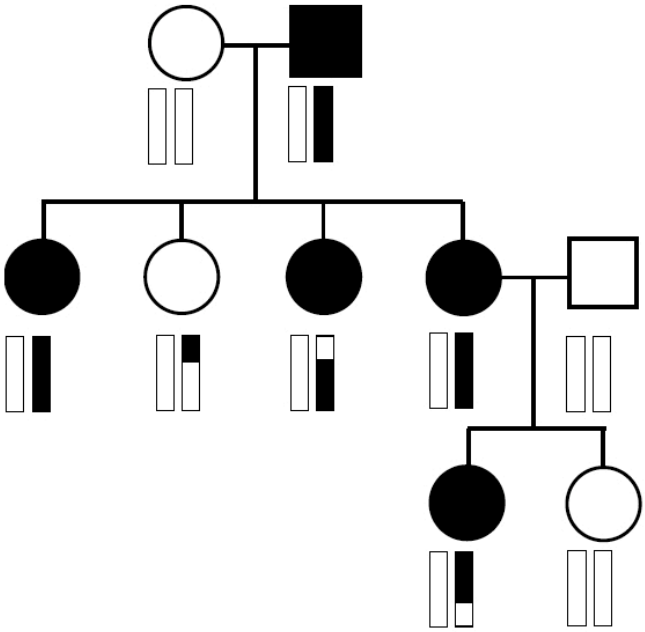
Variation in the human genome & disease

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

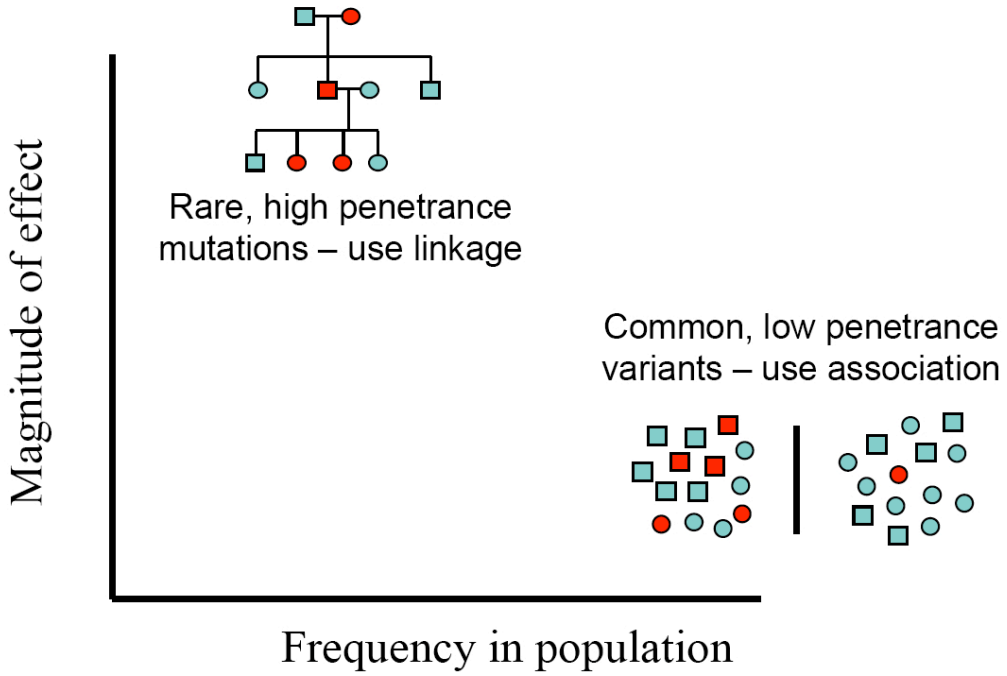
April 7, 2010

Linkage



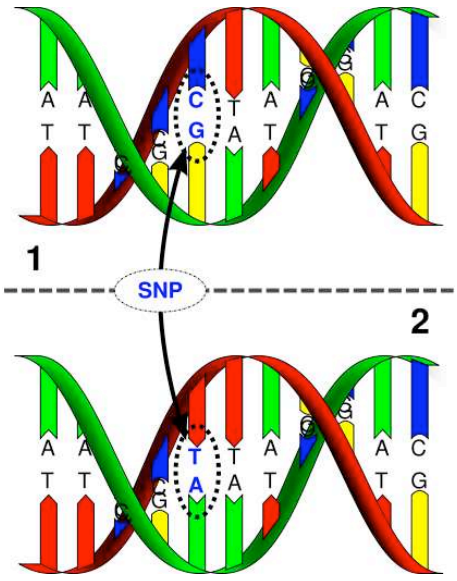
From Goncalo Abecasis

Linkage versus association



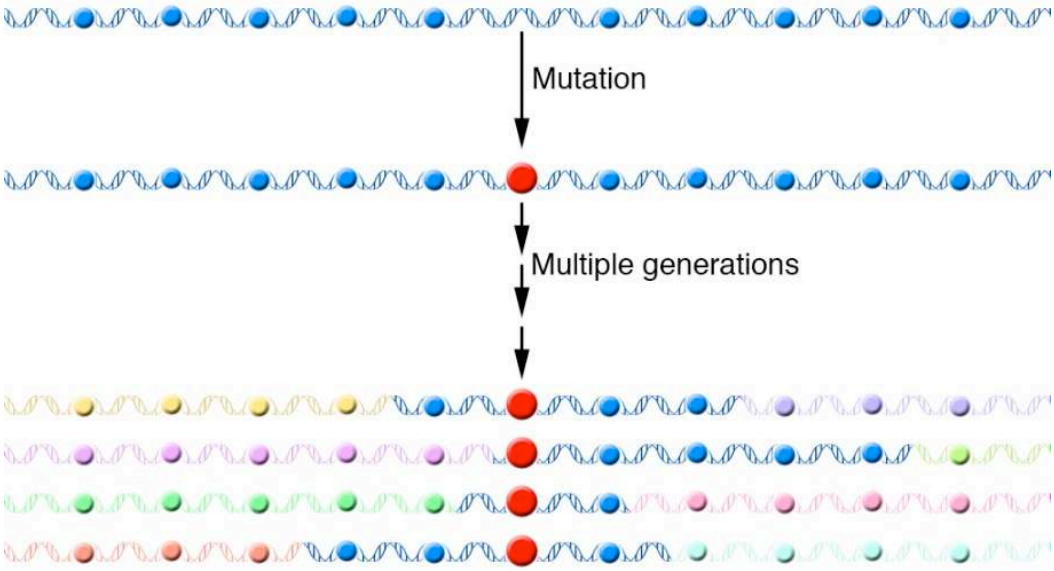
From Goncalo Abecasis

Single nucleotide polymorphisms



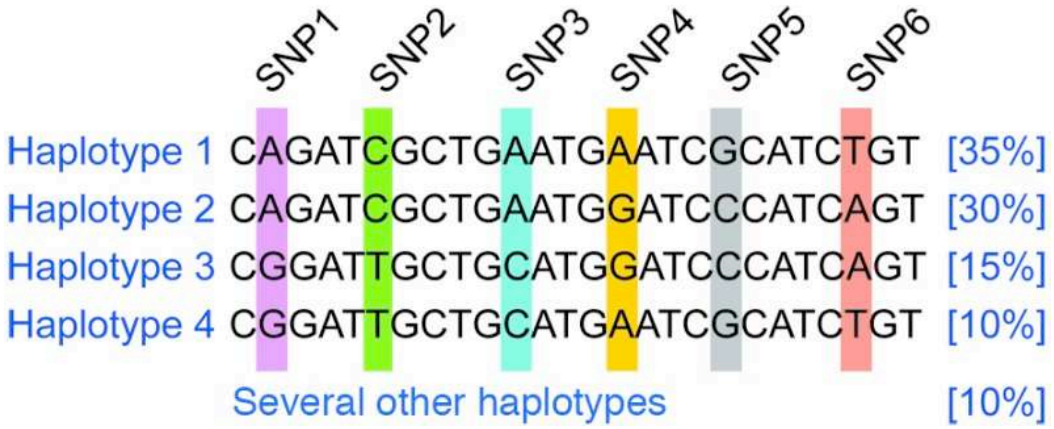
urgi.versailles.inra.fr

Haplotypes



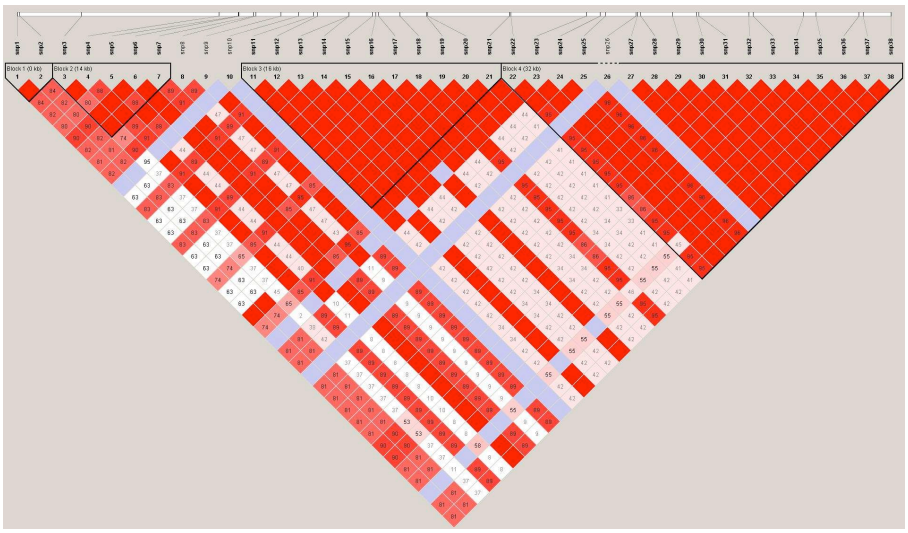
Manolio et al (2008), J Clin Invest 118(5): 1590-605.

Haplotypes



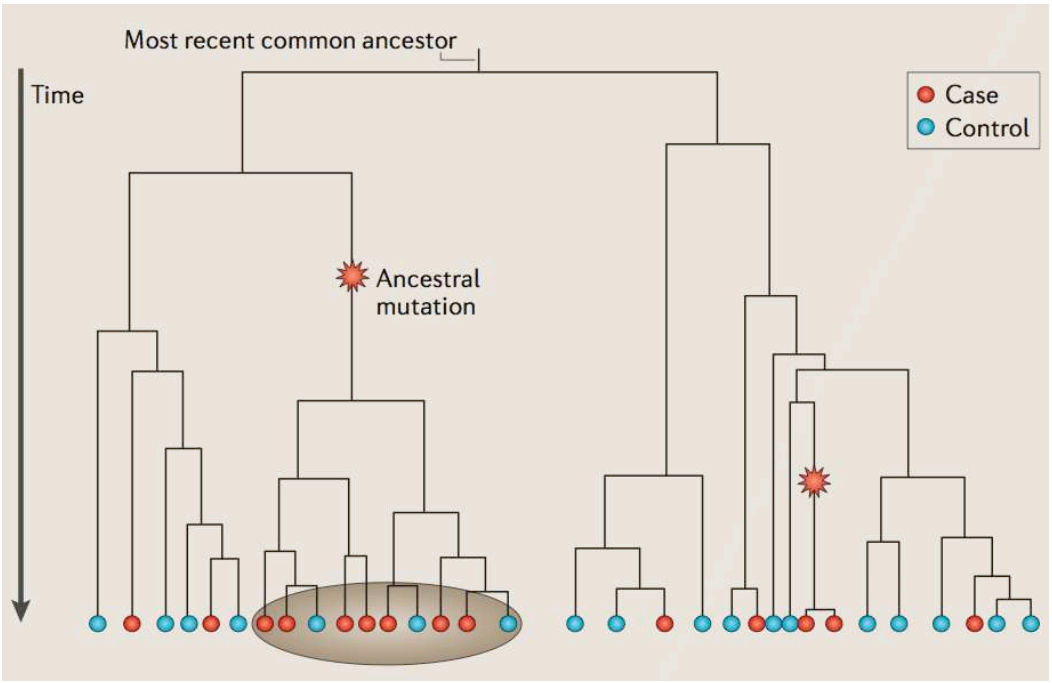
Manolio et al (2008), J Clin Invest 118(5): 1590-605.

Haplotype blocks



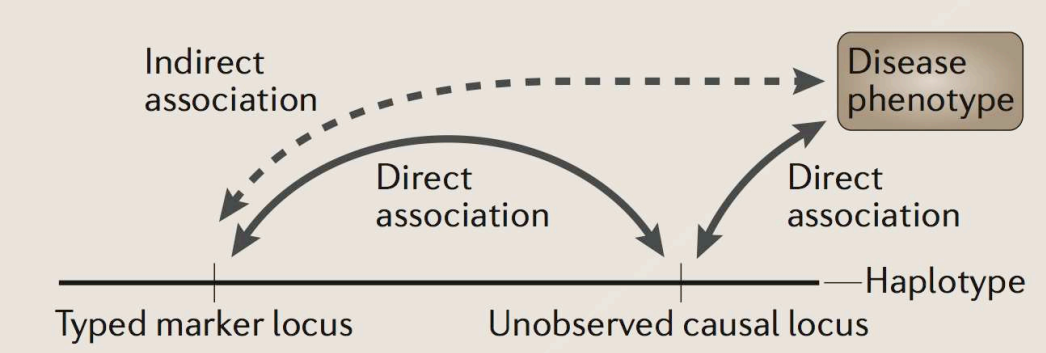
Kim and Dionne (2007)

Case-control design



Balding DJ (2006). Nat Rev Genet. 7(10): 781-91.

Association



Balding DJ (2006). Nat Rev Genet. 7(10): 781-91.

Coverage

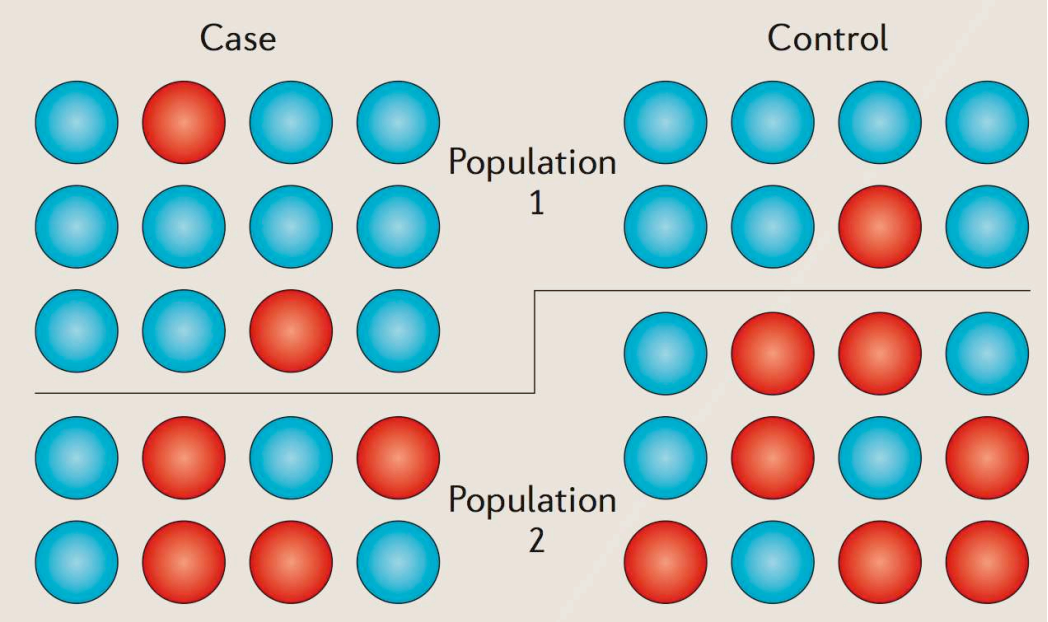
Table 1
Estimated coverage of commercially available fixed marker genotyping platforms

Platform	HapMap population sample		
	YRI	CEU	CHB + JPT
Affymetrix GeneChip 500K	46	68	67
Affymetrix SNP Array 6.0	66	82	81
Illumina HumanHap300	33	77	63
Illumina HumanHap550	55	88	83
Illumina HumanHap650Y	66	89	84
Perlegen 600K	47	92	84

Data represent percent of SNPs tagged at $r^2 \geq 0.8$. Values assume all SNPs on the platform are informative and pass quality control. YRI, Yoruba in Ibadan, Nigeria; CEU, subsample of Utah residents of Northern European ancestry selected from Centre d'Étude du Polymorphisme Humain samples; CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo. From the International HapMap Consortium, 2007 (3).

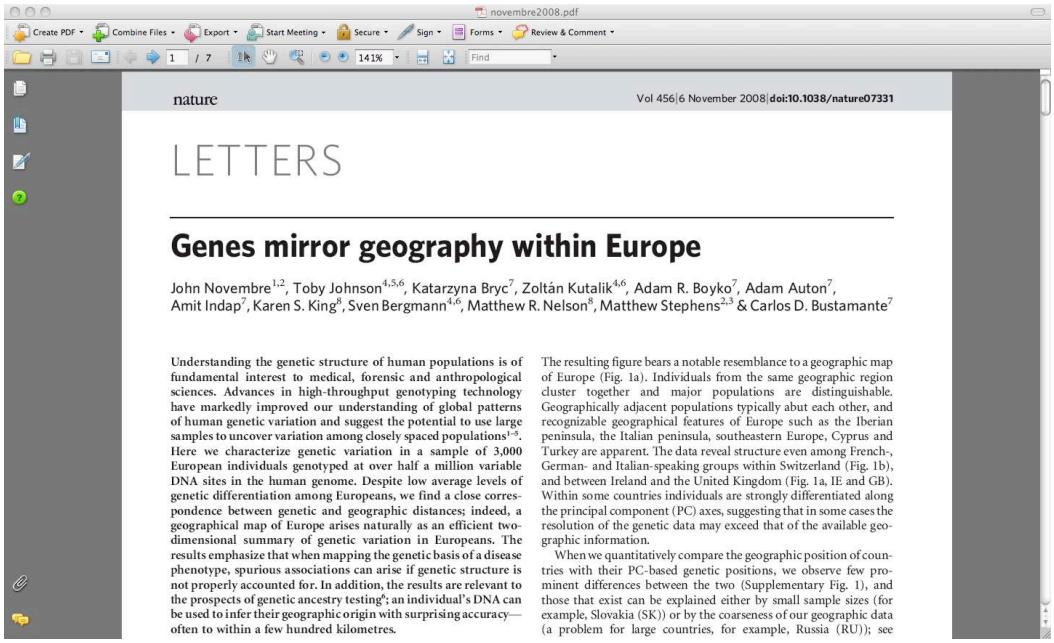
Manolio et al (2008), J Clin Invest 118(5): 1590-605.

Population stratification



Balding DJ (2006). Nat Rev Genet. 7(10): 781-91.

Principal components



novembre2008.pdf

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

LETTERS

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]

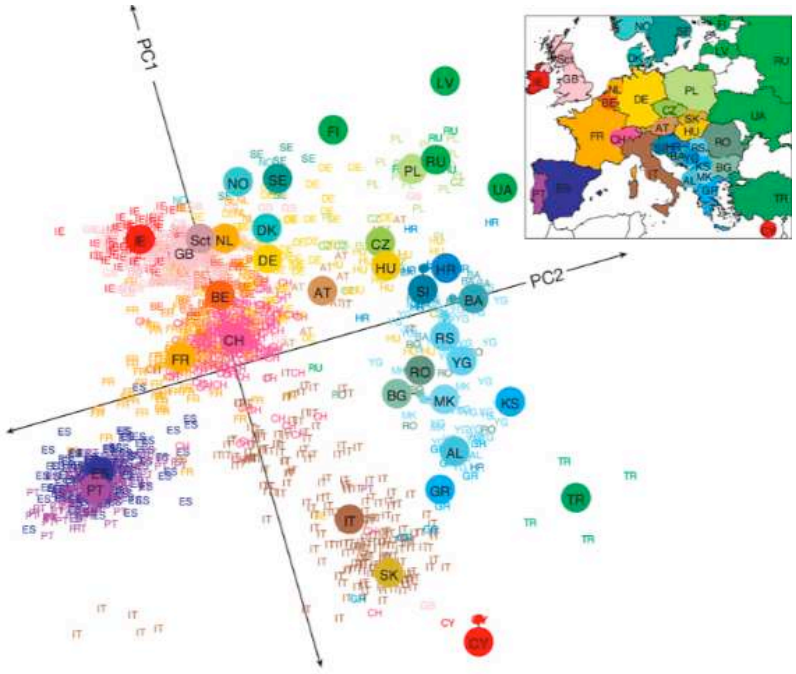
Understanding the genetic structure of human populations is of fundamental interest to medical, forensic and anthropological sciences. Advances in high-throughput genotyping technology have markedly improved our understanding of global patterns of human genetic variation and suggest the potential to use large samples to uncover variation among closely spaced populations[1–3]. Here we characterize genetic variation in a sample of 3,000 European individuals genotyped at over half a million variable DNA sites in the human genome. Despite low average levels of genetic differentiation among Europeans, we find a close correspondence between genetic and geographic distances; indeed, a geographical map of Europe arises naturally as an efficient two-dimensional summary of genetic variation in Europeans. The results emphasize that when mapping the genetic basis of a disease phenotype, spurious associations can arise if genetic structure is not properly accounted for. In addition, the results are relevant to the prospects of genetic ancestry testing[4]; an individual's DNA can be used to infer their geographic origin with surprising accuracy—often to within a few hundred kilometres.

The resulting figure bears a notable resemblance to a geographic map of Europe (Fig. 1a). Individuals from the same geographic region cluster together and major populations are distinguishable. Geographically adjacent populations typically abut each other, and recognizable geographical features of Europe such as the Iberian peninsula, the Italian peninsula, southeastern Europe, Cyprus and Turkey are apparent. The data reveal structure even among French-, German- and Italian-speaking groups within Switzerland (Fig. 1b), and between Ireland and the United Kingdom (Fig. 1a, IE and GB). Within some countries individuals are strongly differentiated along the principal component (PC) axes, suggesting that in some cases the resolution of the genetic data may exceed that of the available geographic information.

When we quantitatively compare the geographic position of countries with their PC-based genetic positions, we observe few prominent differences between the two (Supplementary Fig. 1), and those that exist can be explained either by small sample sizes (for example, Slovakia (SK)) or by the coarseness of our geographic data (a problem for large countries, for example, Russia (RU)); see

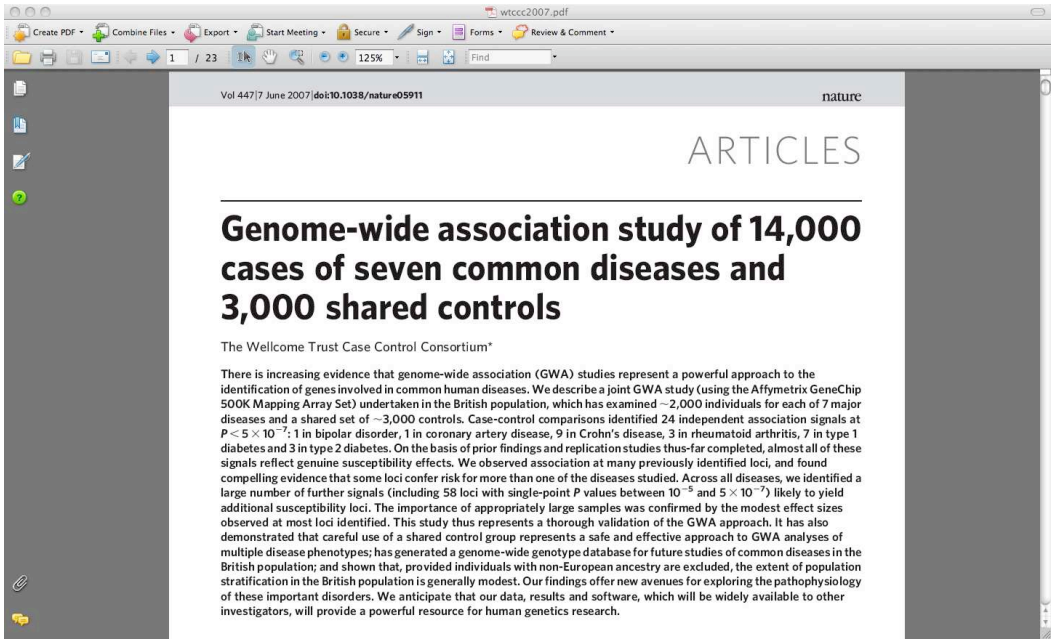
Novembre et al (2008), Nature 456: 98-101.

Principal components



Novembre et al (2008), Nature 456: 98-101.

WTCCC


wtccc2007.pdf

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

ARTICLES

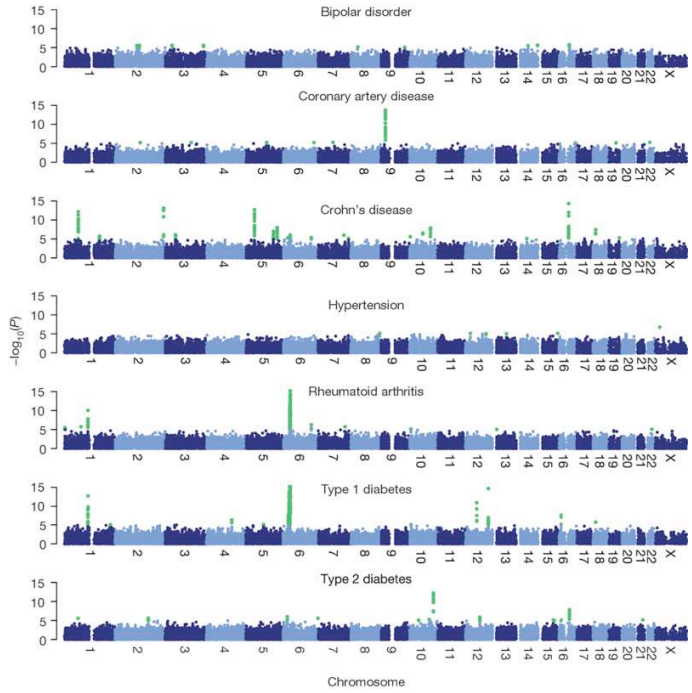
Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

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.

Causality

PNAS PNAS PNAS

Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout

Owen M. Woodward[a,1], Anna Köttgen[b,1], Josef Coresh[b], Eric Boerwinkle[c], William B. Guggino[a], and Michael Köttgen[d,2]

[a]Department of Physiology, Johns Hopkins Medical Institutions, Baltimore, MD 21205; [b]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; [c]Human Genetics Center and Division of Epidemiology, University of Texas, Houston, TX 77030; and [d]Department of Biological Chemistry, Johns Hopkins Medical Institutions, Baltimore, MD 21205

Edited by Maurice B. Burg, National Institutes of Health, Bethesda, MD, and approved May 4, 2009 (received for review February 4, 2009)

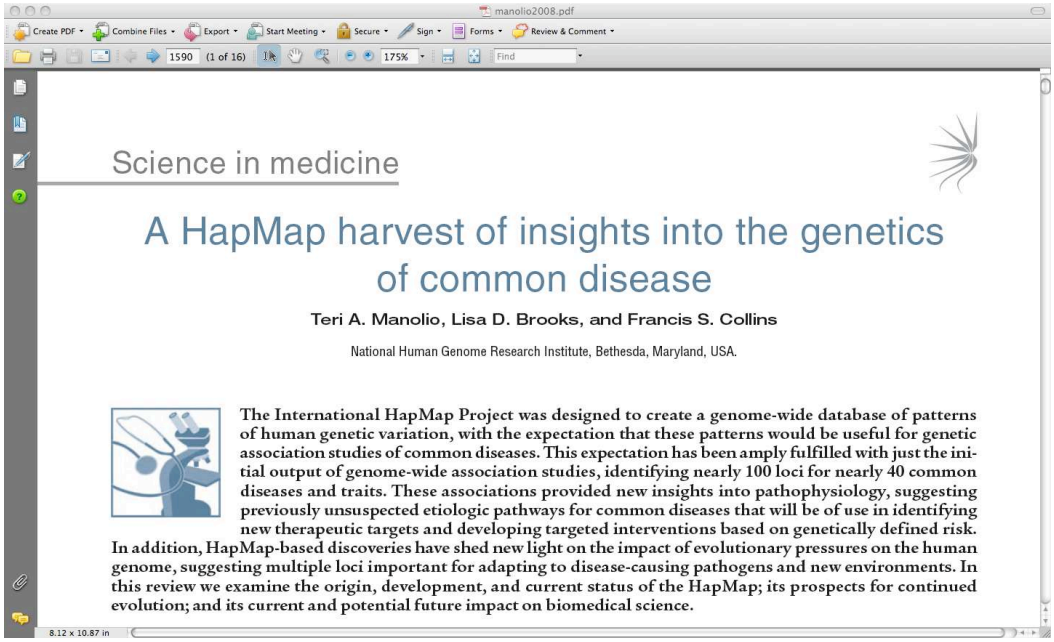
Genome-wide association studies (GWAS) have successfully identified common single nucleotide polymorphisms (SNPs) associated with a wide variety of complex diseases, but do not address gene function or establish causality of disease-associated SNPs. We recently used GWAS to identify SNPs in a genomic region on chromosome 4 that associate with serum urate levels and gout, a consequence of elevated urate levels. Here we show using functional assays that human ATP-binding cassette, subfamily G, 2 (ABCG2), encoded by the *ABCG2* gene contained in this region, is a hitherto unknown urate efflux transporter. We further show that native ABCG2 is located in the brush border membrane of kidney proximal tubule cells, where it mediates renal urate secretion. Introduction of the mutation Q141K encoded by the common SNP rs2231142 by site-directed mutagenesis resulted in 53% reduced urate transport rates compared to wild-type ABCG2 ($P < 0.001$). Data from a population-based study of 14,783 individuals support rs2231142 as the causal variant in the region and show highly significant associations with urate levels [whites: $P = 10^{-30}$, minor allele frequency (MAF) 0.11; blacks $P = 10^{-4}$, MAF 0.03] and gout (adjusted odds ratio 1.68 per risk allele, both races). Our data indicate that at least 10% of all gout cases in whites are attributable to this causal variant. With approximately 3 million US individuals suffering from often insufficiently treated gout, ABCG2 represents an attractive drug target. Our study completes the chain of evidence from association to causation and supports the common disease-common variant hypothesis in the etiology of gout.

causal variant | genome-wide association study | uric acid

in oocytes and urate efflux rates from oocytes were measured. Urate accumulation was significantly decreased by 75.5% in oocytes expressing ABCG2 compared with water-injected control oocytes (Fig. 1*A*). In addition, ABCG2-expressing oocytes showed lower urate concentrations than oocytes expressing the known urate efflux transporter MRP4 (7). The reduced urate accumulation in ABCG2-expressing oocytes was absent in the presence of fumitremorgin C (FTC), a specific ABCG2 inhibitor, or after introduction of a mutation in ABCG2 (187T) that is known to disrupt transport of chemotherapeutic agents (5, 6) (Fig. 1*B*). ABCG2-expressing oocytes showed significantly lower urate accumulation over a wide range of extracellular concentrations compared with control cells or cells expressing the loss-of-function mutation ABCG2 187T (Fig. 1*C*). The reduced urate accumulation was due to ABCG2-mediated export of urate out of the cells as shown in experiments monitoring the decrease of the intracellular urate concentration over time in oocytes preloaded with radiolabeled urate (Fig. 1*D*). Furthermore, ABCG2-mediated urate efflux was dependent on the intracellular urate concentration and significantly higher in ABCG2-expressing oocytes than in control oocytes (Fig. 1*E*). Together, these data show that ABCG2 is a urate efflux transporter.

ABCG2 in Renal Epithelial Cells. In mammals, the proximal tubule is the major site of renal urate handling (1). To study the urate transport capacity of endogenous ABCG2 in polarized renal epithelial cells, we measured urate accumulation in native LLC-PK_1 cells. Inhibition of ABCG2 by FTC resulted in a

Results



manolio2008.pdf

Science in medicine

A HapMap harvest of insights into the genetics of common disease

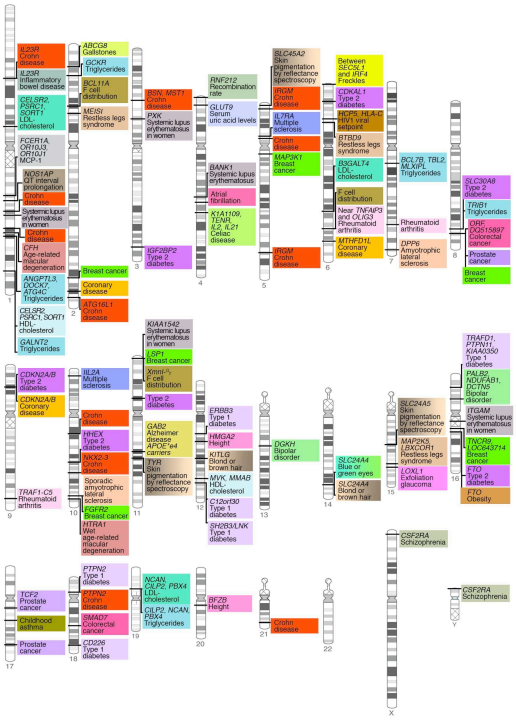
Teri A. Manolio, Lisa D. Brooks, and Francis S. Collins

National Human Genome Research Institute, Bethesda, Maryland, USA.

The International HapMap Project was designed to create a genome-wide database of patterns of human genetic variation, with the expectation that these patterns would be useful for genetic association studies of common diseases. This expectation has been amply fulfilled with just the initial output of genome-wide association studies, identifying nearly 100 loci for nearly 40 common diseases and traits. These associations provided new insights into pathophysiology, suggesting previously unsuspected etiologic pathways for common diseases that will be of use in identifying new therapeutic targets and developing targeted interventions based on genetically defined risk. In addition, HapMap-based discoveries have shed new light on the impact of evolutionary pressures on the human genome, suggesting multiple loci important for adapting to disease-causing pathogens and new environments. In this review we examine the origin, development, and current status of the HapMap; its prospects for continued evolution; and its current and potential future impact on biomedical science.

Manolio et al (2008), J Clin Invest 118(5): 1590-605.

Results



Manolio et al (2008), J Clin Invest 118(5): 1590-605.

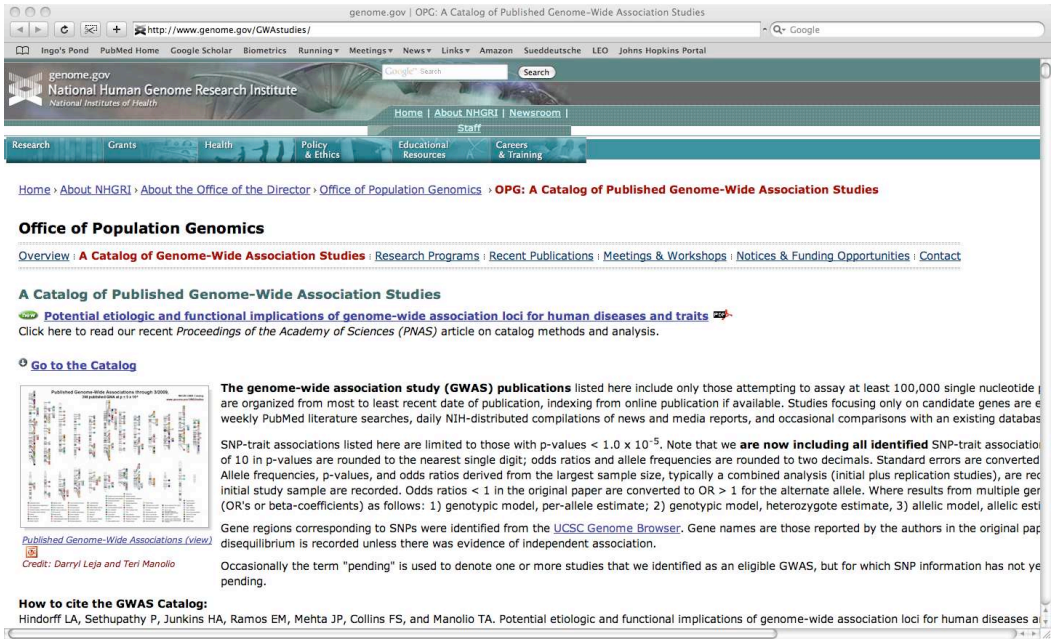
Results

Catalogued findings for the following phenotypes:

1. Eye diseases
2. Diabetes
3. Cancer
4. Gastrointestinal disorders
5. Cardiovascular conditions and lipid metabolism
6. Neuropsychiatric conditions
7. Autoimmune and infectious diseases
8. Various traits

Manolio et al (2008), J Clin Invest 118(5): 1590-605.

Results



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

Results

Phenotypes under investigation in collaborative GWA studies

GAIN[A]	GEI[B]	STAMPEED[C]	CGEMS[D]
Attention deficit hyperactivity disorder	Type 2 diabetes	Early-onset myocardial infarction	Prostate cancer
Major depressive disorder	Maternal metabolism and birth weight	Asthma	Breast cancer
Bipolar I disorder	Preterm birth	Platelet phenotypes	Pancreatic cancer
Schizophrenia	Oral clefts	CHD and other heart, lung and blood disorders	Lung cancer
Type I diabetic nephropathy	Dental caries	Childhood respiratory outcomes	Bladder cancer
Psoriasis	Coronary disease	Hematopoietic cell transplant outcome	Renal cancer
	Lung cancer	Arteriosclerosis in hypertensives	
	Addiction	Asthma and lung function	
		Cardiovascular risk factors	
		Atherosclerosis pathway genes	
		CV events	
		Early coronary artery disease	
		Phenotypic variability in sickle cell anemia	
		Centenarians	

CHD, coronary heart disease; CV, cardiovascular. [A]Genetic Association Information Network (http://www.fnih.org/GAIN2/home_new.shtml). [B]Genes, Environment, and Health Initiative (http://www.gei.nih.gov/). [C]SNP Typing for Association with Multiple Phenotypes from Existing Epidemiologic Data (http://public.nhlbi.nih.gov/GeneticsGenomics/home/stampeed.aspx). [D]Cancer Genetic Markers of Susceptibility (http://cgems.cancer.gov/).

Manolio et al (2008), J Clin Invest 118(5): 1590-605.

Missing heritability

manolio2010.pdf

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 TA et al (2009). Nature. 461: 747-53.

Missing heritability

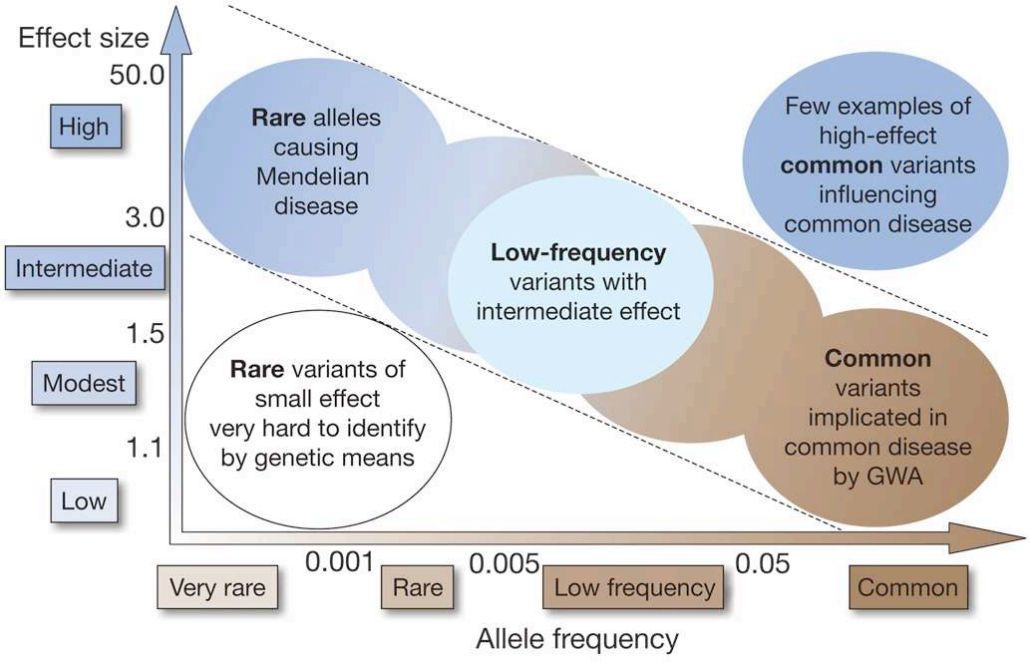
Estimates of heritability and number of loci for several complex traits

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.

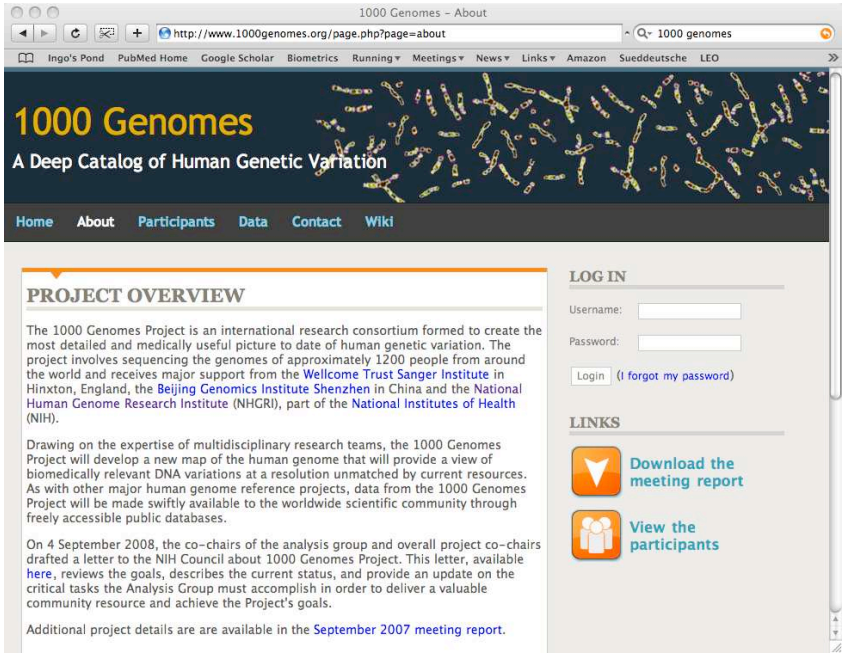
Manolio TA et al (2009). Nature. 461: 747-53.

Missing heritability



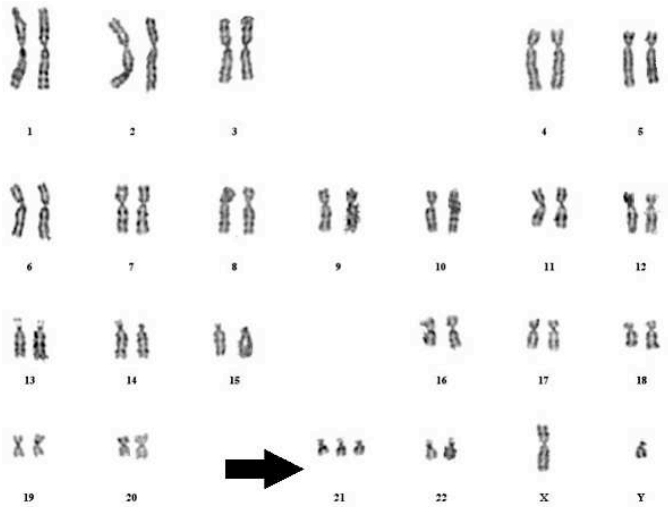
Manolio TA et al (2009). Nature. 461: 747-53.

1000 Genomes

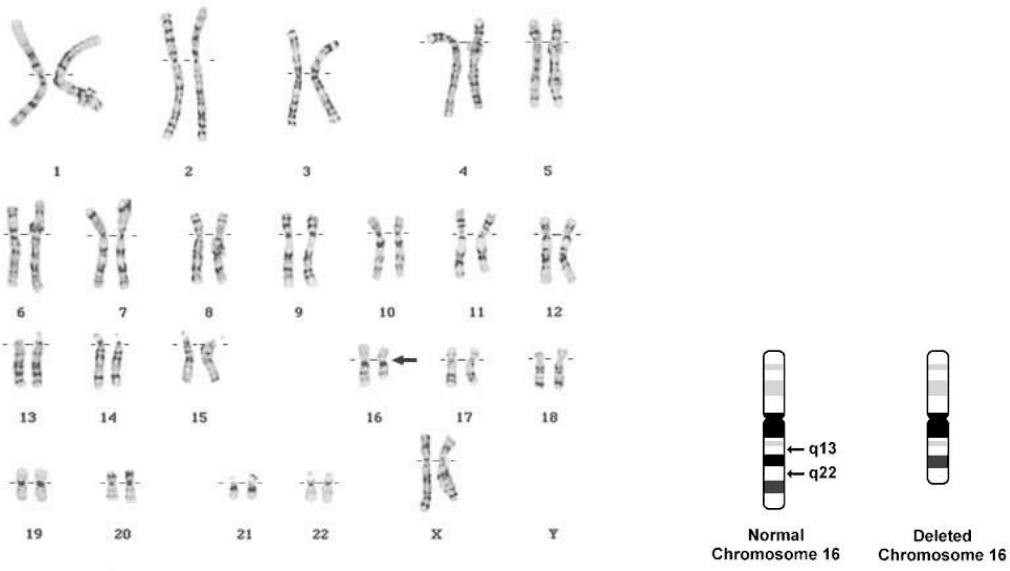


http://www.1000genomes.org

Trisomy

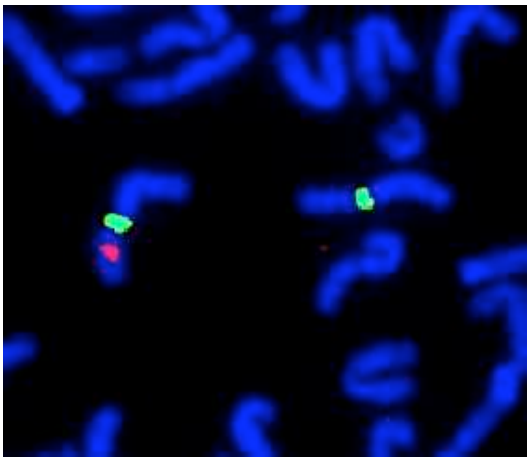


Karyotypes



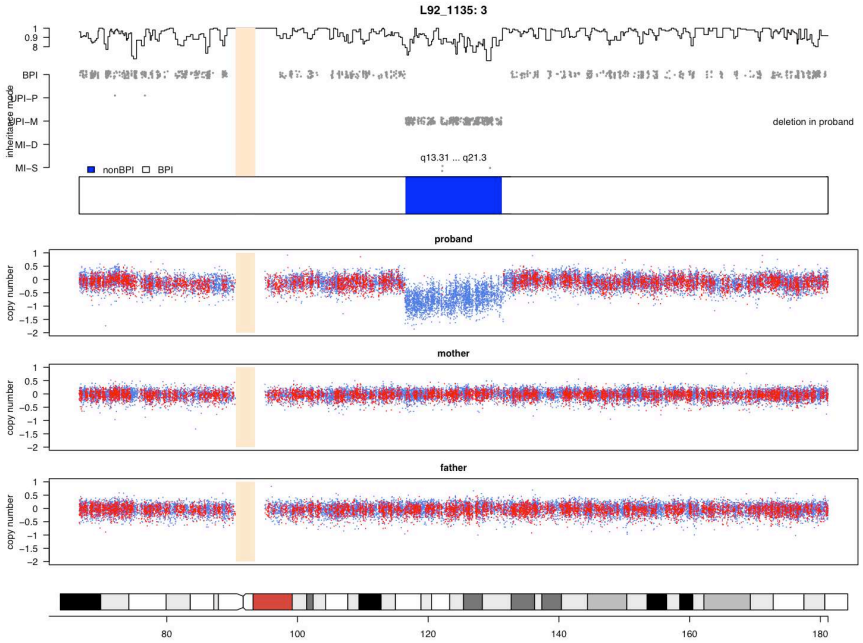
General Cytogenetics Information *http://members.aol.com/chrominfo/*

FISH



Courtesy of the Pevsner Laboratory

De novo deletion



Copy number variants

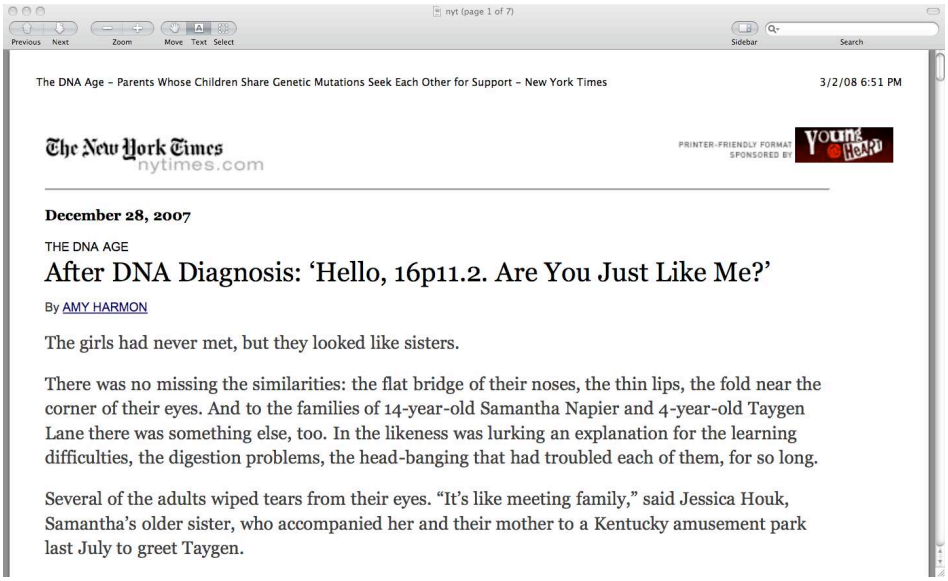
Selected disease associations with rare CNVs and common CNPs

Disease	Locus	Type of CNV	Size (kb)	Population frequency	Case frequency	Effect size (OR)
Rare CNVs						
Autism/IMR[59]	16p11.2	*De novo* deletion	600	1×10^{-4}	1%	100
Autism[59]	16p11.2	Rare duplication	600	3×10^{-4}	0.50%	16
Schizophrenia[60,78]	1q21.1	Rare deletion	1,400	2×10^{-4}	0.30%	15
IMR[79]	1q21.1	Rare deletion	1,400	2×10^{-4}	0.47%	Not observed in 4,737controls
Schizophrenia[60,78]	15q13.3	Rare deletion	1,600	2×10^{-4}	0.20%	12
Epilepsy[80]	15q13.3	Rare deletion	1,600	2×10^{-4}	1.0%	Not observed in 3,699 controls
IMR[79,81]	15q13.3	Rare deletion	1,600	2×10^{-4}	0.30%	Not observed in 960 controls
Schizophrenia[82]	22q11.2	Rare deletion	3,000	2.5×10^{-4}	1%	40
Common CNPs						
Crohn's disease[83]	*IRGM*	Deletion polymorphism	20	7%	11%	1.5
Body mass index[61]	*NEGR1*	Deletion polymorphism	45	65%	Quantitative trait	<1 kg
Psoriasis[84]	*LCE3C*	Deletion polymorphism	30	55%	65%	1.3

IMR, idiopathic mental retardation.

Manolio TA et al (2009). Nature. 461: 747-53.

Clinical practice



The DNA Age - Parents Whose Children Share Genetic Mutations Seek Each Other for Support - New York Times 3/2/08 6:51 PM

The New York Times nytimes.com

PRINTER-FRIENDLY FORMAT SPONSORED BY Young Heart

December 28, 2007

THE DNA AGE

After DNA Diagnosis: 'Hello, 16p11.2. Are You Just Like Me?'

By AMY HARMON

The girls had never met, but they looked like sisters.

There was no missing the similarities: the flat bridge of their noses, the thin lips, the fold near the corner of their eyes. And to the families of 14-year-old Samantha Napier and 4-year-old Taygen Lane there was something else, too. In the likeness was lurking an explanation for the learning difficulties, the digestion problems, the head-banging that had troubled each of them, for so long.

Several of the adults wiped tears from their eyes. "It's like meeting family," said Jessica Houk, Samantha's older sister, who accompanied her and their mother to a Kentucky amusement park last July to greet Taygen.

New York Times, December 28, 2007

More information

→ The confidence in genotype calls can differ substantially between SNPs!

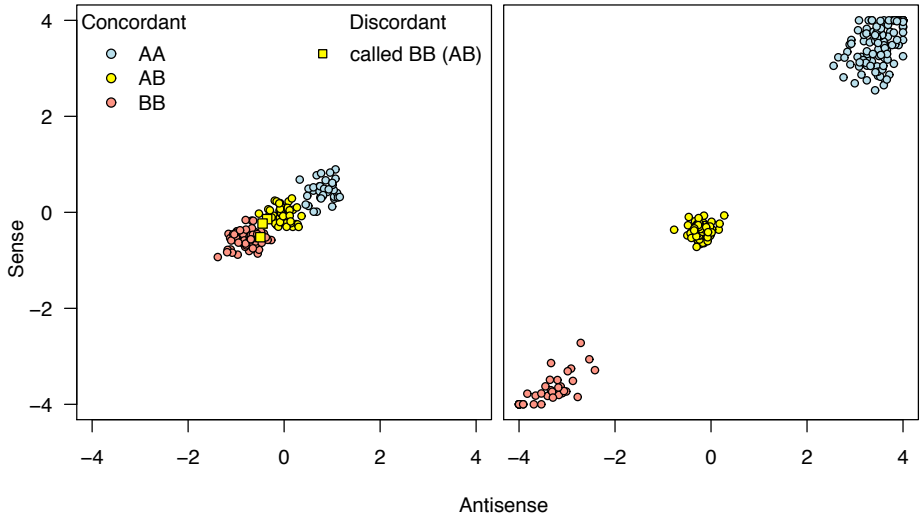
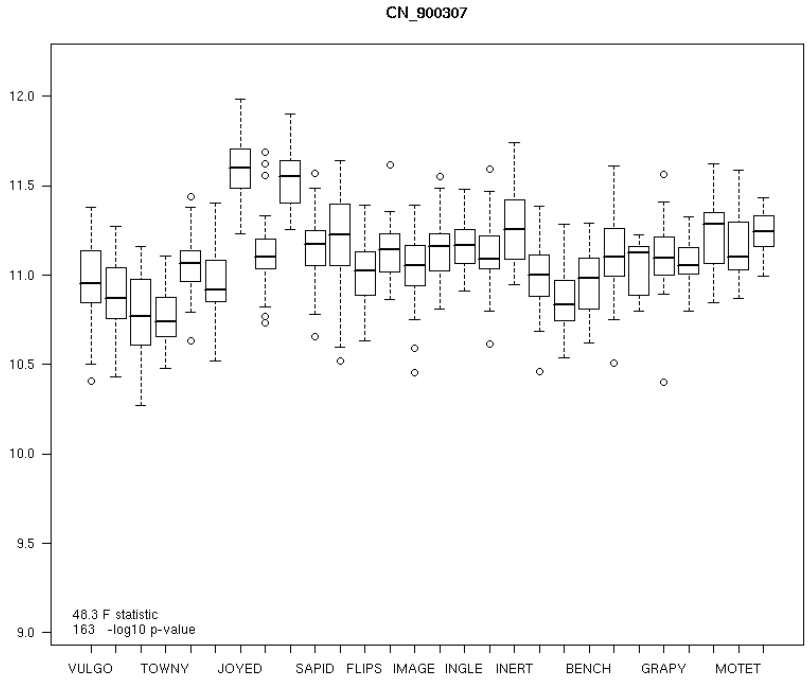
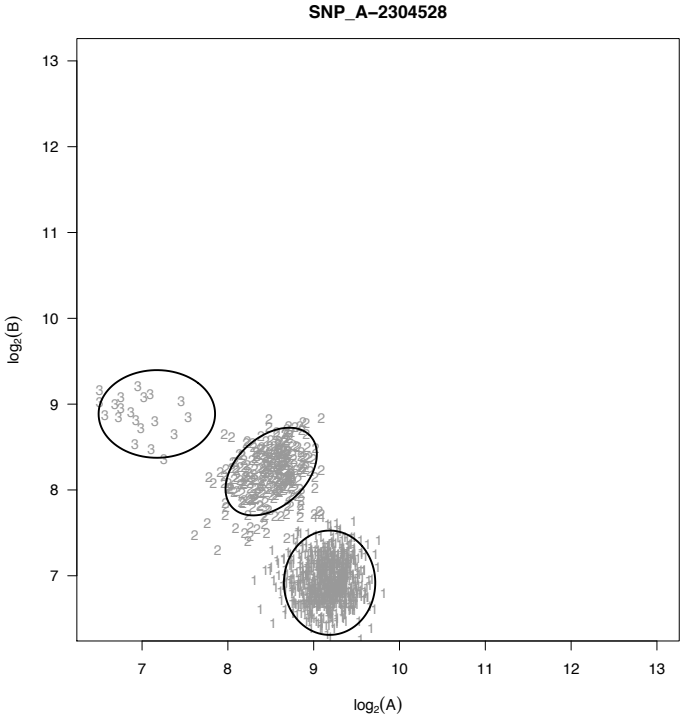


Plate effects

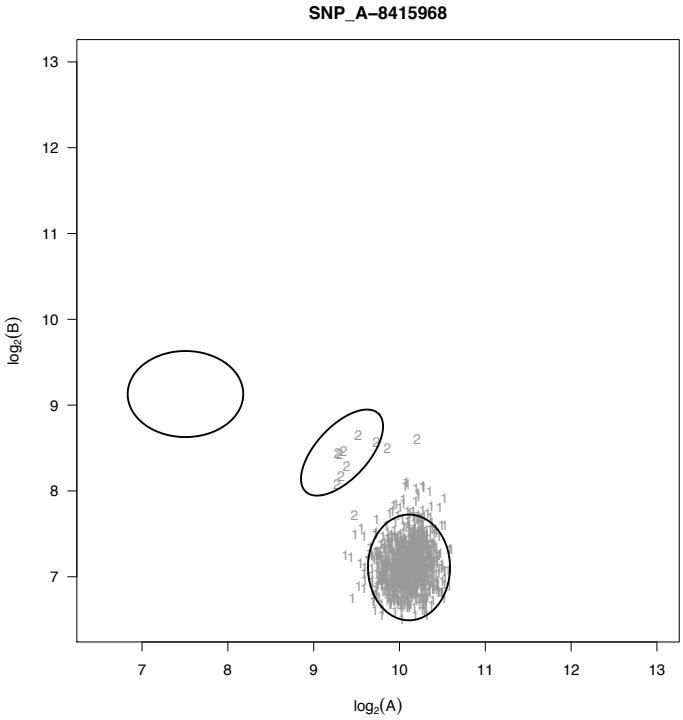


Bipolar GWAS (EA controls) from dbGap

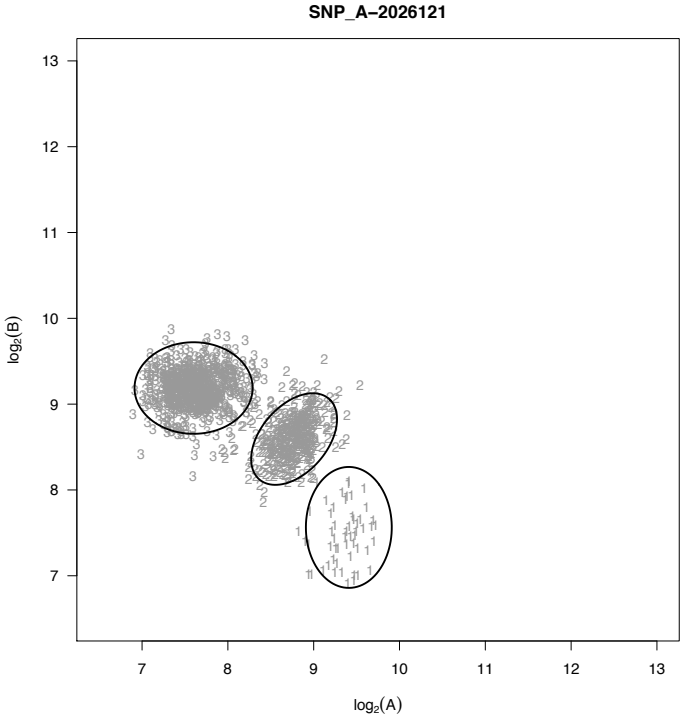
A versus B plots



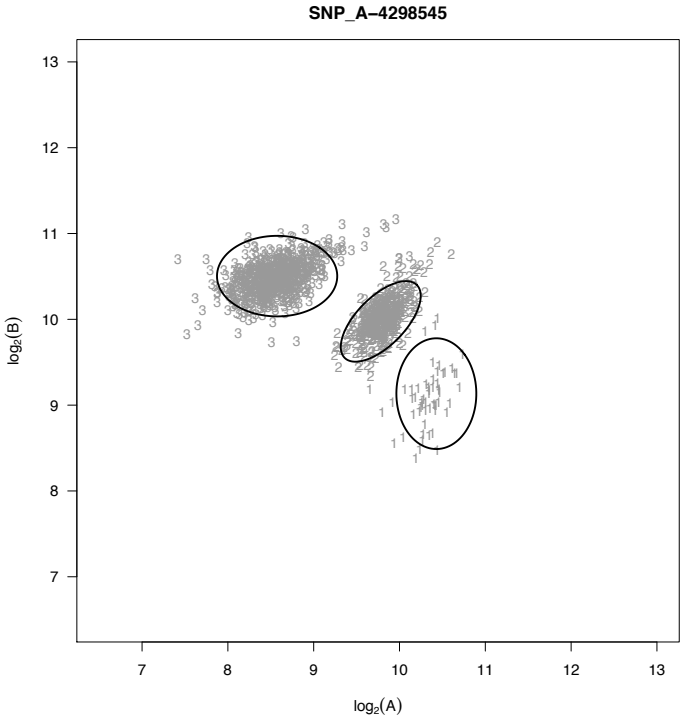
A versus B plots



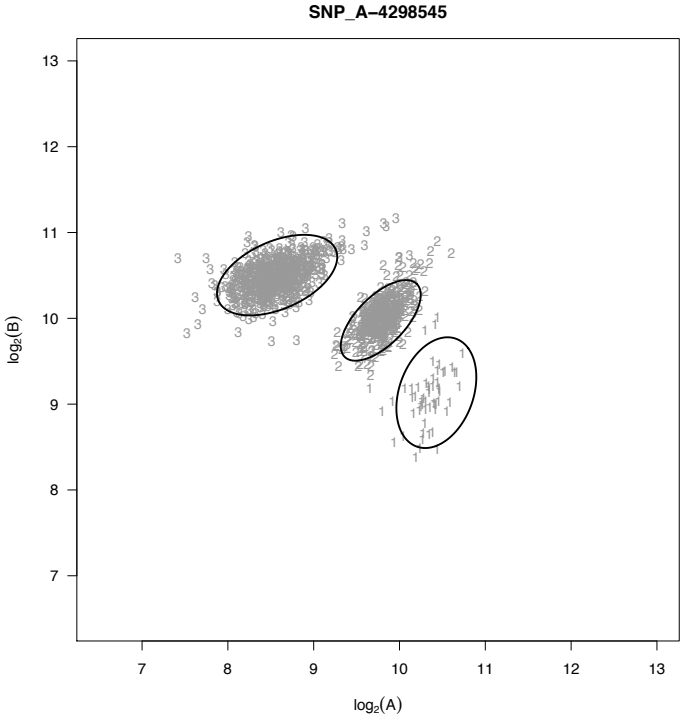
A versus B plots



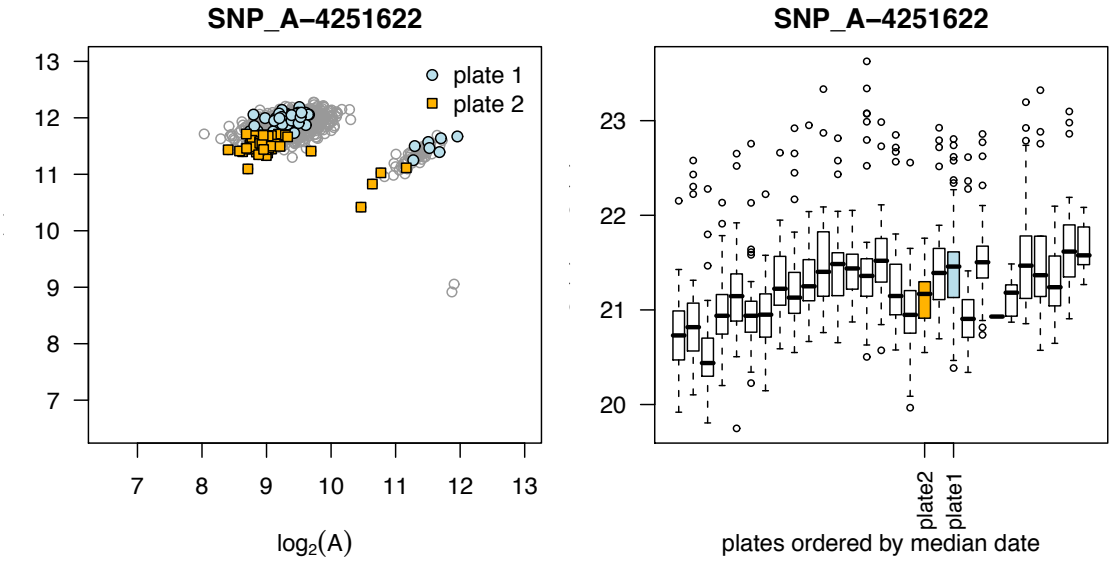
A versus B plots



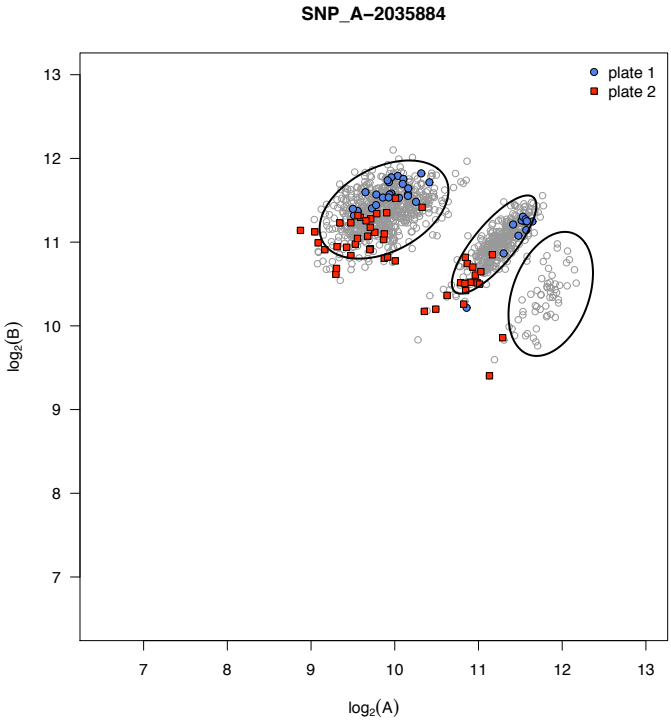
A versus B plots



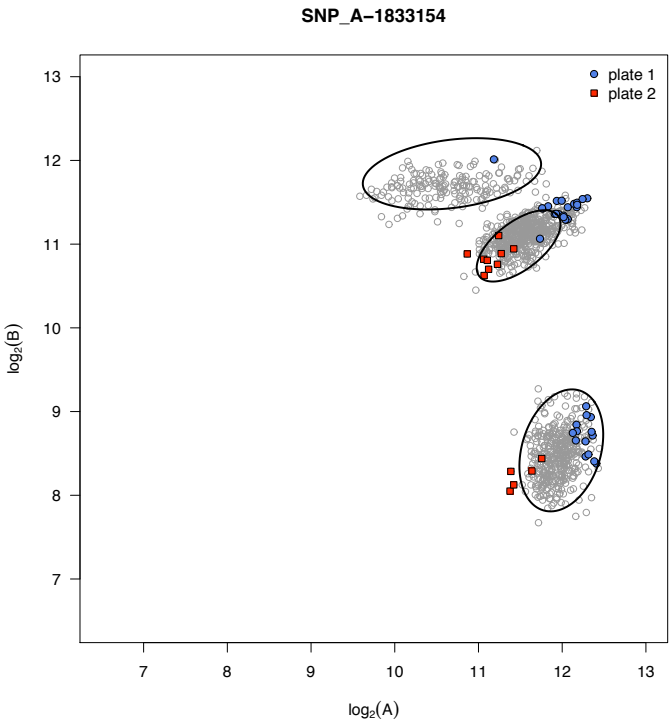
A versus B plots



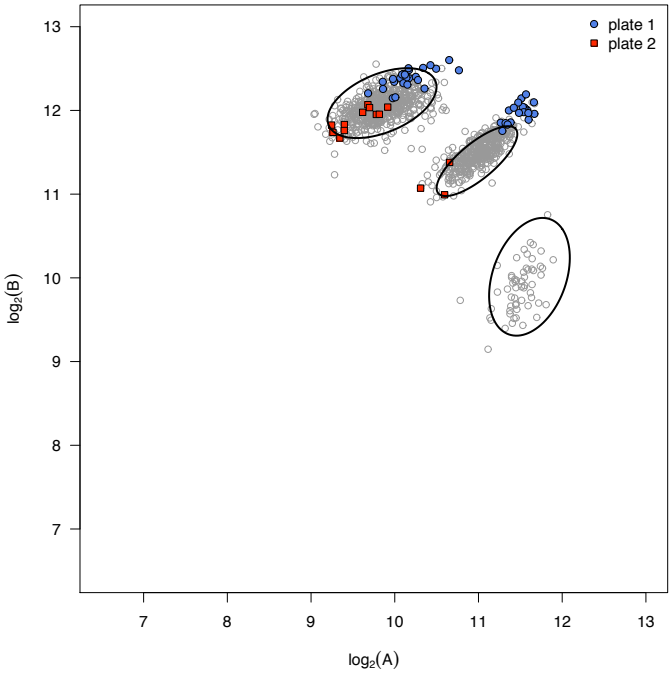
A versus B plots



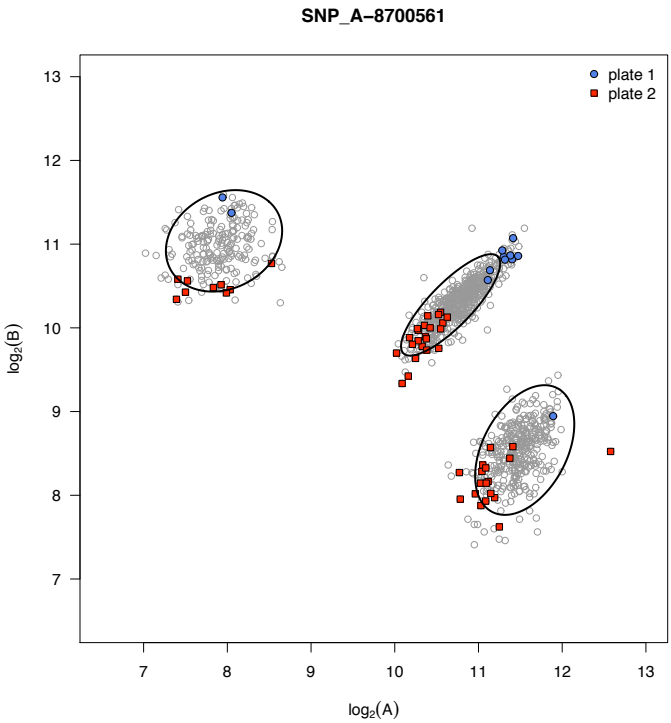
A versus B plots



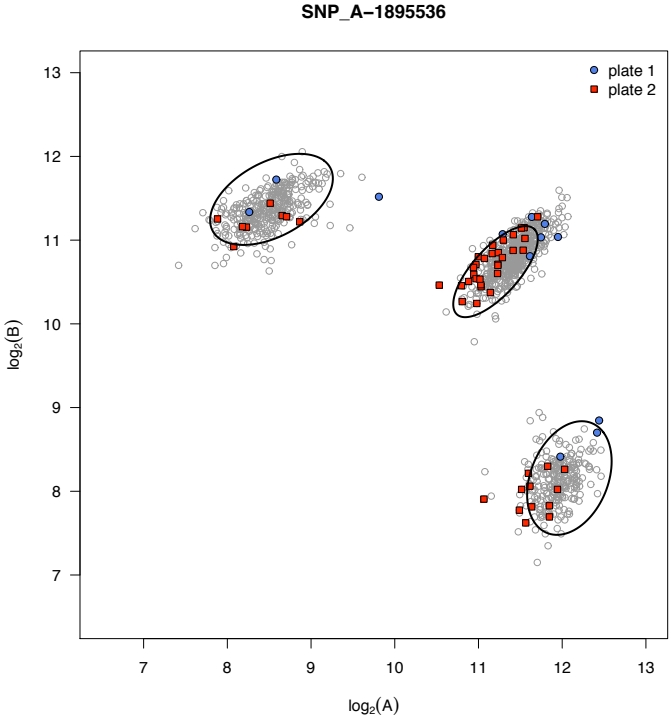
A versus B plots



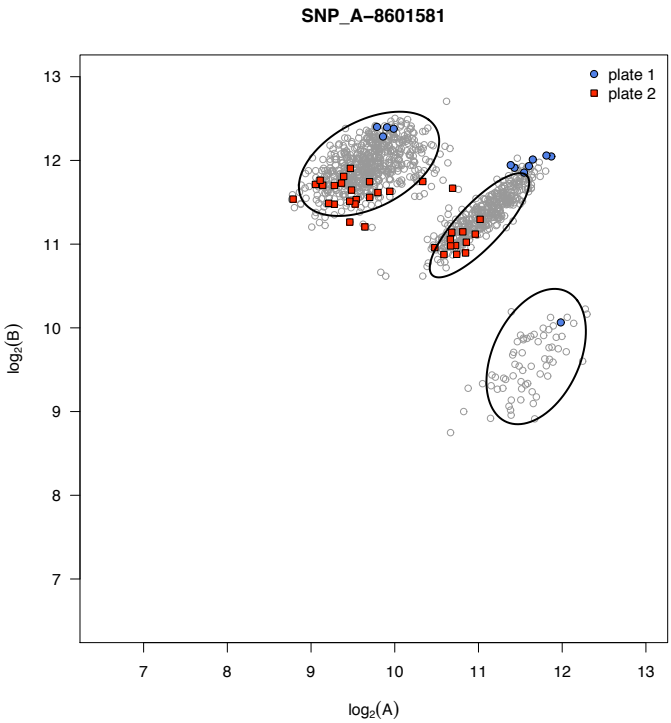
A versus B plots



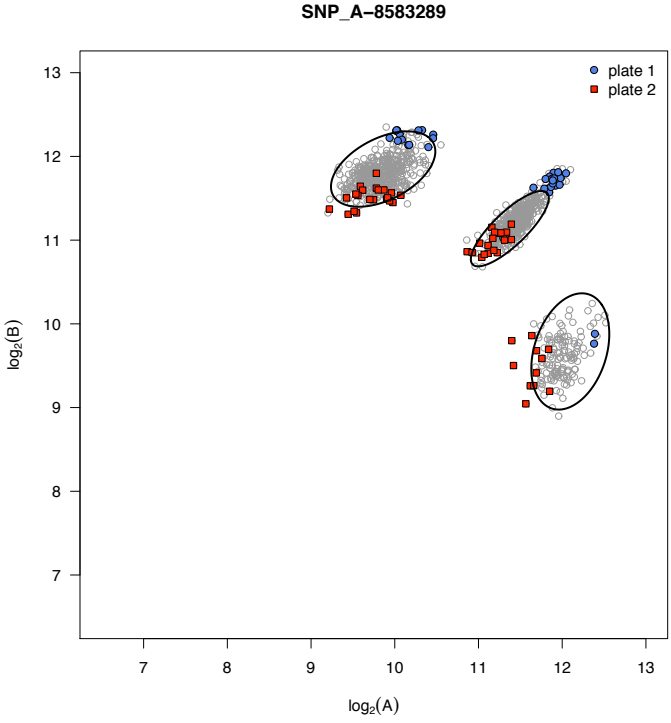
A versus B plots



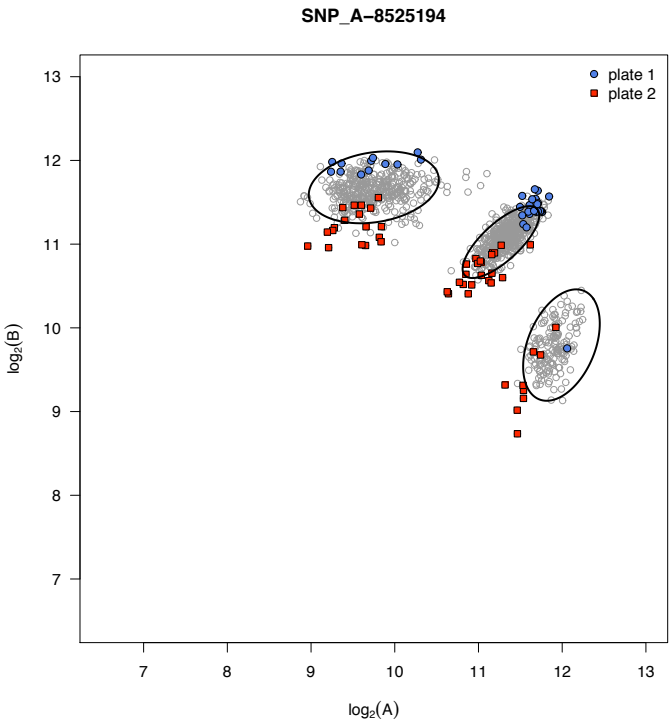
A versus B plots



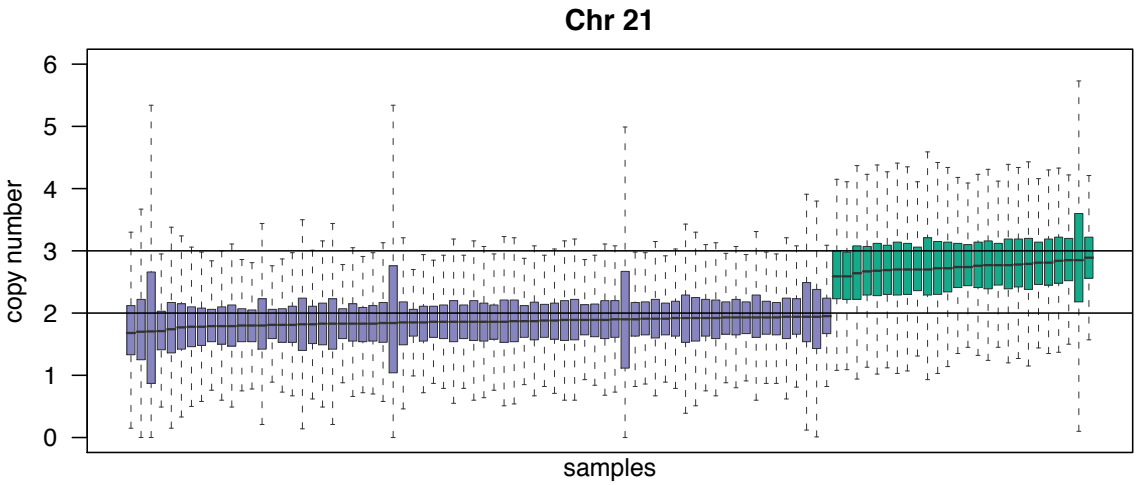
A versus B plots



A versus B plots

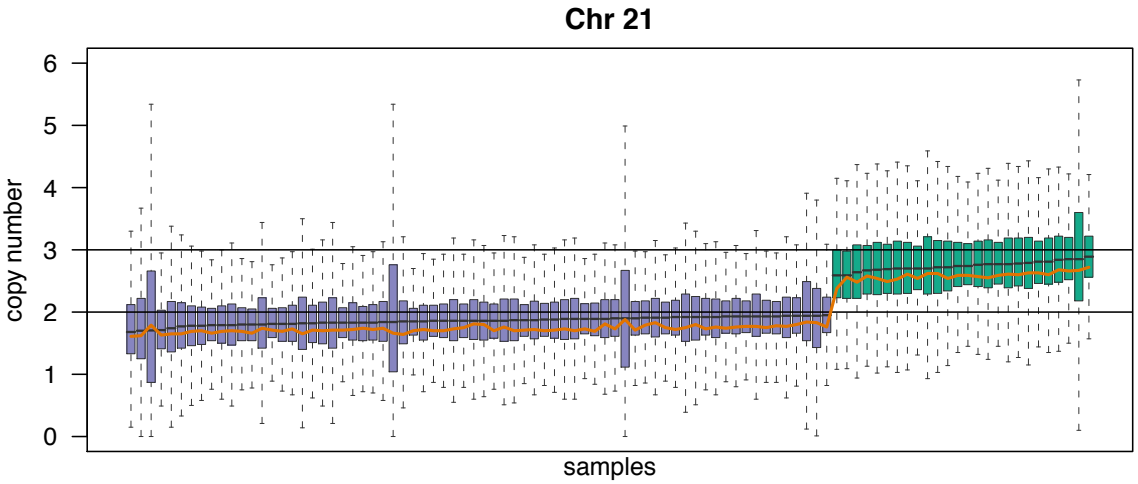


Trisomy 21



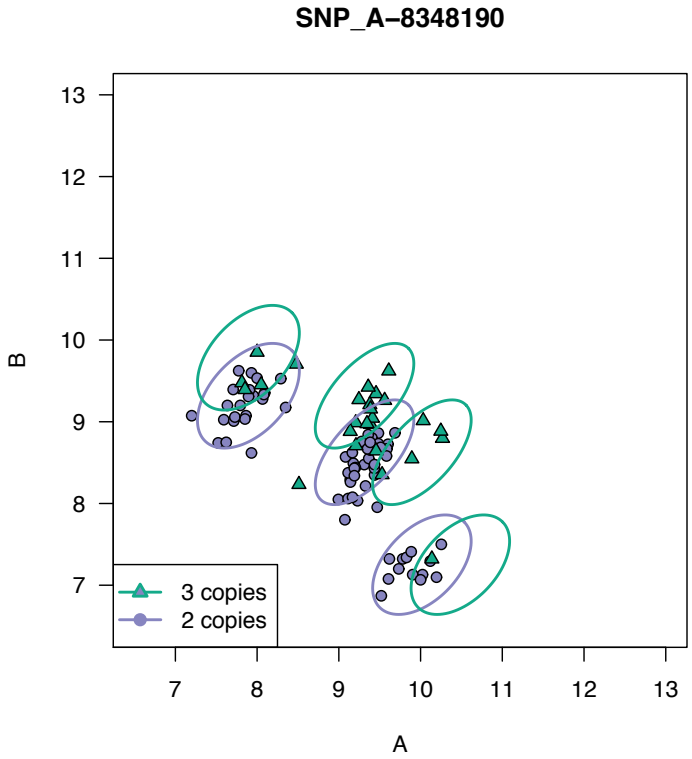
Samples from Aravinda Chakravarti and Betty Doan

Trisomy 21

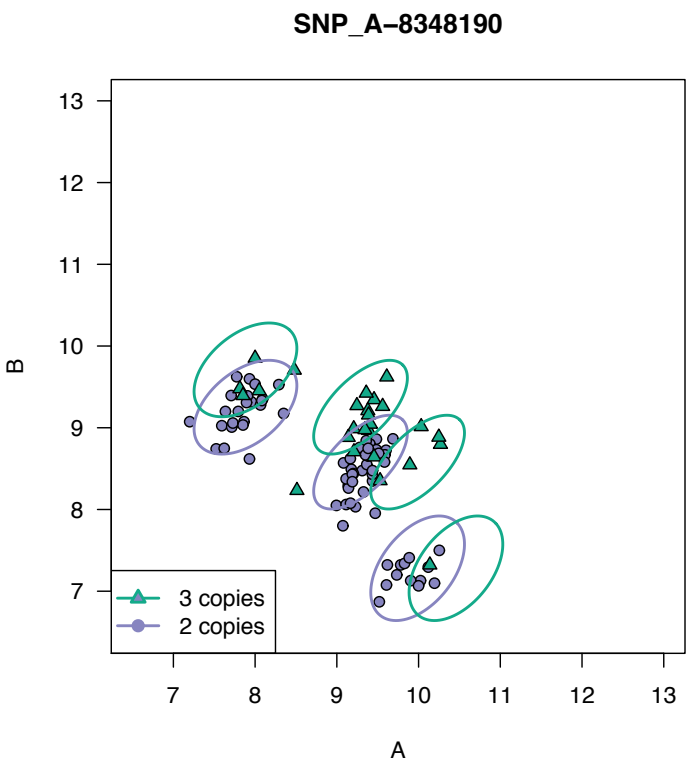


Samples from Aravinda Chakravarti and Betty Doan

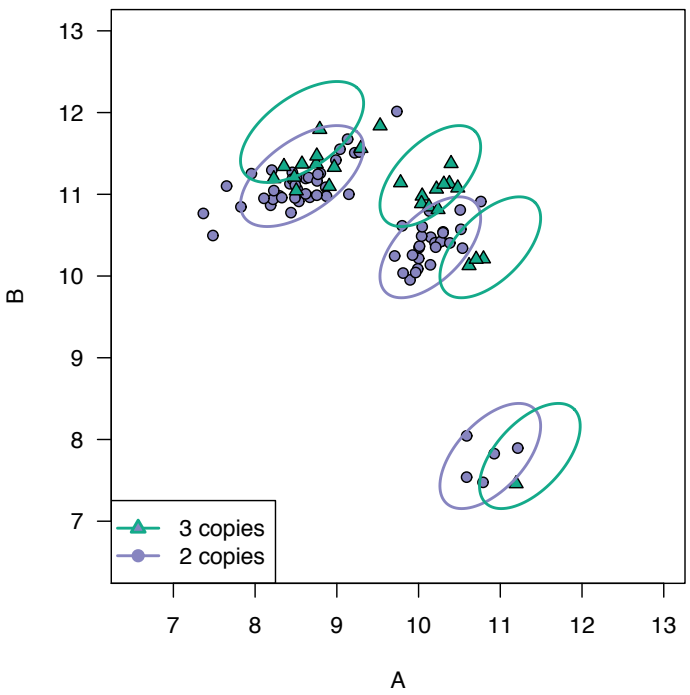
A versus B plots



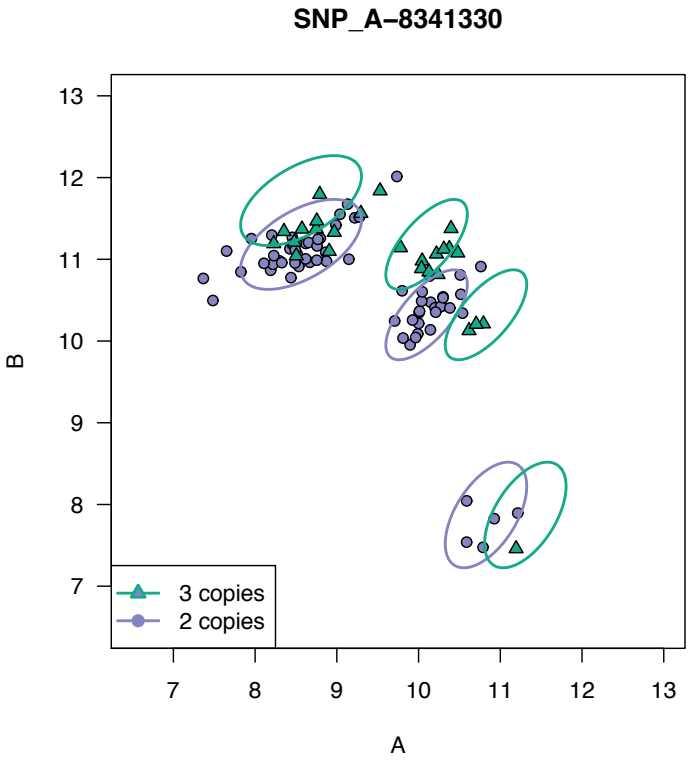
A versus B plots



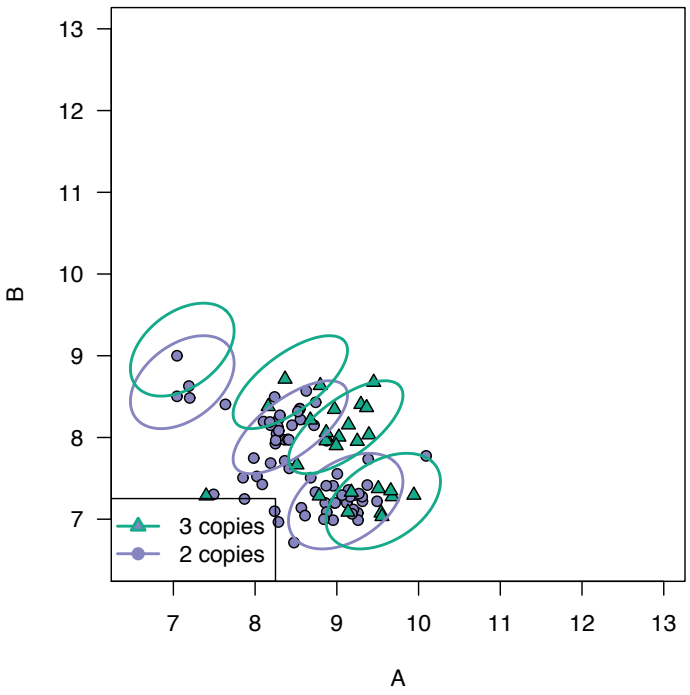
A versus B plots



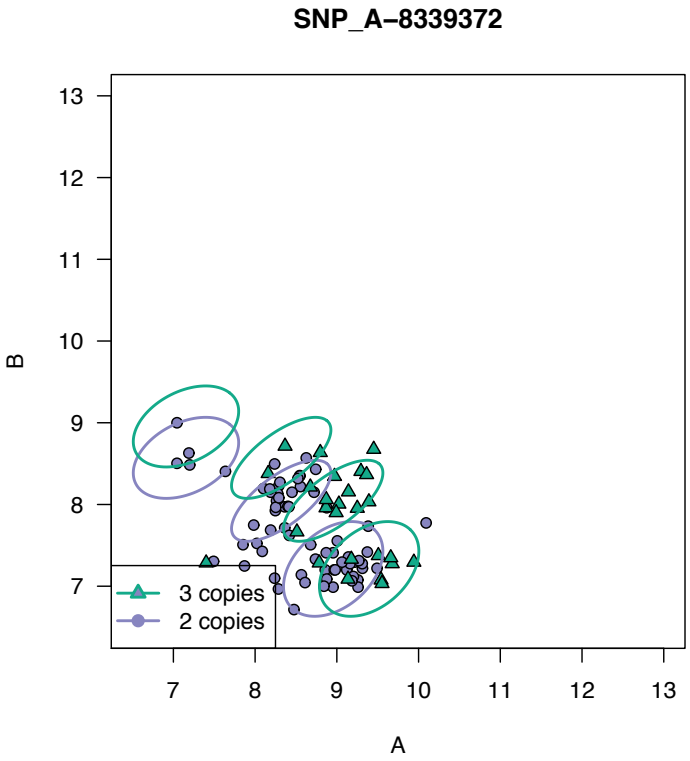
A versus B plots



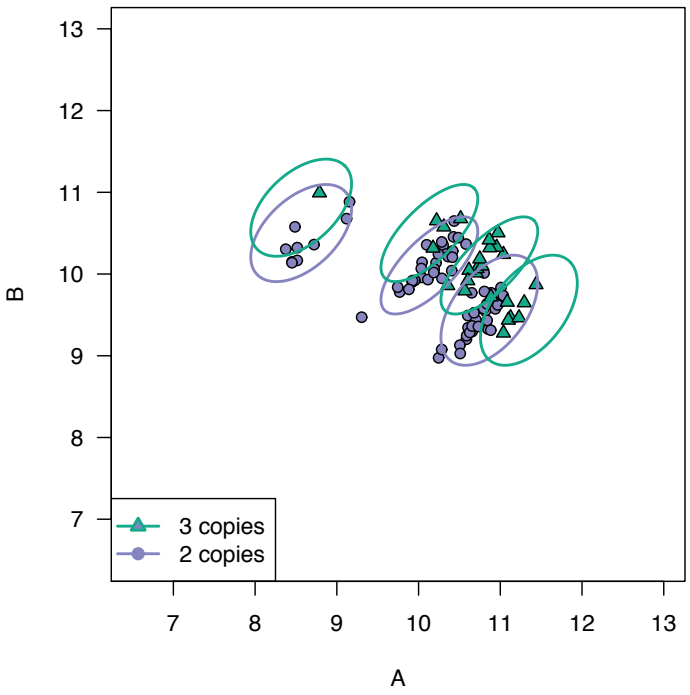
A versus B plots



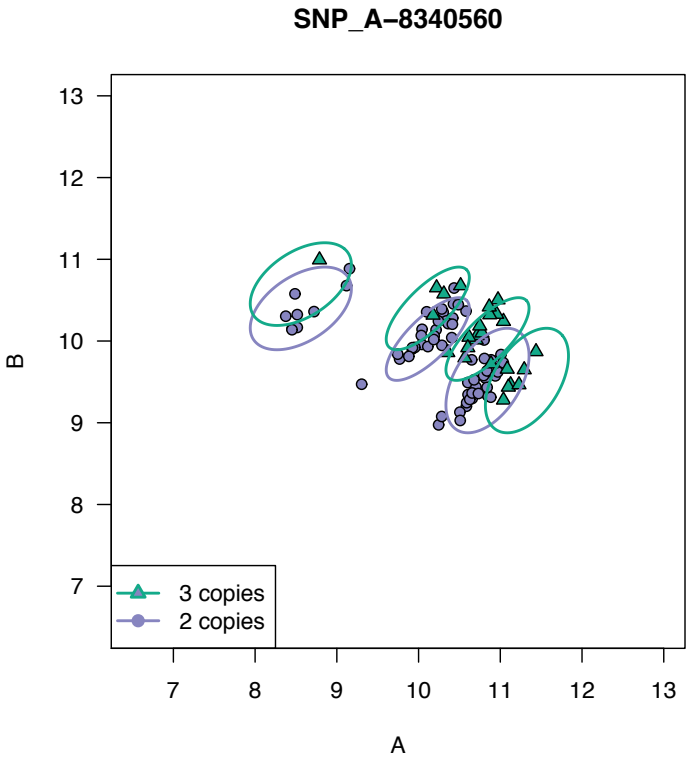
A versus B plots



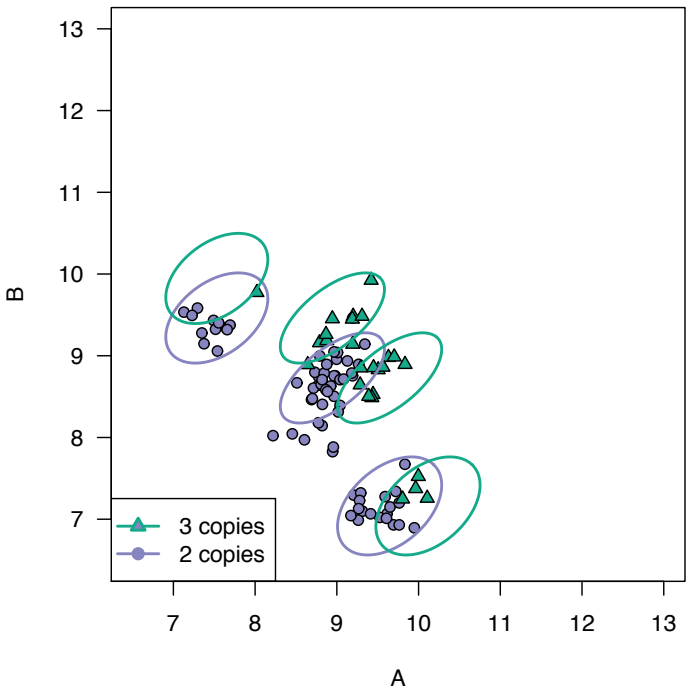
A versus B plots



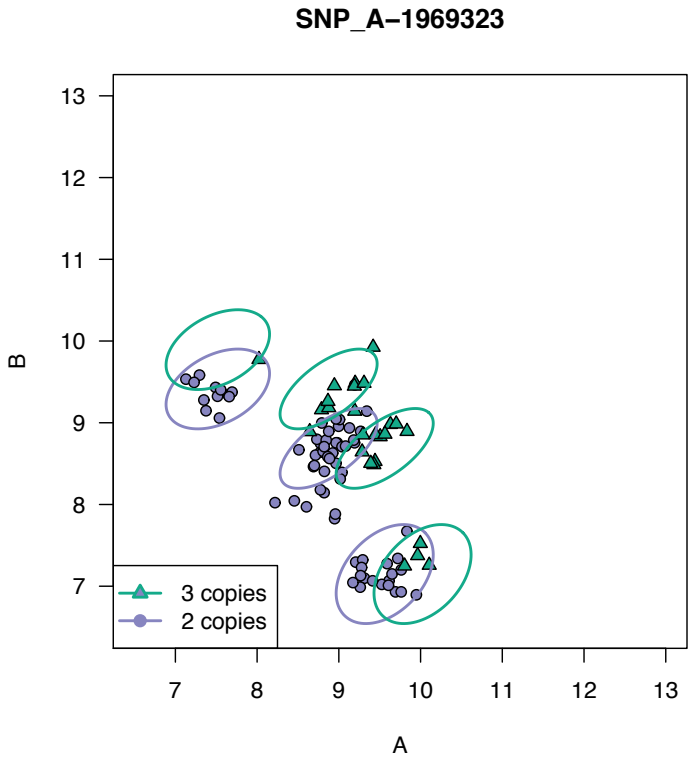
A versus B plots



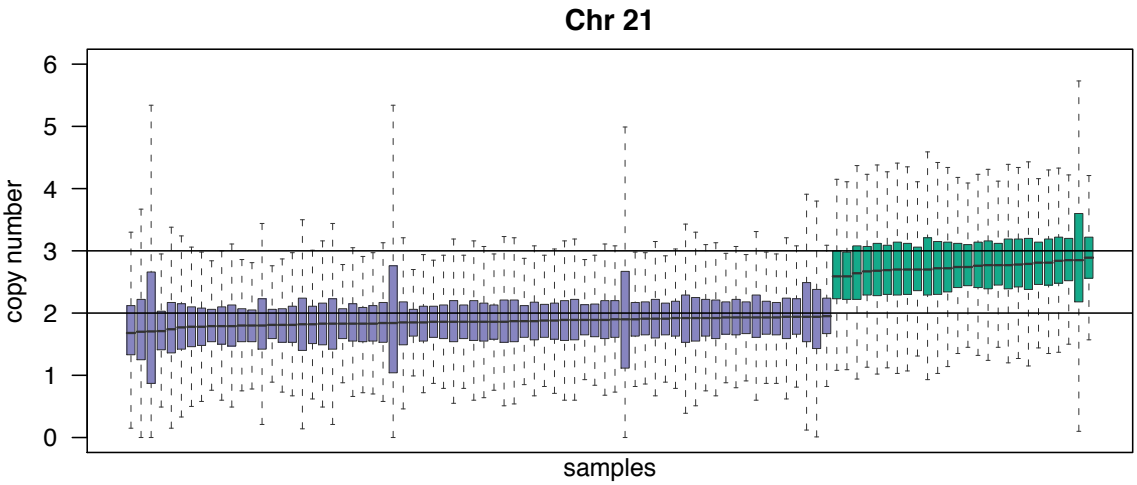
A versus B plots



A versus B plots

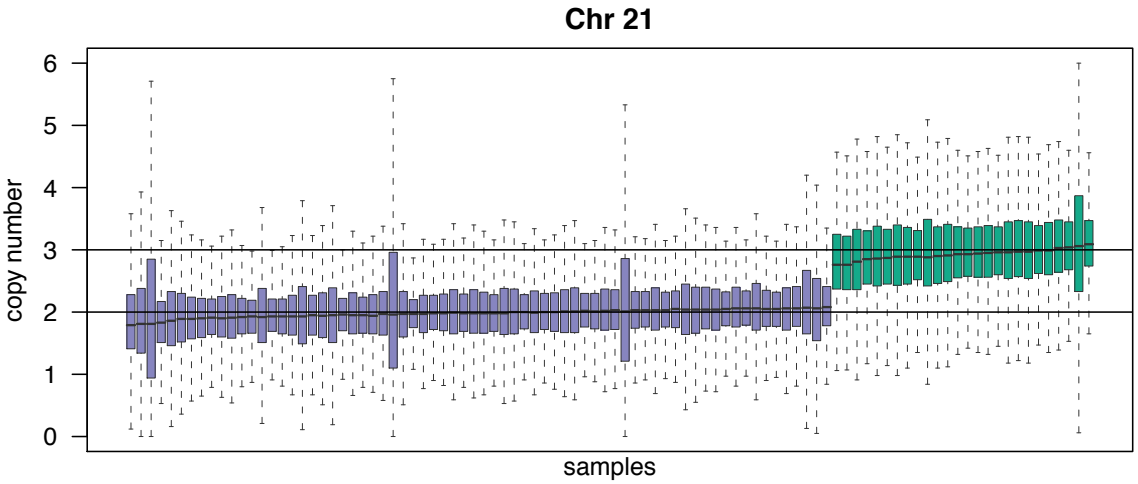


Trisomy 21



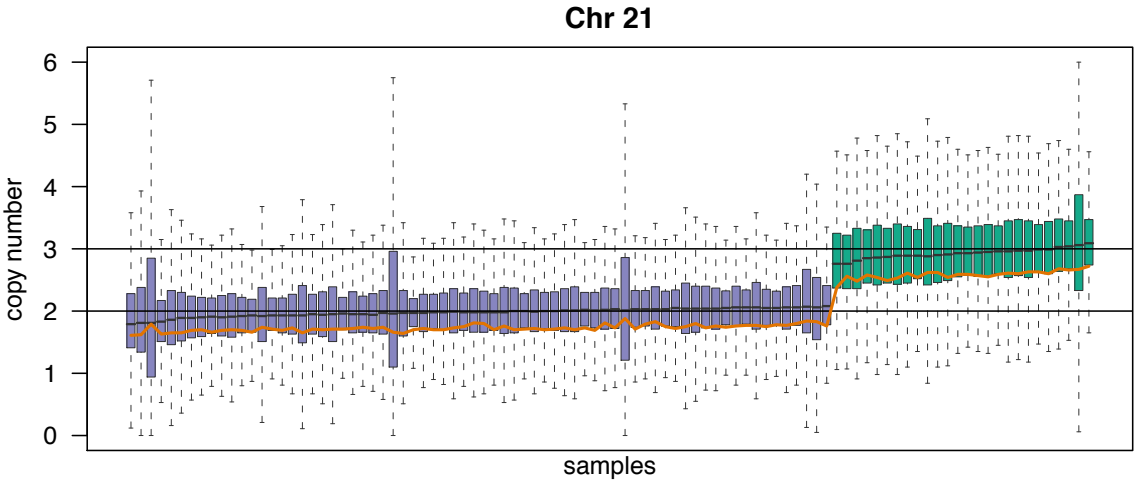
Samples from Aravinda Chakravarti and Betty Doan

Trisomy 21



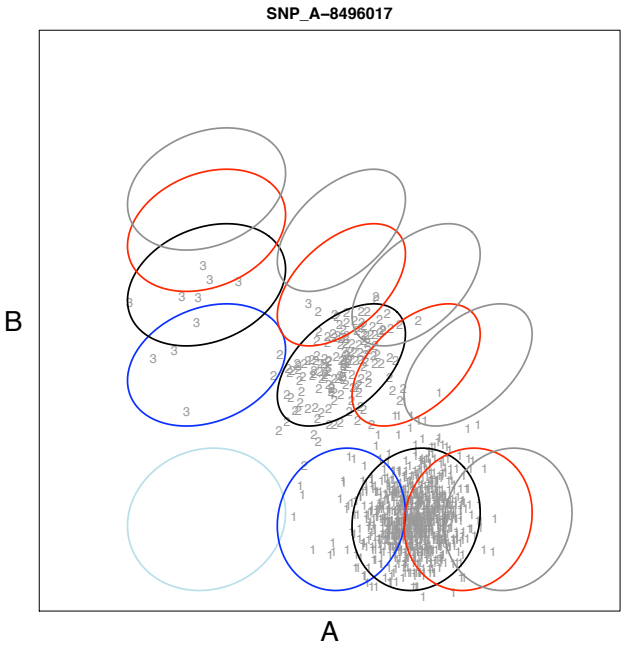
Samples from Aravinda Chakravarti and Betty Doan

Trisomy 21



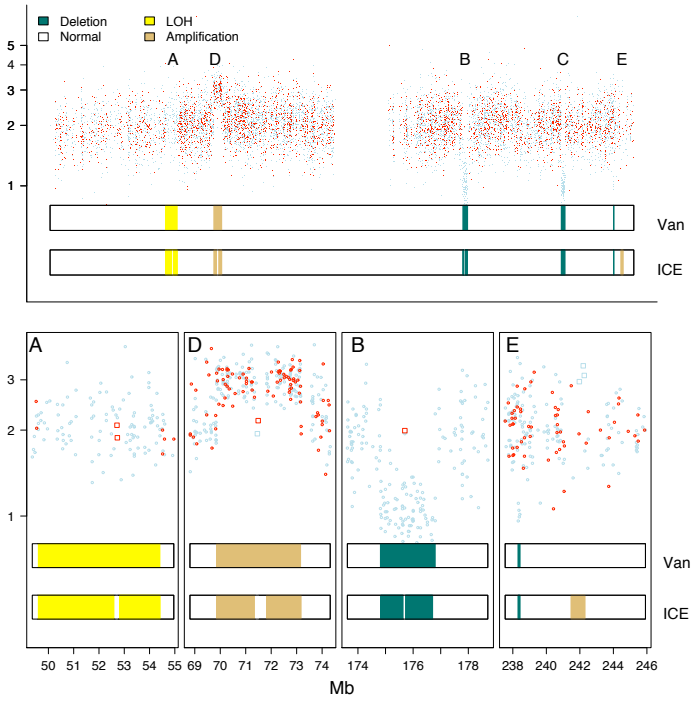
Samples from Aravinda Chakravarti and Betty Doan

Prediction regions for copy number



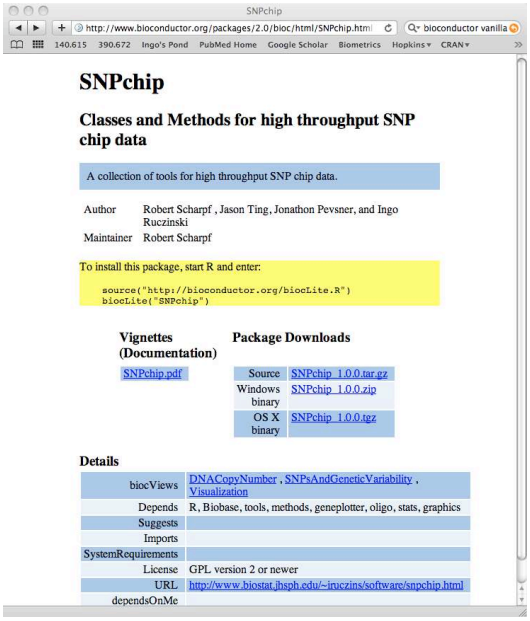
Scharpf et al (2010), in revision.

Vanilla and ICE HMMs for genotype and copy number



Scharpf et al (2008). Ann Appl Stat 2(2): 687-713.

Open source software



SNPchip

Classes and Methods for high throughput SNP chip data

A collection of tools for high throughput SNP chip data.

Author	Robert Scharpf , Jason Ting, Jonathon Pevsner, and Ingo Ruczinski
Maintainer	Robert Scharpf

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("SNPchip")
```

Vignettes (Documentation)

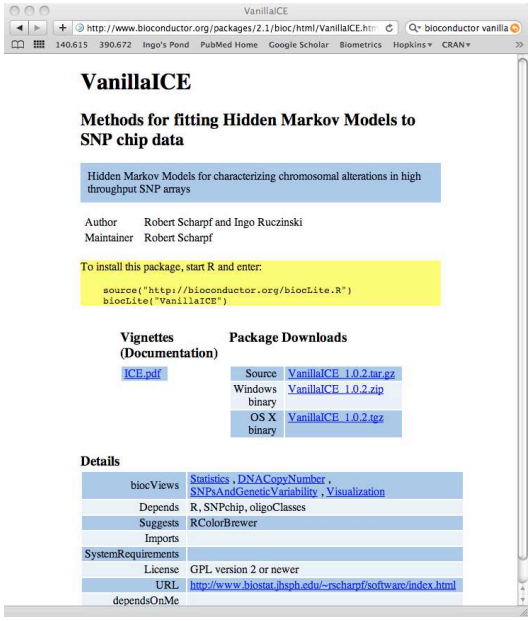
SNPchip.pdf

Package Downloads

Source	SNPchip_1.0.0.tar.gz
Windows binary	SNPchip_1.0.0.zip
OS X binary	SNPchip_1.0.0.tgz

Details

biocViews	DNACopyNumber , SNPsAndGeneticVariability , Visualization
Depends	R, Biobase, tools, methods, geneplotter, oligo, stats, graphics
Suggests	
Imports	
SystemRequirements	
License	GPL version 2 or newer
URL	http://www.biostat.jhsph.edu/~iruczins/software/snpchip.html
dependsOnMe	



VanillaICE

Methods for fitting Hidden Markov Models to SNP chip data

Hidden Markov Models for characterizing chromosomal alterations in high throughput SNP arrays

Author	Robert Scharpf and Ingo Ruczinski
Maintainer	Robert Scharpf

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("VanillaICE")
```

Vignettes (Documentation)

ICE.pdf

Package Downloads

Source	VanillaICE_1.0.2.tar.gz
Windows binary	VanillaICE_1.0.2.zip
OS X binary	VanillaICE_1.0.2.tgz

Details

biocViews	Statistics , DNACopyNumber , SNPsAndGeneticVariability , Visualization
Depends	R, SNPchip, oligoClasses
Suggests	RColorBrewer
Imports	
SystemRequirements	
License	GPL version 2 or newer
URL	http://www.biostat.jhsph.edu/~rscharpf/software/index.html
dependsOnMe	

Scharpf et al (2007). Bioinformatics. 23(5): 627-8.

Scharpf and Ruczinski (2010). Methods Mol Biol 593: 67-79.

Acknowledgments

Collaborators: Rob Scharpf

Benilton Carvalho, Rafael Irizarry, Giovanni Parmigiani, Holger Schwender.

Terri Beaty & group, Kathleen Barnes & group.

Computing support: Marvin Newhouse, Jiong Yang.

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

http://biostat.jhsph.edu/∼iruczins/