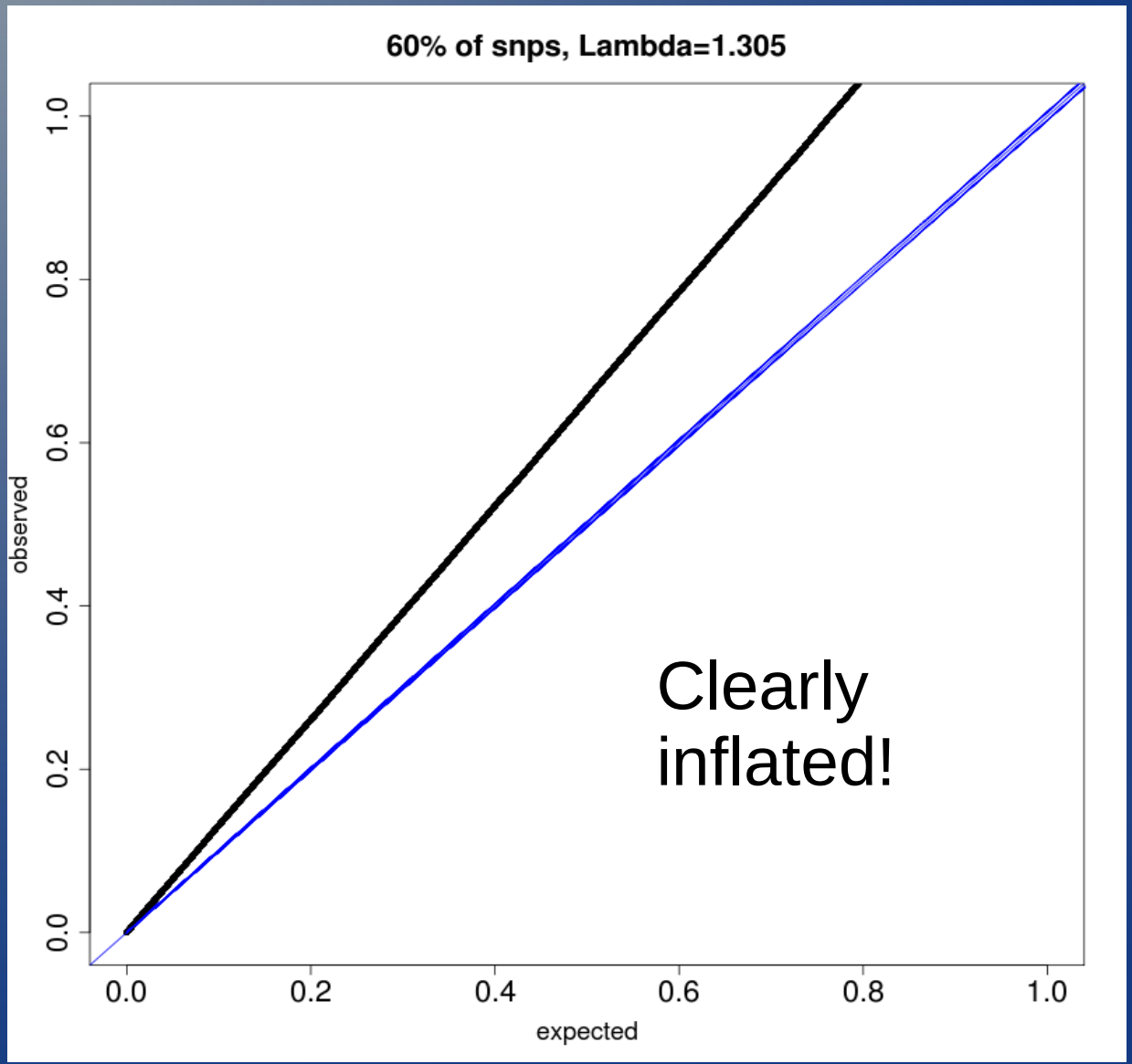
QQ-plot

- R EXAMPLE: 3 QQ-plots
  - NULL, inflated and exciting
- Inflation (measure lambda) arises because of
  - Polygenic effects AND / OR
  - Ascertainment bias (in case-control studies)

Population structure in case-control association studies

- Are there differences in genotype frequencies between cases and controls?
  - If yes, then locus is possibly interesting
  - But could also reflect ascertainment scheme if cases and controls are not well matched w.r.t. genetic background!
  - Example: Let's look at analysis where 5,000 UK controls are compared against 1,800 UK + 500 Irish Psoriasis cases

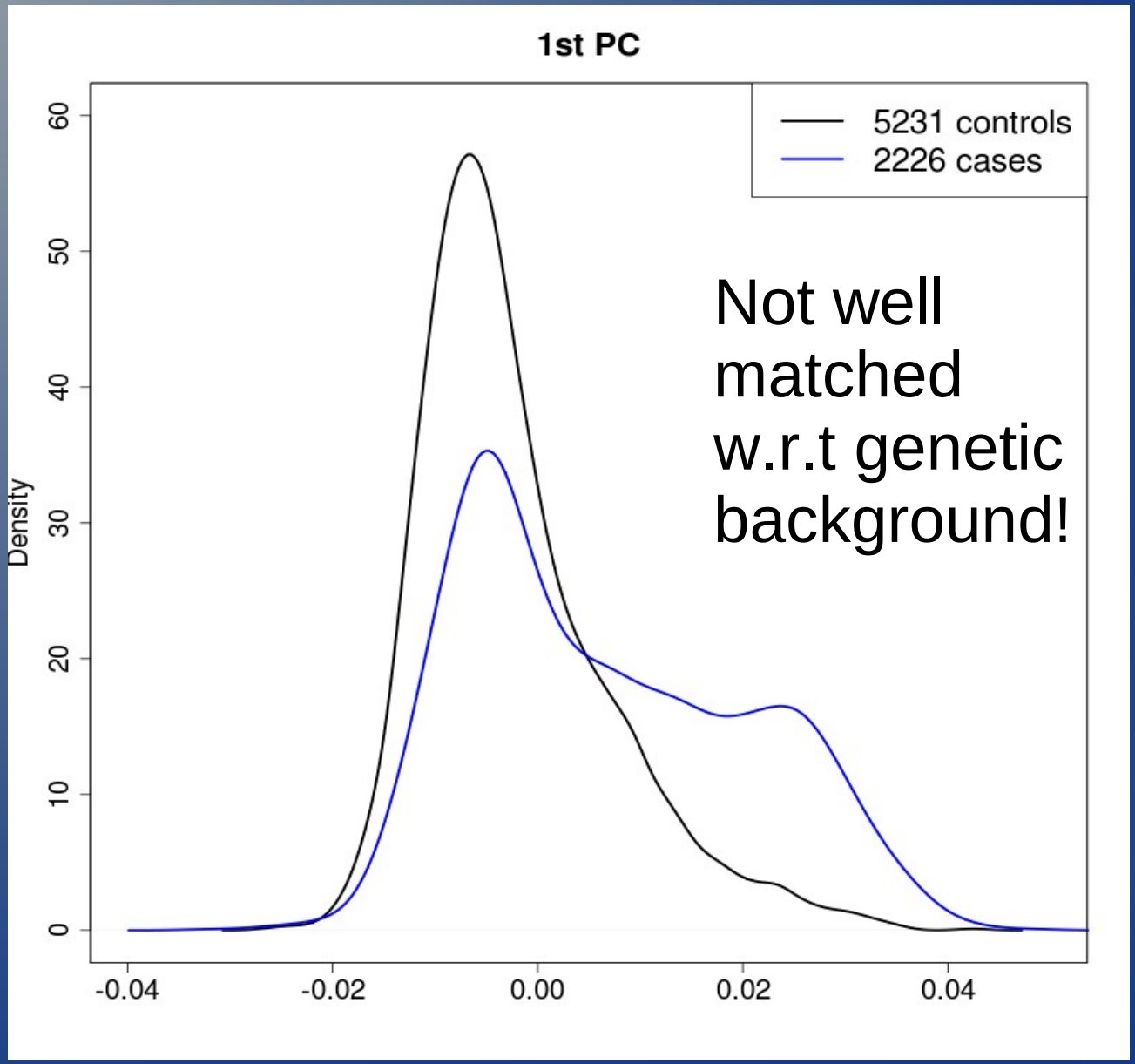
Plot association statistics over the genome and compare to the expectation under the null



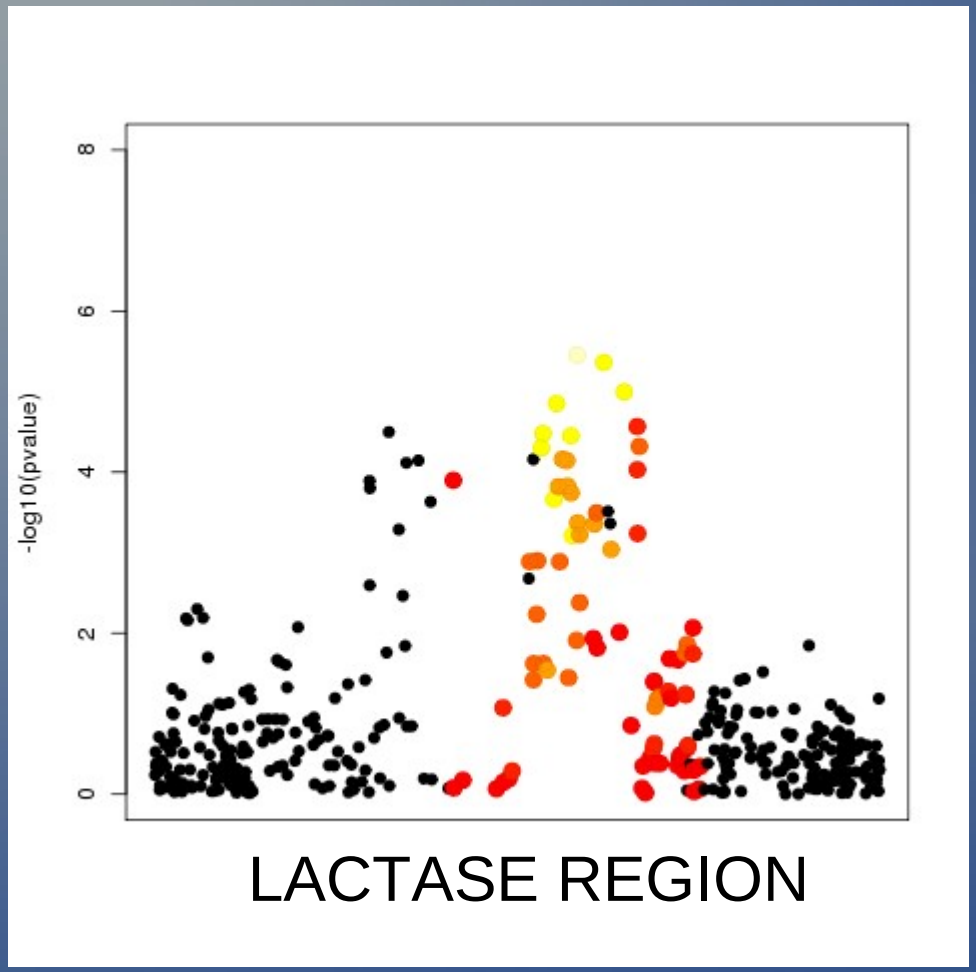
Principal components of population structure

- Find linear combination of SNPs that has the largest sample variance (1st PC)
  - Sequentially, find further PCs ORTHOGONAL to the previous ones
- Novembre 2008: European PC plot
- R EXAMPLE

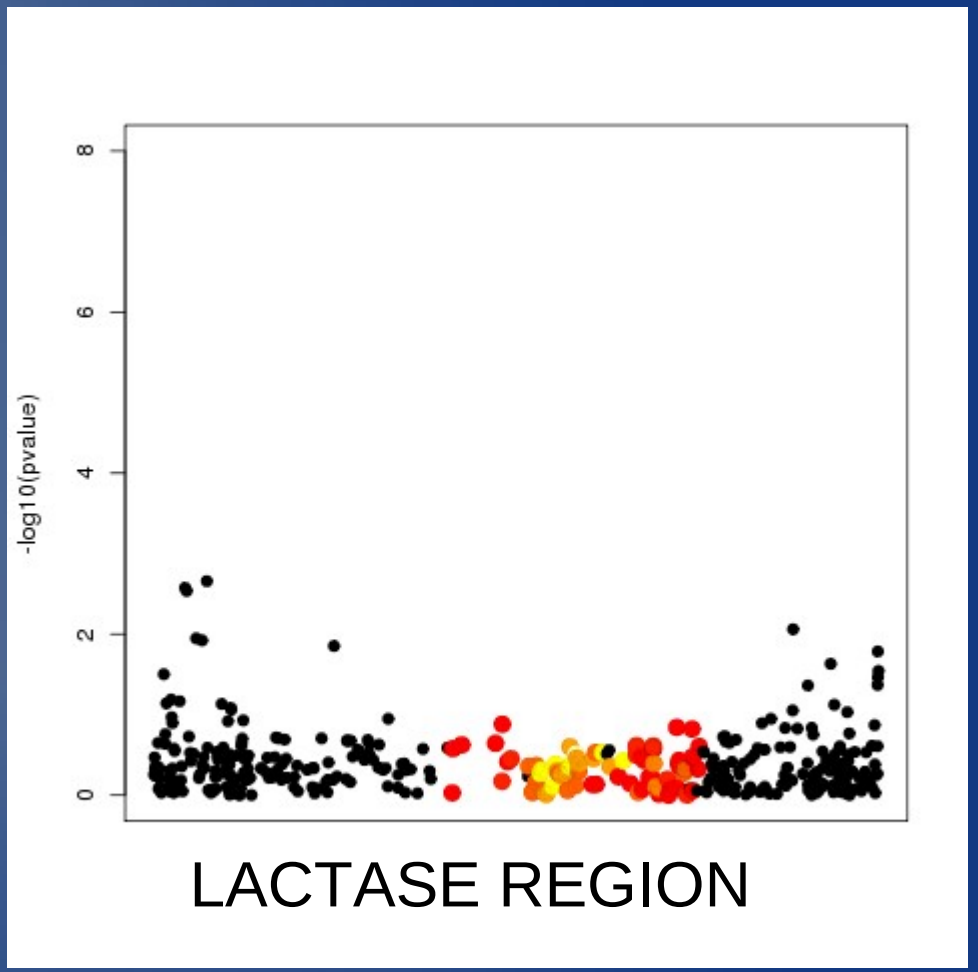
PCA on UK+Irish Psoriasis study



Basic analysis



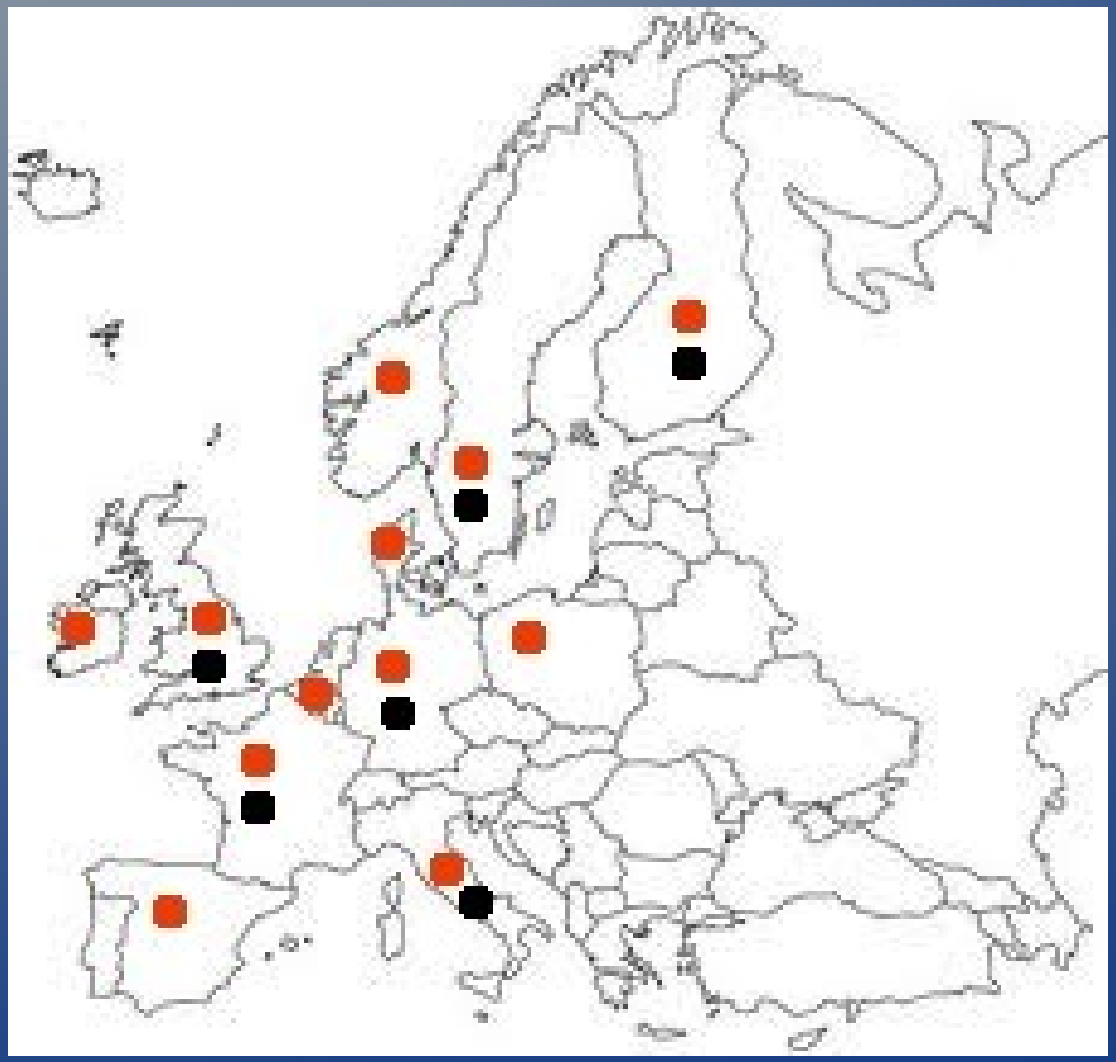
First PC as covariate



SNP - - - ? - - - → Phenotype

Population

Multiple sclerosis: 10,000 cases/ 17,000 controls

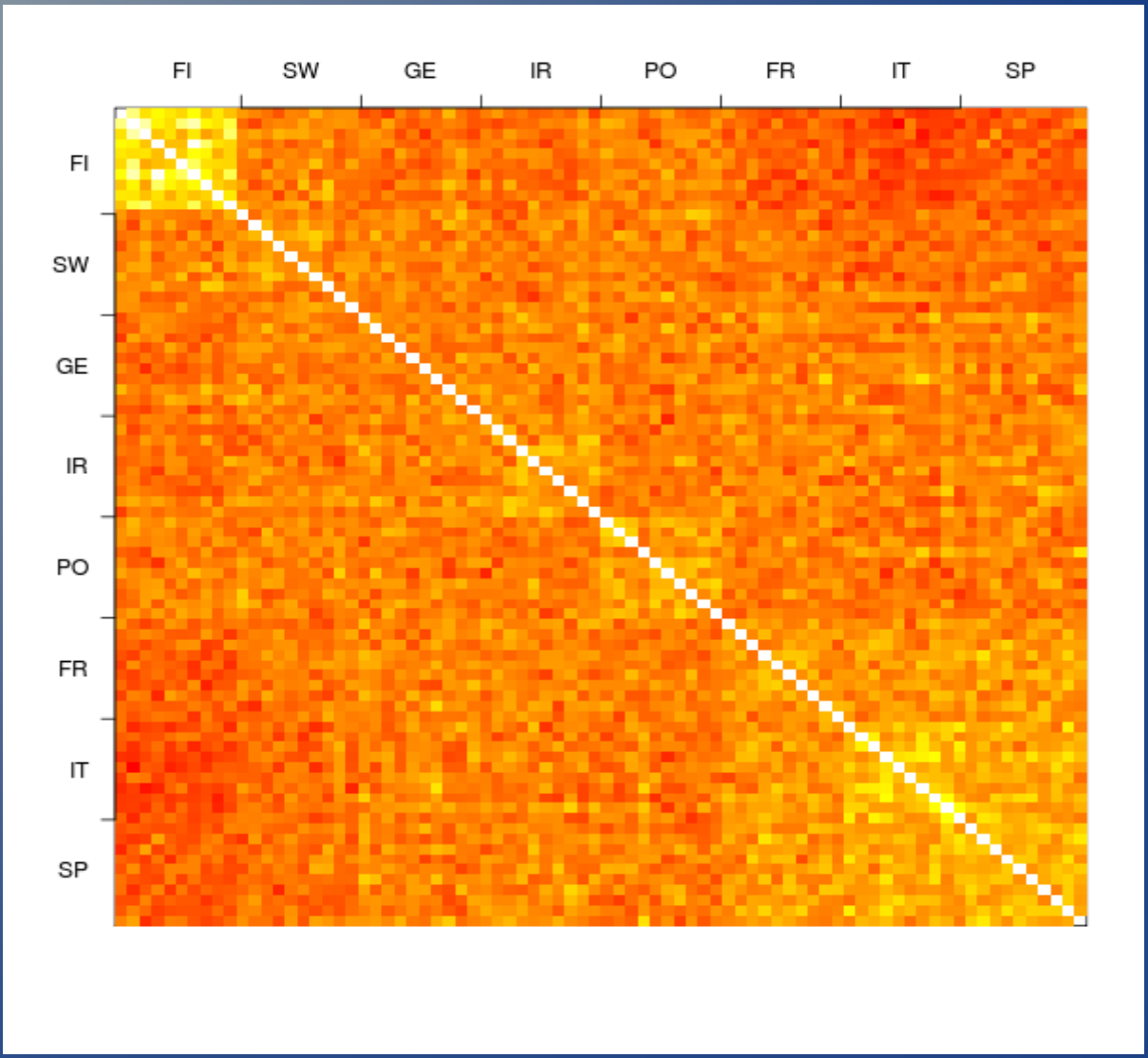


Pop	Cases	Controls
Spain	205	--
Norway	953	121
UK	1854	5175
Sweden	685	1928
France	479	347
Poland	58	--
Denmark	332	--
Belgium	544	--
Italy	745	571
Germany	1100	1699
Ireland	61	--
Finland	581	2165
Australia	648	--
NZ	146	--
USA	1383	5370

Possible strategies

- Each country analysed separately and results combined (loses samples and power due to differing case-control ratios)
- Clustering individuals into genetically homogeneous groups which are analysed separately and results combined
- Regression models on whole data with population structure included in the model
  - A few leading principal components of the genetic structure
  - Linear mixed model

Genome-wide correlation matrix (R)



LMM: $Y \sim \mu + G\beta + z$, where $E(z)=0$ and $var(z) \propto R$

PC: $Y \sim \mu + G\beta + P\gamma$, where P is the leading eigenvector of R

Linear Mixed model

- In the standard linear model it is computationally possible to introduce individual specific effects that follow empirical genetic correlations (R matrix)
- $Y = \alpha + G\beta + Z + \varepsilon$, where
    - $Z \sim N(0, h\sigma^2 R)$ $\varepsilon \sim N(0, (1-h)\sigma^2 I)$
- Is this appropriate for case-control data?

Linear model with case-control data

- Likelihood ratio statistic approaches the Armitage trend test
- Least-squares applied to binary data approximates the ML-estimates of logistic regression when effects are small
  - Relative error in log-odds estimates is less than 0.5% in typical GWAS application

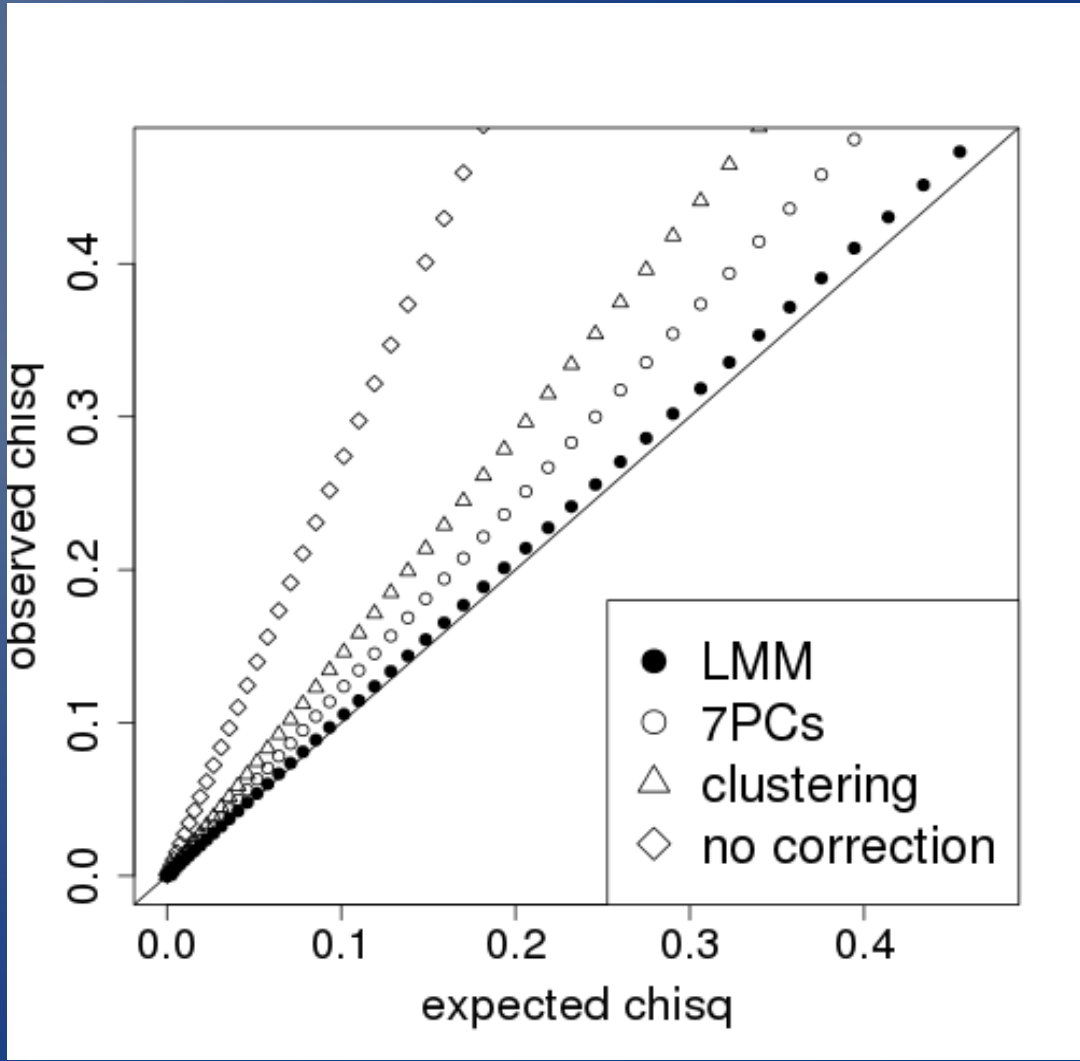
QQ-plots

λ =1.04
Linear mixed model

λ =1.22
Logistic regression with 7 PCs as covariates

λ =1.44
Clustering individuals to homogeneous groups

λ =2.7
No structure corrections



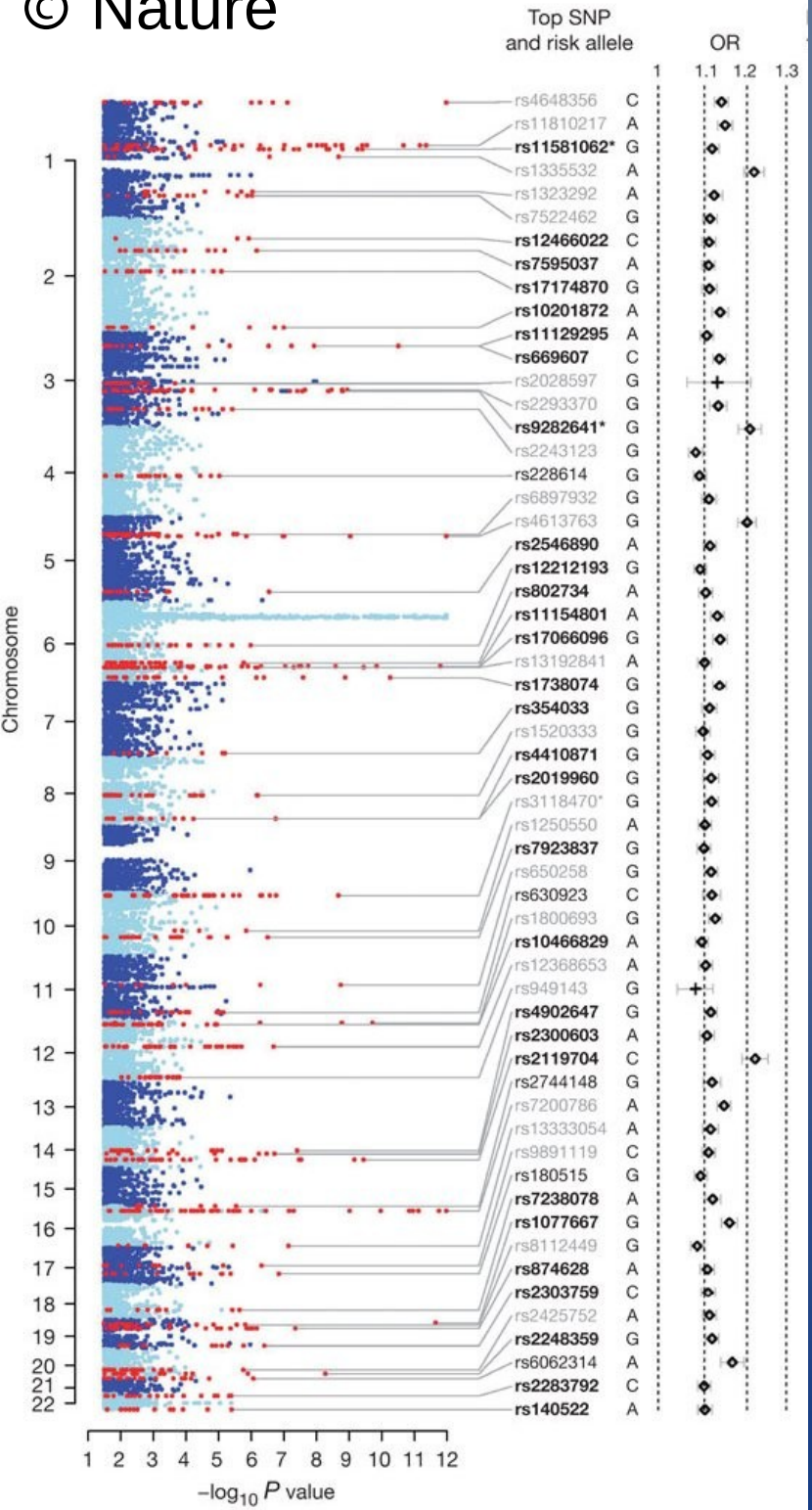
We also applied

- A beta-binomial model to exclude variants that differ among control data sets more than expected under no selection
- Empirical allele variation scores
- 1000 Genomes database to see whether our hits tag strongly differentiated variants
- Bayesian hierarchical model to assess whether the effect sizes differ between the populations

Replication

- We took 102 SNPs to replication based on LMM analysis
- Replication attempted in silico from meta-analysis of 6 previous MS-studies having together 4,200 cases and 7,300 controls
- Logistic regression with 10 PCs
- Carried out at the Broad Institute

© Nature



- 98 /102 SNPs have consistent effects in replication data
- 29 novel: p<5e-8 (combined), p<1e-4.5 (discovery), p<0.05 (replication, 1-sided)
- plus 5 novel: p<5e-7 (combined), p<1e-4.5 (discovery), p<0.05 (replication, 1-sided)
- 23 / 26 previously reported MS loci had p<1e-3 in our discovery data
- Immunological genes are overrepresented among the hits; in particular, T-helper cell differentiation pathway

EXAMPLE GWAS: Multiple sclerosis

- WTCCC2 2011
- http://wattle.well.ox.ac.uk/wtccc2/external/ms/
- Over 50 associations now known
  - Primary role for immune mechanism
  - T-helper cell differentiation
  - 2 GWAS hits close to known drug targets, what about remaining ~50 loci!?

Further study on one SNP

- Reported before: Tumour necrosis factor (TNF) blocking drugs are bad for MS patients
- Gregory et al. (July 2012 NATURE) “TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis”
  - MS risk variant at rs1800693 (one of MS GWAS hits) promote TNF-blocking effect
  - Consistent effect between GWAS findings and empirical experience with a drug

WTCCC2 Oxford

Chris Spencer

Garrett Hellenthal

Peter Donnelly

Stephen Sawcer[1]*, Garrett Hellenthal[2]*, Matti Pirinen[2]*, Chris C. A. Spencer[2]*, Nikolaos A. Patsopoulos[3,4,5], Loukas Moutsianas[6], Alexander Dilthey[6], Zhan Su[2], Colin Freeman[2], Sarah E. Hunt[7], Sarah Edkins[7], Emma Gray[7], David R. Booth[8], Simon C. Potter[7], An Goris[9], Gavin Band[2], Annette Bang Oturai[10], Amy Strange[2], Janna Saarela[11], Céline Bellenguez[2], Bertrand Fontaine[12], Matthew Gillman[7], Bernhard Hemmer[13], Rhian Gwilliam[7], Frauke Zipp[14,15], Alagurevathi Jayakumar[7], Roland Martin[16], Stephen Leslie[17], Stanley Hawkins[18], Eleni Giannoulatou[2], Sandra D'alfonso[19], Hannah Blackburn[7], Filippo Martinelli Boneschi[20], Jennifer Liddle[7], Hanne F. Harbo[21,22], Marc L. Perez[7], Anne Spurkland[23], Matthew J. Waller[7], Marcin P. Mycko[24], Michelle Ricketts[7], Manuel Comabella[25], Naomi Hammond[7], Ingrid Kockum[26], Owen T. McCann[7], Maria Ban[1], Pamela Whittaker[7], Anu Kemppinen[1], Paul Weston[7], Clive Hawkins[27], Sara Widaa[7], John Zajicek[28], Serge Dronov[7], Neil Robertson[29], Suzannah J. Bumpstead[7], Lisa F. Barcellos[30,31], Rathi Ravindrarajah[7], Roby Abraham[27], Lars Alfredsson[32], Kristin Ardlie[4], Cristin Aubin[4], Amie Baker[1], Katharine Baker[29], Sergio E. Baranzini[33], Laura Bergamaschi[19], Roberto Bergamaschi[34], Allan Bernstein[31], Achim Berthele[13], Mike Boggild[35], Jonathan P. Bradfield[36], David Brassat[37], Simon A. Broadley[38], Dorothea Buck[13], Helmut Butzkueven[39,40,41,42], Ruggero Capra[43], William M. Carroll[44], Paola Cavalla[45], Elisabeth G. Celius[21], Sabine Cepok[13], Rosetta Chiavacci[36], Françoise Clerget-Darpoux[46], Katleen Clysters[9], Giancarlo Comi[20], Mark Cossburn[29], Isabelle Cournu-Rebeix[12], Mathew B. Cox[47], Wendy Cozen[48], Bruce A. C. Cree[33], Anne H. Cross[49], Daniele Cusi[50], Mark J. Daly[4,51,52], Emma Davis[53], Paul I.W. de Bakker[3,4,54,55], Marc Debouverie[56], Marie Beatrice D'hooghe[57], Katherine Dixon[53], Rita Dobosi[9], Bénédicte Dubois[9], David Ellinghaus[58], Irina Elovaara[59,60], Federica Esposito[20], Claire Fontenille[12], Simon Foote[61], Andre Franke[58], Daniela Galimberti[62], Angelo Ghezzi[63], Joseph Glessner[36], Refujia Gomez[33], Olivier Gout[64], Colin Graham[65], Struan F. A. Grant[36,66,67], Franca Rosa Guerini[68], Hakon Hakonarson[36,66,67], Per Half[69], Anders Hamsten[70], Hans-Peter Hartung[71], Rob N. Heard[8], Simon Heath[72], Jeremy Hobart[28], Muna Hoshi[13], Carmen Infante-Duarte[73], Gillian Ingram[29], Wendy Ingram[28], Talat Islam[48], Maja Jagodic[26], Michael Kabesch[74], Allan G. Kermode[44], Trevor J. Kilpatrick[39,40,75], Cecilia Kim[36], Norman Klopp[76], Keijo Koivisto[77], Malin Larsson[70], Mark Lathrop[72], Jeannette S. Lechner-Scott[47,78], Maurizio A. Leone[79], Virpi Leppä[11,80], Ulrika Liljedahl[81], Izaura Lima Bomfim[26], Robin R. Lincoln[33], Jenny Link[26], Jianjun Liu[82], Aslaug R. Lorentzen[22,83], Sara Lupoli[50,84], Fabio Macciardi[50,85], Thomas Mack[48], Mark Marriott[39,40], Vittorio Martinelli[20], Deborah Mason[86], Jacob L. McCauley[87], Frank Mentch[36], Inger-Lise Mero[21,83], Tania Mihalova[27], Xavier Montalban[25], John Mottershead[88,89], Kjell-Morten Myhr[90,91], Paola Naldi[79], William Ollier[53], Alison Page[92], Aarno Palotie[7,11,93,94], Jean Pelletier[95], Laura Piccio[49], Trevor Pickersgill[29], Fredrik Piehl[26], Susan Pobywajlo[5], Hong L. Quach[30], Patricia P. Ramsay[30], Mauri Reunanen[96], Richard Reynolds[97], John D. Rioux[98], Mariaemma Rodegher[20], Sabine Roesner[16], Justin P. Rubio[39], Ina-Maria Rückert[76], Marco Salvetti[99], Erika Salvi[50,100], Adam Santaniello[33], Catherine A. Schaefer[31], Stefan Schreiber[58,101], Christian Schulze[102], Rodney J. Scott[47], Finn Sellebjerg[10], Krzysztof W. Selmaj[24], David Sexton[103], Ling Shen[31], Brigid Simms-Acuna[31], Sheila Skidmore[1], Patrick M. A. Sleiman[36,66], Cathrine Smestad[21], Per Soelberg Sørensen[10], Helle Bach Søndergaard[10], Jim Stankovich[61], Richard C. Strange[27], Anna-Maija Sulonen[11,80], Emilie Sundqvist[26], Ann-Christine Syvänen[81], Francesca Taddeo[100], Bruce Taylor[61], Jenefer M. Blackwell[104,105], Pentti Tienari[106], Elvira Bramon[107], Ayman Tourbah[108], Matthew A. Brown[109], Ewa Tronczynska[24], Juan P. Casas[110], Niall Tubridy[40,111], Aiden Corvin[112], Jane Vickery[28], Janusz Jankowski[113], Pablo Villoslada[114], Hugh S. Markus[115], Kai Wang[36,66], Christopher G. Mathew[116], James Wason[117], Colin N. A. Palmer[118], H-Erich Wichmann[76,119,120], Robert Plomin[121], Ernest Willoughby[122], Anna Rautanen[2], Juliane Winkelmann[13,123,124], Michael Wittig[58,125], Richard C. Trembath[116], Jacqueline Yaouanq[126], Ananth C. Viswanathan[127], Haitao Zhang[36,66], Nicholas W. Wood[128], Rebecca Zuvich[103], Panos Deloukas[7], Cordelia Langford[7], Audrey Duncanson[129], Jorge R. Oksenberg[33], Margaret A. Pericak-Vance[87], Jonathan L. Haines[103], Tomas Olsson[26], Jan Hillert[26], Adrian J. Ivinson[51,130], Philip L. De Jager[4,5,51], Leena Peltonen†, Graeme J. Stewart[8], David A. Hafler[4,131], Stephen L. Hauser[33], Gil McVean[2], Peter Donnelly[2,6]* & Alastair Compston[1]*

Thank you

- 245 authors with 140 affiliations around the world
  - MS disease groups (including Finland)
  - Genotyping at Sanger institute
  - Replication analyses at the Broad institute
- Stephen Sawcer, and Alastair Compston, Clinical Neuroscience, Cambridge, UK

References

- **IMSGC & WTCCC2.** (Aug 2011) “Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis.” **Nature.**
- **Pirinen, Donnelly, Spencer. (2012)** “Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies.” (available online) **Annals of Applied Statistics.**
  - **Binary data**
  - **Algorithmic improvement from $O(n^3)$ to $O(n^2)$**
  - **Bayesian computation**

Summaries of GWAS

- YOUTUBE: “Genome-Wide Association Studies - Karen Mohlke (2012)” ~90 minutes
- Paper: Pearson, Manolio (2008) “How to Interpret a Genome-wide Association Study?” JAMA 299(11):1335-1344