MSc GBE Course:

Genes: from sequence to function

Genome-wide Association Studies

Sven Bergmann
Department of Medical Genetics
University of Lausanne
Rue de Bugnon 27 - DGM 328
CH-1005 Lausanne
Switzerland

work: ++41-21-692-5452
cell: ++41-78-663-4980
http://serverdgm.unil.ch/bergmann

Overview

- Population stratification
- Associations: Basics
- Whole genome associations
- Genotype imputation
- Uncertain genotypes
- New Methods

Overview

- Population stratification
- Associations: Basics
- Whole genome associations
- Genotype imputation
- Uncertain genotypes
- New Methods

CoLaus = Cohort Lausanne

6'189 individuals

Genotypes

Phenotypes

500.000 SNPs

159 measurement
144 questions

Collaboration with:
Vincent Mooser (GSK), *Peter Vollenweider & Gerard Waeber (CHUV)*

Genetic variation in SNPs (**S**ingle **N**ucleotide **P**olymorphisms)

ATTGCAATCCGTGG...ATCGAGCCA...TACGATTGCACGCCG...

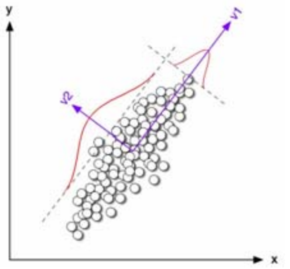
ATTGCAAGCCGTGG...ATCTAGCCA...TACGATTGCAAGCCG...

ATTGCAAGCCGTGG...ATCTAGCCA...TACGATTGCAAGCCG...

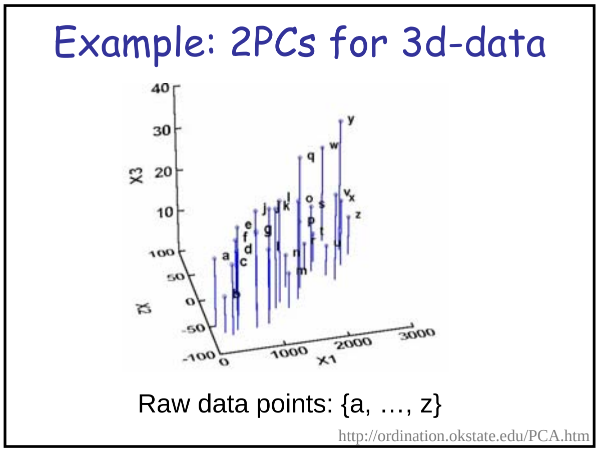
ATTGCAATCCGTGG...ATCGAGCCA...TACGATTGCACGCCG...

ATTGCAAGCCGTGG...ATCTAGCCA...TACGATTGCAAGCCG...

Analysis of Genotypes only



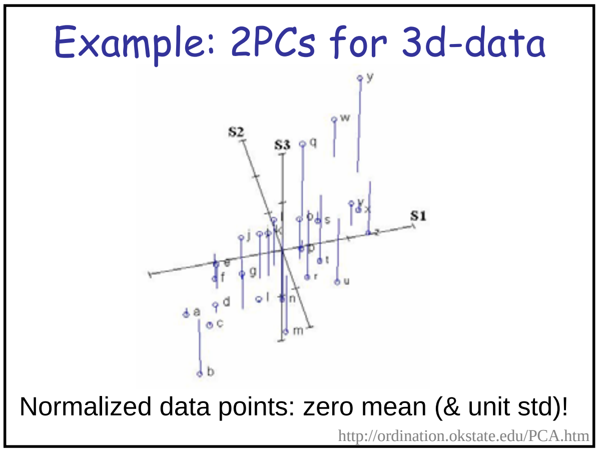
Principle Component Analysis reveals SNP-vectors explaining largest variation in the data



Example: 2PCs for 3d-data

Raw data points: {a, …, z}

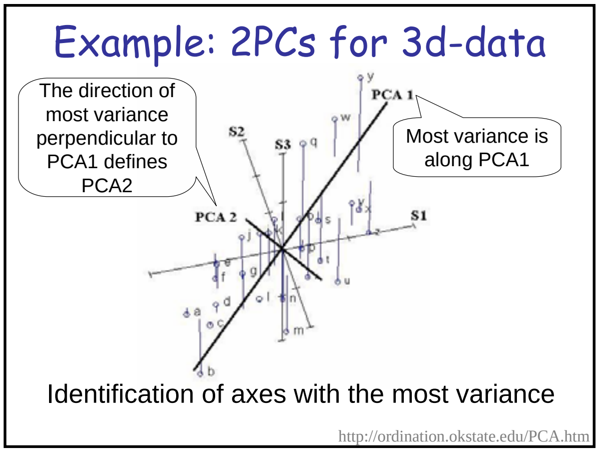
http://ordination.okstate.edu/PCA.htm



Example: 2PCs for 3d-data

Normalized data points: zero mean (& unit std)!

http://ordination.okstate.edu/PCA.htm



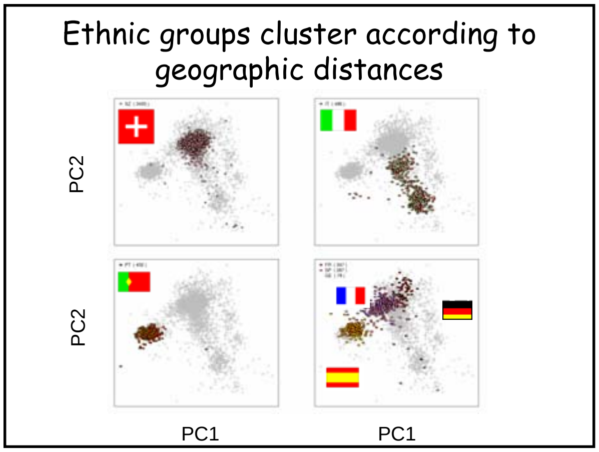
Example: 2PCs for 3d-data

The direction of most variance perpendicular to PCA1 defines PCA2

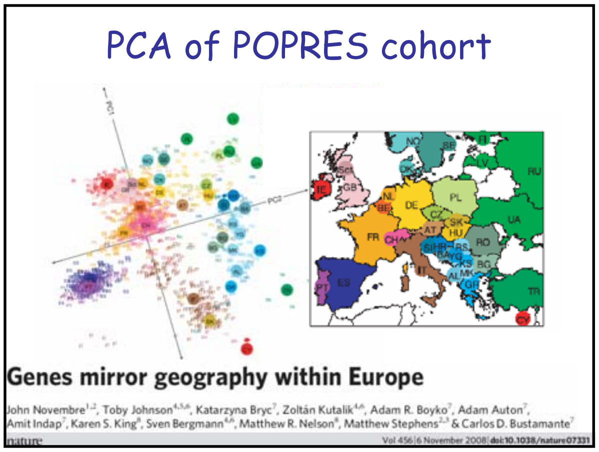
Most variance is along PCA1

Identification of axes with the most variance

http://ordination.okstate.edu/PCA.htm



Ethnic groups cluster according to geographic distances



PCA of POPRES cohort

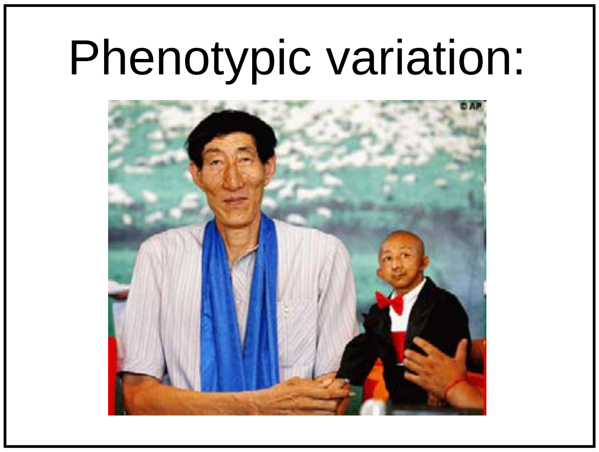
Genes mirror geography within Europe

John Novembre[1,2], Toby Johnson[4,5,6], Katarzyna Bryc[7], Zoltán Kutalik[4,6], Adam R. Boyko[7], Adam Auton[7], Amit Indap[7], Karen S. King[8], Sven Bergmann[4,6], Matthew R. Nelson[8], Matthew Stephens[2,3] & Carlos D. Bustamante[7]

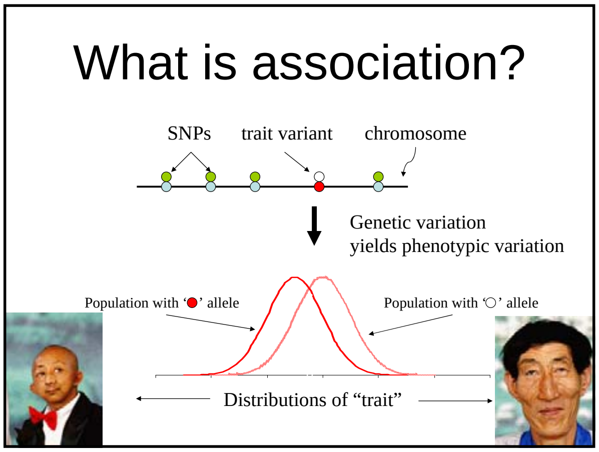
nature Vol 456|6 November 2008|doi:10.1038/nature07331

Overview

- Population stratification
- Associations: Basics
- Whole genome associations
- Genotype imputation
- Uncertain genotypes
- New Methods

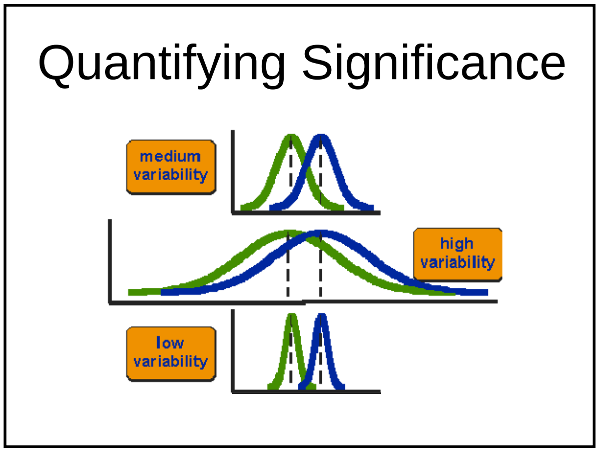


Phenotypic variation:

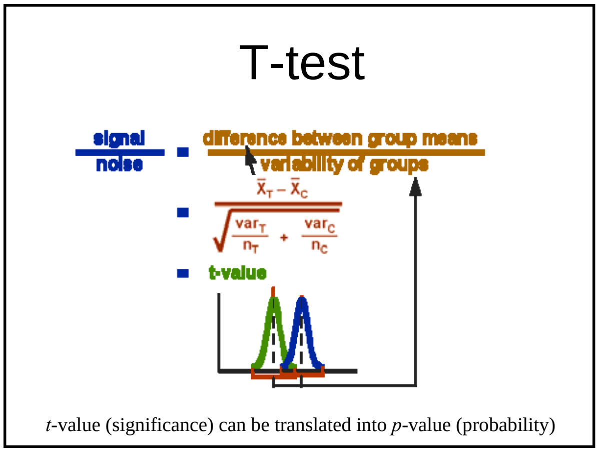


What is association?

SNPs
trait variant
chromosome
Genetic variation
yields phenotypic variation
Population with ‘●’ allele
Population with ‘○’ allele
Distributions of “trait”



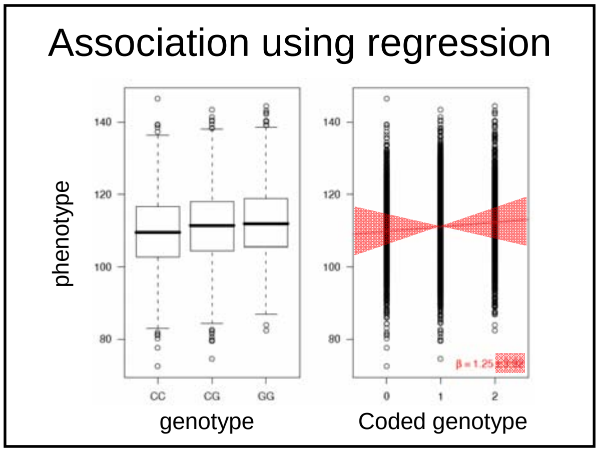
Quantifying Significance



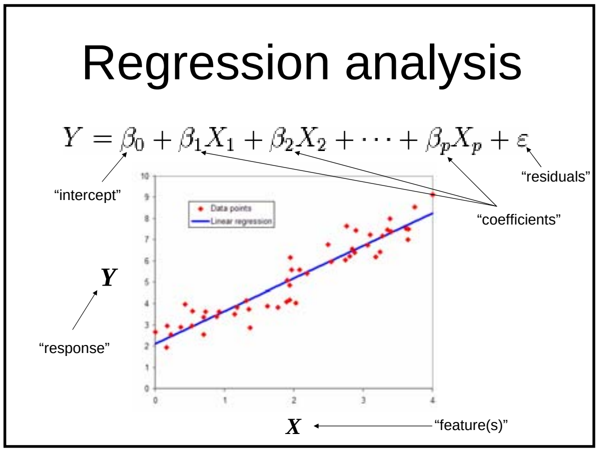
T-test

$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}} = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}} = \text{t-value}$$

t-value (significance) can be translated into *p*-value (probability)

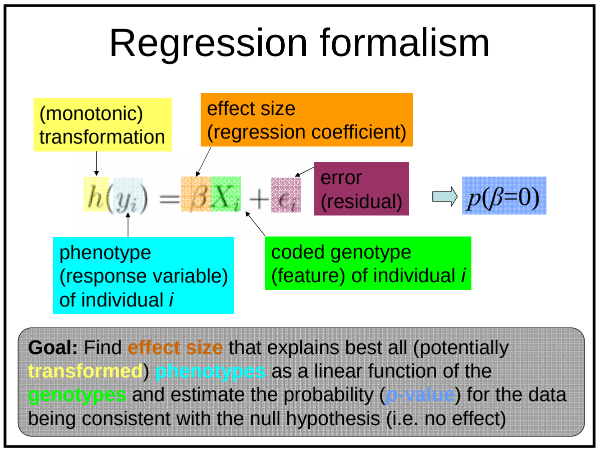


Association using regression



Regression analysis

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$



Regression formalism

(monotonic) transformation

effect size (regression coefficient)

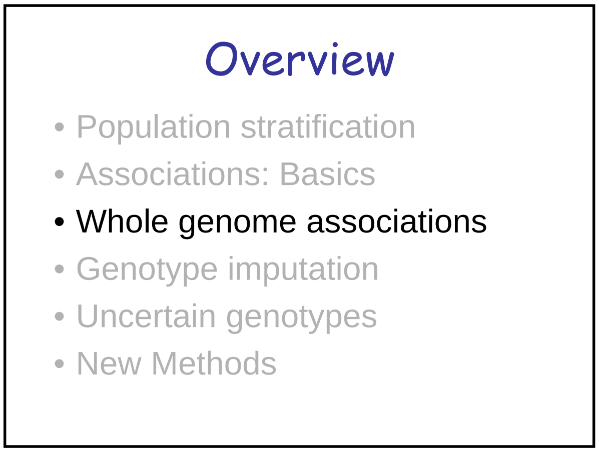
error (residual)

$h(y_i) = \beta X_i + \epsilon_i$ ⇨ $p(\beta=0)$

phenotype (response variable) of individual *i*

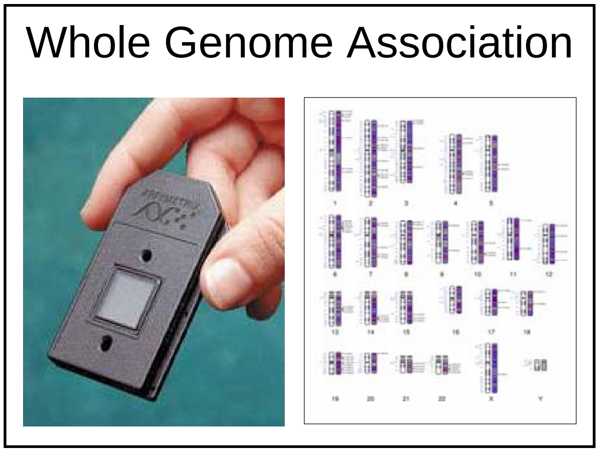
coded genotype (feature) of individual *i*

Goal: Find effect size that explains best all (potentially transformed) phenotypes as a linear function of the genotypes and estimate the probability (p-value) for the data being consistent with the null hypothesis (i.e. no effect)

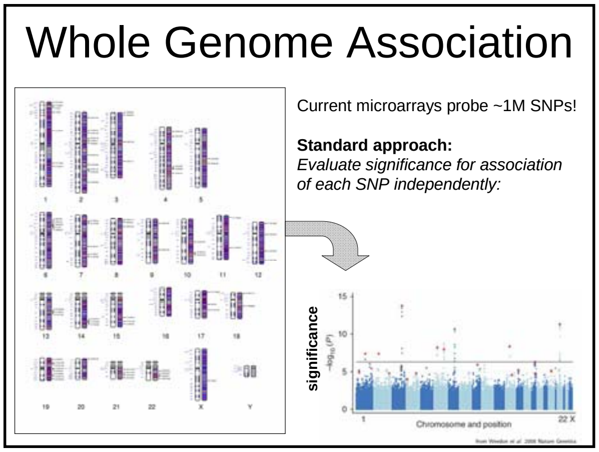


Overview

- Population stratification
- Associations: Basics
- Whole genome associations
- Genotype imputation
- Uncertain genotypes
- New Methods



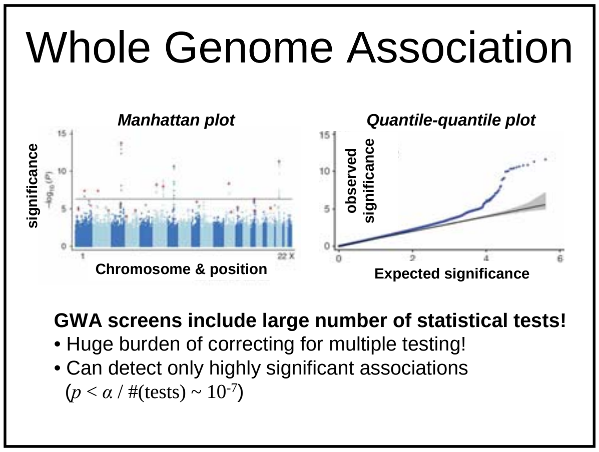
Whole Genome Association



Whole Genome Association

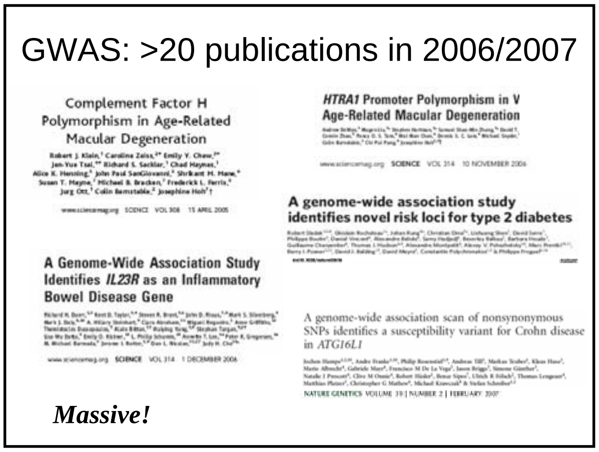
Current microarrays probe ~1M SNPs!

Standard approach:
Evaluate significance for association of each SNP independently:



Whole Genome Association

GWA screens include large number of statistical tests!
- Huge burden of correcting for multiple testing!
- Can detect only highly significant associations ($p < \alpha$ / #(tests) ~ 10^{-7})



GWAS: >20 publications in 2006/2007

Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein, Caroline Zeiss, Emily Y. Chew, Jen-Yue Tsai, Richard S. Sackler, Chad Haynes, Alice K. Henning, John Paul SanGiovanni, Shrikant M. Mane, Susan T. Mayne, Michael B. Bracken, Frederick L. Ferris, Jurg Ott, Colin Barnstable, Josephine Hoh†

www.sciencemag.org SCIENCE VOL 308 15 APRIL 2005

HTRA1 Promoter Polymorphism in W
Age-Related Macular Degeneration

Andrew DeWan, Mugen Liu, Stephen Hartman, Samuel Shao-Min Zhang, David T. L. Liu, Connie Zhao, Pancy O. S. Tam, Wai Man Chan, Dennis S. C. Lam, Michael Snyder, Colin Barnstable, Chi Pui Pang, Josephine Hoh

www.sciencemag.org SCIENCE VOL 314 10 NOVEMBER 2006

A genome-wide association study identifies novel risk loci for type 2 diabetes

Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley Balkau, Barbara Heude, Guillaume Charpentier, Thomas J. Hudson, Alexandre Montpetit, Alexey V. Pshezhetsky, Marc Prentki, Barry I. Posner, David J. Balding, David Meyre, Constantin Polychronakos & Philippe Froguel

doi:10.1038/nature05616 nature

A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

Richard H. Duerr, Kent D. Taylor, Steven R. Brant, John D. Rioux, Mark S. Silverberg, Mark J. Daly, A. Hillary Steinhart, Clara Abraham, Miguel Regueiro, Anne Griffiths, Themistocles Dassopoulos, Alain Bitton, Huiying Yang, Stephan Targan, Lisa Wu Datta, Emily O. Kistner, L. Philip Schumm, Annette T. Lee, Peter K. Gregersen, M. Michael Barmada, Jerome I. Rotter, Dan L. Nicolae, Judy H. Cho

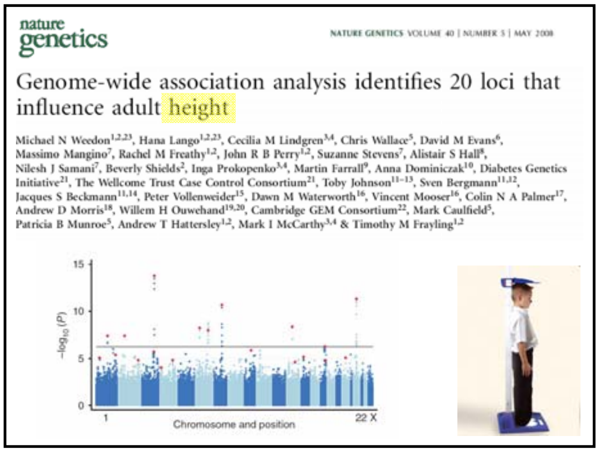
www.sciencemag.org SCIENCE VOL 314 1 DECEMBER 2006

A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*

Jochen Hampe, Andre Franke, Philip Rosenstiel, Andreas Till, Markus Teuber, Klaus Huse, Mario Albrecht, Gabriele Mayr, Francisco M De La Vega, Jason Briggs, Simone Günther, Natalie J Prescott, Clive M Onnie, Robert Häsler, Bence Sipos, Ulrich R Fölsch, Thomas Lengauer, Matthias Platzer, Christopher G Mathew, Michael Krawczak & Stefan Schreiber

NATURE GENETICS VOLUME 39 | NUMBER 2 | FEBRUARY 2007

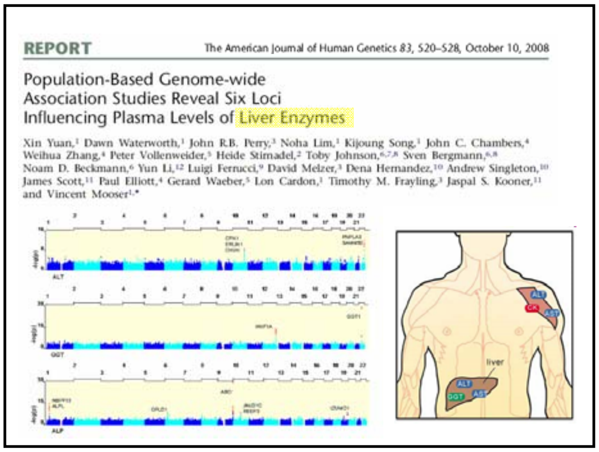
Massive!


nature genetics

NATURE GENETICS VOLUME 40 | NUMBER 5 | MAY 2008

Genome-wide association analysis identifies 20 loci that influence adult height

Michael N Weedon[1,2,23], Hana Lango[1,2,23], Cecilia M Lindgren[3,4], Chris Wallace[5], David M Evans[6], Massimo Mangino[7], Rachel M Freathy[1,2], John R B Perry[1,2], Suzanne Stevens[7], Alistair S Hall[8], Nilesh J Samani[7], Beverly Shields[2], Inga Prokopenko[3,4], Martin Farrall[9], Anna Dominiczak[10], Diabetes Genetics Initiative[21], The Wellcome Trust Case Control Consortium[21], Toby Johnson[11–13], Sven Bergmann[11,12], Jacques S Beckmann[11,14], Peter Vollenweider[15], Dawn M Waterworth[16], Vincent Mooser[16], Colin N A Palmer[17], Andrew D Morris[18], Willem H Ouwehand[19,20], Cambridge GEM Consortium[22], Mark Caulfield[5], Patricia B Munroe[5], Andrew T Hattersley[1,2], Mark I McCarthy[3,4] & Timothy M Frayling[1,2]


REPORT

The American Journal of Human Genetics *83*, 520–528, October 10, 2008

Population-Based Genome-wide Association Studies Reveal Six Loci Influencing Plasma Levels of Liver Enzymes

Xin Yuan,[1] Dawn Waterworth,[1] John R.B. Perry,[3] Noha Lim,[1] Kijoung Song,[1] John C. Chambers,[4] Weihua Zhang,[4] Peter Vollenweider,[5] Heide Stirnadel,[2] Toby Johnson,[6,7,8] Sven Bergmann,[6,8] Noam D. Beckmann,[6] Yun Li,[12] Luigi Ferrucci,[9] David Melzer,[3] Dena Hernandez,[10] Andrew Singleton,[10] James Scott,[11] Paul Elliott,[4] Gerard Waeber,[5] Lon Cardon,[1] Timothy M. Frayling,[3] Jaspal S. Kooner,[11] and Vincent Mooser[1,*]

ALT
GGT
ALP
ALT
CK
AST
liver
ALT
GGT
AST

nature genetics

VOLUME 40 | NUMBER 6 | JUNE 2008 NATURE GENETICS

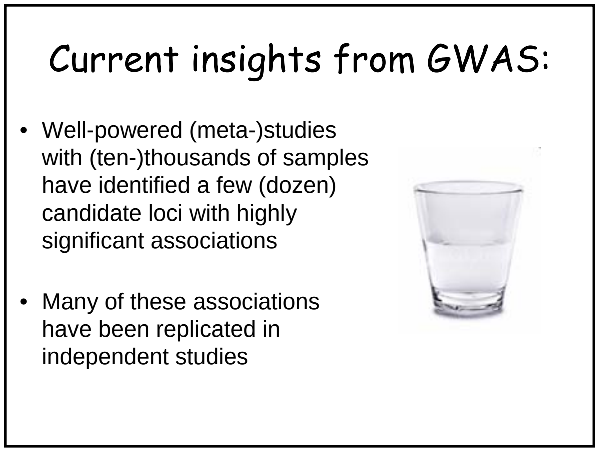
Common variants near *MC4R* are associated with fat mass, weight and risk of obesity

Ruth J F Loos[*,1,2,73], Cecilia M Lindgren[3,4,73], Shengxu Li[1,2,73], Eleanor Wheeler[5], Jing Hua Zhao[1,2], Inga Prokopenko[3,4], Michael Inouye[5], Rachel M Freathy[6,7], Antony P Attwood[5,8], Jacques S Beckmann[9,10], Sonja I Berndt[11], The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial[71], Sven Bergmann[9,12], Amanda J Bennett[3,4], Sheila A Bingham[13], Murielle Bochud[14], Morris Brown[15], Stéphane Cauchi[16], John M Connell[17], Cyrus Cooper[18], George Davey Smith[19], Ian Day[18], Christian Dina[16], Subhajyoti De[20], Emmanouil T Dermitzakis[5], Alex S F Doney[21], Katherine S Elliott[3], Paul Elliott[22,23], David M Evans[3,19], I Sadaf Farooqi[2,24], Philippe Froguel[16,25], Jilur Ghori[5], Christopher J Groves[3,4], Rhian Gwilliam[5], David Hadley[26], Alistair S Hall[27], Andrew T Hattersley[6,7], Johannes Hebebrand[28], Iris M Heid[29,30], KORA[71], Blanca Herrera[3,4], Anke Hinney[28], Sarah E Hunt[5], Marjo-Riitta Jarvelin[22,23,31], Toby Johnson[9,12,14], Jennifer D M Jolley[8], Fredrik Karpe[4], Andrew Keniry[5], Kay-Tee Khaw[32], Robert N Luben[32], Massimo Mangino[33], Jonathan Marchini[34], Wendy L McArdle[35], Ralph McGinnis[5], David Meyre[16], Patricia B Munroe[36], Andrew D Morris[21], Andrew R Ness[37], Matthew J Neville[4], Alexandra C Nica[5], Ken K Ong[1,2], Stephen O'Rahilly[2,24], Katharine R Owen[4], Colin N A Palmer[38], Konstantinos Papadakis[26], Simon Potter[5], Anneli Pouta[31,39], Lu Qi[40], Nurses' Health Study[71], Joshua C Randall[3,4], Nigel W Rayner[3,4], Susan M Ring[35], Manjinder S Sandhu[1,32], André Scherag[41], Matthew A Sims[1,2], Kijoung Song[42], Nicole Soranzo[5], Elizabeth K Speliotes[43,44], Diabetes Genetics Initiative[71], Holly E Syddall[18], Sarah A Teichmann[20], Nicholas J Timpson[3,19], Jonathan H Tobias[45], Manuela Uda[46], The SardiNIA Study[71], Carla I Ganz Vogel[28], Chris Wallace[36], Dawn M Waterworth[42], Michael N Weedon[6,7], The Wellcome Trust Case Control Consortium[72], Cristen J Willer[47], FUSION[71], Vicki L Wraight[2,24], Xin Yuan[42], Eleftheria Zeggini[3], Joel N Hirschhorn[44,48–51], David P Strachan[26], Willem H Ouwehand[8], Mark J Caulfield[36], Nilesh J Samani[33], Timothy M Frayling[6,7], Peter Vollenweider[52], Gerard Waeber[52], Vincent Mooser[42], Panos Deloukas[5], Mark I McCarthy[3,4,73], Nicholas J Wareham[1,2,73] & Inês Barroso[5,73]

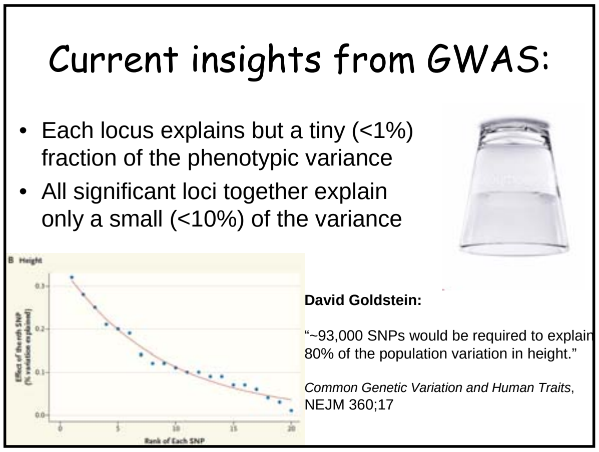
nature genetics

Genome-wide association study identifies eight loci associated with blood pressure

Christopher Newton-Cheh[1–3,94], Toby Johnson[4,5,6,94], Vesela Gateva[7,94], Martin D Tobin[8,94], Murielle Bochud[5], Lachlan Coin[9], Samer S Najjar[10], Jing Hua Zhao[11,12], Simon C Heath[13], Susana Eyheramendy[14,15], Konstantinos Papadakis[16], Benjamin F Voight[1,3], Laura J Scott[7], Feng Zhang[17], Martin Farrall[18,19], Toshiko Tanaka[20,21], Chris Wallace[22–24], John C Chambers[9], Kay-Tee Khaw[12,25], Peter Nilsson[26], Pim van der Harst[27], Silvia Polidoro[28], Diederick E Grobbee[29], N Charlotte Onland-Moret[29,30], Michiel L Bots[29], Louise V Wain[8], Katherine S Elliott[19], Alexander Teumer[31], Jian'an Luan[11], Gavin Lucas[32], Johanna Kuusisto[33], Paul R Burton[8], David Hadley[16], Wendy L McArdle[34], Wellcome Trust Case Control Consortium[93], Morris Brown[35], Anna Dominiczak[36], Stephen J Newhouse[22,23], Nilesh J Samani[37], John Webster[38], Eleftheria Zeggini[19,39], Jacques S Beckmann[4,40], Sven Bergmann[4,6], Noha Lim[41], Kijoung Song[41], Peter Vollenweider[42], Gerard Waeber[42], Dawn M Waterworth[41], Xin Yuan[41], Leif Groop[43,44], Marju Orho-Melander[26], Alessandra Allione[28], Alessandra Di Gregorio[28,45], Simonetta Guarrera[28], Salvatore Panico[46], Fulvio Ricceri[28], Valeria Romanazzi[28,45], Carlotta Sacerdote[47], Paolo Vineis[9,28], Inês Barroso[12,39], Manjinder S Sandhu[11,12,25], Robert N Luben[12,25], Gabriel J Crawford[3], Pekka Jousilahti[48], Markus Perola[48,49], Michael Boehnke[7], Lori L Bonnycastle[50], Francis S Collins[50], Anne U Jackson[7], Karen L Mohlke[51], Heather M Stringham[7], Timo T Valle[52], Cristen J Willer[7], Richard N Bergman[53], Mario A Morken[50], Angela Döring[15], Christian Gieger[15], Thomas Illig[15], Thomas Meitinger[54,55], Elin Org[56], Arne Pfeufer[54,55], H Erich Wichmann[15,57], Sekar Kathiresan[1–3], Jaume Marrugat[32], Christopher J O'Donnell[58,59], Stephen M Schwartz[60,61], David S Siscovick[60,61], Isaac Subirana[32,62], Nelson B Freimer[63], Anna-Liisa Hartikainen[64], Mark I McCarthy[19,65,66], Paul F O'Reilly[9], Leena Peltonen[39,49], Anneli Pouta[64,67], Paul E de Jong[68], Harold Snieder[69], Wiek H van Gilst[27], Robert Clarke[70], Anuj Goel[18,19], Anders Hamsten[71], John F Peden[18,19], Udo Seedorf[72], Ann-Christine Syvänen[73], Giovanni Tognoni[74], Edward G Lakatta[10], Serena Sanna[75], Paul Scheet[76], David Schlessinger[77], Angelo Scuteri[78], Marcus Dörr[79], Florian Ernst[31], Stephan B Felix[79], Georg Homuth[31], Roberto Lorbeer[80], Thorsten Reffelmann[79], Rainer Rettig[81], Uwe Völker[31], Pilar Galan[82], Ivo G Gut[13], Serge Hercberg[82], G Mark Lathrop[13], Diana Zelenika[13], Panos Deloukas[12,39], Nicole Soranzo[17,39], Frances M Williams[17], Guangju Zhai[17], Veikko Salomaa[48], Markku Laakso[33], Roberto Elosua[32,62], Nita G Forouhi[11], Henry Völzke[80], Cuno S Uiterwaal[29], Yvonne T van der Schouw[29], Mattijs E Numans[29], Giuseppe Matullo[28,45], Gerjan Navis[68], Göran Berglund[26], Sheila A Bingham[12,83], Jaspal S Kooner[84], John M Connell[36], Stefania Bandinelli[85], Luigi Ferrucci[21], Hugh Watkins[18,19], Tim D Spector[17], Jaakko Tuomilehto[52,86,87], David Altshuler[1,3,88,89], David P Strachan[16], Maris Laan[56], Pierre Meneton[90], Nicholas J Wareham[11,12], Manuela Uda[75], Marjo-Riitta Jarvelin[9,67,91], Vincent Mooser[41], Olle Melander[26], Ruth JF Loos[11,12], Paul Elliott[9,94], Gonçalo R Abecasis[92,94], Mark Caulfield[22,23,94] & Patricia B Munroe[22,23,94]


Current insights from GWAS:

- Well-powered (meta-)studies with (ten-)thousands of samples have identified a few (dozen) candidate loci with highly significant associations
- Many of these associations have been replicated in independent studies

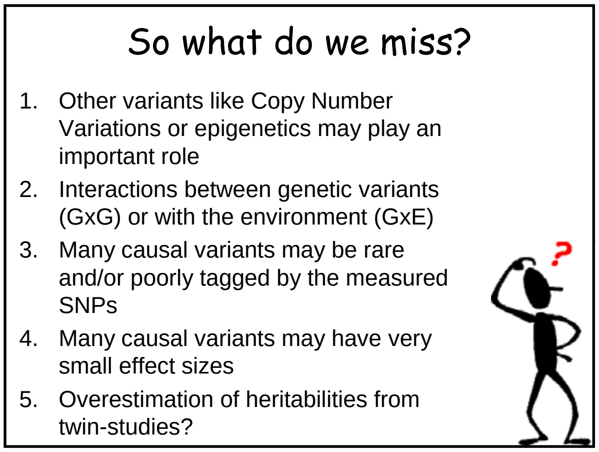

Current insights from GWAS:

- Each locus explains but a tiny (<1%) fraction of the phenotypic variance
- All significant loci together explain only a small (<10%) of the variance

David Goldstein:

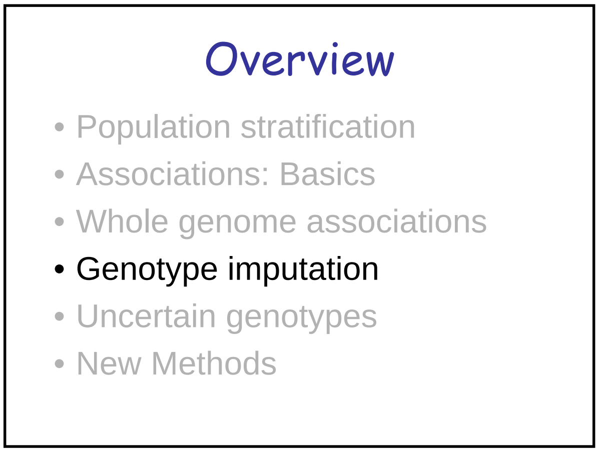
"~93,000 SNPs would be required to explain 80% of the population variation in height."

Common Genetic Variation and Human Traits, NEJM 360;17



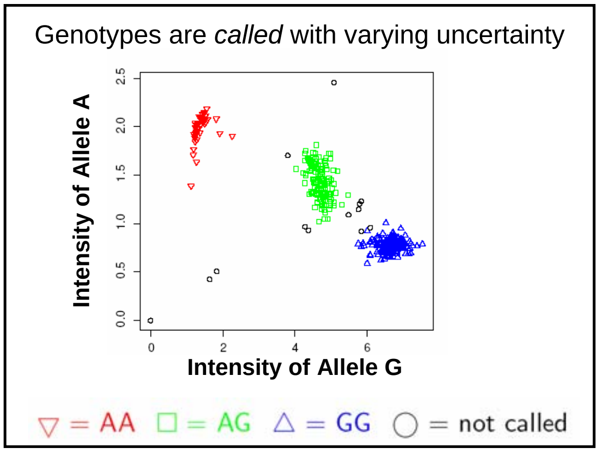
So what do we miss?

1. Other variants like Copy Number Variations or epigenetics may play an important role
2. Interactions between genetic variants (GxG) or with the environment (GxE)
3. Many causal variants may be rare and/or poorly tagged by the measured SNPs
4. Many causal variants may have very small effect sizes
5. Overestimation of heritabilities from twin-studies?



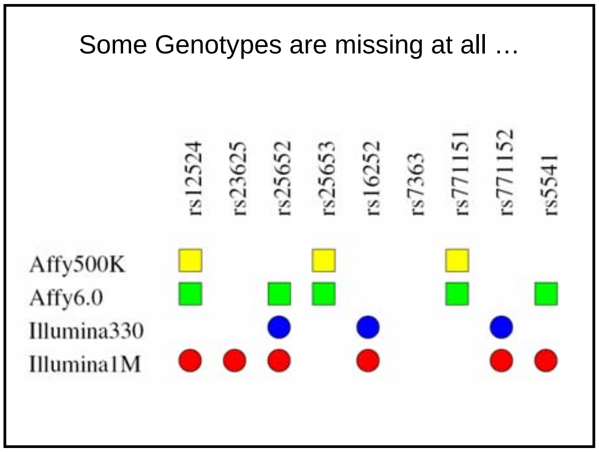
Overview

- Population stratification
- Associations: Basics
- Whole genome associations
- Genotype imputation
- Uncertain genotypes
- New Methods



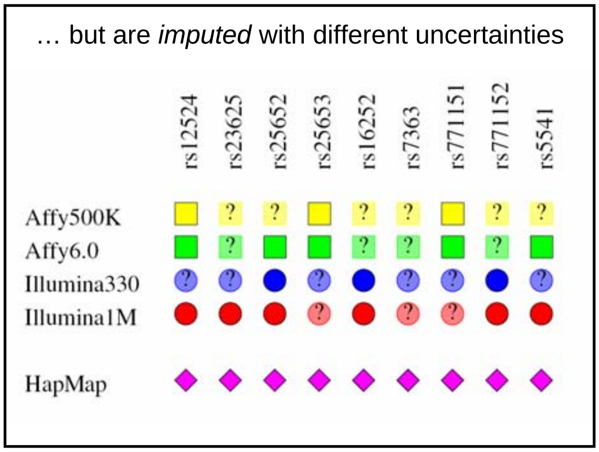
Genotypes are *called* with varying uncertainty

▽ = AA □ = AG △ = GG ○ = not called



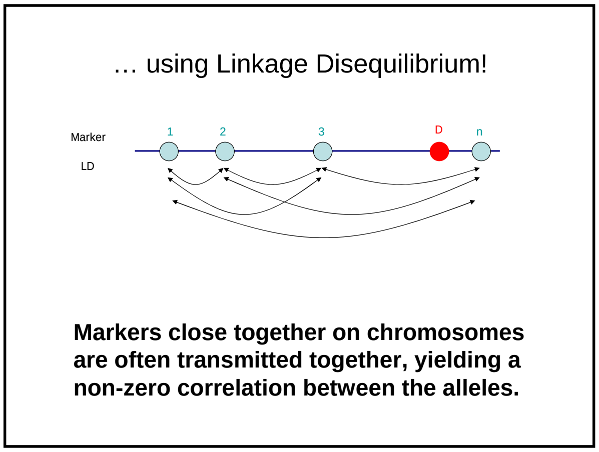
Some Genotypes are missing at all …

	rs12524	rs23625	rs25652	rs25653	rs16252	rs7363	rs771151	rs771152	rs5541
Affy500K	🟨			🟨			🟨		
Affy6.0	🟩		🟩	🟩			🟩		🟩
Illumina330			🔵		🔵			🔵	
Illumina1M	🔴	🔴	🔴		🔴			🔴	🔴



… but are *imputed* with different uncertainties

	rs12524	rs23625	rs25652	rs25653	rs16252	rs7363	rs771151	rs771152	rs5541
Affy500K	🟨	?	?	🟨	?	?	🟨	?	?
Affy6.0	🟩	?	🟩	🟩	?	?	🟩	?	🟩
Illumina330	?	?	🔵	?	🔵	?	?	🔵	?
Illumina1M	🔴	🔴	🔴	?	🔴	?	?	🔴	🔴
HapMap	◆	◆	◆	◆	◆	◆	◆	◆	◆



… using Linkage Disequilibrium!

Markers close together on chromosomes are often transmitted together, yielding a non-zero correlation between the alleles.

Conclusion

- Genotypic markers are *always* measured or inferred with *some* degree of uncertainty
- Association methods should take into account this uncertainty

Two easy ways dealing with uncertain genotypes

1. ***Genotype Calling:***
   Choose the most likely genotype and continue as if it is true
   (p_{11}=10%, p_{12}=20% p_{22}=70% => G=2)

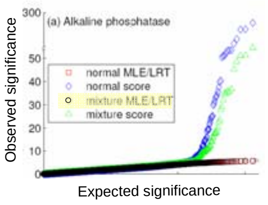
2. ***Mean genotype:***
   Use the weighted average genotype
   (p_{11}=10%, p_{12}=20% p_{22}=70% => G=1.6)

Overview

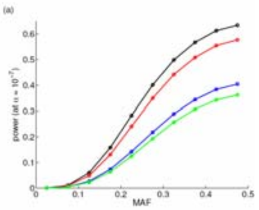
- Associations: Basics
- Whole genome associations
- Population stratification
- Genotype imputation
- Uncertain genotypes
- New Methods

New Method

based on a mixture model *both* for phenotypes and uncertain genotypes

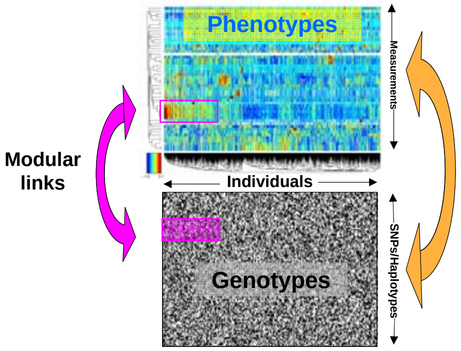


Better control of false positives

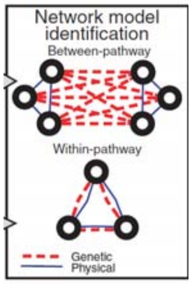


More power

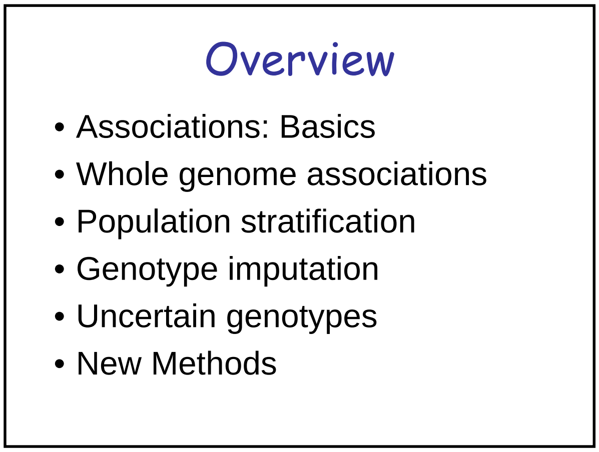
Modular Approach for Integrative Analysis of Genotypes and Phenotypes



Network Approaches for Integrative Association Analysis



Using knowledge on physical gene-interactions or pathways to prioritize the search for functional interactions



Overview

- Associations: Basics
- Whole genome associations
- Population stratification
- Genotype imputation
- Uncertain genotypes
- New Methods