Assessing variants in the human genome

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

November 19, 2010

Very large data sets

Printable Version

Statistical Methods for Very Large Datasets Conference 2011

Register Now

Wednesday, June 01, 2011 8:00 AM -
Friday, June 03, 2011 5:30 PM (Eastern Time)

Add to Calendar

77 days left to take advantage of early price.

InterContinental Harbor Court Hotel
800-824-0076
550 Light Street
Baltimore, Maryland 21202
Map and Directions

Contact Information

Phone: 410-955-3067
email: rzuckerm@jhsph.edu
Send Email

SHARE

Welcome!

The Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health invites you to a 3-day conference on Very Large Data Sets. The conference is scheduled from June 1-3, 2011 and will be hosted in beautiful, downtown Baltimore, Maryland, USA at the InterContinental Harbor Court Hotel.

The conference has a one-track session for invited presentations and a high profile session for contributed poster presentations. A panel discussion will attempt to define what large data sets are, anticipate new challenges, and identify possible solutions.

Conference Overview

There is an acute and increasing need to adapt standard statistical methods and to develop new approaches for the analysis of very large data sets. A data set is very large if it raises difficult or insurmountable computational problems for standard data analysis using available computing systems. The continuous increase in size and complexity of data sets is due in part to increased computational and storage capabilities, new measurement technologies and study designs, and an increasing number of study "units".

Very large data sets

Printable Version

Important Dates:

June 1, 2010: Call for poster presentation abstracts
February 1, 2011: Final date for a early bird registration fee
March 1, 2011: Final date for submission of poster abstracts
April 1, 2011 : Notification for poster abstract acceptance
May 3, 2011: Final date for reduced conference rate at hotel
June 1-3, 2011: Conference dates

Register Now

Add to Calendar
77 days left to take advantage of early price.

Contact Information
Phone: 410-955-3067
email:
rzuckerm@jhsph.edu
Send Email

SHARE

Confirmed Speakers include:

Goncalo Abecasis, *University of Michigan*
DuBois Bowman, *Emory University*
Brian Caffo, *Johns Hopkins University*
Raymond Carroll, *Texas A&M University*
Ciprian Crainiceanu, *Johns Hopkins University*
Francesca Dominici, *Harvard University*
William DuMouchel, *Phase Forward Lincoln Safety Group*
Sandrine Dudoit,*University of California at Berkeley*
Jay Emerson, *Yale University*
Stephen Eubank, *Virginia Tech*
Montse Fuentes, *North Carolina State University*
Robert Gentleman, *Fred Hutchinson Cancer Research Center*
Rafael Irizarry, *Johns Hopkins University*
Hongkai Ji, *Johns Hopkins University*
Nicole Lazar, *University of Illinois at Chicago*
Jeffrey Morris, *MD Anderson Cancer Center*
Hans-Georg Muller, *University of California at Davis*
Doug Nychka, *National Center for Atmospheric Research*
Todd Ogden, *Columbia University*
Roger Peng, *Johns Hopkins University*
James Ramsay, *McGill University*
Ingo Ruczinski, *Johns Hopkins University*
Steven Salzberg, *University of Maryland*
Terry Speed, *University of California at Berkeley*
John Storey, *Princeton University*
Alex Szalay, *Johns Hopkins University*
Jonathan Taylor, *Stanford University*
Chris Volinsky, *AT&T Labs-Research*

http: //biostat.jhsph.edu/∼iruczins/

ingo@jhu.edu

Ingo Ruczinski | Assessing variants in the human genome

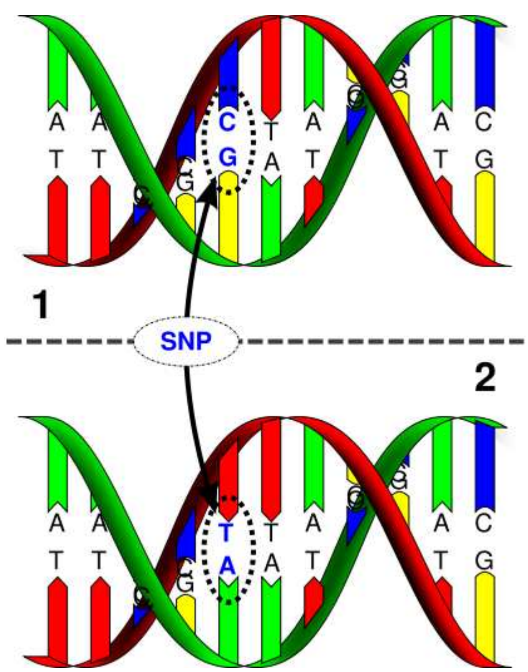
Acknowledgments

Collaborators: Kathleen Barnes, Terri Beaty, Benilton Carvalho, Bob Cole, Rafael Irizarry, Tom Louis, Rasika Mathias, Matt Ritchie, Rob Scharpf, Holger Schwender, Keith West.

Computing support: Marvin Newhouse, Jiong Yang.

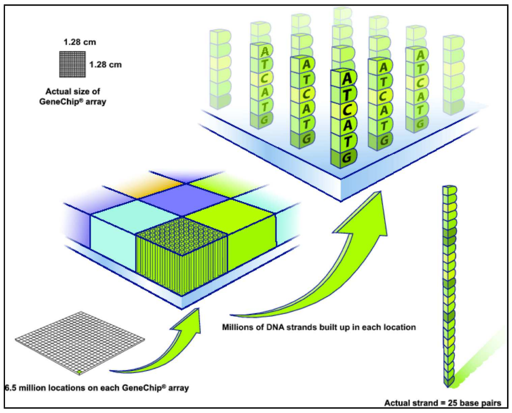
Funding: NIH R01 DK061662, GM083084, HL090577, and a CTSA grant to the Johns Hopkins Medical Institutions.

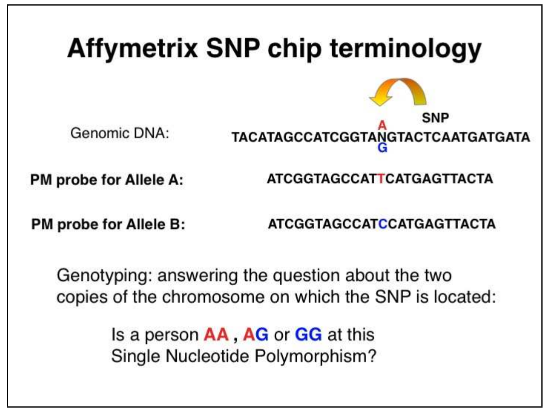
Single nucleotide polymorphisms



urgi.versailles.inra.fr

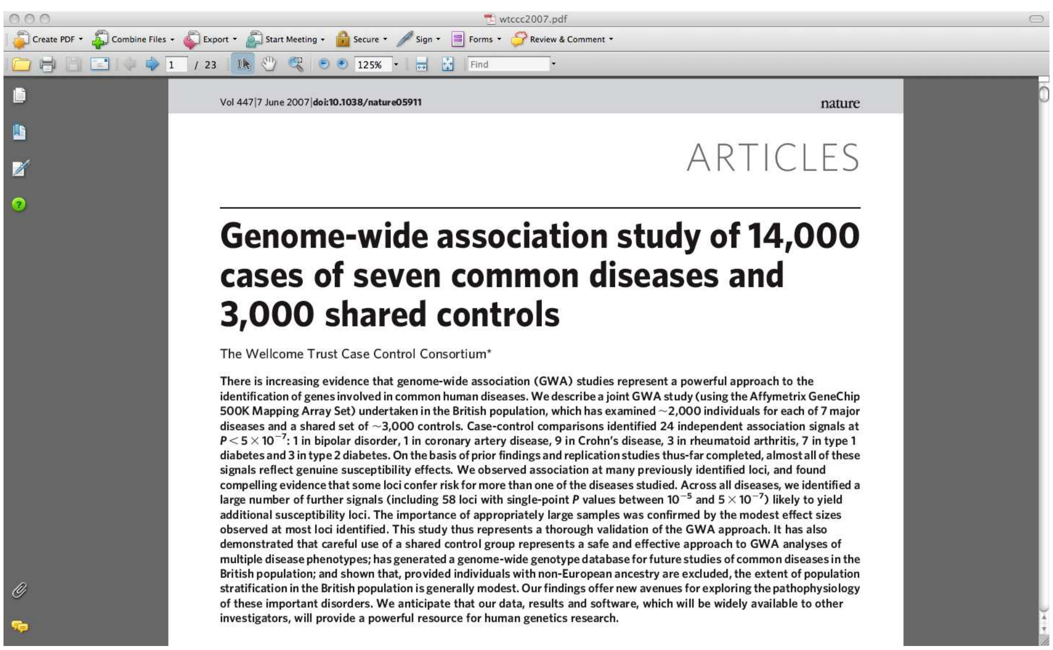
Genomic arrays





http://www.affymetrix.com

WTCCC



Vol 447|7 June 2007|doi:10.1038/nature05911 nature

ARTICLES

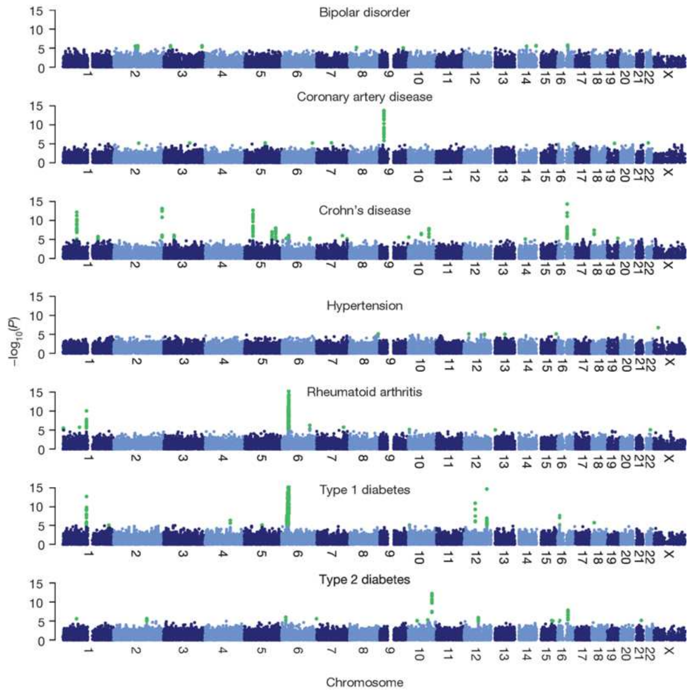
Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We describe a joint GWA study (using the Affymetrix GeneChip 500K Mapping Array Set) undertaken in the British population, which has examined ~2,000 individuals for each of 7 major diseases and a shared set of ~3,000 controls. Case-control comparisons identified 24 independent association signals at $P < 5 \times 10^{-7}$: 1 in bipolar disorder, 1 in coronary artery disease, 9 in Crohn's disease, 3 in rheumatoid arthritis, 7 in type 1 diabetes and 3 in type 2 diabetes. On the basis of prior findings and replication studies thus-far completed, almost all of these signals reflect genuine susceptibility effects. We observed association at many previously identified loci, and found compelling evidence that some loci confer risk for more than one of the diseases studied. Across all diseases, we identified a large number of further signals (including 58 loci with single-point P values between 10^{-5} and 5×10^{-7}) likely to yield additional susceptibility loci. The importance of appropriately large samples was confirmed by the modest effect sizes observed at most loci identified. This study thus represents a thorough validation of the GWA approach. It has also demonstrated that careful use of a shared control group represents a safe and effective approach to GWA analyses of multiple disease phenotypes; has generated a genome-wide genotype database for future studies of common diseases in the British population; and shown that, provided individuals with non-European ancestry are excluded, the extent of population stratification in the British population is generally modest. Our findings offer new avenues for exploring the pathophysiology of these important disorders. We anticipate that our data, results and software, which will be widely available to other investigators, will provide a powerful resource for human genetics research.

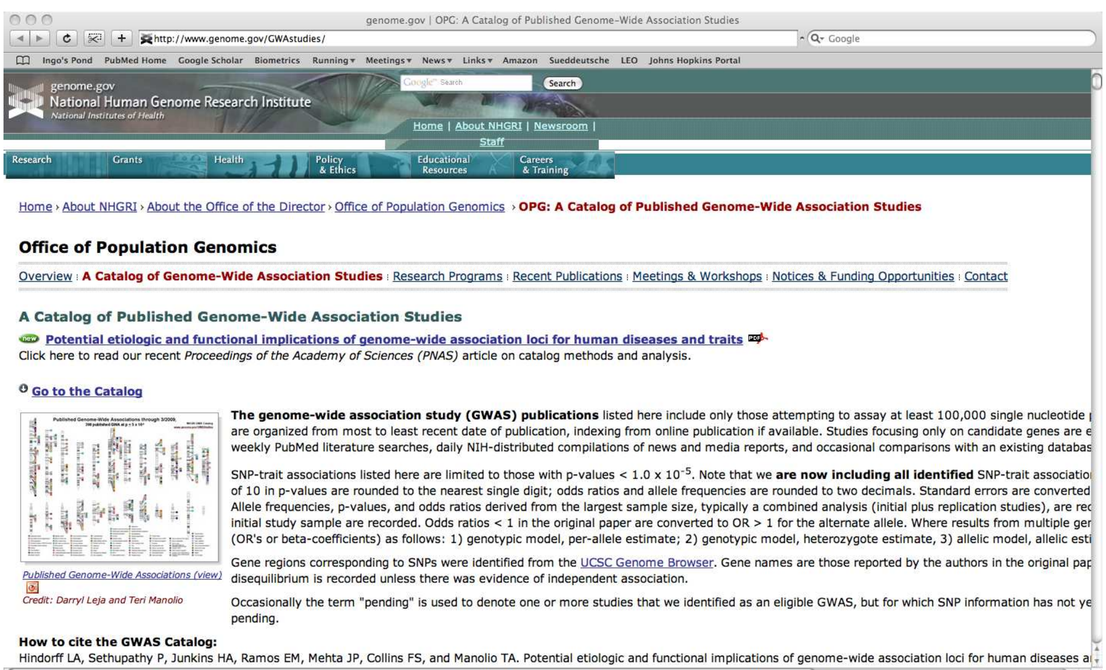
Wellcome Trust Case Control Consortium (2007). Nature 447(7145): 661-78.

WTCCC



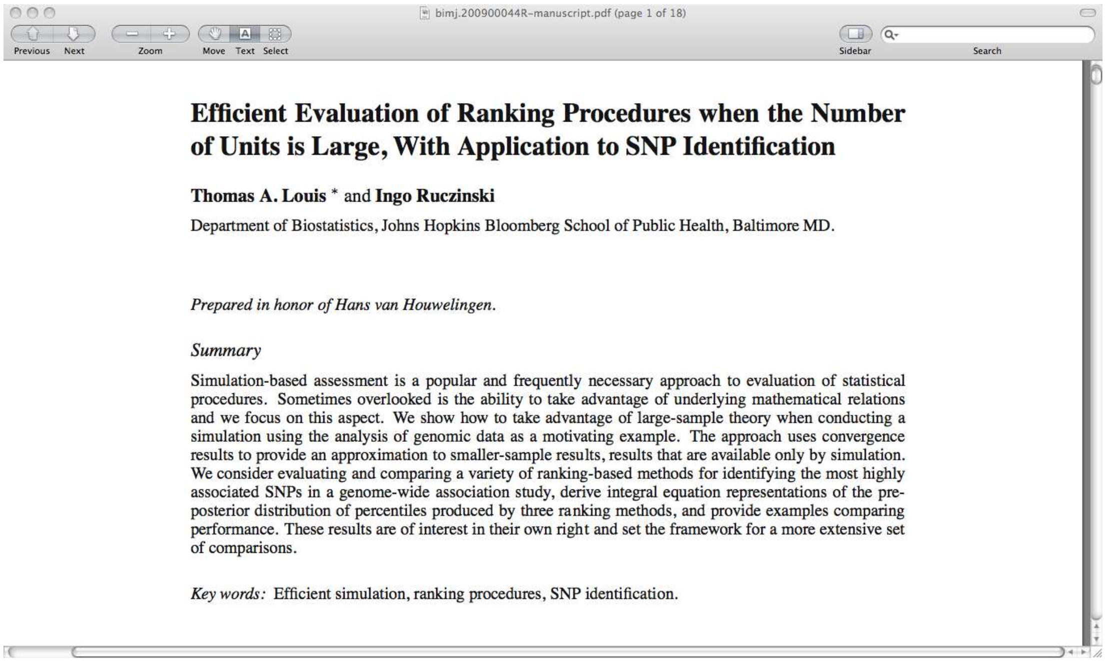
Wellcome Trust Case Control Consortium (2007). Nature 447(7145): 661-78.

Results



http://www.genome.gov/GWAstudies/

Ranking


bimj.200900044R-manuscript.pdf (page 1 of 18)

Efficient Evaluation of Ranking Procedures when the Number of Units is Large, With Application to SNP Identification

Thomas A. Louis * and **Ingo Ruczinski**

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore MD.

Prepared in honor of Hans van Houwelingen.

Summary

Simulation-based assessment is a popular and frequently necessary approach to evaluation of statistical procedures. Sometimes overlooked is the ability to take advantage of underlying mathematical relations and we focus on this aspect. We show how to take advantage of large-sample theory when conducting a simulation using the analysis of genomic data as a motivating example. The approach uses convergence results to provide an approximation to smaller-sample results, results that are available only by simulation. We consider evaluating and comparing a variety of ranking-based methods for identifying the most highly associated SNPs in a genome-wide association study, derive integral equation representations of the pre-posterior distribution of percentiles produced by three ranking methods, and provide examples comparing performance. These results are of interest in their own right and set the framework for a more extensive set of comparisons.

Key words: Efficient simulation, ranking procedures, SNP identification.

[Lou · Ruc | Biom · J 2010] • [Ruc · · · Lou | Bayes · Stat 2010]

Case-parent trios

Recruitment Site	CL	CLP	CP	Total
Utah	68	96	52	216
Norway	106	174	107	387
Korea	19	40	5	64
Maryland	19	71	25	115
Pittsburgh	26	70	11	107
Singapore	15	45	53	113
Taiwan	42	176	74	292
Iowa	16	29	24	69
Denmark	6	15	5	26
Philippines	0	94	0	94
WuHan	39	136	42	217
Shandong Province	54	129	30	213
Western China	43	63	38	144
Total	453	1138	466	2057

Ingo Ruczinski Assessing variants in the human genome

Case-parent trios

F : 12 ┬ M : 12 — C : 11

F : 12 ┬ M : 12 — C : 12

F : 12 ┬ M : 12 — C : 22

F : 11 ┬ M : 12 — C : 11

F : 11 ┬ M : 12 — C : 12

F : 12 ┬ M : 22 — C : 12

F : 12 ┬ M : 22 — C : 22

Genotypic TDT

Assume that at a certain locus the father has alleles 11 and the mother has alleles 12. The four *Mendelian children* thus have alleles 11, 12, 11, and 12.

Assume the affected proband has genotype 11.

The three *Pseudo controls* then have the genotypes 11, 12, and 12.

	Y	X
Affected proband	1	11
Pseudo control #1	0	11
Pseudo control #2	0	12
Pseudo control #3	0	12

We can use coonditional logistic regression to analyze the data.

Allelic TDT

The transmission disequilibrium test measures the over-transmission of an allele from parents to affected offsprings. For a set of n parents with alleles 1 and 2 at a genetic locus, each parent can be summarized by the transmitted and the non-transmitted allele:

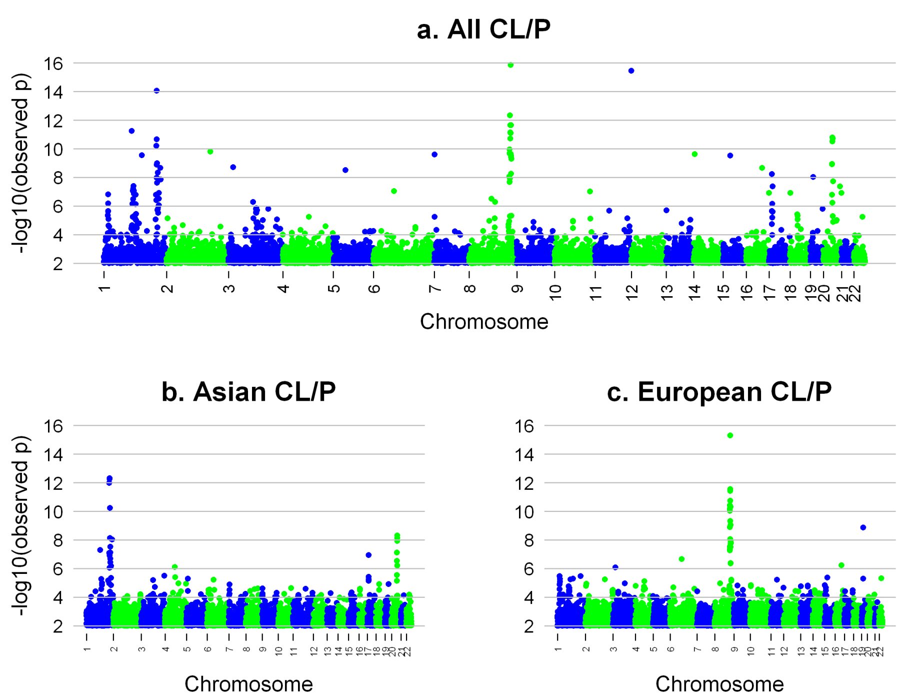
		Non-TA		$\sum$
		1	2	
TA	1	a	b	a + b
	2	c	d	c + d
$\sum$		a + c	b + d	2n

Only the heterozygous parents contribute information!

Under the null of no association, $\frac{(b-c)^2}{b+c} \sim \chi^2_1$

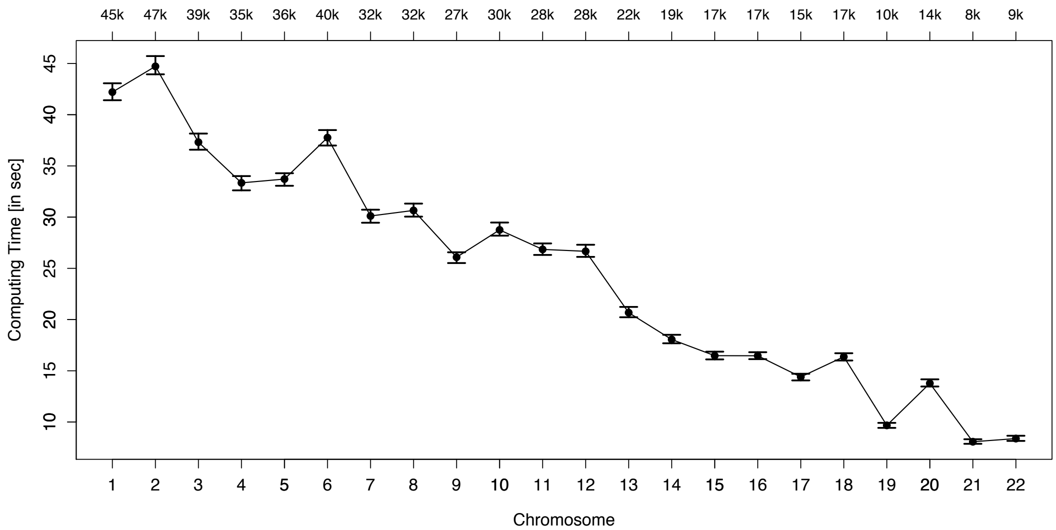
$\rightarrow$ Even better, use `binom.test()` in R.

GWAs results



[BEA · · · RUC · · · SCO | NAT·GEN 2010]

Fast genotypic TDT



[SCH · · · RUC | TECH · REP 2010]

Candidate genes

OPEN ACCESS Freely available online PLoS one

Genetic Determinants of Facial Clefting: Analysis of 357 Candidate Genes Using Two National Cleft Studies from Scandinavia

Astanand Jugessur[1,9], Min Shi[2,9], Håkon Kristian Gjessing[3,4], Rolv Terje Lie[4,5], Allen James Wilcox[6], Clarice Ring Weinberg[2], Kaare Christensen[7], Abee Lowman Boyles[6], Sandra Daack-Hirsch[8], Truc Nguyen Trung[5], Camilla Bille[7], Andrew Carl Lidral[9], Jeffrey Clark Murray[7,9]*

1 Craniofacial Development, Musculoskeletal Disorders, Murdoch Childrens Research Institute, Royal Children's Hospital, Parkville, Australia, **2** Biostatistics Branch, National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, Durham, North Carolina, United States of America, **3** Department of Epidemiology (EPAM), Norwegian Institute of Public Health, Oslo, Norway, **4** Section for Epidemiology and Medical Statistics, Department of Public Health and Primary Health Care, University of Bergen, Bergen, Norway, **5** Medical Birth Registry of Norway, Norwegian Institute of Public Health, Bergen, Norway, **6** Epidemiology Branch, National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, Durham, North Carolina, United States of America, **7** Department of Epidemiology, University of Southern Denmark, Odense, Denmark, **8** College of Nursing, University of Iowa, Iowa City, Iowa, United States of America, **9** Departments of Pediatrics, Epidemiology and Biological Sciences, University of Iowa, Iowa City, Iowa, United States of America

Abstract

Background: Facial clefts are common birth defects with a strong genetic component. To identify fetal genetic risk factors for clefting, 1536 SNPs in 357 candidate genes were genotyped in two population-based samples from Scandinavia (Norway: 562 case-parent and 592 control-parent triads; Denmark: 235 case-parent triads).

Methodology/Principal Findings: We used two complementary statistical methods, TRIMM and HAPLIN, to look for associations across these two national samples. TRIMM tests for association in each gene by using multi-SNP genotypes from case-parent triads directly without the need to infer haplotypes. HAPLIN on the other hand estimates the full haplotype distribution over a set of SNPs and estimates relative risks associated with each haplotype. For isolated cleft lip with or without cleft palate (I-CL/P), TRIMM and HAPLIN both identified significant associations with *IRF6* and *ADH1C* in both populations, but only HAPLIN found an association with *FGF12*. For isolated cleft palate (I-CP), TRIMM found associations with *ALX3*, *MKX*, and *PDGFC* in both populations, but only the association with *PDGFC* was identified by HAPLIN. In addition, HAPLIN identified an association with *ETV5* that was not detected by TRIMM.

Conclusion/Significance: Strong associations with seven genes were replicated in the Scandinavian samples and our approach effectively replicated the strongest previously known association in clefting—with *IRF6*. Based on two national cleft cohorts of similar ancestry, two robust statistical methods and a large panel of SNPs in the most promising cleft candidate genes to date, this study identified a previously unknown association with clefting for *ADH1C* and provides additional candidates and analytic approaches to advance the field.

Citation: Jugessur A, Shi M, Gjessing HK, Lie RT, Wilcox AJ, et al. (2009) Genetic Determinants of Facial Clefting: Analysis of 357 Candidate Genes Using Two National Cleft Studies from Scandinavia. PLoS ONE 4(4): e5385. doi:10.1371/journal.pone.0005385

Parent-of-origin effects

		Paternal				Maternal					
		TAT				TAT				PO-LRT[b]	
No.	SNP name	T	NT	*P*-value	OR[c]	T	NT	*P*-value	OR[c]	OR[d]	*P*-value
1	rs7771980	9	8	0.808	1.13	16	18	0.732	0.89	0.79	0.692
2	rs2677104	25	30	0.500	0.83	22	24	0.768	0.92	1.10	0.811
3	rs2819855	36	34	0.811	1.06	37	25	0.128	1.48	1.40	0.342
4	rs2819854	35	36	0.906	0.97	37	29	0.325	1.28	1.32	0.417
5	rs910586	15	13	0.705	1.15	20	5	**0.003**	4.00	**3.59**	**0.036**
6	rs2819853	14	12	0.695	1.17	18	5	**0.007**	3.60	3.19	0.063
7	rs765724	15	13	0.705	1.15	20	6	**0.006**	3.33	2.97	0.065
8	rs1343799	14	12	0.695	1.17	18	5	**0.007**	3.60	3.19	0.063
9	rs2819861	13	12	0.841	1.08	19	5	**0.004**	3.80	**3.73**	**0.036**
10	rs2790103	16	11	0.336	1.45	20	5	**0.003**	4.00	2.86	0.092
11	rs2790093	15	12	0.564	1.25	18	5	**0.007**	3.60	2.99	0.079
12	rs2790098	15	12	0.564	1.25	19	6	**0.009**	3.17	2.60	0.110
13	rs4714854	15	12	0.564	1.25	19	6	**0.009**	3.17	2.60	0.110
14	rs9472494	15	14	0.853	1.07	22	7	**0.005**	3.14	2.99	0.051
15	rs2396442	17	14	0.590	1.21	24	8	**0.005**	3.00	2.51	0.086
16	rs1934328	41	17	**0.002**	2.41	35	33	0.808	1.06	**0.44**	**0.029**
17	rs7773875	33	21	0.102	1.57	32	32	1.000	1.00	0.65	0.245
18	rs7771889	36	18	0.014	2.00	40	31	0.285	1.29	0.64	0.238
19	rs10485422	15	13	0.705	1.15	17	6	0.022	2.83	2.42	0.135
20	rs6904353	13	14	0.847	0.93	18	11	0.194	1.64	1.78	0.294
21	rs13207392	16	15	0.857	1.07	19	7	0.019	2.71	2.50	0.102
22	rs7748231	13	13	1.000	1.00	18	11	0.194	1.64	1.64	0.373
23	rs10948237	13	14	0.847	0.93	18	11	0.194	1.64	1.78	0.294
24	rs1928533	12	13	0.841	0.92	15	13	0.705	1.15	1.27	0.671

[SUL · · · RUC · · · BEA | GEN·EPI 2008]

Copy number estimates are noisy

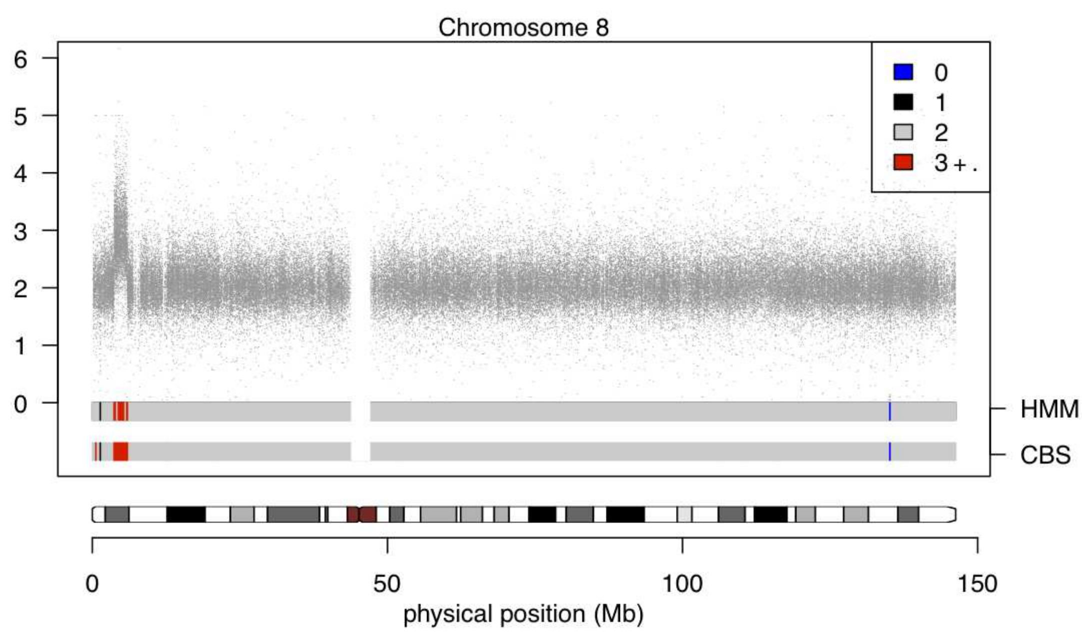
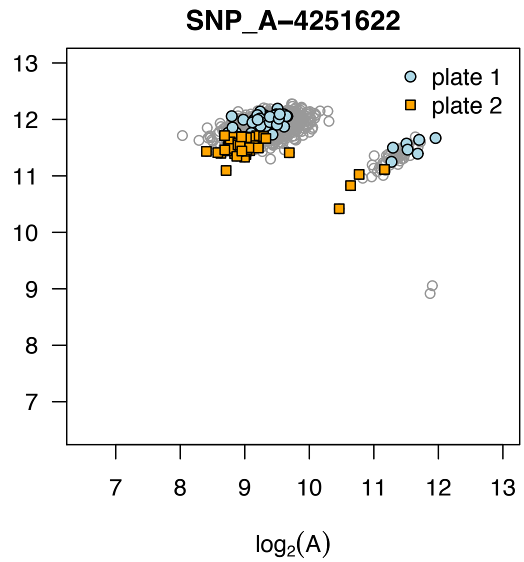
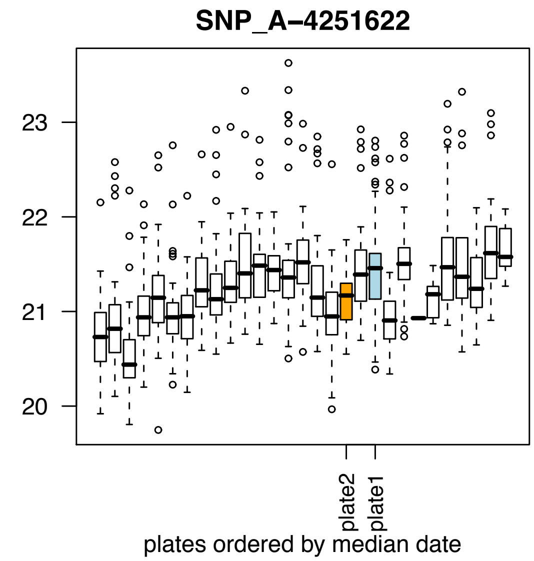
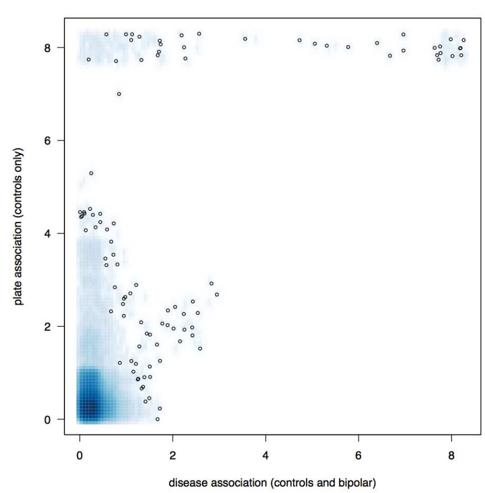


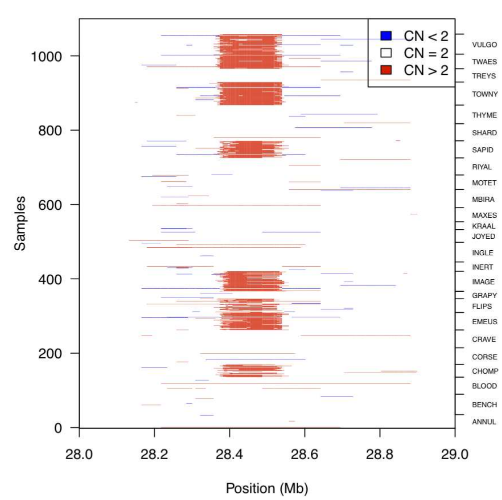
Plate effects





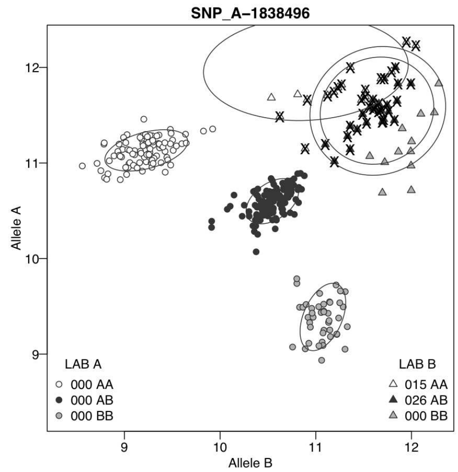
Confounding of plate and disease



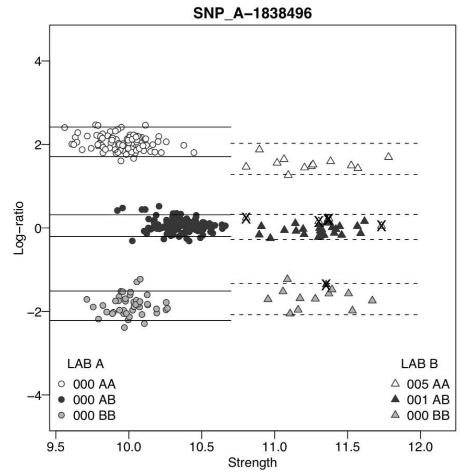


Genotype estimates are more robust

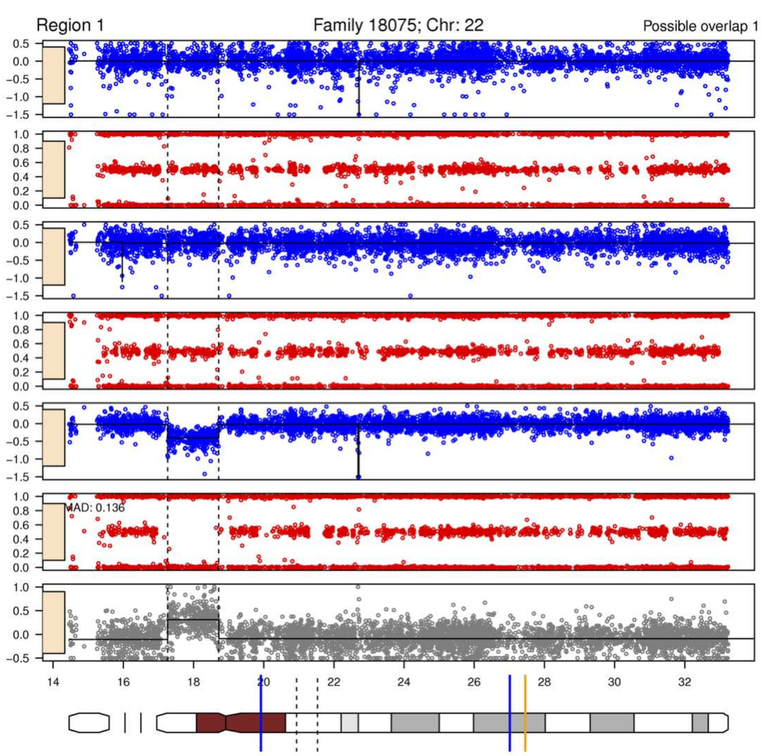
Birdseed



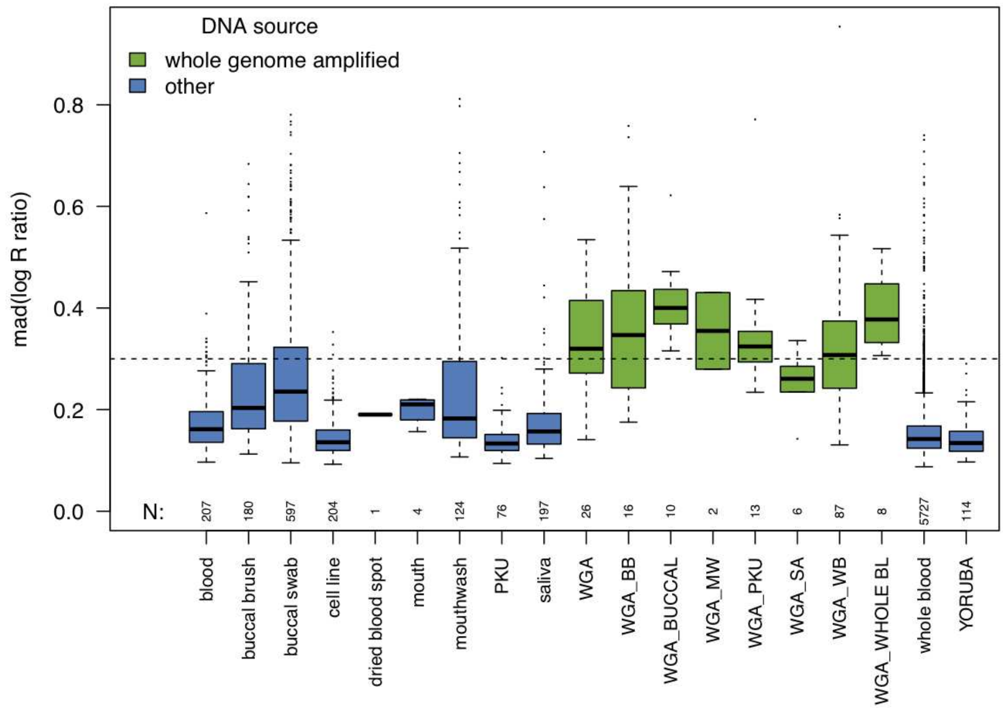
CRLMM



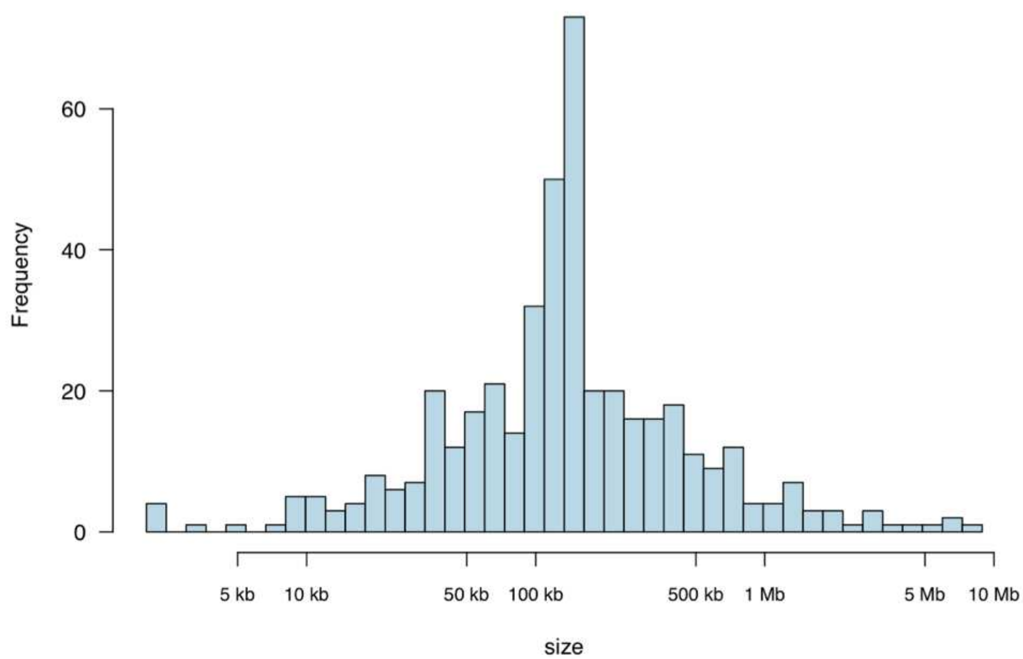
De-novo deletions



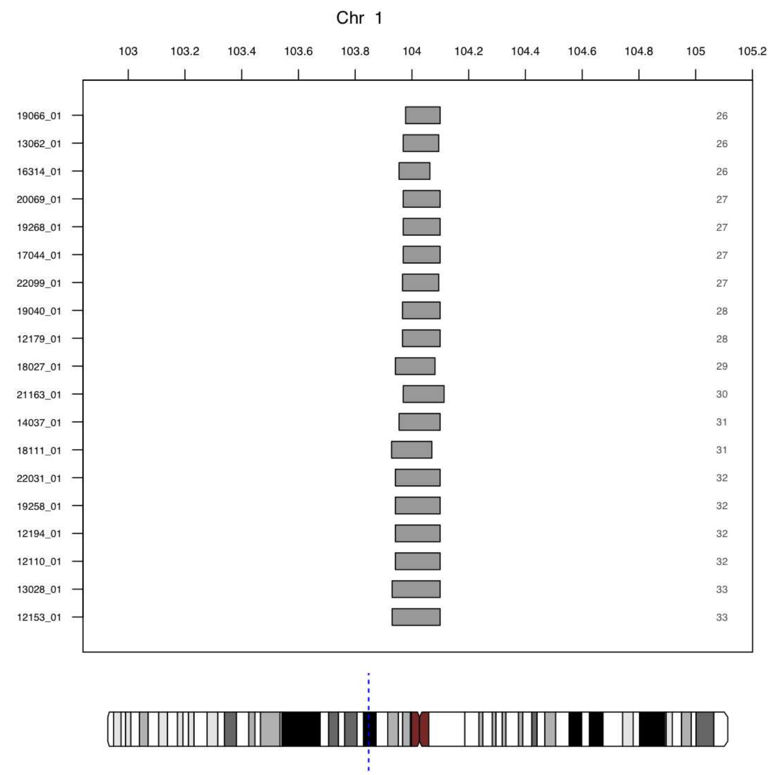
DNA source



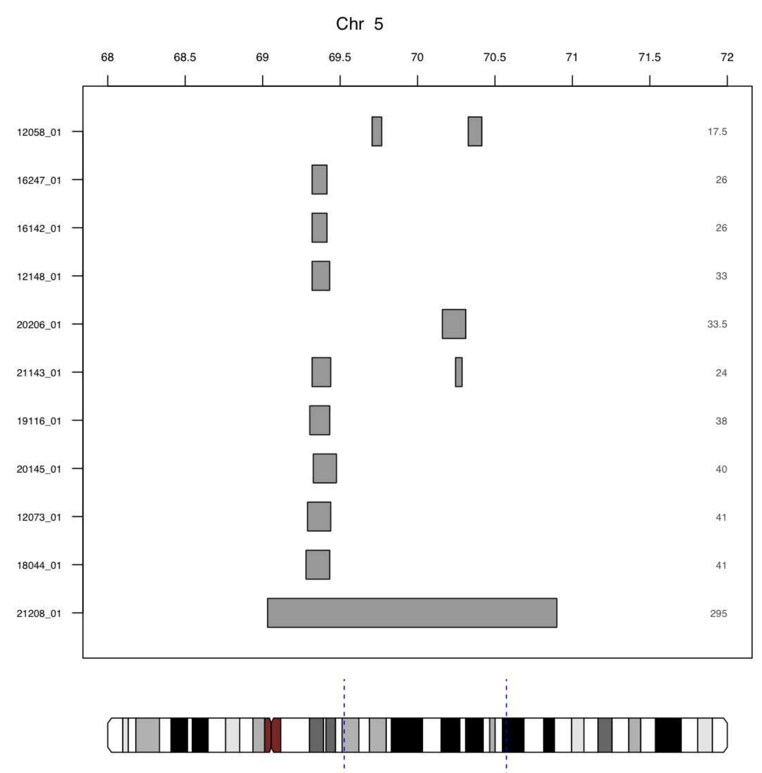
De-novo deletions



De-novo deletions

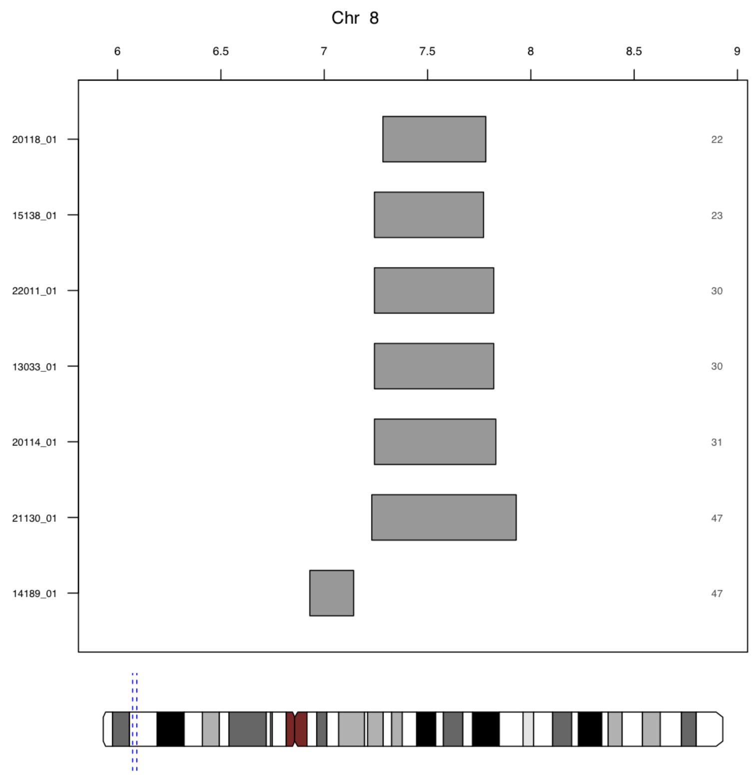


De-novo deletions

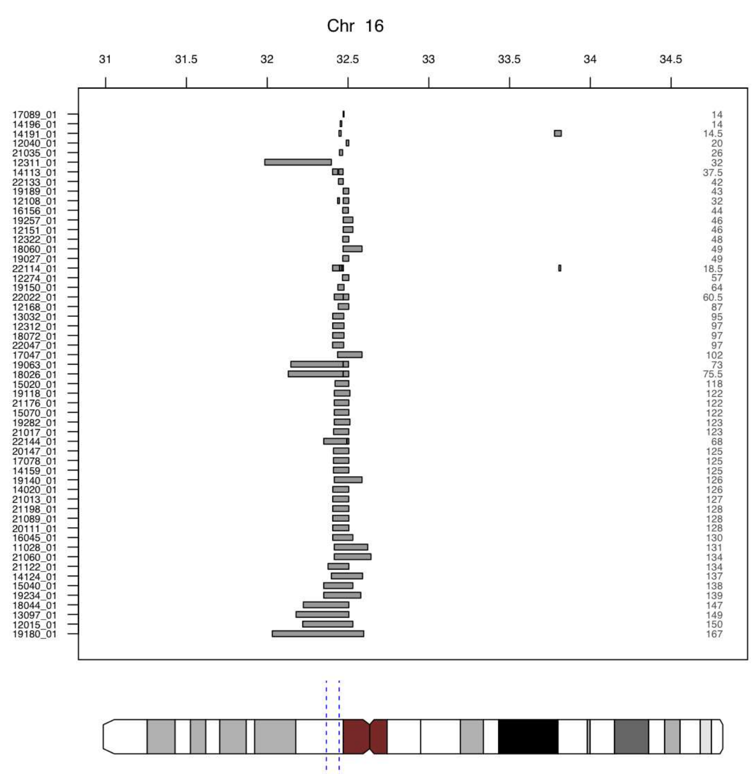


Ingo Ruczinski Assessing variants in the human genome

De-novo deletions

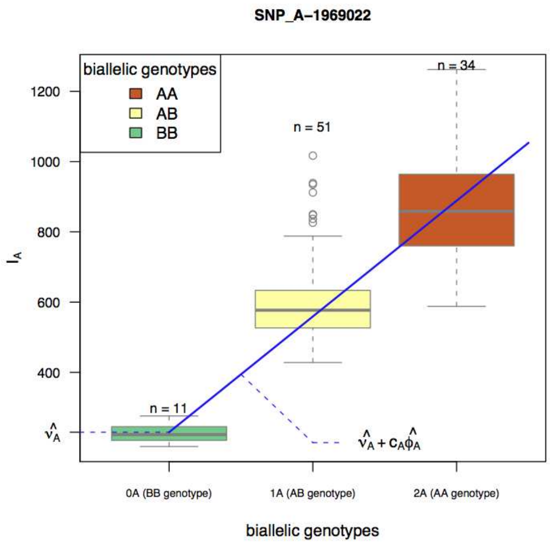


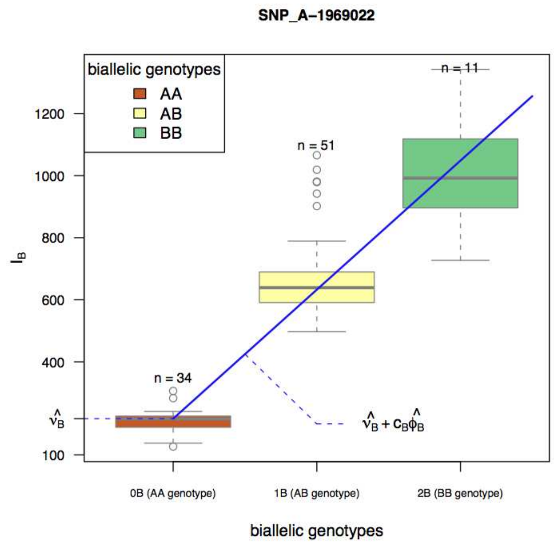
De-novo deletions



Ingo Ruczinski Assessing variants in the human genome

Allele specific copy numbers

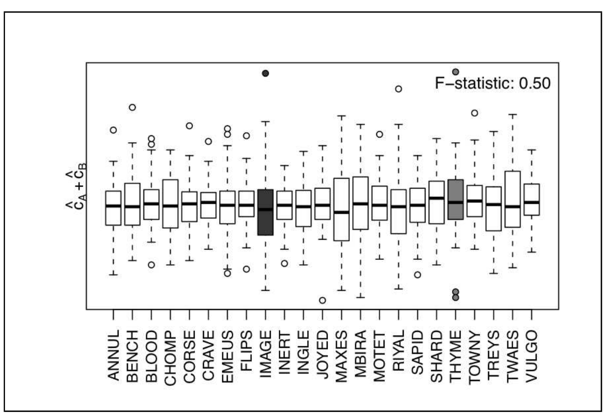


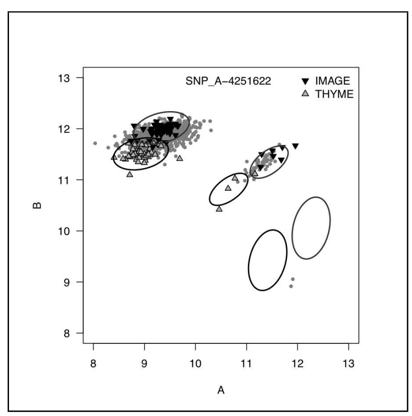


Allele specific copy numbers

At locus i, for subject j in plate p, we have for allele $k \in \{A, B\}$

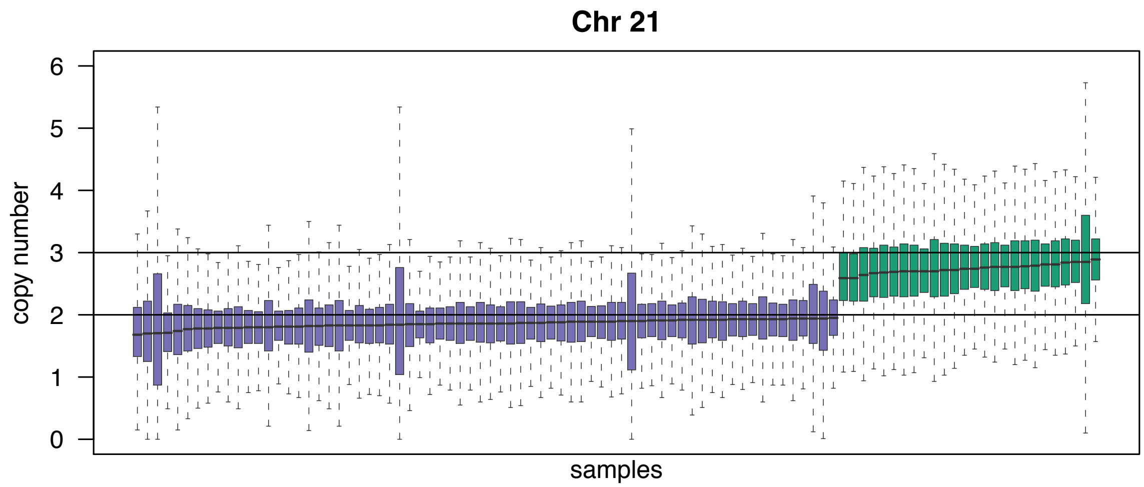
$$I_{kijp} = \nu_{kip}\delta_{kijp} + \phi_{kip}c_{kijp}\epsilon_{kijp} \implies \hat{c}_{kijp} = \max\left\{\frac{1}{\hat{\phi}_{kip}}\left(I_{kijp} - \hat{\nu}_{kip}\right),\ 0\right\}$$





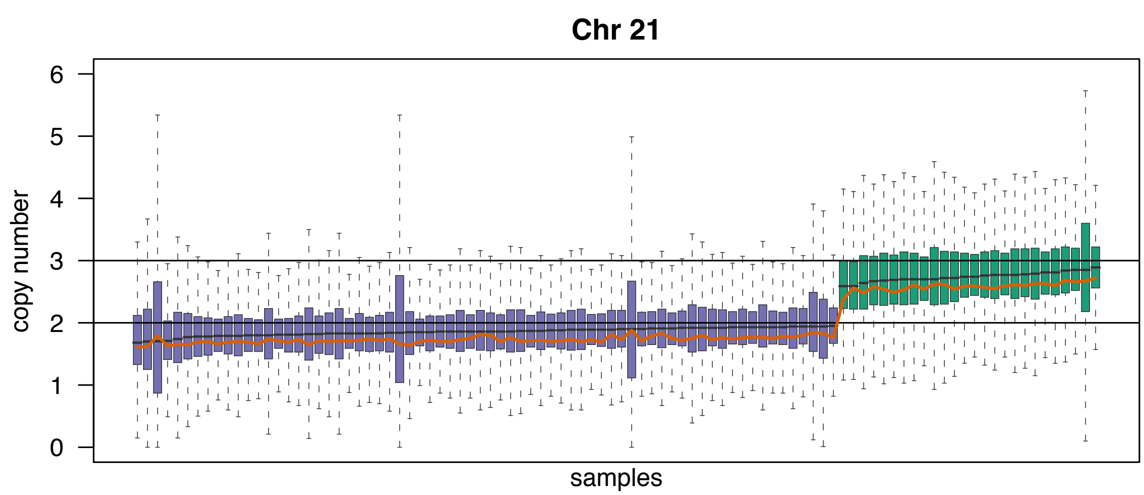
[SCH · IRI · RIT · CAR · RUC | TECH·REP 2010] • [SCH · RUC · CAR · DOA · CHA · IRI | BIOSTAT 2010]

Trisomy 21



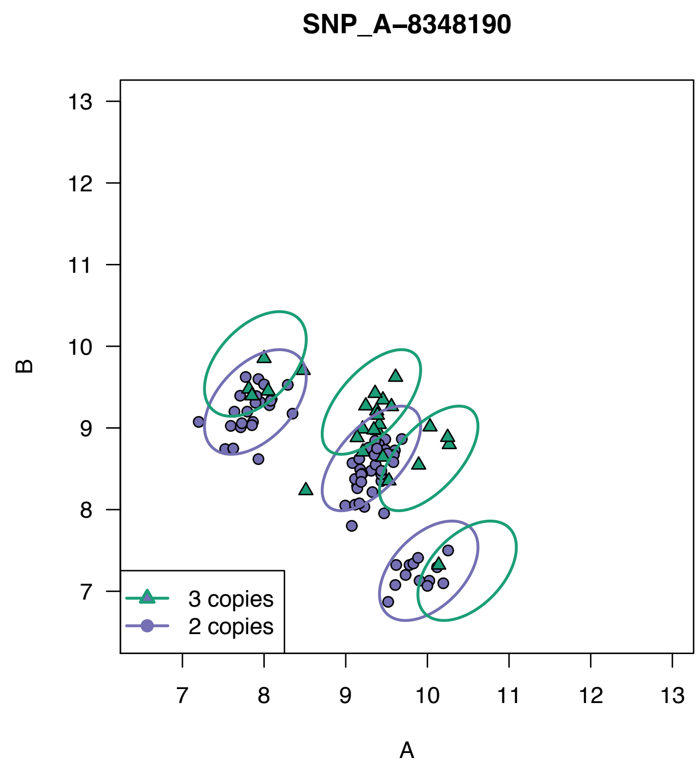
Samples from Aravinda Chakravarti and Betty Doan

Trisomy 21

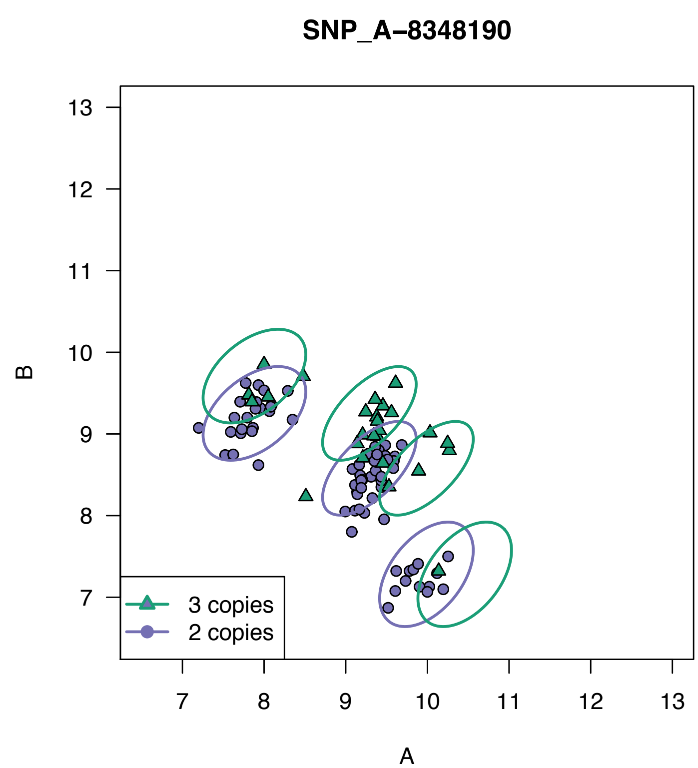


Samples from Aravinda Chakravarti and Betty Doan

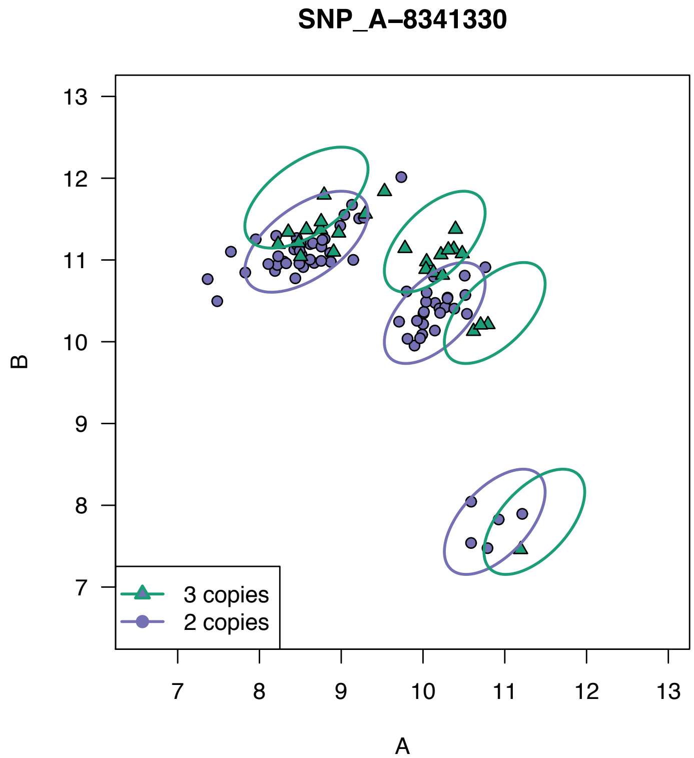
A versus B plots



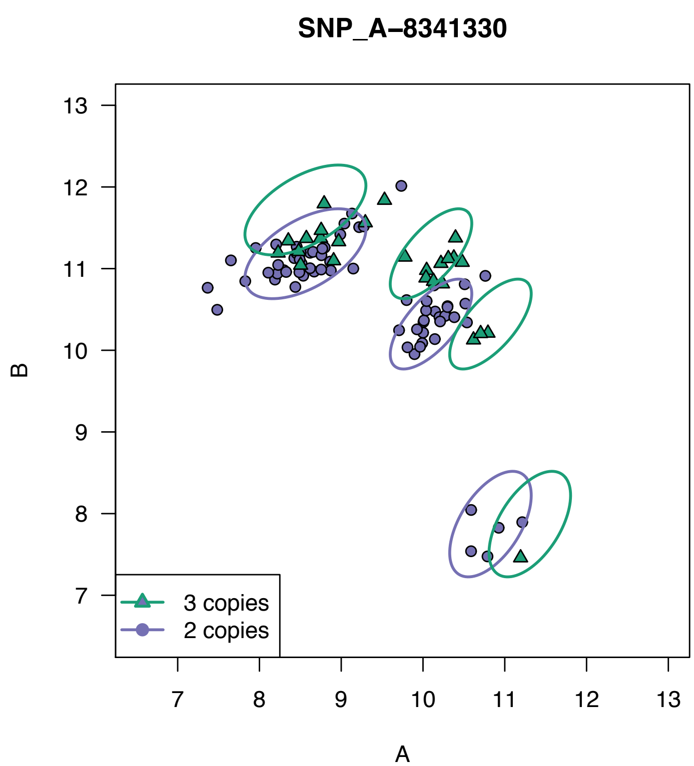
A versus B plots



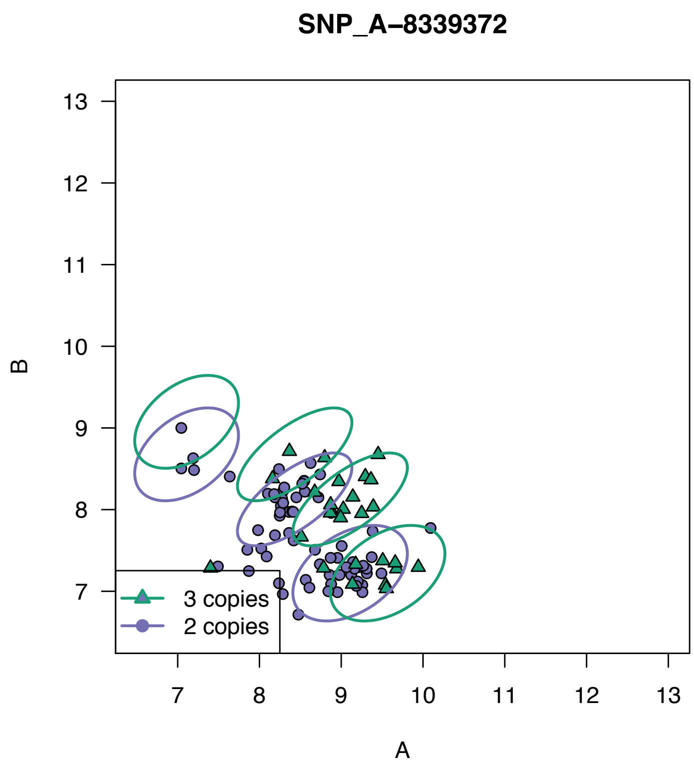
A versus B plots



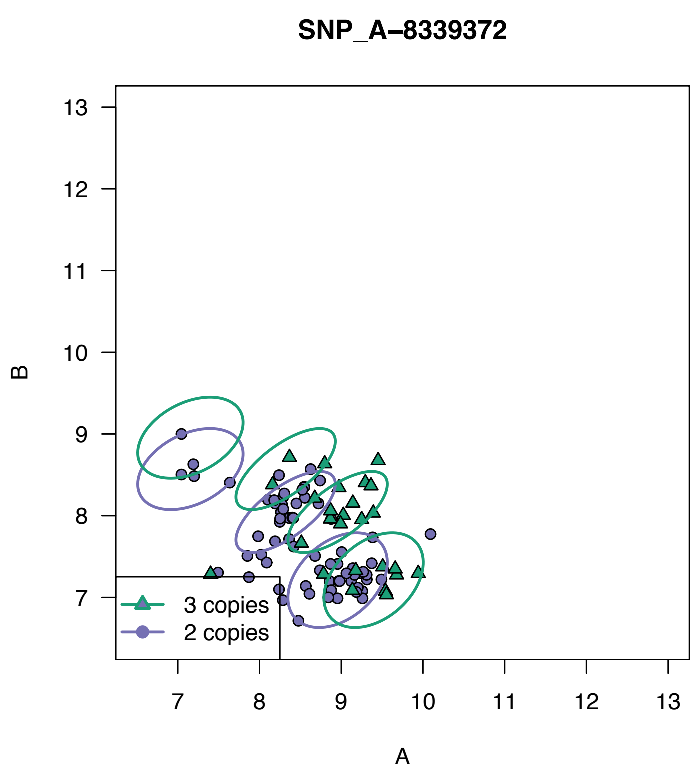
A versus B plots



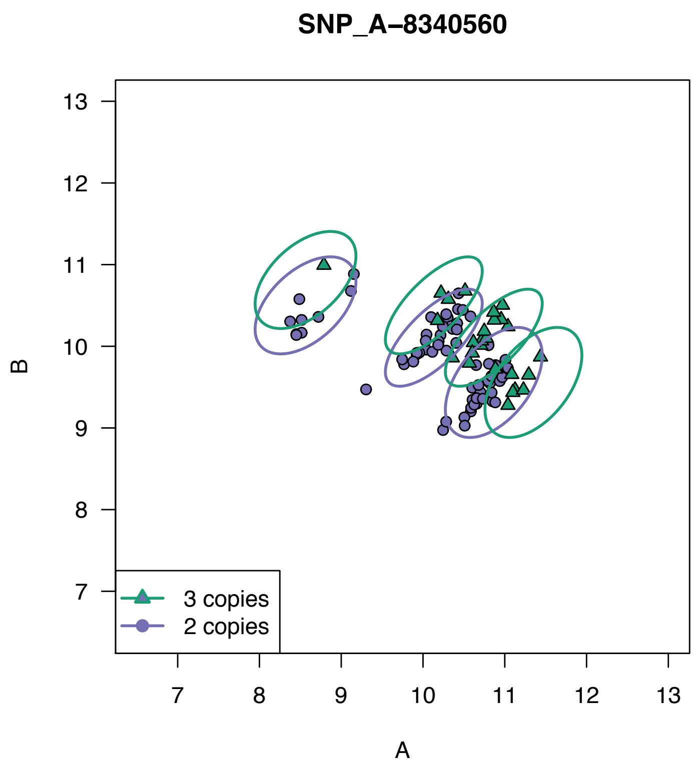
A versus B plots



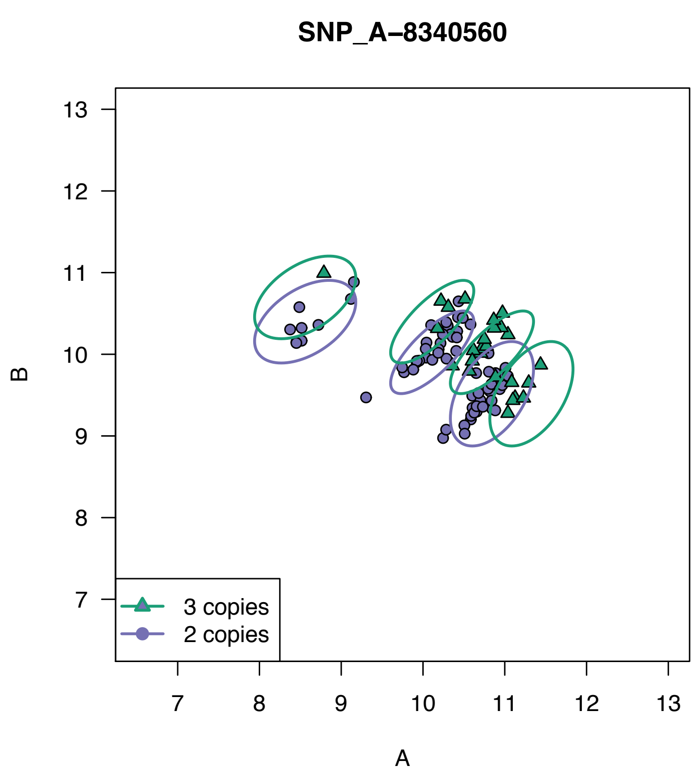
A versus B plots



A versus B plots

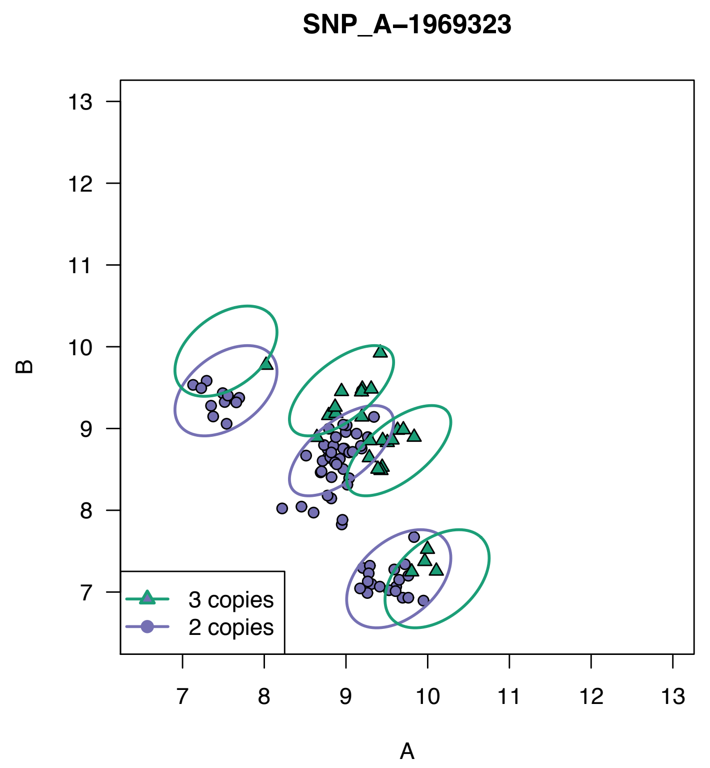


A versus B plots

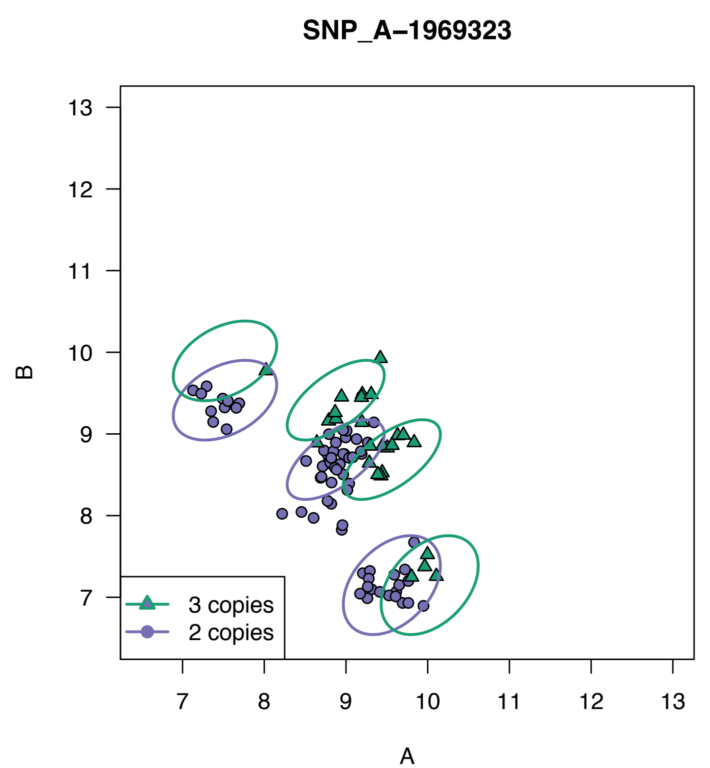


Ingo Ruczinski Assessing variants in the human genome

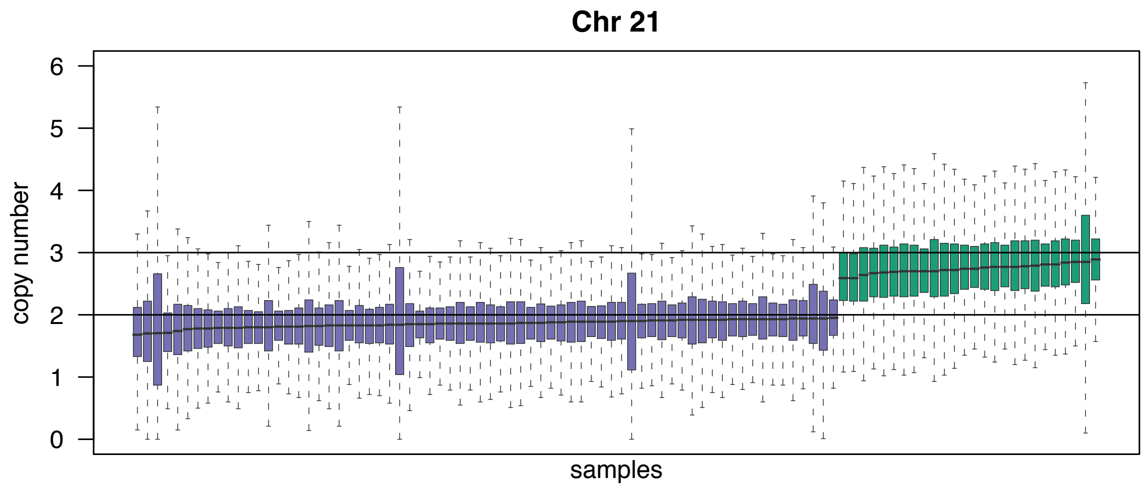
A versus B plots



A versus B plots

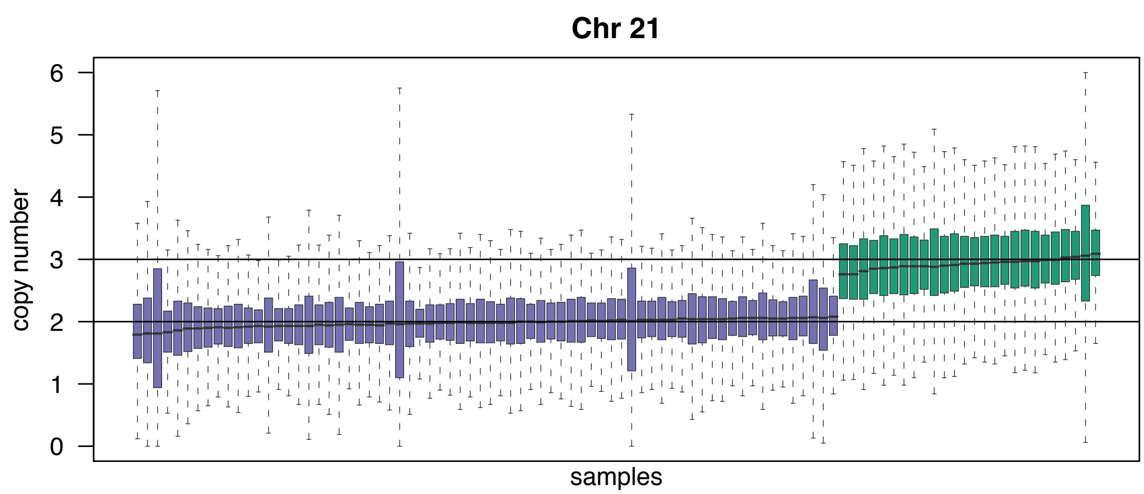


Trisomy 21



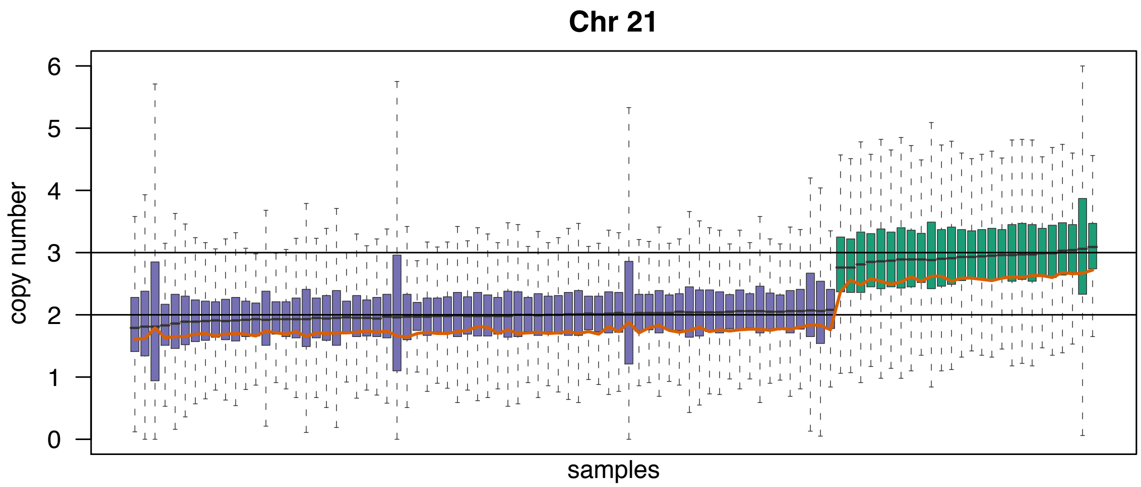
Samples from Aravinda Chakravarti and Betty Doan

Trisomy 21



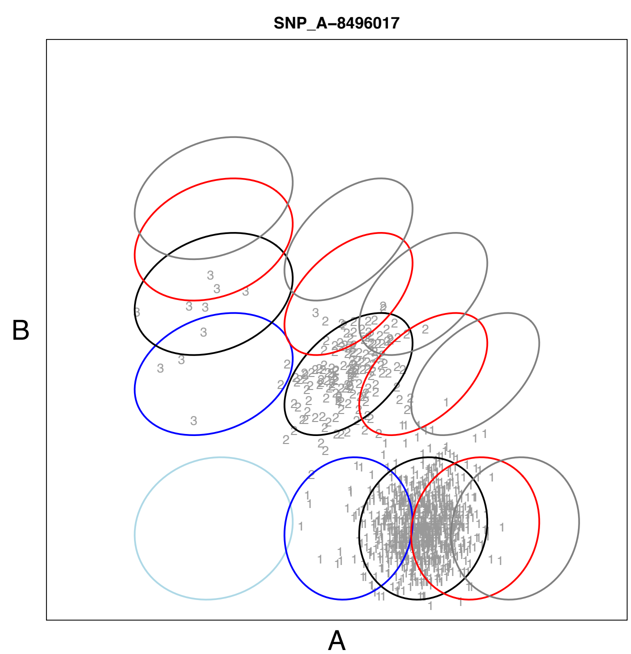
Samples from Aravinda Chakravarti and Betty Doan

Trisomy 21



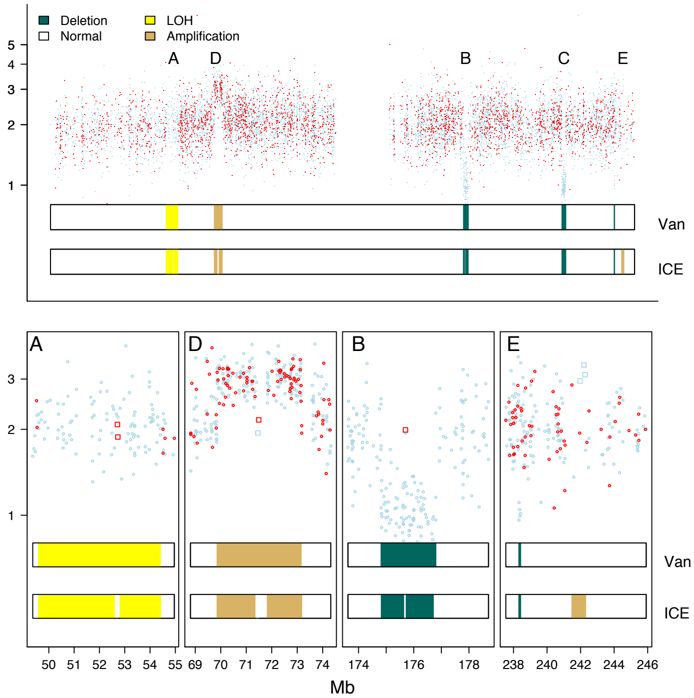
Samples from Aravinda Chakravarti and Betty Doan

Prediction regions for copy number



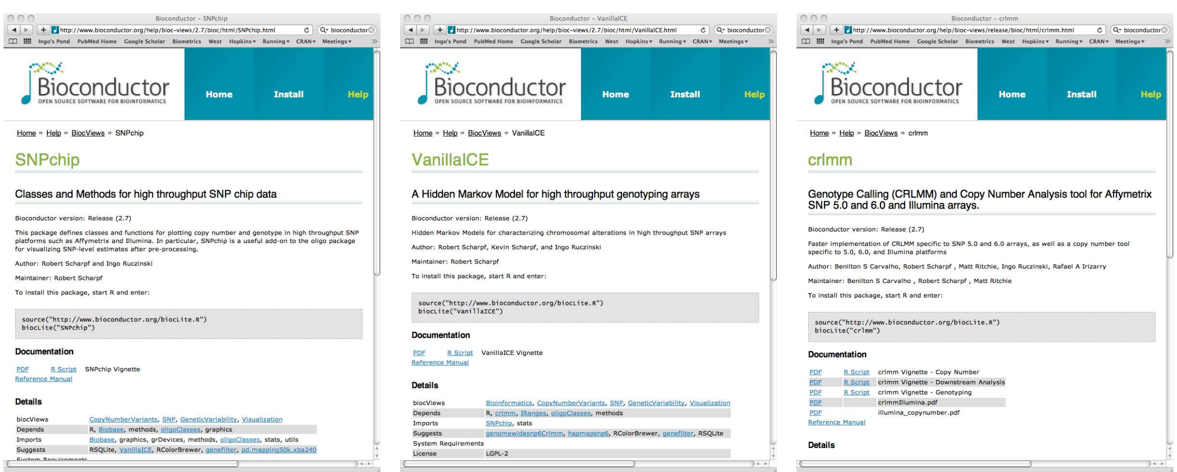
[SCH · RUC · CAR · DOA · CHA · IRI | BIOSTAT 2010]

Vanilla and ICE HMMs for genotype and copy number



[Sch · Par · Pev · Ruc | AoAS 2008]

Open source software



[Sch · · · Ruc | Bioinf 2007] • [Sch · Ruc | M·Mol·Bio 2010] • [Sch · Ruc · · · Iri | Biostat 2010]

A software vignette

Using the R Package crlmm for Genotyping and Copy Number Estimation

Robert B Scharpf
Johns Hopkins University

Rafael A Irizarry
Johns Hopkins University

Matthew E Ritchie
Walter+Eliza Hall Institute of Medical Research

Benilton Carvalho
University of Cambridge

Ingo Ruczinski
Johns Hopkins University

Abstract

Genotyping platforms such as Affymetrix can be used to assess genotype-phenotype as well as copy number-phenotype associations at millions of markers. While genotyping algorithms are largely concordant when assessed on HapMap samples, tools to assess copy number changes are more variable and often discordant. One explanation for the discordance is that copy number estimates are susceptible to systematic differences between groups of samples that were processed at different times or by different labs. Analysis algorithms that do not adjust for batch effects are prone to spurious measures of association. The R package **crlmm** implements a multilevel model that adjusts for batch effects and provides allele-specific estimates of copy number. This paper illustrates a workflow for the estimation of allele-specific copy number, develops marker- and study-level summaries of batch effects, and demonstrates how the marker-level estimates can be integrated with complimentary Bioconductor software for inferring regions of copy number gain or loss. All analyses are performed in the statistical environment R. A compendium for reproducing the analysis is available from the author's website (`http://www.biostat.jhsph.edu/~rscharpf/crlmmCompendium/index.html`).

Keywords: copy number, batch effects, robust, multilevel model, high-throughput, oligonucleotide array.

[SCH · IRI · RIT · CAR · RUC | TECH·REP 2010]

Compendium

Compendium for "Using the R Package crlmm for Genotyping and Copy Number Estimation" by Scharpf, et al. (2010)

http://www.biostat.jhsph.edu/~rscharpf/crlmmCompendium/index.html

Compendium Ingo's Pond PubMed Home Google Scholar Biometrics West Hopkins Running CRAN Meetings News Links Amazon SZ LEO

2.1 Reproducing the Figures

The `crlmmCompendium` package contains the text, data, and R functions used to make the figures in this paper. Users should be able to reproduce the figures upon successfull installation of the compendium. The compendium requires `R >= 2.12`. To install the compendium and its dependencies you will need an internet connection.

```
source("http://www.bioconductor.org/biocLite.R")
pkgs <- c("crlmm", "DNAcopy", "SNPchip", "RColorBrewer", "VanillaICE")
biocLite(pkgs)
install.packages("crlmmCompendium_1.0.4.tar.gz", repos=NULL)
```

To install the crlmmCompendium, download the tarball of the latest build:

R package	build
crlmmCompendium	1.0.4

The package can be installed from the command line by `R CMD INSTALL crlmmCompendium_1.0.4.tar.gz`, or from an R session in the same directory by:

```
install.packages("crlmmCompendium_1.0.4.tar.gz", repos=NULL)
```

Windows users would first need to install the appropriate \Rpackage{Rtools} executable.

R code extracted from the manuscript.Rnw vignette for reproducing the figures is available from the `Code` links adjacent to the figures below. To reproduce the figures, simply copy the code into R.

2.2 Reproducing the Manuscript

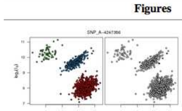
The complete analysis of the HapMap phase III data is contained in the manuscript.Rnw `Sweave` file. This document is located in the `inst/scripts` subdirectory of the `crlmmCompendium` package. Three additional steps are required for the complete analysis. First, one must download and install the HapMap Phase 3 CEL files for the Affymetrix 6.0 platform. Secondly, one must change the following lines in the manuscript vignette as appropriate:

```
pathToCels <- "/your/path/to/CEL/files"
outdir <- "/directory/to/store/results"
```

Finally, one must install additional package dependencies that were not required for installing the `crlmmCompendium`. In particular, the packages `ff`, `genefilter`, `ellipse`, and `MASS`. Note that the genotyping and copy number estimation steps in the manuscript.Rnw `Sweave` file involve long computations. We suggest submitting the code using `R CMD batch`. Provided that LaTeX is installed, a pdf version of the manuscript can be generated by issuing the following commands from R:

```
library(tools)
texi2dvi("manuscript.tex", pdf=TRUE)
```

3 Figures and Code

Figures	R code
	Code

Missing heritability

Vol 461|8 October 2009|doi:10.1038/nature08494 nature

REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable insights into their genetic architecture. Most variants identified so far confer relatively small increments in risk, and explain only a small proportion of familial clustering, leading many to question how the remaining, 'missing' heritability can be explained. Here we examine potential sources of missing heritability and propose research strategies, including and extending beyond current genome-wide association approaches, to illuminate the genetics of complex diseases and enhance its potential to enable effective disease prevention or treatment.

[Manolio et al | Nature 461: 2009]

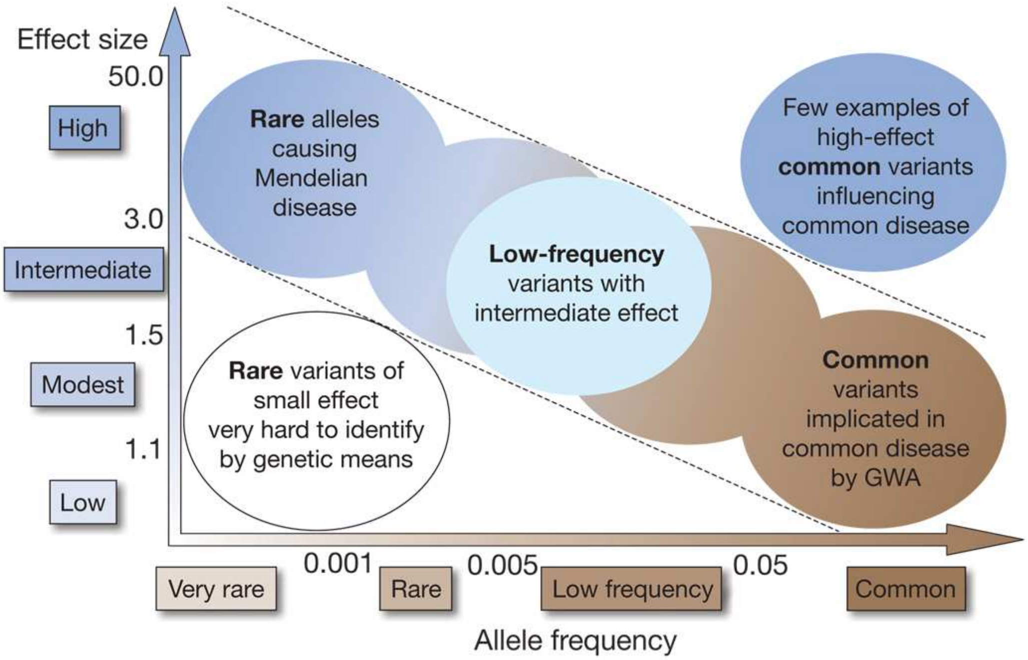
Missing heritability

Estimates of heritability and number of loci for several complex traits

Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration[72]	5	50%	Sibling recurrence risk
Crohn's disease[21]	32	20%	Genetic risk (liability)
Systemic lupus erythematosus[73]	6	15%	Sibling recurrence risk
Type 2 diabetes[74]	18	6%	Sibling recurrence risk
HDL cholesterol[75]	7	5.2%	Residual* phenotypic variance
Height[15]	40	5%	Phenotypic variance
Early onset myocardial infarction[76]	9	2.8%	Phenotypic variance
Fasting glucose[77]	4	1.5%	Phenotypic variance

*Residual is after adjustment for age, gender, diabetes.

Missing heritability



Genetic heterogeneity

Cell

Leading Edge
Essay

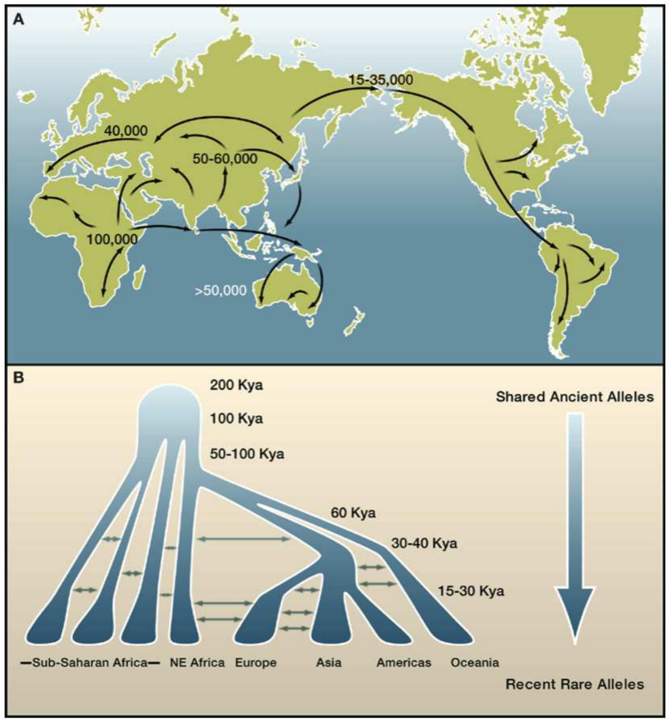
Genetic Heterogeneity in Human Disease

Jon McClellan[1,*] and Mary-Claire King[2,*]
[1]Department of Psychiatry
[2]Departments of Medicine and Genome Sciences
University of Washington, Seattle, WA 98195-7720, USA
*Correspondence: drjack@uw.edu (J.M.), mcking@uw.edu (M.-C.K.)
DOI 10.1016/j.cell.2010.03.032

Strong evidence suggests that rare mutations of severe effect are responsible for a substantial portion of complex human disease. Evolutionary forces generate vast genetic heterogeneity in human illness by introducing many new variants in each generation. Current sequencing technologies offer the possibility of finding rare disease-causing mutations and the genes that harbor them.

[McClellan and King | Cell 141: 2010]

Genetic heterogeneity



From the New Yorker

"O.K., let's slowly lower in the grant money."
Todd Bearson
Arlington, Mass.

NHLBI exome sequencing project

National Heart Lung and Blood Institute
National Institutes of Health

Program Officer: Deborah Applebaum-Bowden

Heart GO	Lung GO	Women's Health Initiative Sequencing Program
PI: Stephen Rich University of Virginia	PIs: Michael Bamshad U. of Washington Kathleen Barnes, Johns Hopkins University	PI: Rebecca Jackson Ohio State University

UNIVERSITY of WASHINGTON

PIs: Debbie Nickerson,
Mark Rieder,
Jay Shendure, & Phil Green

BROAD INSTITUTE

PIs: Stacey Gabriel &
David Altshuler