

Assessing Genomic Variability using High-throughput SNP Arrays

Ingo Ruczinski

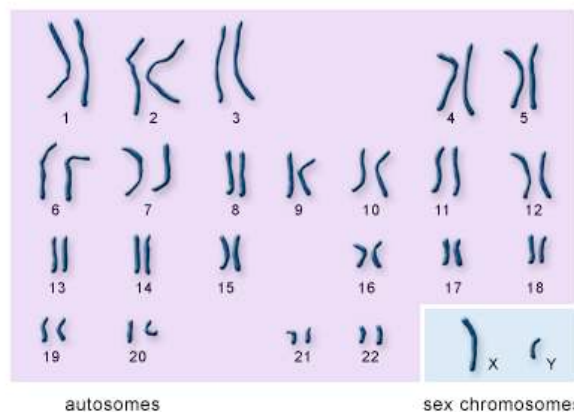
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

February 17, 2010

Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays

Karyotypes

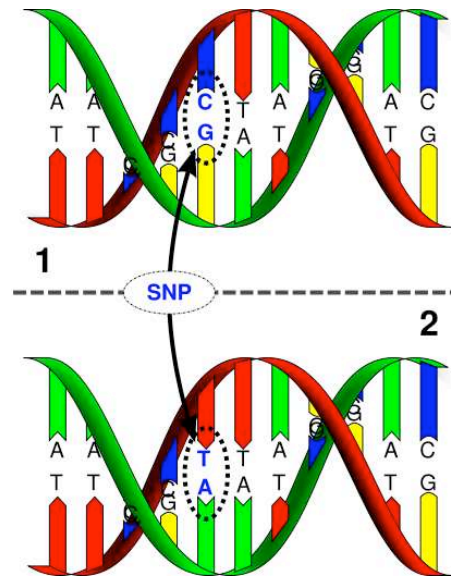


<http://ghr.nlm.nih.gov/>

Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays

Single Nucleotide Polymorphisms



urgi.versailles.inra.fr

Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays

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.

Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays

Results

genome.gov | OPG: A Catalog of Published Genome-Wide Association Studies

genome.gov
National Human Genome Research Institute
National Institutes of Health

Home | About NHGRI | Newsroom | Staff

Research Grants Health Policy & Ethics Educational Resources Careers & Training

Home > About NHGRI > About the Office of the Director > Office of Population Genomics > **OPG: A Catalog of Published Genome-Wide Association Studies**

Office of Population Genomics

Overview | **A Catalog of Genome-Wide Association Studies** | Research Programs | Recent Publications | Meetings & Workshops | Notices & Funding Opportunities | Contact

A Catalog of Published Genome-Wide Association Studies

[Potential etiologic and functional implications of genome-wide association loci for human diseases and traits](#)

Click here to read our recent *Proceedings of the Academy of Sciences (PNAS)* article on catalog methods and analysis.

Go to the Catalog

The genome-wide association study (GWAS) publications listed here include only those attempting to assay at least 100,000 single nucleotide polymorphisms (SNPs) and are organized from most to least recent date of publication, indexing from online publication if available. Studies focusing only on candidate genes are excluded. We update the catalog weekly from PubMed literature searches, daily NIH-distributed compilations of news and media reports, and occasional comparisons with an existing database.

SNP-trait associations listed here are limited to those with p -values $< 1.0 \times 10^{-5}$. Note that we **are now including all identified** SNP-trait associations with p -values rounded to the nearest single digit; odds ratios and allele frequencies are rounded to two decimals. Standard errors are converted to standard deviations. Allele frequencies, p -values, and odds ratios derived from the largest sample size, typically a combined analysis (initial plus replication studies), are recorded. Odds ratios < 1 in the original paper are converted to $OR > 1$ for the alternate allele. Where results from multiple genetic models (OR's or beta-coefficients) are reported, we list: 1) genotypic model, per-allele estimate; 2) genotypic model, heterozygote estimate, 3) allelic model, allelic estimate.

Gene regions corresponding to SNPs were identified from the [UCSC Genome Browser](#). Gene names are those reported by the authors in the original paper. Linkage disequilibrium is recorded unless there was evidence of independent association.

Occasionally the term "pending" is used to denote one or more studies that we identified as an eligible GWAS, but for which SNP information has not yet been published.

How to cite the GWAS Catalog:
Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS* 2009;106:9514-9519.

<http://www.genome.gov/GWastudies/>

Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays

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.

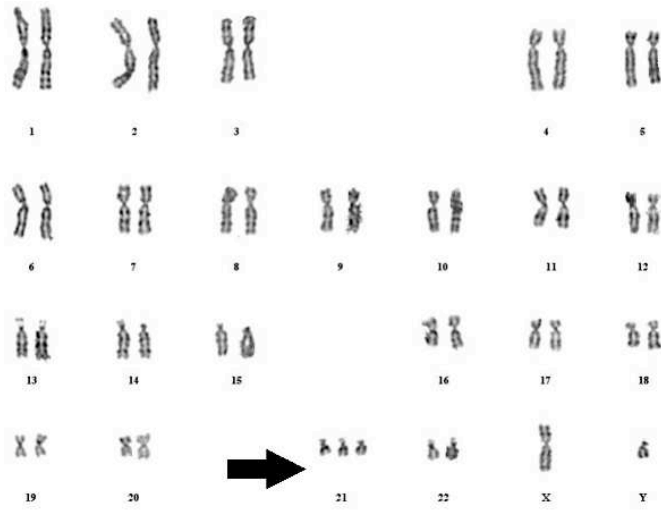
Key words: Efficient simulation, ranking procedures, SNP identification.

Louis and Ruczinski (2010). *Biometrical Journal* 52(1), 1-16.

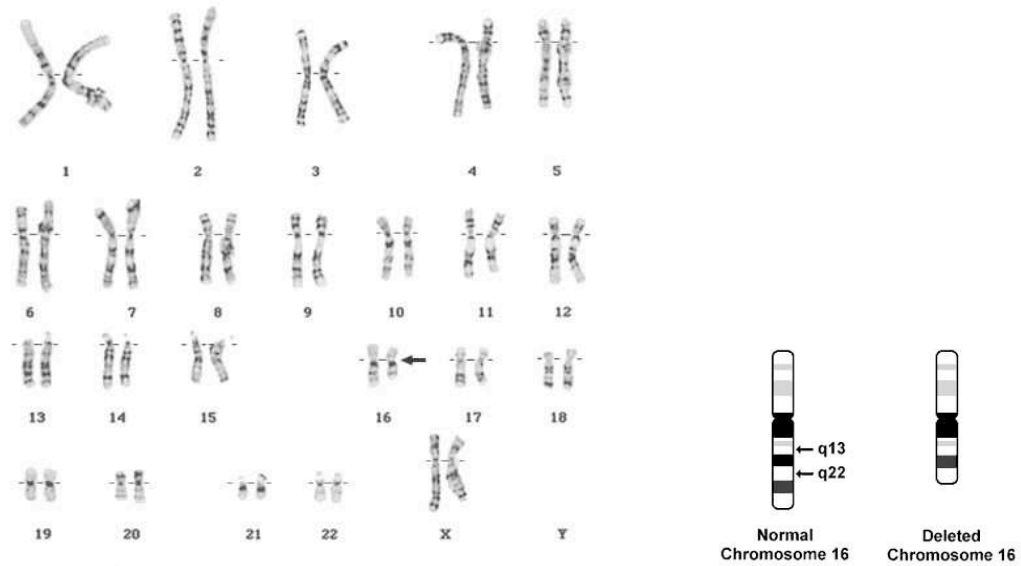
Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays

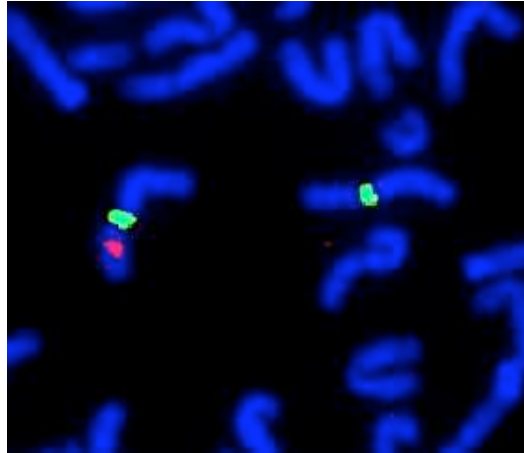
Trisomy



Karyotypes



FISH



Courtesy of the Pevsner Laboratory

Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays

Clinical practice

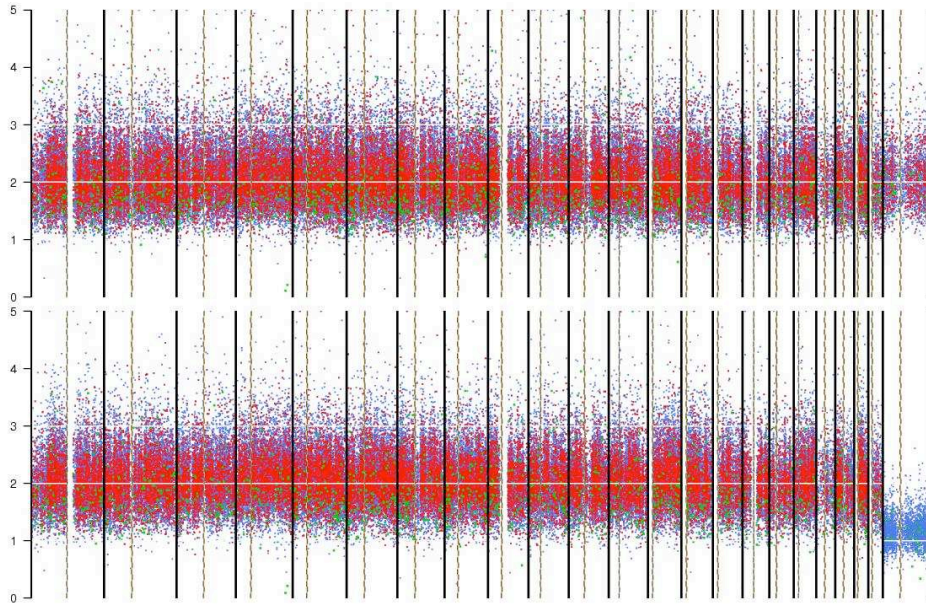
A screenshot of a web browser displaying a New York Times article. The browser window shows the URL 'nytimes.com' and the page title 'The DNA Age - Parents Whose Children Share Genetic Mutations Seek Each Other for Support - New York Times'. The article is dated December 28, 2007, and is titled 'After DNA Diagnosis: 'Hello, 16p11.2. Are You Just Like Me?'' by Amy Harmon. The article text describes the experience of parents whose children share a genetic mutation on chromosome 16p11.2, highlighting the similarities in physical features and learning difficulties among the children. The article mentions that several adults wiped tears from their eyes, and one parent, Jessica Houk, accompanied her daughter to a Kentucky amusement park to greet another child with the same mutation.

New York Times, December 28, 2007

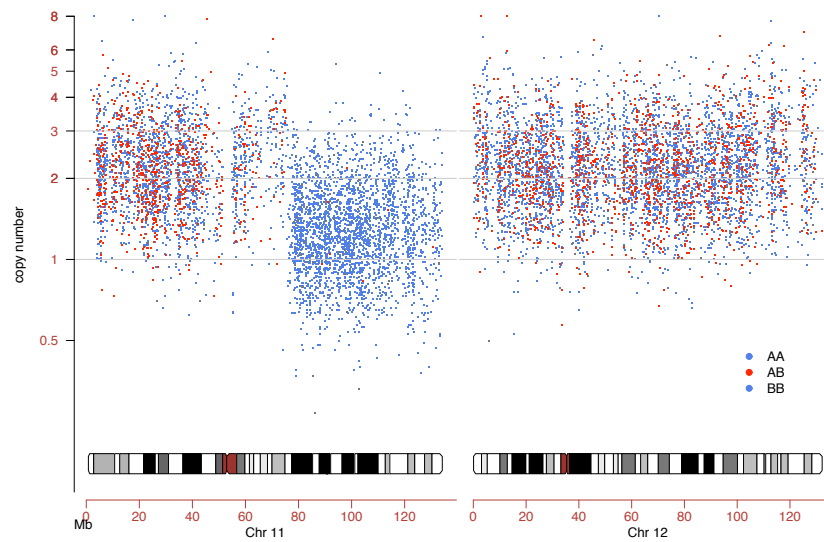
Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays

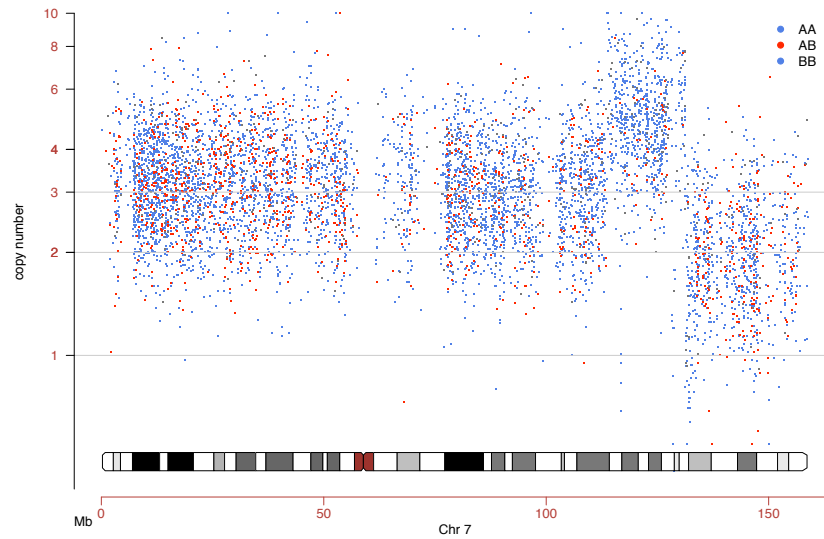
SNP chip data



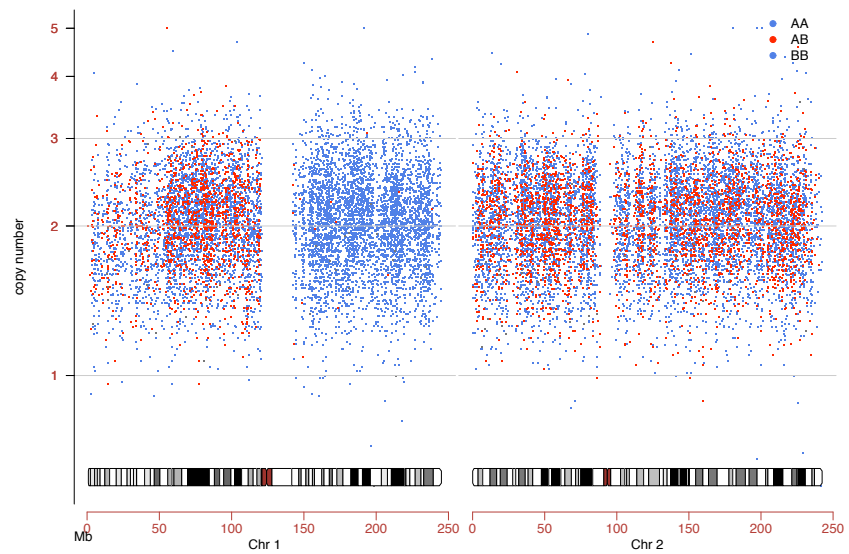
Deletion

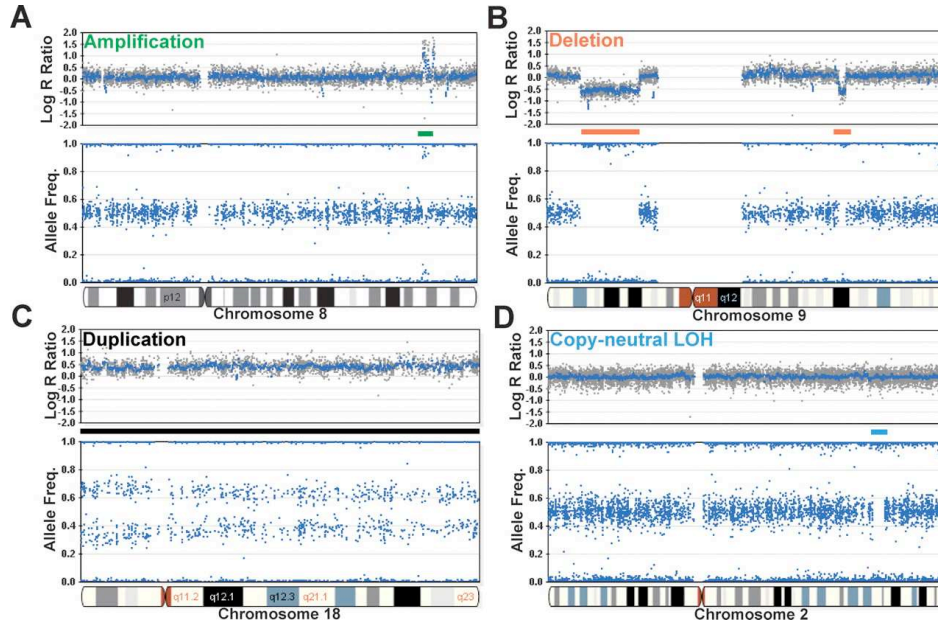


Amplification



Uniparental Isodisomy





Pfeiffer et al (2006), *Genome Res.* 2006. 16: 1136-1148.

Structural variation

The screenshot shows the 'Copy Number Variation (CNV) Project' page on the Sanger Institute website. The page includes a navigation menu on the left, a main content area with text about CNV, and several plots. The text describes the project's goals, such as identifying CNVs associated with diseases and developing high-resolution discovery methods. The plots show genome-wide CNV data for two donors, with one plot highlighting a deletion on chromosome 8.

<http://www.sanger.ac.uk/humgen/cnv/>

Structural variation

Database of Genomic Variants

A curated catalogue of structural variation in the human genome

Hosted by: The Centre for Applied Genomics

Copy Number Variation Project Database of Genomic Variants

Please select genome assembly: Build 36 (Mar. 2006)

View Data by Chromosome
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y All

Keyword Search
Exact Match? Yes No
Examples: clone name, accession number, cytoband, gene

BLAT Search
Enter sequence in FASTA format here:

Summary Statistics
Total entries: 49988 (hg18)
CNVs: 29133
Inversions: 914
InDels (100bp-1Kb): 19941
Total CNV loci: 8410
Articles cited: 35
Last updated: Aug 05, 2009
Join our mailing list

CNV genotype data from the Genome Structural Variation Consortium now available - Aug 21, 2009

Read about the updates in our [newsletter](#).

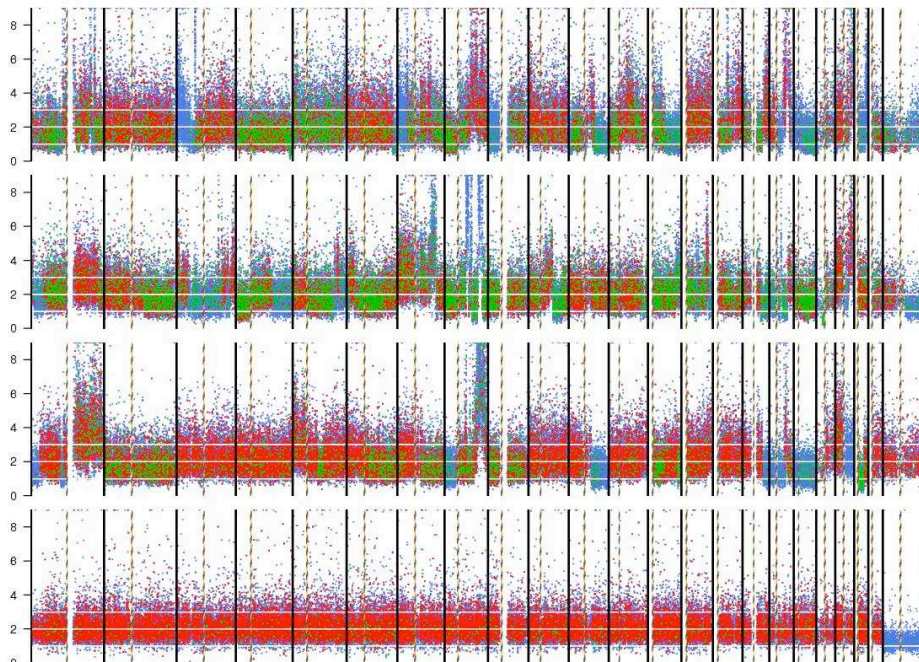
In citing the Database of Genomic Variants please refer to: Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: Detection of large-scale variation in the human genome. *Nat Genet.* 2004 Sep;36(9):949-51.

<http://projects.tcag.ca/variation/>

Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays

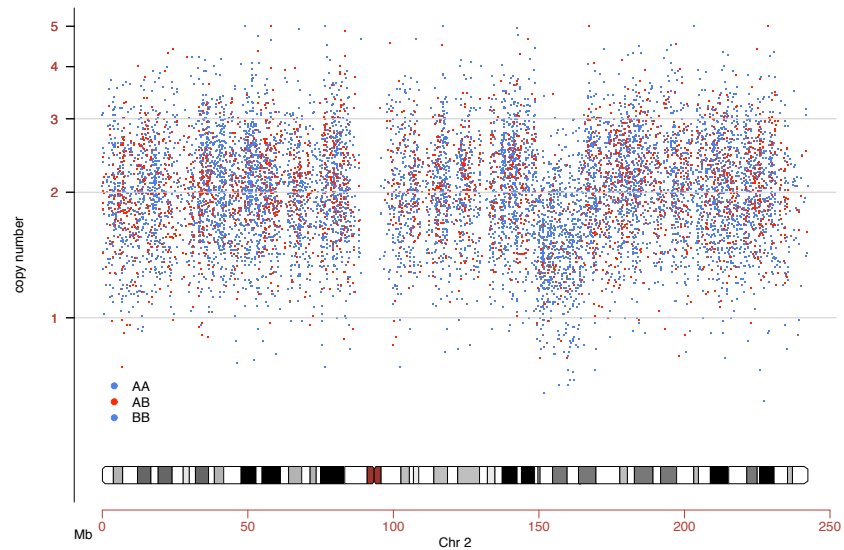
Cancer samples



Ingo Ruczinski

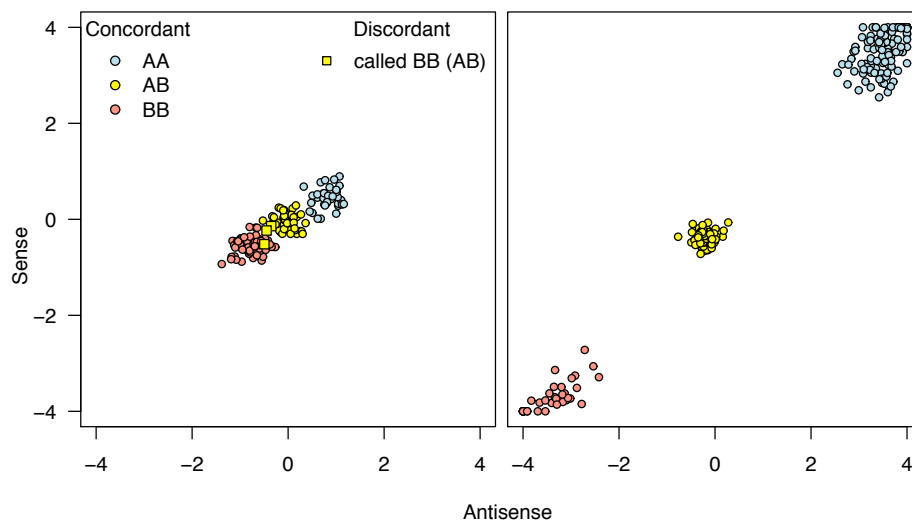
Assessing Genomic Variability with SNP Arrays

Mosaicism

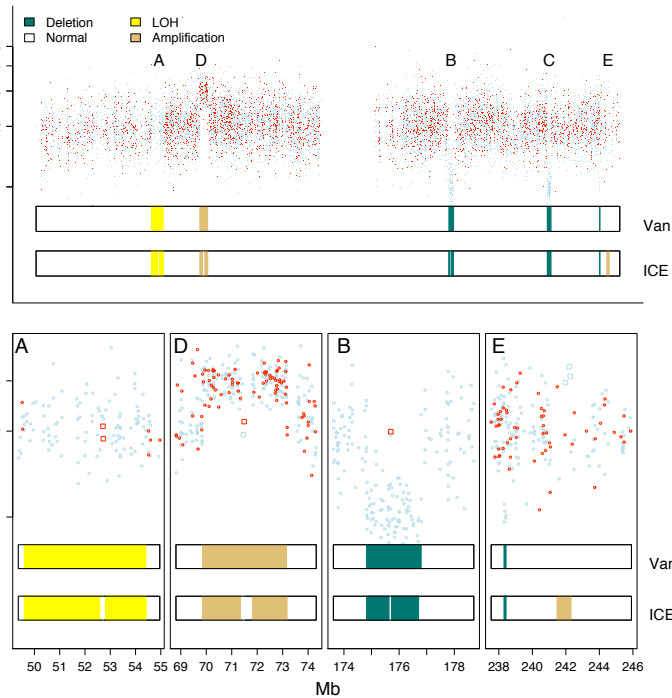


More information

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



Vanilla and ICE HMMs for genotype and copy number



Scharf 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 Scharf, Jason Ting, Jonathon Pevsner, and Ingo Ruczinski
Maintainer Robert Scharf

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 Scharf and Ingo Ruczinski
Maintainer Robert Scharf

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](#)

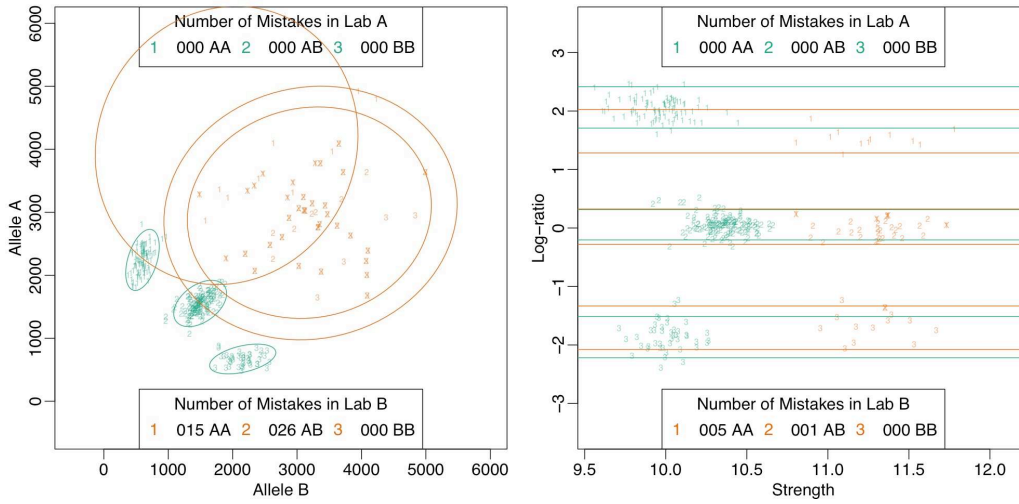
Details

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	

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

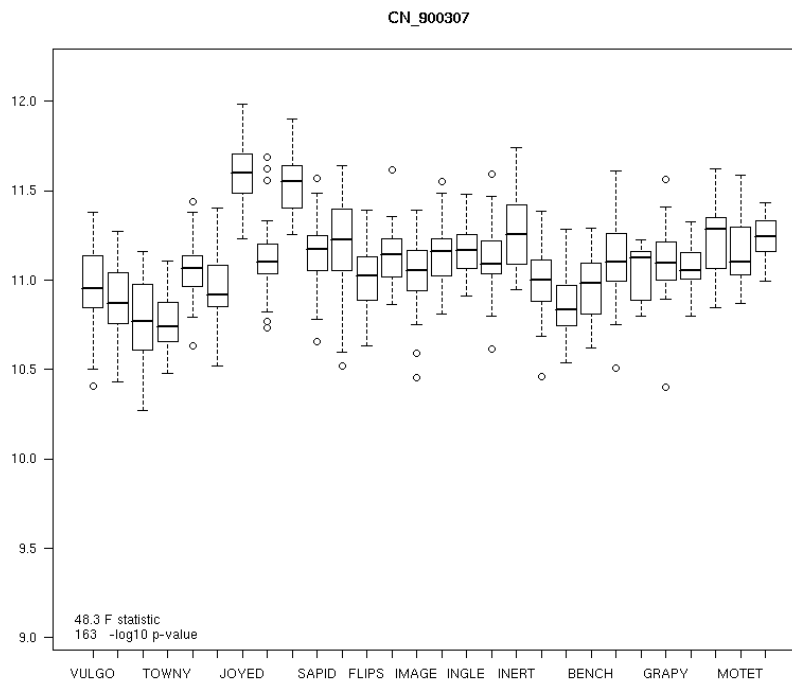
Scharf and Ruczinski (2010). *Methods Mol Biol* 593: 67-79.

Genotypes and copy numbers



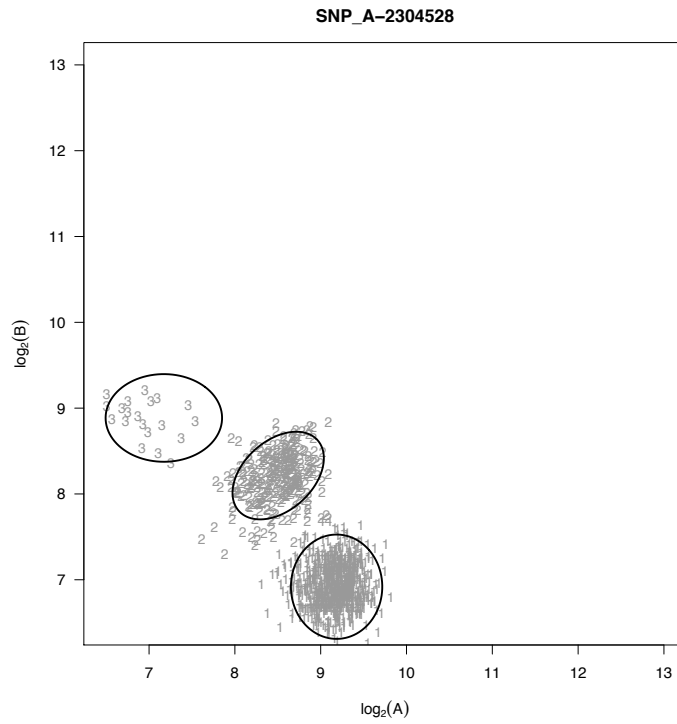
From Benilton Carvalho and Rafa Irizarry

Plate effects

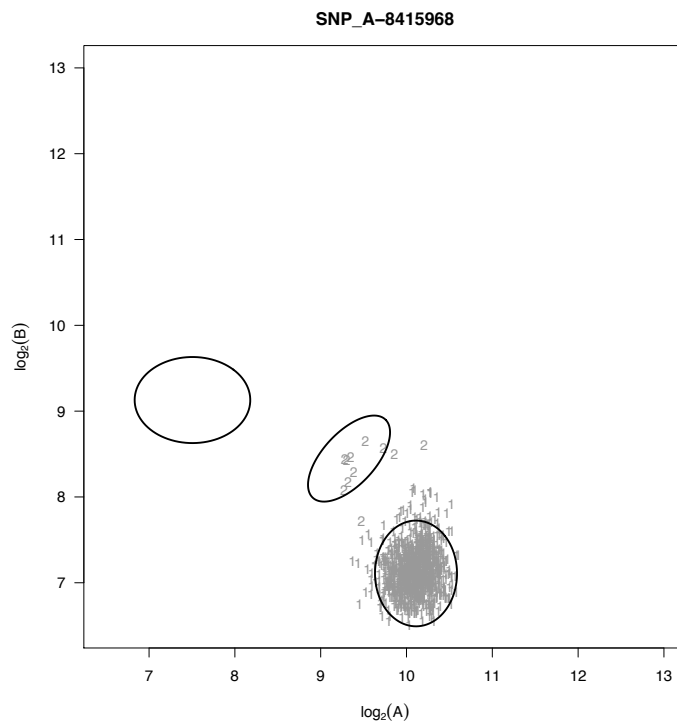


Bipolar GWAS (EA controls) from dbGap

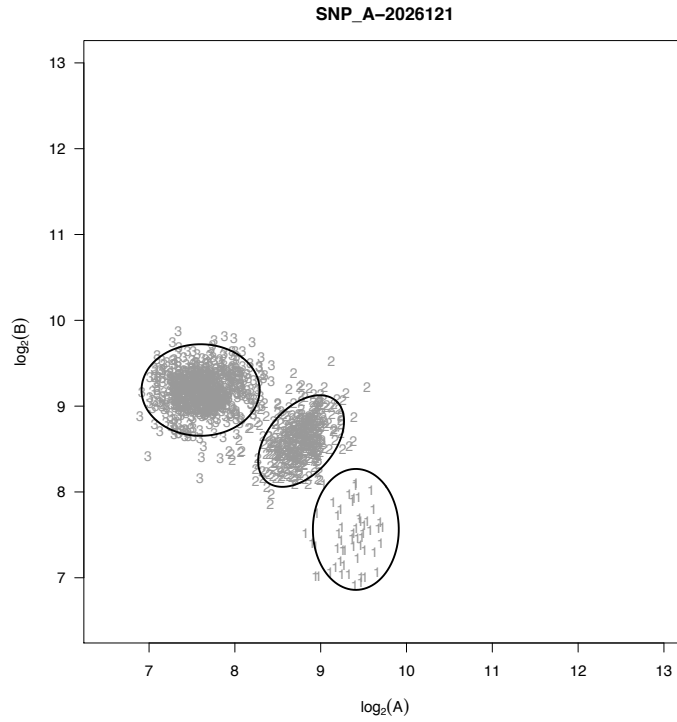
A versus B plots



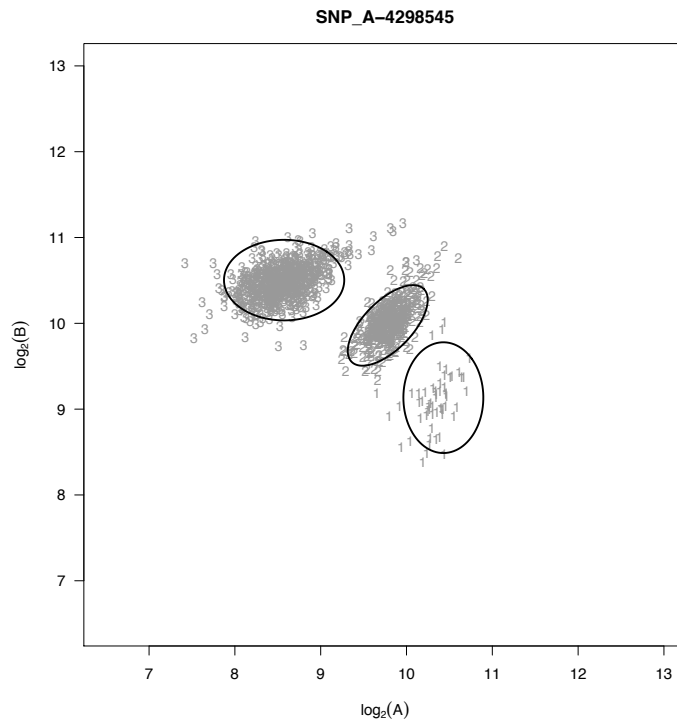
A versus B plots



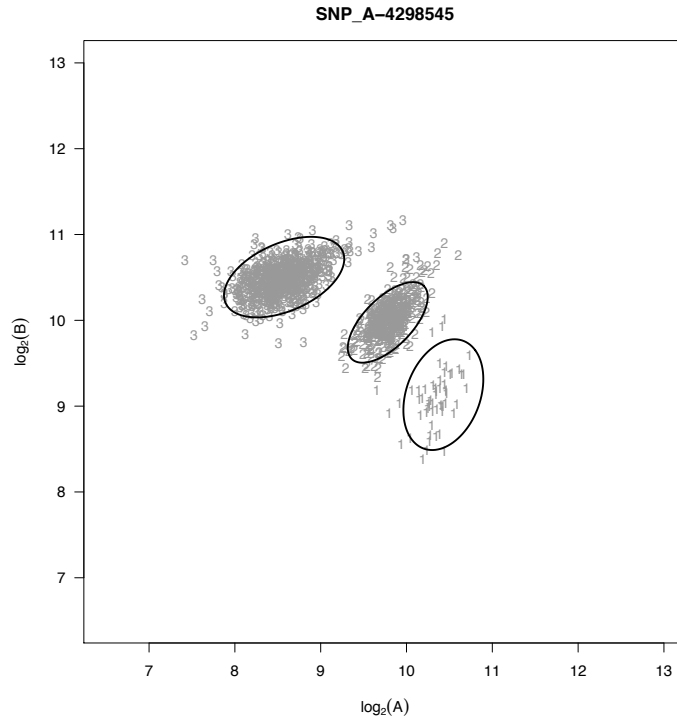
A versus B plots



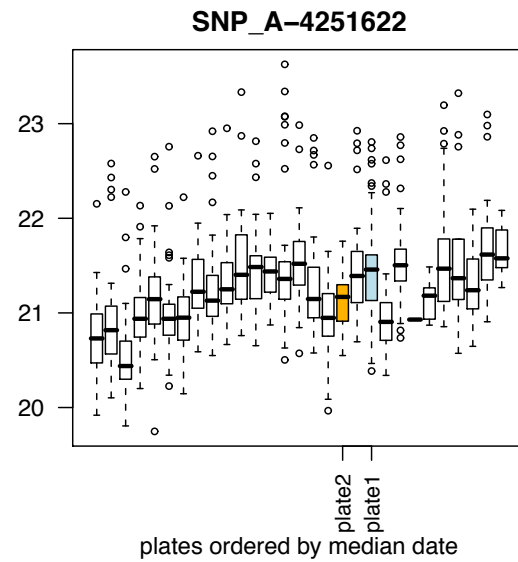
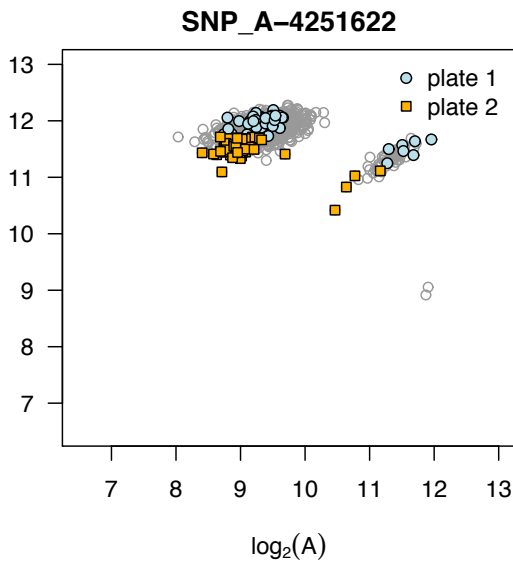
A versus B plots



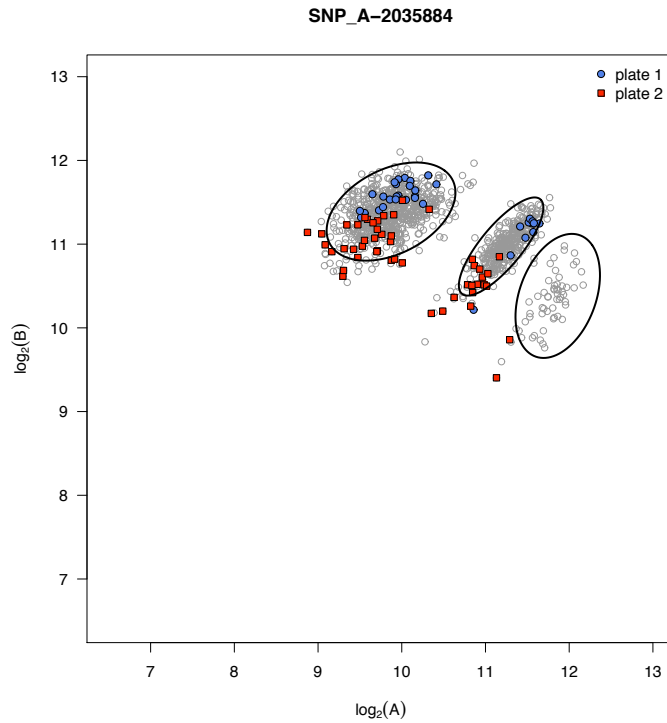
A versus B plots



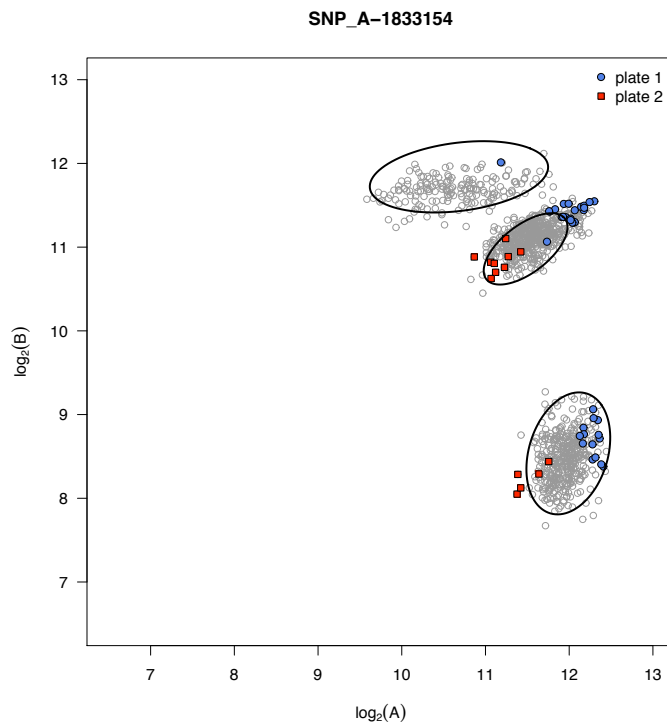
A versus B plots



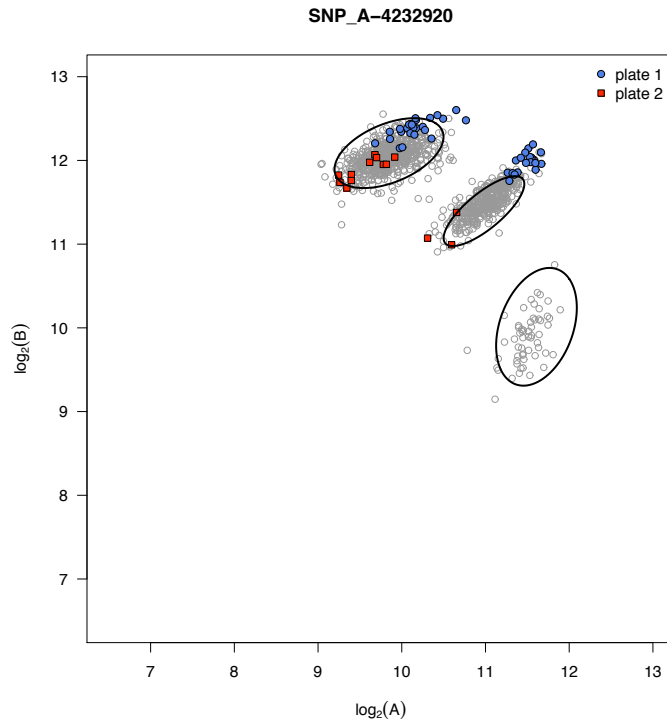
A versus B plots



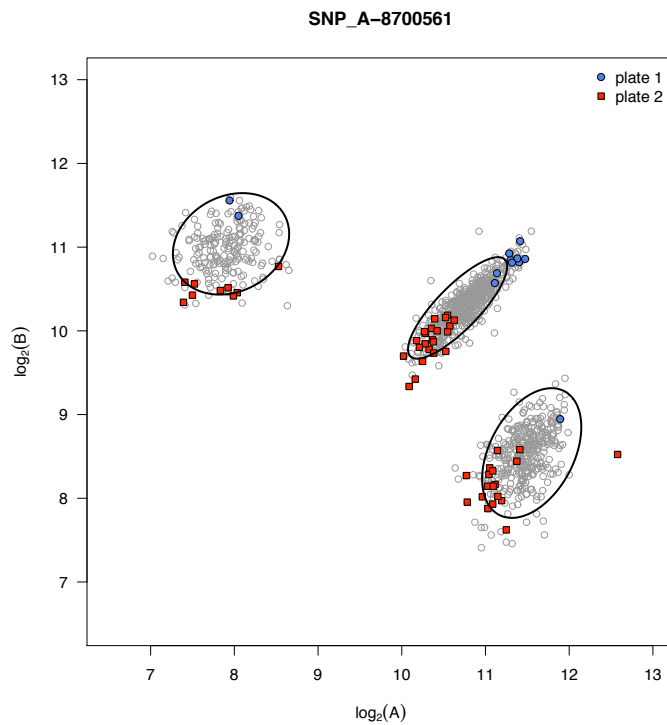
A versus B plots



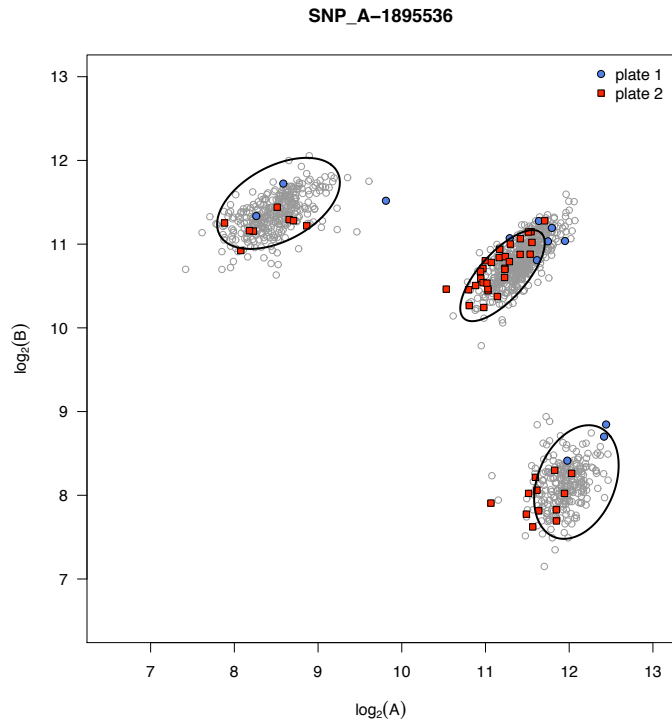
A versus B plots



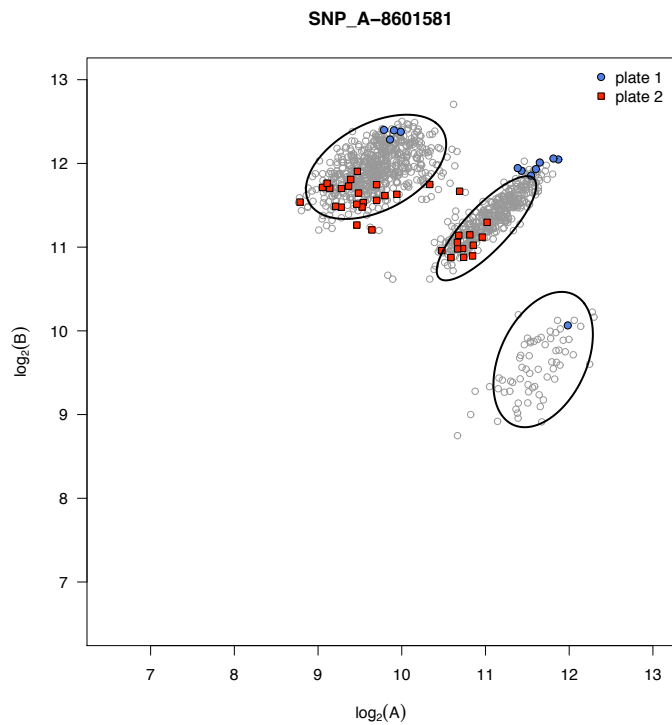
A versus B plots



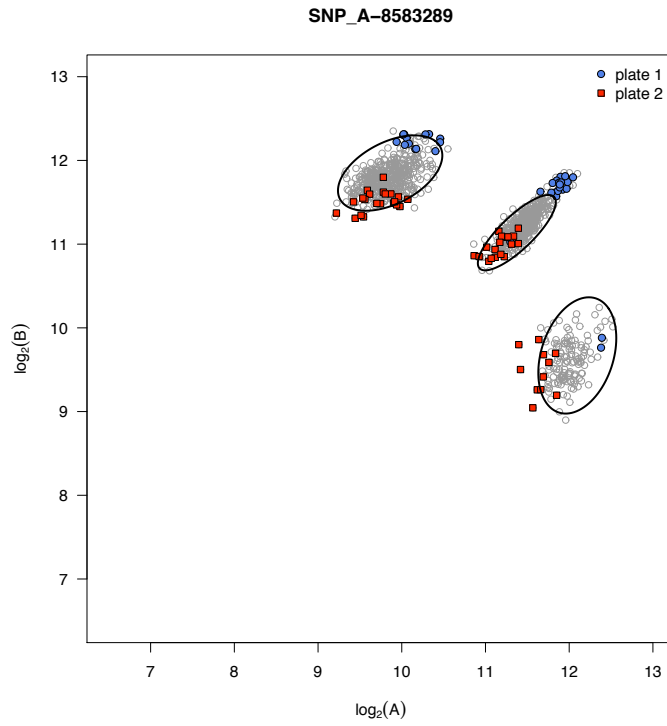
A versus B plots



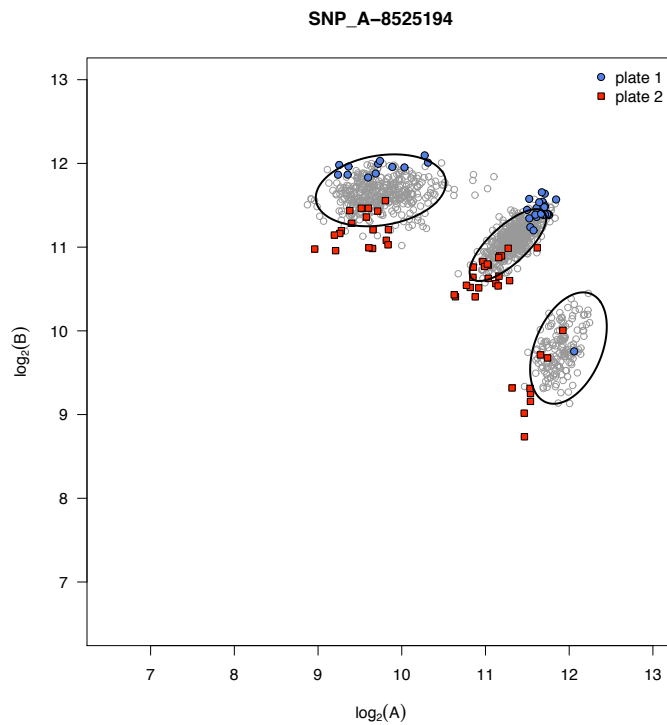
A versus B plots



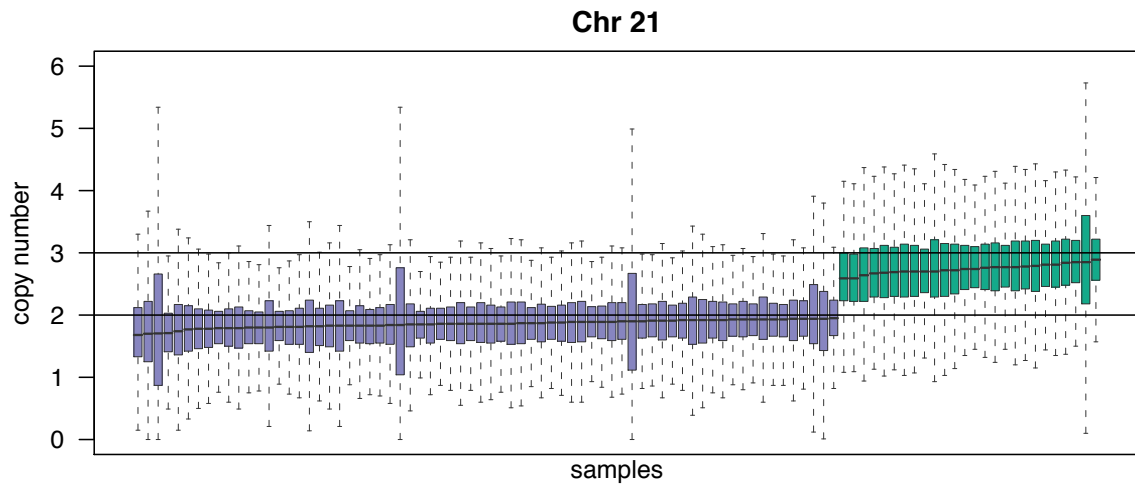
A versus B plots



A versus B plots

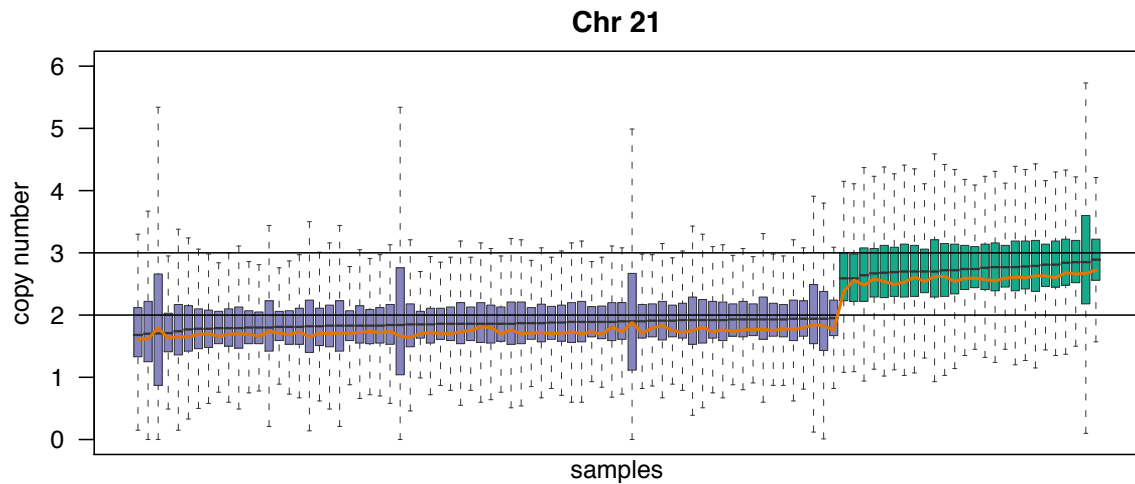


Trisomy 21



Samples from Aravinda Chakravarti and Betty Doan

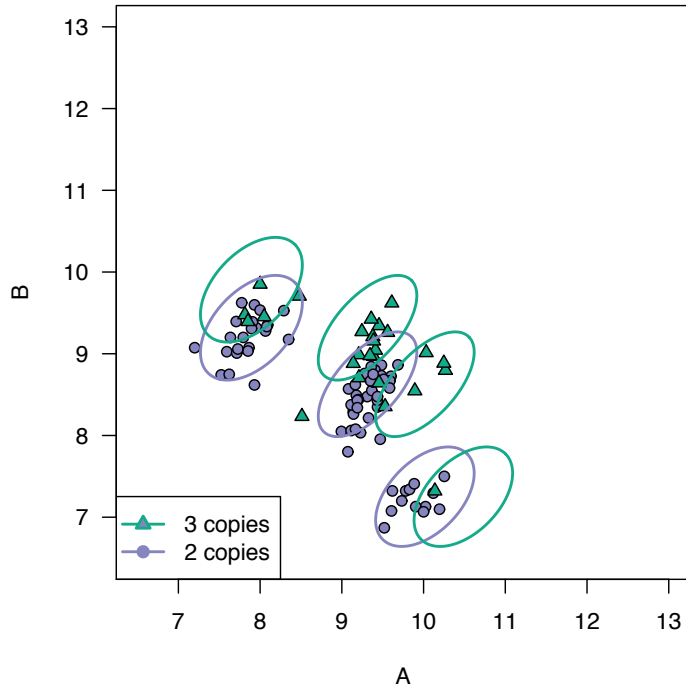
Trisomy 21



Samples from Aravinda Chakravarti and Betty Doan

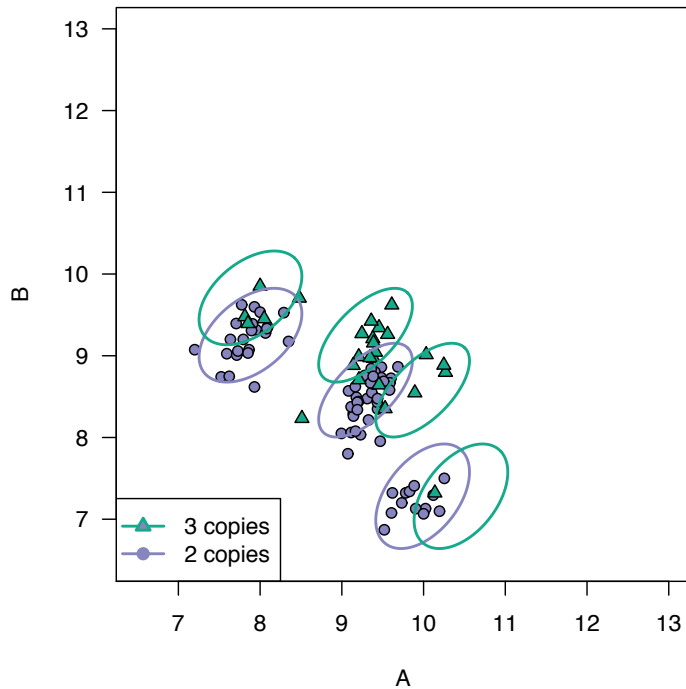
A versus B plots

SNP_A-8348190



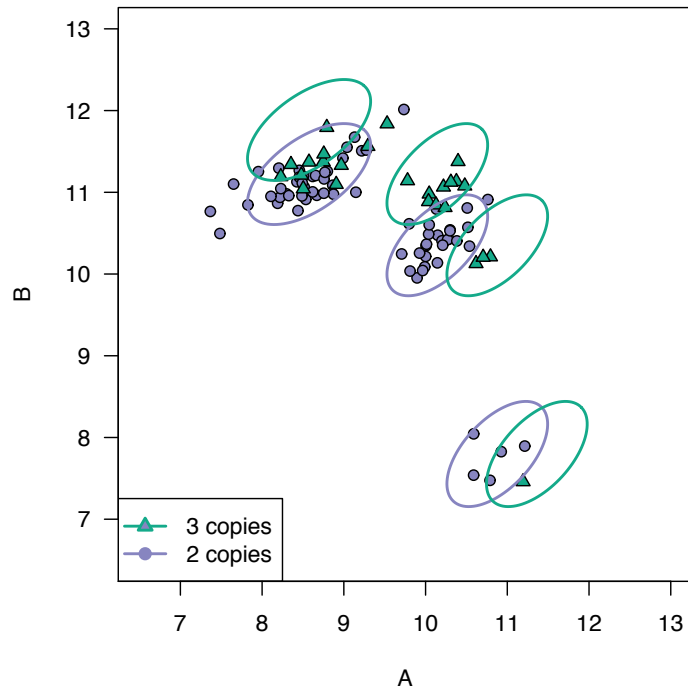
A versus B plots

SNP_A-8348190



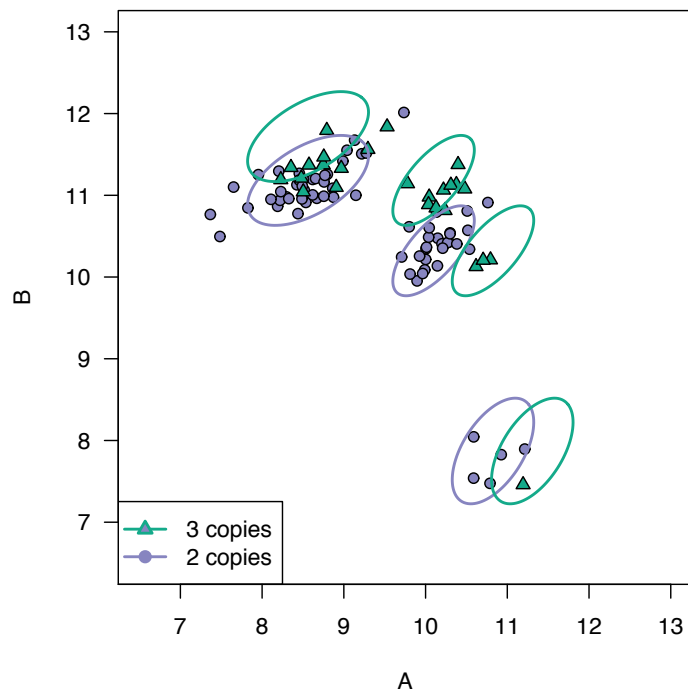
A versus B plots

SNP_A-8341330



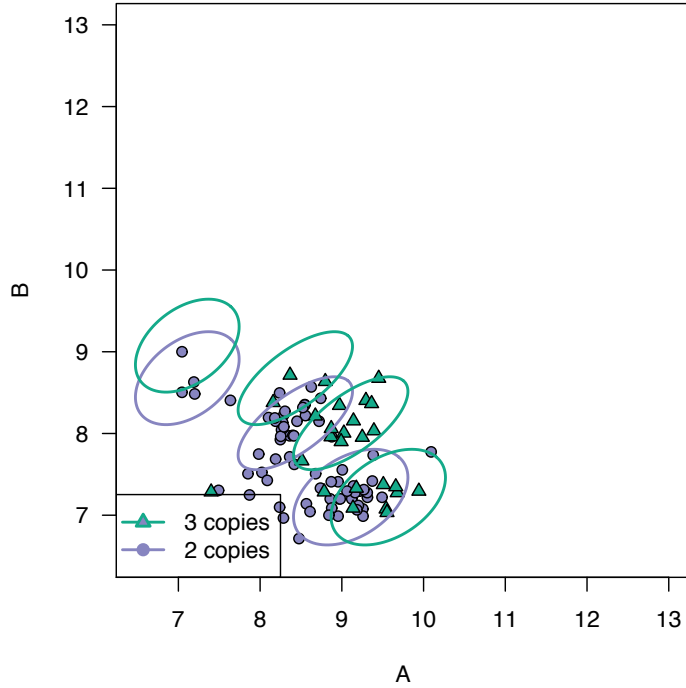
A versus B plots

SNP_A-8341330



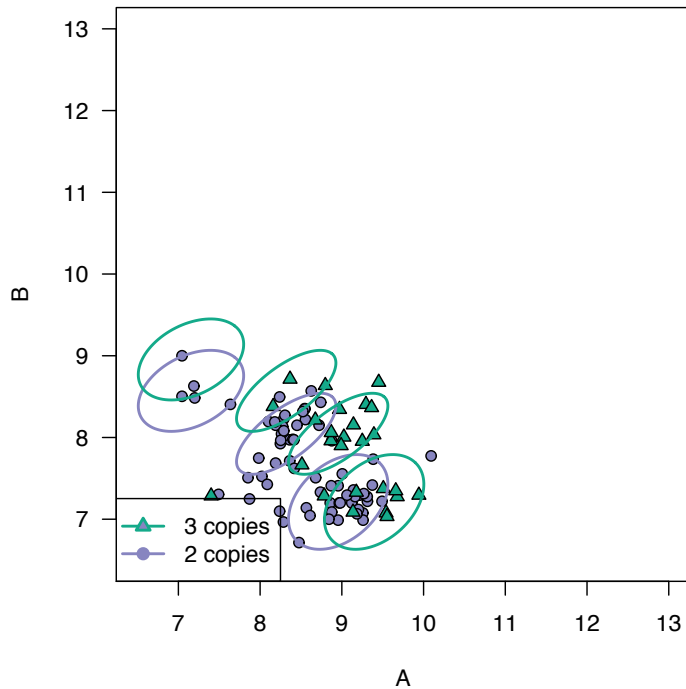
A versus B plots

SNP_A-8339372



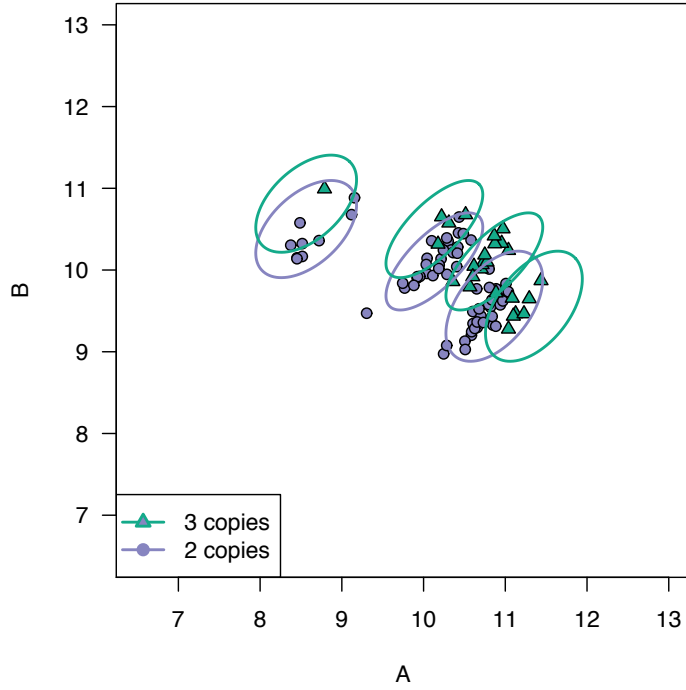
A versus B plots

SNP_A-8339372



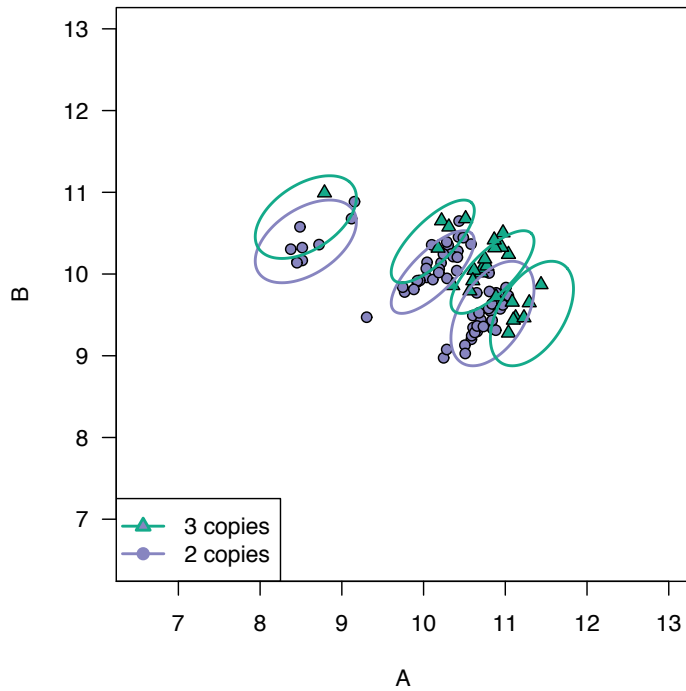
A versus B plots

SNP_A-8340560



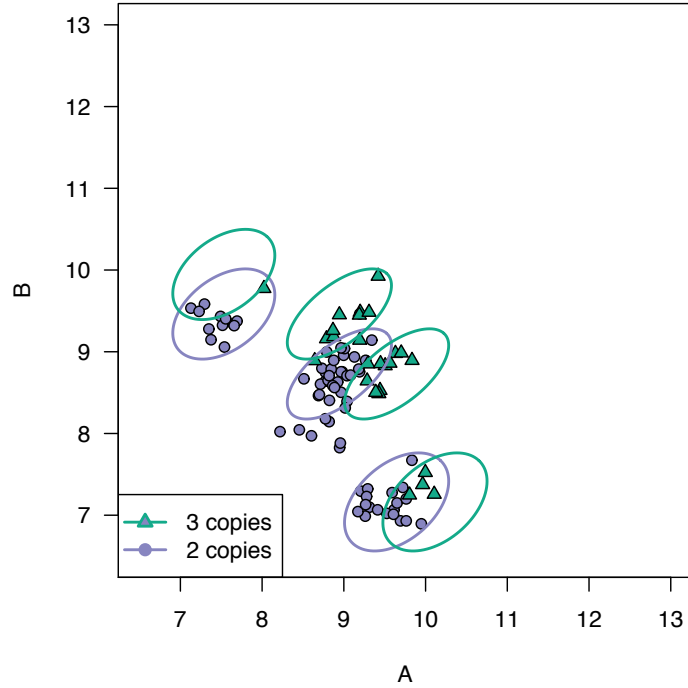
A versus B plots

SNP_A-8340560



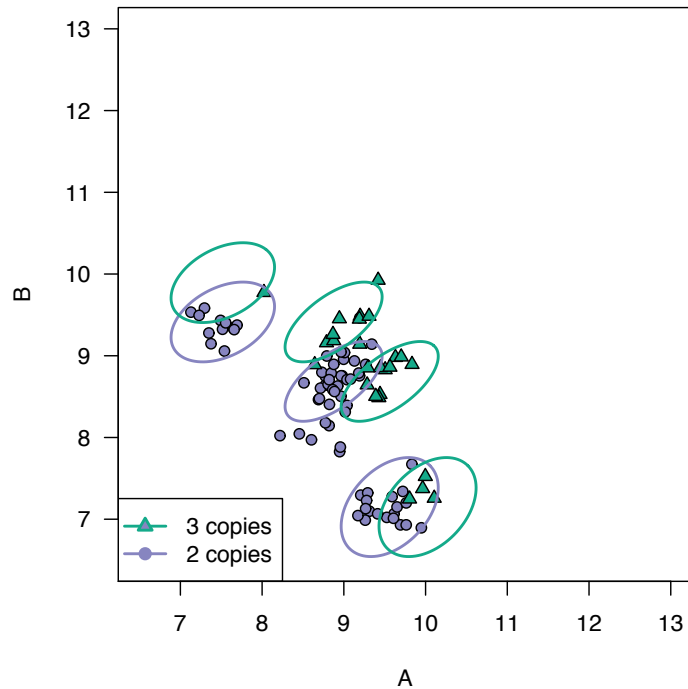
A versus B plots

SNP_A-1969323

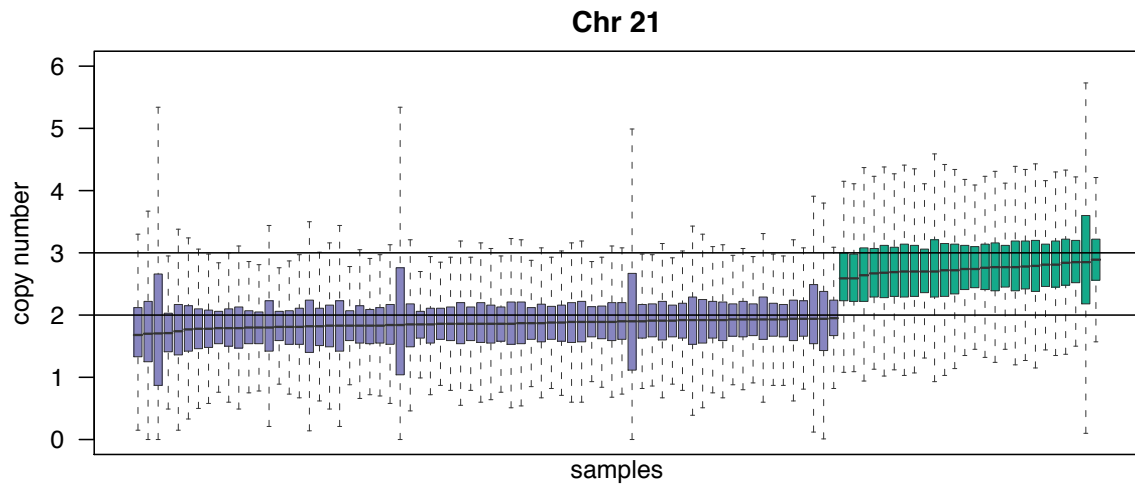


A versus B plots

SNP_A-1969323

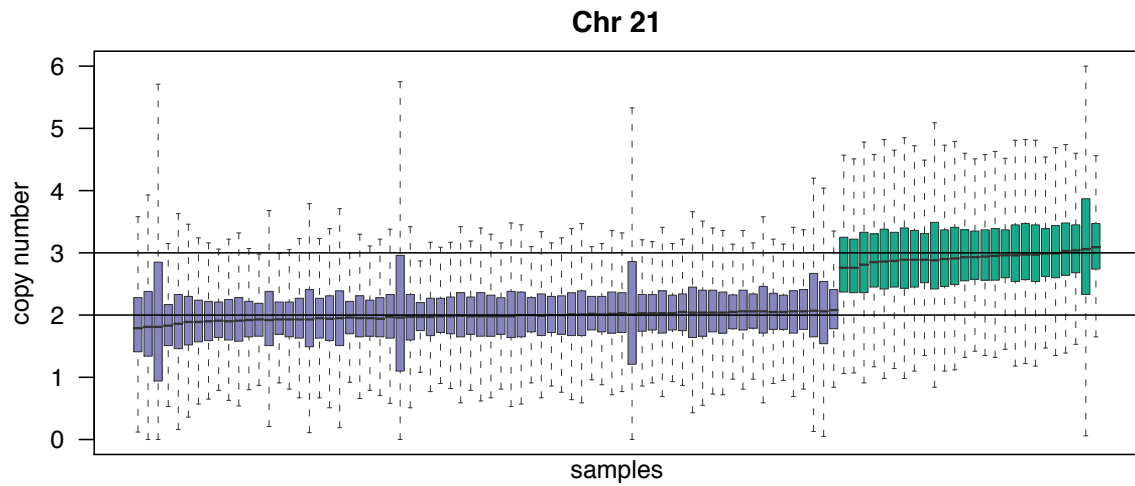


Trisomy 21



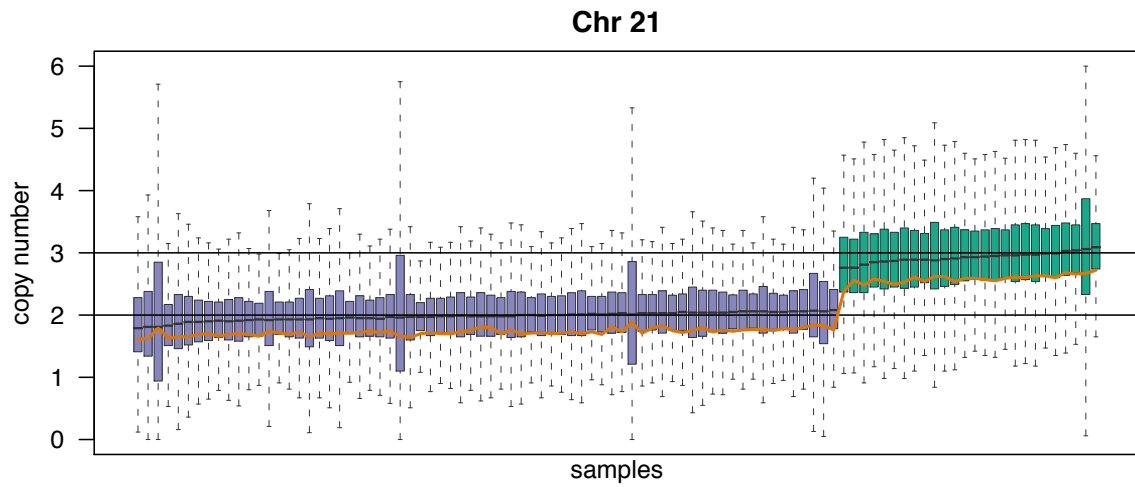
Samples from Aravinda Chakravarti and Betty Doan

Trisomy 21



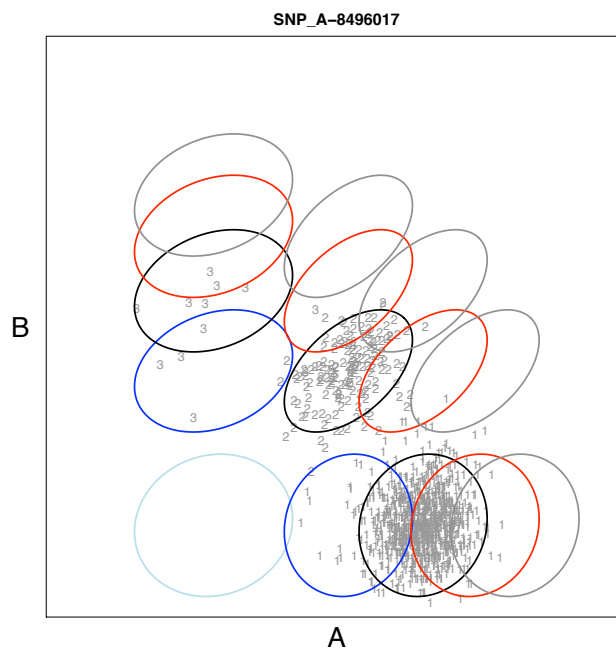
Samples from Aravinda Chakravarti and Betty Doan

Trisomy 21



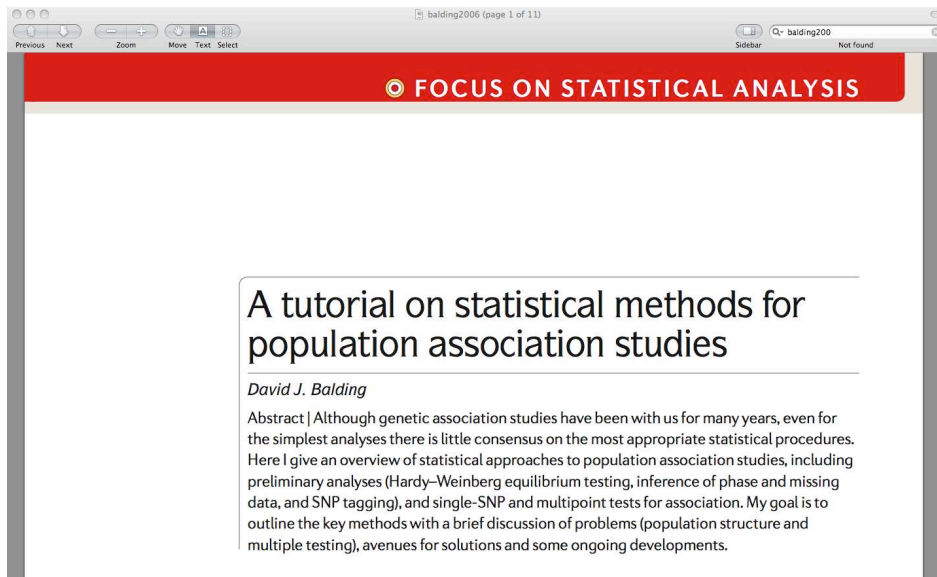
Samples from Aravinda Chakravarti and Betty Doan

Prediction regions for copy number



Scharpf et al (2010), in revision.

Population-based association studies



Balding (2006). Nature Reviews Genetics 7(10): 781-91.

Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays

Family-based designs



Laird and Lange (2006), Nature Reviews Genetics 7, 385-94.

Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays

Summary

Table 1. Study Designs Used in Genome-wide Association Studies

	Case-Control	Cohort	Trio
Assumptions	Case and control participants are drawn from the same population Case participants are representative of all cases of the disease, or limitations on diagnostic specificity and representativeness are clearly specified Genomic and epidemiologic data are collected similarly in cases and controls Differences in allele frequencies relate to the outcome of interest rather than differences in background population between cases and controls	Participants under study are more representative of the population from which they are drawn Diseases and traits are ascertained similarly in individuals with and without the gene variant	Disease-related alleles are transmitted in excess of 50% to affected offspring from heterozygous parents
Advantages	Short time frame Large numbers of case and control participants can be assembled Optimal epidemiologic design for studying rare diseases	Cases are incident (developing during observation) and free of survival bias Direct measure of risk Fewer biases than case-control studies Continuum of health-related measures available in population samples not selected for presence of disease	Controls for population structure; immune to population stratification Allows checks for Mendelian inheritance patterns in genotyping quality control Logistically simpler for studies of children's conditions Does not require phenotyping of parents
Disadvantages	Prone to a number of biases including population stratification Cases are usually prevalent cases, may exclude fatal or short episodes, or mild or silent cases Overestimate relative risk for common diseases	Large sample size needed for genotyping if incidence is low Expensive and lengthy follow-up Existing consent may be insufficient for GWA genotyping or data sharing Requires variation in trait being studied Poorly suited for studying rare diseases	May be difficult to assemble both parents and offspring, especially in disorders with older ages of onset Highly sensitive to genotyping error

Pearson and Manolio (2008). JAMA. 299(11): 1335-44.

Family-based designs —

- Typically less power per SNP typed than case-control studies.
- Pedigrees maybe hard to get except for childhood diseases, and may not be feasible for late-onset diseases.
- Can be a lot more expensive.
- Highly sensitive to genotyping errors.
- Might be computationally more demanding, especially for studies with large pedigrees.
- Software may be an issue.

Family-based designs +

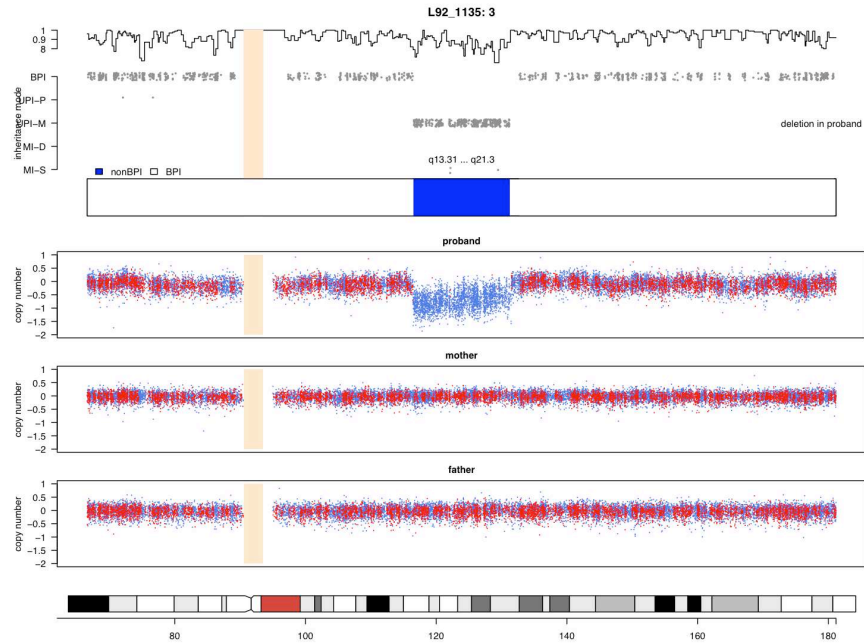
- Robust to possible effects of population stratification and genetic heterogeneity.
- Parent-of-origin effects (imprinting) can be assessed.
- Data quality control is usually more thorough (e. g. genotyping errors and sample swaps are easier to catch).
- Distinction between de-novo and inherited events (copy number changes) is possible.
- Logistically easier for childhood diseases.
- In case-parent data, low minor allele frequencies are of less worry (genotyping errors are still possible).
- Case-parent designs do not require phenotyping parents.
- Linkage information from previous family studies can be employed in association studies.

Parent-of-origin

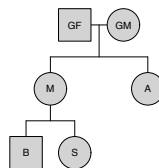
No.	SNP name	Paternal				Maternal				PO-LRT ^b	
		TAT				TAT				OR ^d	P-value
		T	NT	P-value	OR ^c	T	NT	P-value	OR ^c		
1	rs7771980	9	8	0.808	1.13	16	18	0.732	0.89	0.79	0.692
2	rs2677104	25	30	0.500	0.83	22	24	0.768	0.92	1.10	0.811
3	rs2819855	36	34	0.811	1.06	37	25	0.128	1.48	1.40	0.342
4	rs2819854	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

Sull et al (2008), *Genetic Epidemiology* 32: 505-12.

De-novo deletion

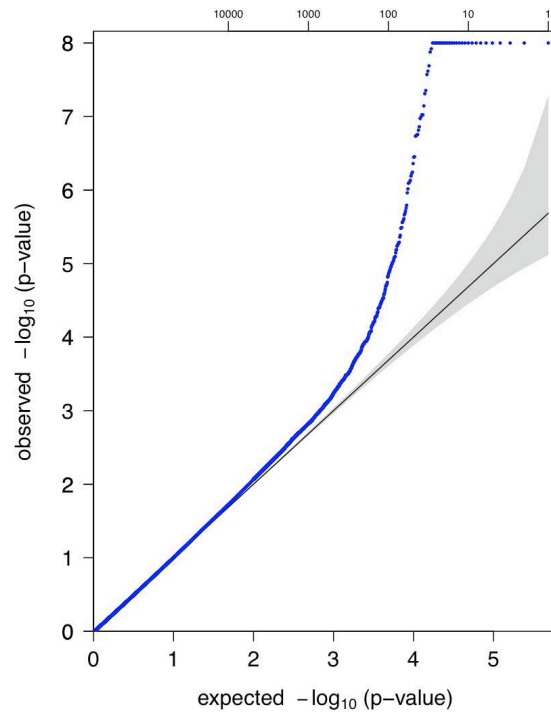


Homozygous and hemizygous deletions



Grandfather	Grandmother	Mother	Aunt	Brother	Sister
AB 0.01	BB 0.05	AB -0.11	AB -0.32	AB 0.06	BB -0.02
AA 0.27	NC -5.52	AA -0.48	AA -0.45	AB 0.12	BB -0.42
AB 0.15	NC -5.04	AA -0.20	AA -0.24	AB -0.09	BB -0.49
AB -0.03	NC -4.59	AA -0.40	AA -0.24	AA 0.30	AA -0.72
BB 0.20	NC -2.46	NC -0.38	BB -0.28	BB 0.22	BB -0.45
AB 0.03	NC -6.14	BB -0.28	BB -0.42	AB 0.09	AA -0.70
AB -0.05	NC -5.02	BB -0.17	BB -0.34	AB -0.22	AA -1.06
BB 0.01	NC -4.04	BB 0.04	BB -0.68	BB 0.14	NC -0.98
AB 0.17	NC -4.06	AA -0.27	AA -0.33	AA -0.03	AA -0.76
AB 0.01	NC -4.70	AA -0.67	AA -0.52	AA 0.16	AA -0.80
AB -0.10	NC -4.42	BB -0.25	BB -0.62	AB 0.13	AA -0.58
AB 0.01	NC -8.29	BB -0.17	BB -0.15	AB -0.15	AA -0.29
BB 0.16	NC -5.73	BB -0.64	BB -0.46	BB 0.10	BB -0.52
AB 0.06	NC -7.48	AA -0.23	AA -0.33	AB 0.07	BB -0.47
AA 0.17	NC -3.70	AA -0.50	AA -0.52	AB -0.06	BB -0.48
BB 0.02	NC -5.00	BB -0.34	BB -0.45	AB 0.13	AA -0.55
AA 0.21	NC -6.10	AA -0.43	AA -0.40	AB 0.20	BB -0.40
BB 0.05	BB 0.11	BB 0.15	BB 0.29	AB 0.13	AB -0.01

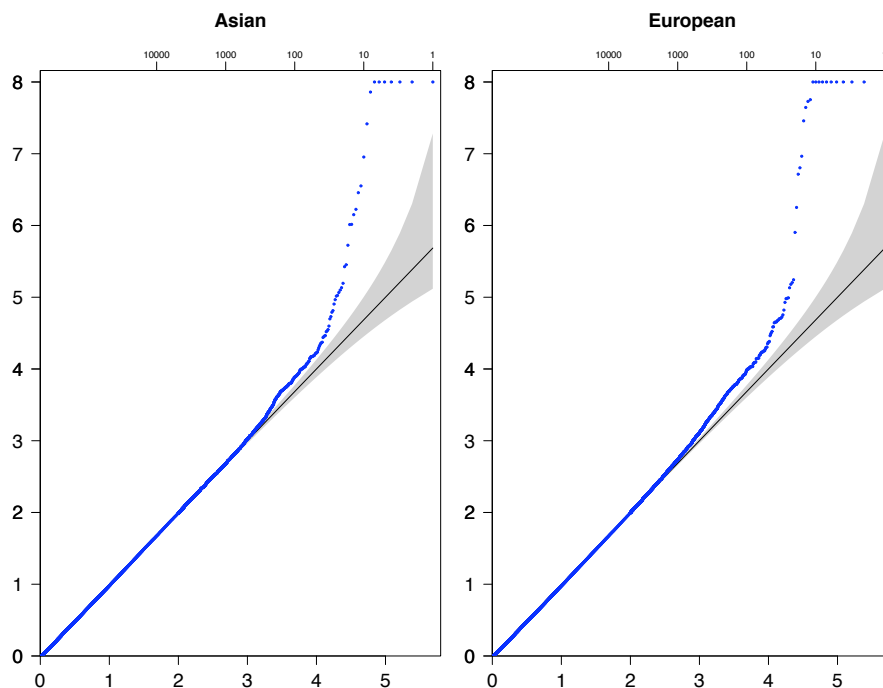
International Cleft Consortium



Ingo Ruczinski

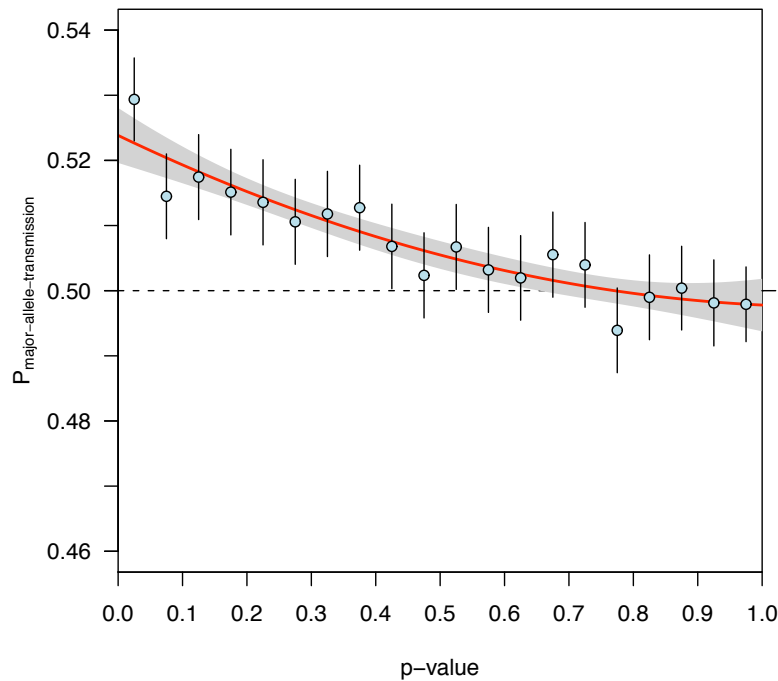
Assessing Genomic Variability with SNP Arrays

International Cleft Consortium



Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays



Ingo Ruczinski

Assessing Genomic Variability with SNP Arrays

Principal components

novembre2008.pdf

Create PDF • Combine Files • Export • Start Meeting • Secure • Sign • Forms • Review & Comment

1 / 7 141% Find

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 Bryc⁷, Zoltán Kutalik^{4,6}, Adam R. Boyko⁷, Adam Auton⁷, Amit Indap⁷, Karen S. King⁸, Sven Bergmann^{4,6}, Matthew R. Nelson⁹, Matthew Stephens^{2,3} & Carlos D. Bustamante⁷

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^{1–5}. 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; 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.

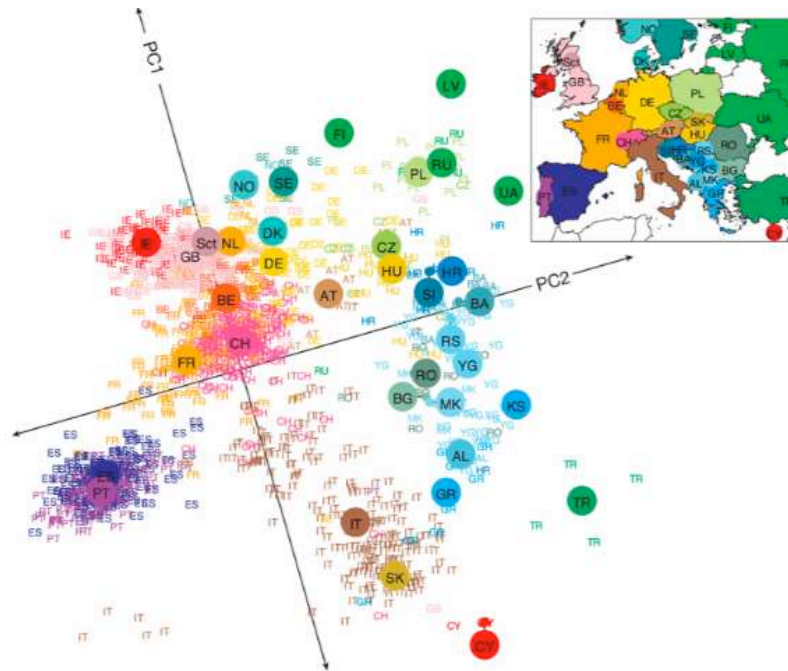
When we quantitatively compare the geographic position of countries with their PC-based genetic positions, we observe few prominent differences between the two (Supplementary Fig. 1), and those that exist can be explained either by small sample sizes (for example, Slovakia (SK)) or by the coarseness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

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

Ingo Ruczinski

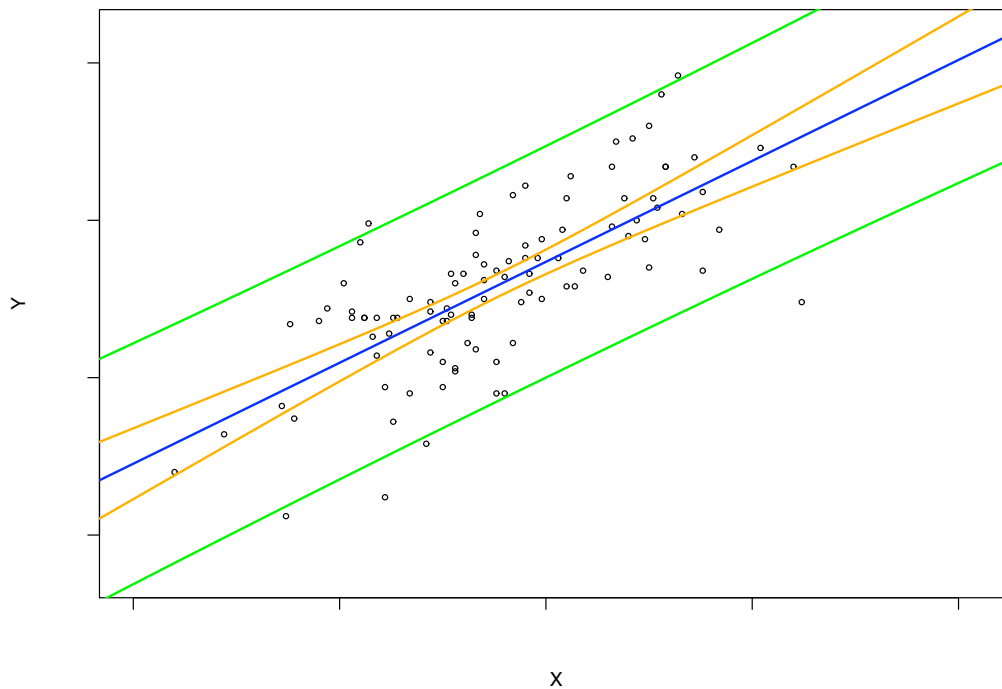
Assessing Genomic Variability with SNP Arrays

Principal components

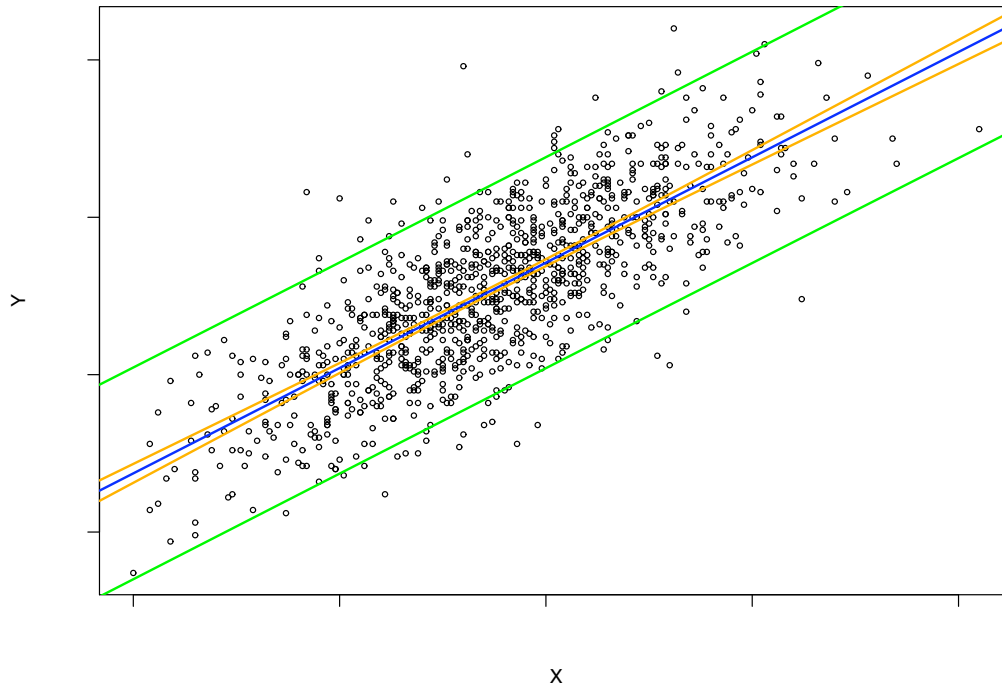


Novembre et al (2008), Nature 456: 98-101.

Prediction



Prediction



Acknowledgments

Collaborators: Rob Scharpf

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

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/>