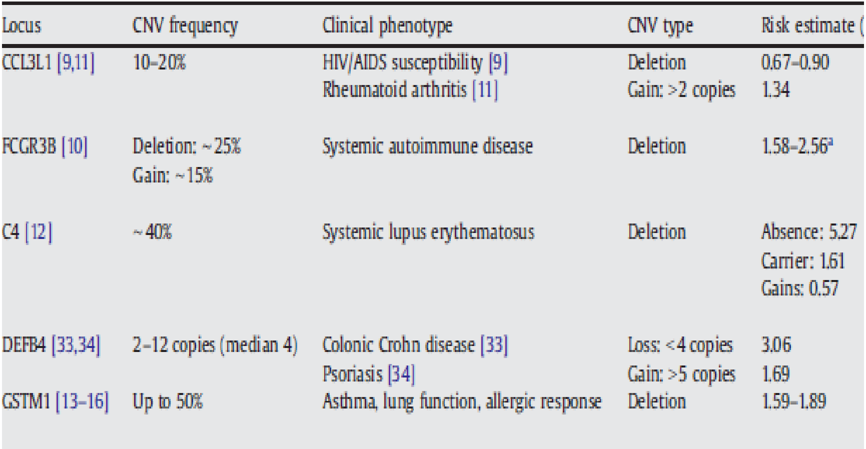
Finding structural variants associated with disease

Lachlan Coin

Department of Epidemiology and Public Health,
Imperial College London, UK

January 2010

Limited association between common CNVs and common disease

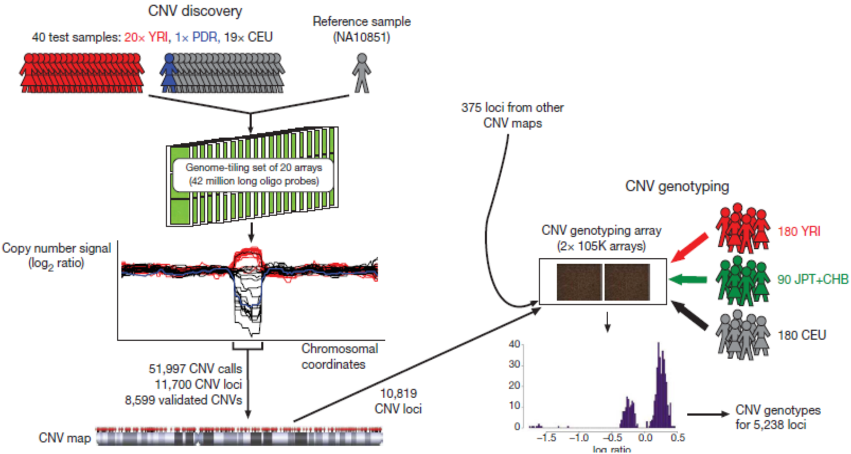


Locus	CNV frequency	Clinical phenotype	CNV type	Risk estimate (
CCL3L1 [9,11]	10–20%	HIV/AIDS susceptibility [9]	Deletion	0.67–0.90
		Rheumatoid arthritis [11]	Gain: >2 copies	1.34
FCGR3B [10]	Deletion: ~25% Gain: ~15%	Systemic autoimmune disease	Deletion	1.58–2.56[a]
C4 [12]	~40%	Systemic lupus erythematosus	Deletion	Absence: 5.27 Carrier: 1.61 Gains: 0.57
DEFB4 [33,34]	2–12 copies (median 4)	Colonic Crohn disease [33]	Loss: <4 copies	3.06
		Psoriasis [34]	Gain: >5 copies	1.69
GSTM1 [13–16]	Up to 50%	Asthma, lung function, allergic response	Deletion	1.59–1.89

IonitaLaza et. al. *Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis*

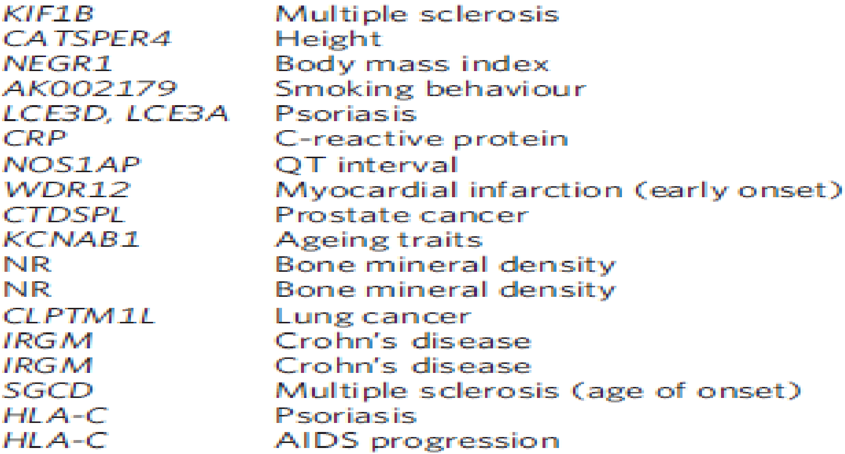
Genomics 2008

WTCCC scan for common CNVs associated with disease found ‘few’ new signals


CNV discovery
40 test samples: 20× YRI, 1× PDR, 19× CEU
Reference sample (NA10851)
Genome-tiling set of 20 arrays (42 million long oligo probes)
Copy number signal ($\log_2$ ratio)
Chromosomal coordinates
51,997 CNV calls
11,700 CNV loci
8,599 validated CNVs
CNV map
10,819 CNV loci
375 loci from other CNV maps
CNV genotyping
CNV genotyping array (2× 105K arrays)
180 YRI
90 JPT+CHB
180 CEU
40
30
20
10
0
−1.5
−1.0
−0.5
0.0
0.5
log ratio
CNV genotypes for 5,238 loci

Conrad et al. *Origins and functional impact of copy number variation in the human genome* Nature 2009

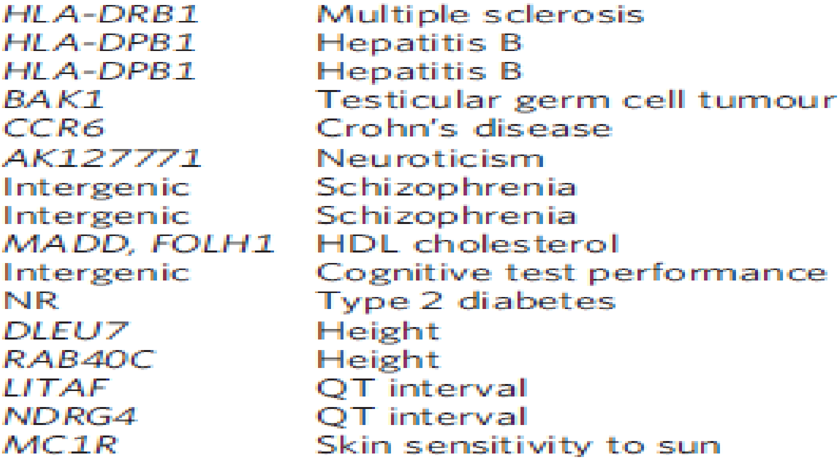
Few CNVs correlated with trait-associated SNPs (I)



KIF1B	Multiple sclerosis
CATSPER4	Height
NEGR1	Body mass index
AK002179	Smoking behaviour
LCE3D, LCE3A	Psoriasis
CRP	C-reactive protein
NOS1AP	QT interval
WDR12	Myocardial infarction (early onset)
CTDSPL	Prostate cancer
KCNAB1	Ageing traits
NR	Bone mineral density
NR	Bone mineral density
CLPTM1L	Lung cancer
IRGM	Crohn's disease
IRGM	Crohn's disease
SGCD	Multiple sclerosis (age of onset)
HLA-C	Psoriasis
HLA-C	AIDS progression

Conrad et al. *Origins and functional impact of copy number variation in the human genome* Nature 2009

Few CNVs correlated with trait-associated SNPs (II)



HLA-DRB1	Multiple sclerosis
HLA-DPB1	Hepatitis B
HLA-DPB1	Hepatitis B
BAK1	Testicular germ cell tumour
CCR6	Crohn's disease
AK127771	Neuroticism
Intergenic	Schizophrenia
Intergenic	Schizophrenia
MADD, FOLH1	HDL cholesterol
Intergenic	Cognitive test performance
NR	Type 2 diabetes
DLEU7	Height
RAB40C	Height
LITAF	QT interval
NDRG4	QT interval
MC1R	Skin sensitivity to sun

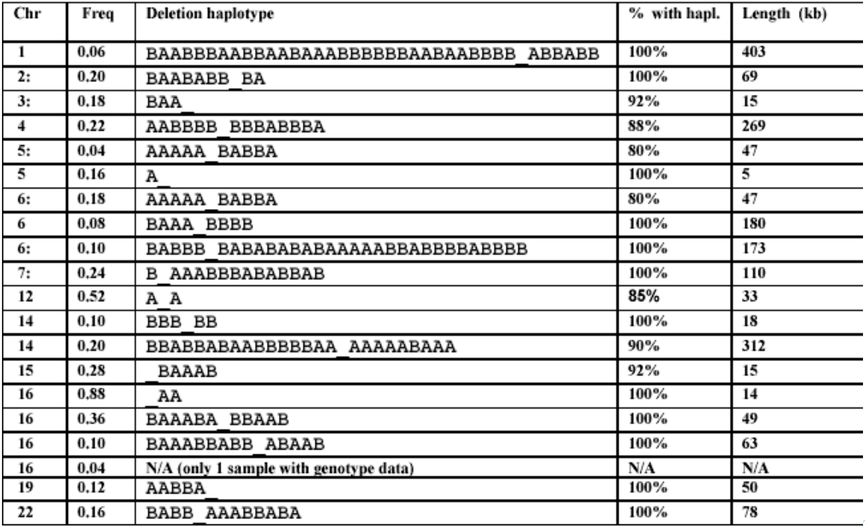
Conrad et al. *Origins and functional impact of copy number variation in the human genome* Nature 2009

WTCCC conclude common CNVs do not account for missing heritability

- 77% of ‘genotypeable’ CNVs well-tagged ($r^2 > 0.5$) by SNPs
- Conclude that GWAS have already screened for SNP effects
- Estimate they have genotyped $25 - 35\%$ of common CNVs $>$ 1kb

Conrad et al. *Origins and functional impact of copy number variation in the human genome* Nature 2009

50% of 20 sequenced deletions were part of extended haplotypes $> 50kb$

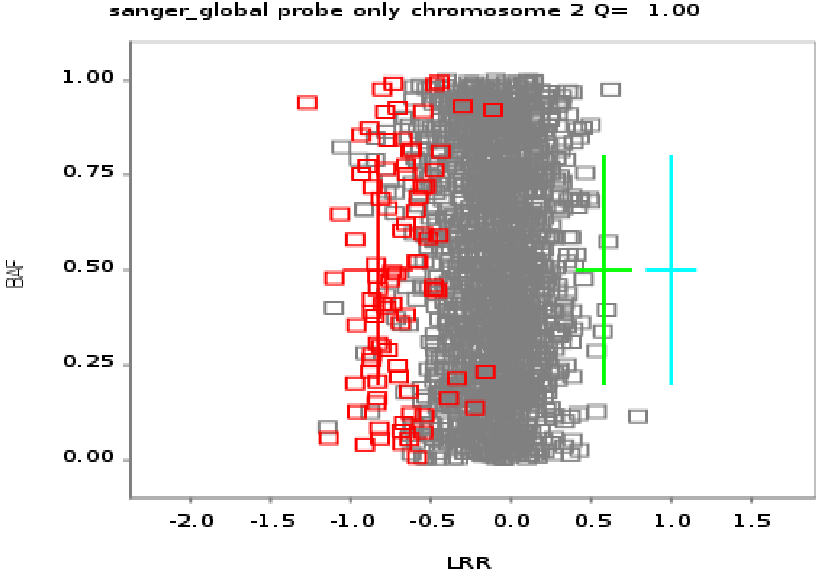


Chr	Freq	Deletion haplotype	% with hapl.	Length (kb)
1	0.06	BAABBBAABBAABAAABBBBBBAABAABBBB_ABBABB	100%	403
2:	0.20	BAABABB_BA	100%	69
3:	0.18	BAA_	92%	15
4	0.22	AABBBB_BBBABBBA	88%	269
5:	0.04	AAAAA_BABBA	80%	47
5	0.16	A_	100%	5
6:	0.18	AAAAA_BABBA	80%	47
6	0.08	BAAA_BBBB	100%	180
6:	0.10	BABBB_BABABABABAAAAABBABBBBABBBB	100%	173
7:	0.24	B_AAABBBABABBAB	100%	110
12	0.52	A_A	85%	33
14	0.10	BBB_BB	100%	18
14	0.20	BBABBABAABBBBBAA_AAAAABAAA	90%	312
15	0.28	_BAAAB	92%	15
16	0.88	_AA	100%	14
16	0.36	BAAABA_BBAAB	100%	49
16	0.10	BAAABBABB_ABAAB	100%	63
16	0.04	N/A (only 1 sample with genotype data)	N/A	N/A
19	0.12	AABBA_	100%	50
22	0.16	BABB_AAABBABA	100%	78

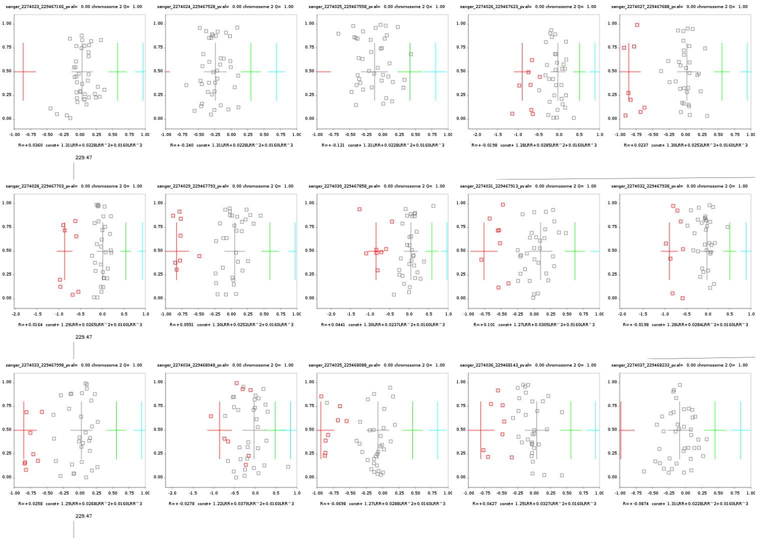
Are CNVs still worth pursuing?

- Amplifications are less well-tagged than deletions
- Tagging efficiency of 0.5 will require many more samples to detect weak effect
- Conclusions not applicable to complex multi-allelic CNVs
- Conclusions only for common CNVs which were discovereable in cohort of 20YRI+20CEU

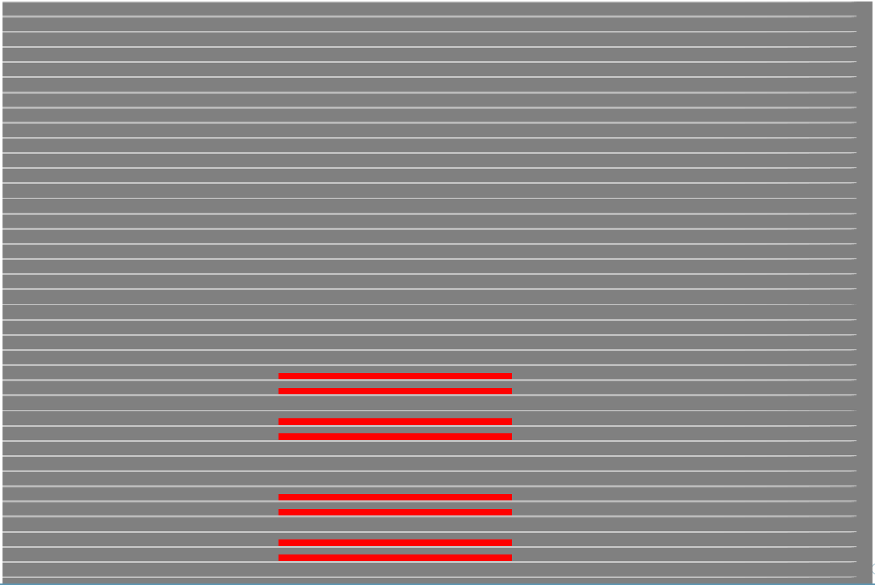
WTCCC discovery data is challenging



WTCCC discovery data is challenging (I)



cnvHap identified sequenced deletion with $< 100bp$ resolution using WTCCC discovery data



CNV-phenotype association strategies

1. Identify ‘Genotypable CNV regions’
   - CNV discovery (typically using HMM, or circular segmentation) per-sample
   - Known CNV regions
2. Genotype CNV pointwise across samples in fixed CNV regions $\Rightarrow$ association of integer CN state with phenotype
3. Association of continuous intensity signal with phenotype in fixed CNV regions

CNV association beset by various technical difficulties

- Different plates have different intensity response at each probe $\Rightarrow$ need for between plate normalisation $\Rightarrow$ particularly problematic if plates are case/control specific
- Probe binding efficiency varies according to GC-content, which results in wave-like effects of intensity across genome
- High variance of intensity measurements
- $\Rightarrow$ CNV genotyping accuracy is still low
- Difficult to combine results in meta-analyses across different chips and different populations $\Rightarrow$ wide variety of chips and platforms in use

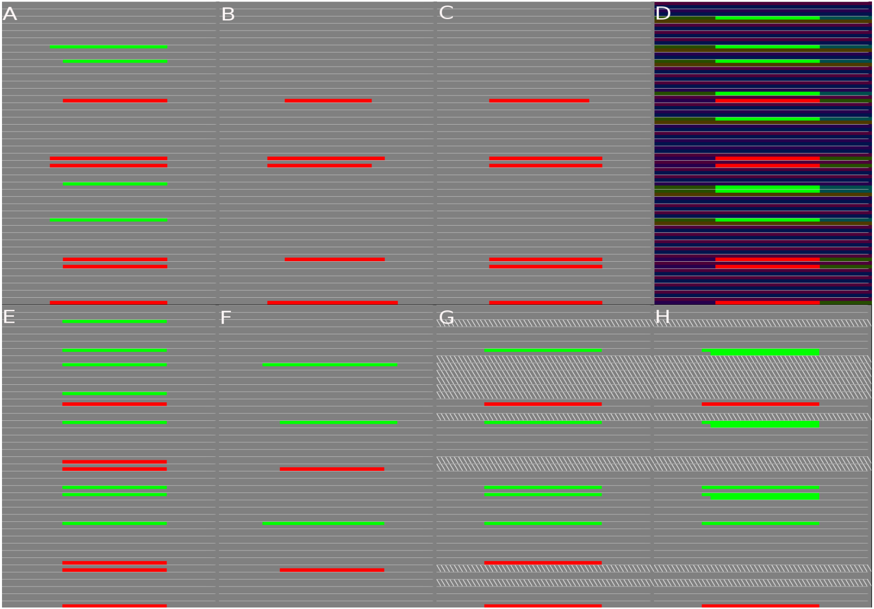
Benchmarking study to ascertain sensitivity/specifity genomewide

- 50 French individuals genotyped on
  - Illumina Human1M BeadArray
  - Agilent 244k CGH array
- 35 of these genotyped on
  - Illumina 317k BeadArray
  - Agilent 185k CGH array

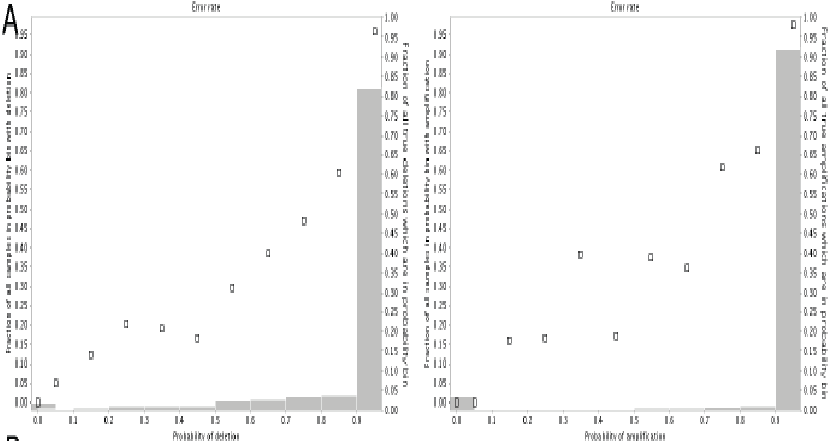
We performed two comparisons

- Run PennCNV, cnvPartition, cnvHap on 1M data
  - Map predictions to 244k probes using imputation
  - Compare with direct CNV annotation on aCGH 244k probeset according to ADM2
- Reverse experiment to compare cnvHap to ADM2 on aCGH data using 1M annotation as benchmark

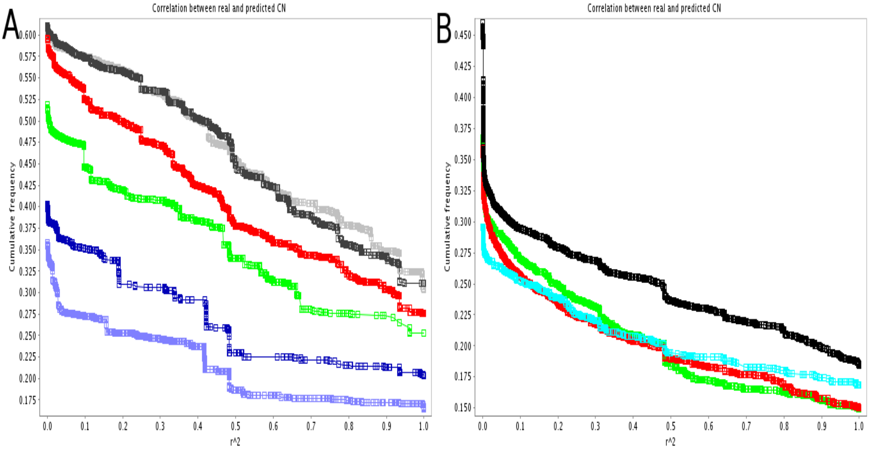
Performance of algorithms on test region



Feasibility of CNV Imputation



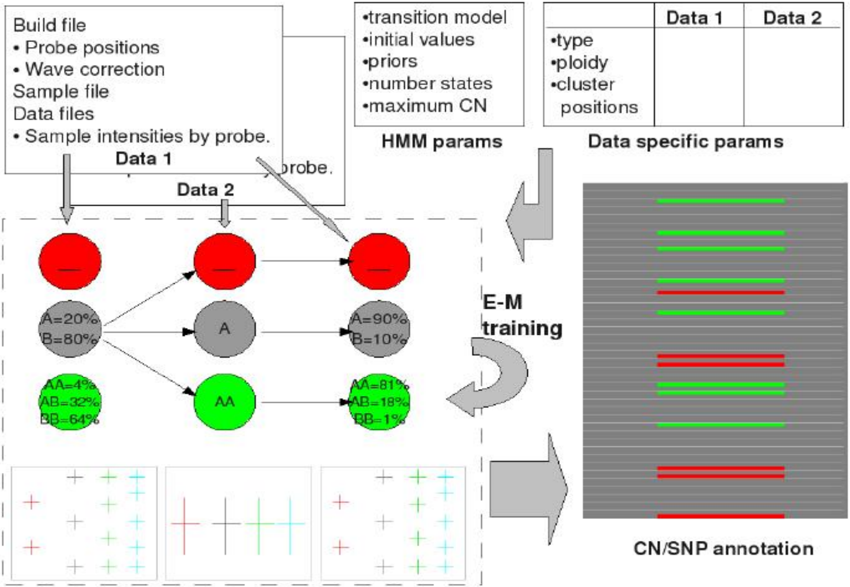
Correlation between real and predicted CN genomewide



cnvHap: Population haplotype model for multi-platform CNV prediction and imputation

- Integrates information from multiple chips into a single consistent CNV annotation
- Models CNVs at the single chromosome level → improves sensitivity by integrating LD information (between SNPS and CNVS, and also between SNPS/SNPS and CNVS/CNVS) into CNV prediction
- Models all samples simultaneously
- Updates cluster positions as part of maximisation procedure
- Also imputes CN genotype at unmeasured loci, and estimates the uncertainty in this estimation, so can be used to map CNV prediction from one probeset onto another

cnvHap model



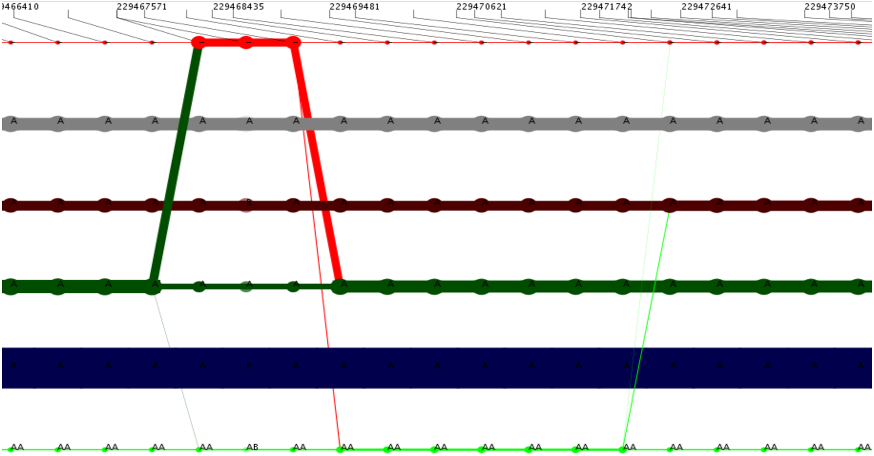
Build file
- Probe positions
- Wave correction

Sample file

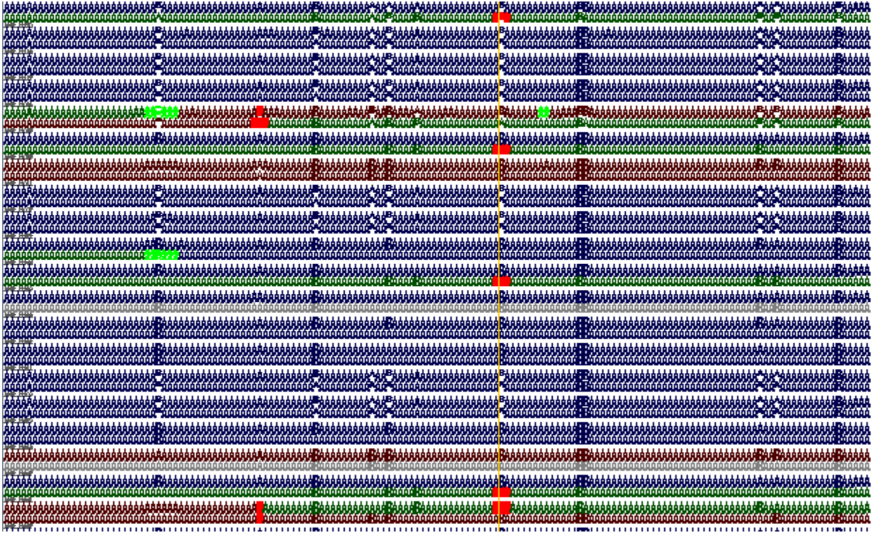
Data files
- Sample intensities by probe.

	Data 1	Data 2
•type •ploidy •cluster positions		

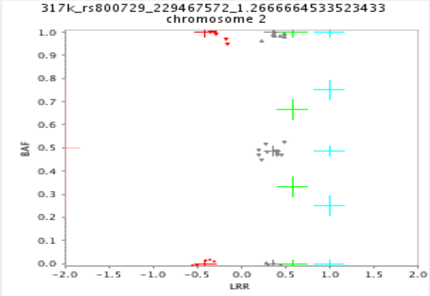
Haplotype model found common deletion haplotype

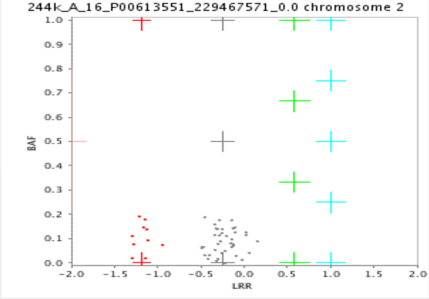

9466410
229467571
229468435
229469481
229470621
229471742
229472641
229473750
A A A A A A A A A A A A A A A A A A A
A A A A A B A A A A A A A A A A A A A
A A A A A A A A A A A A A A A A A A A
A A A A A B A A A A A A A A A A A A A
AA AA AA AA AA AB AA AA AA AA AA AA AA AA AA AA AA AA AA

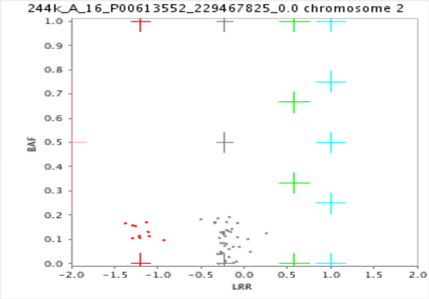
Identifying CNV haplotypes



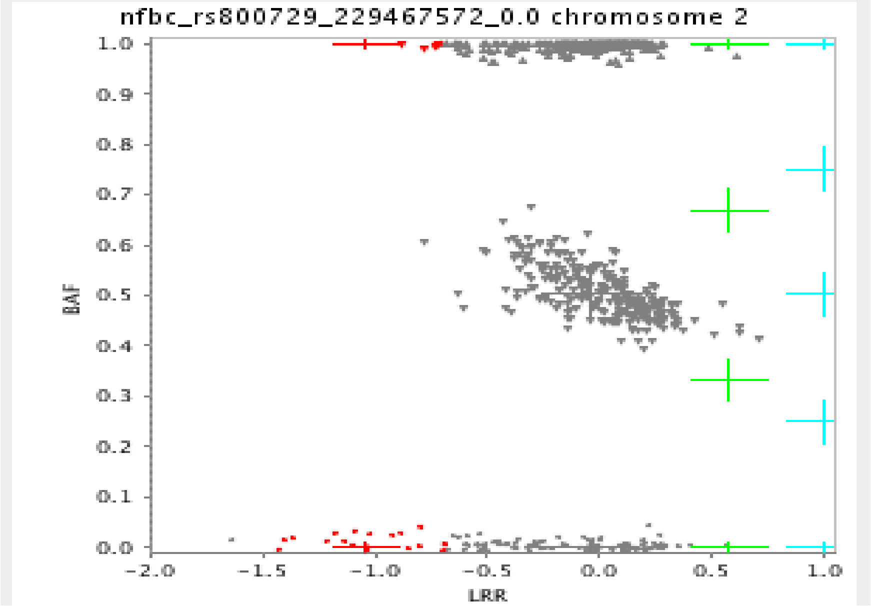
Considering population improves accuracy . . .



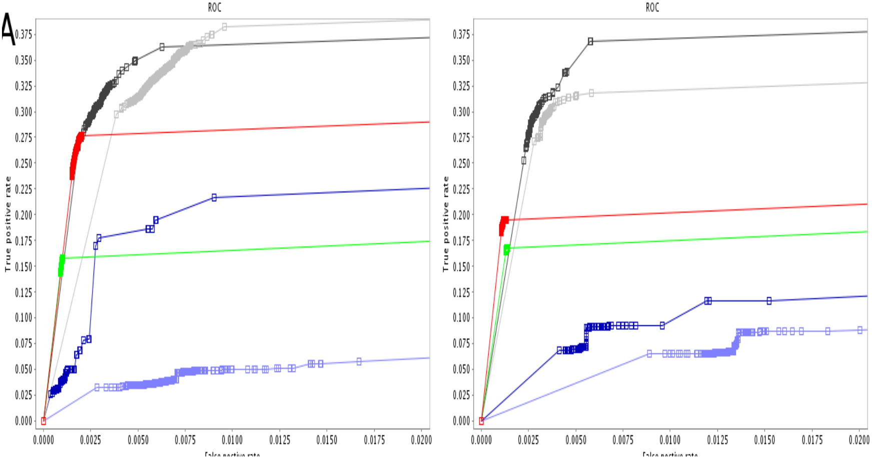




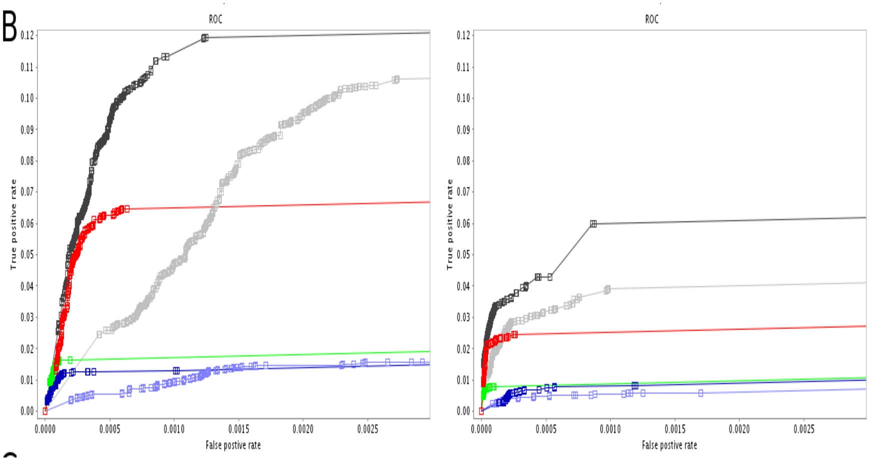
. . . particularly in a larger population



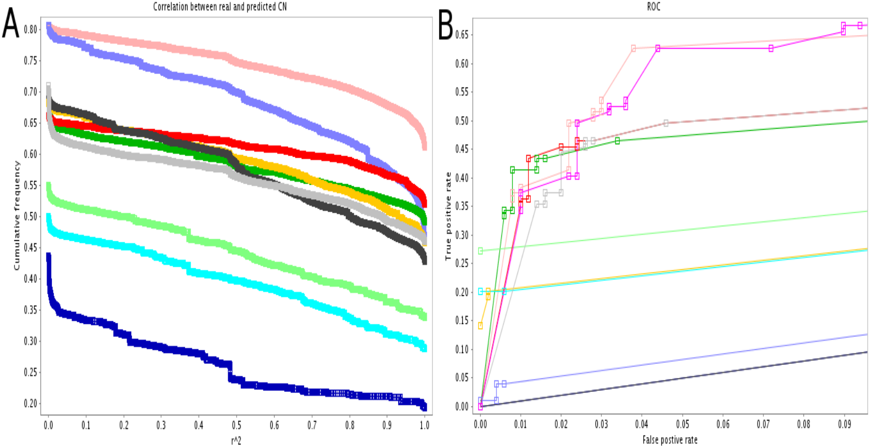
ROC curves for detecting CNVs by individual



ROC curves for detecting CNV break points

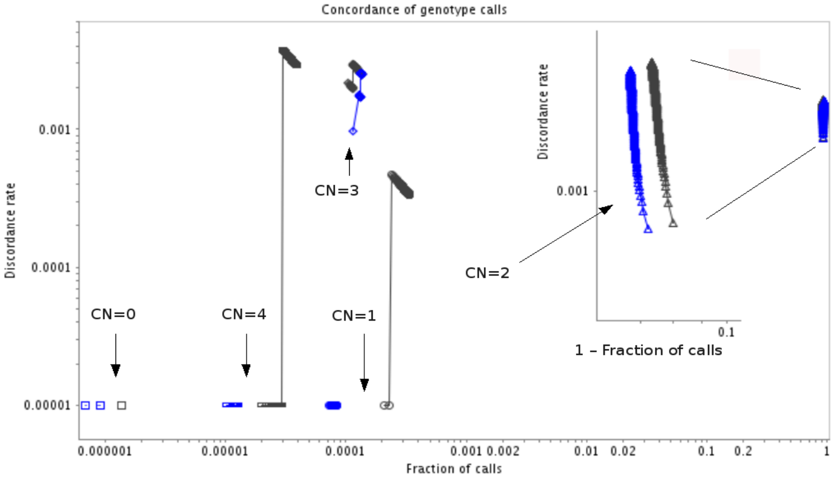


Combining datasets improves accuracy

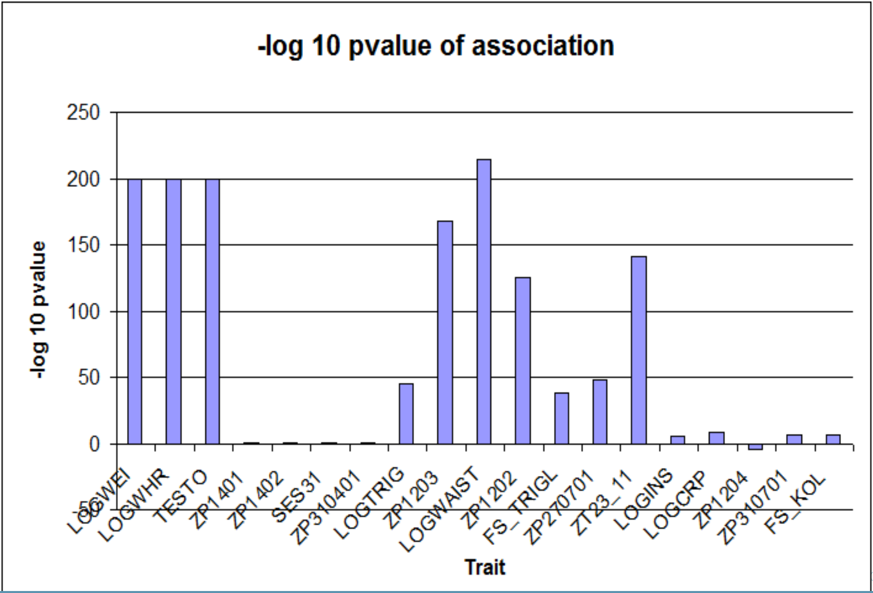


Magenta=244k+1M+185k+317k; pink=244k+1M; red=185k+244k; dark green=244k+317k;light blue =317k+1M;orange=1M+185k; dark grey=1M; light-grey=244k;light green=185k+317k; cyan = 185k; dark blue= 317k.

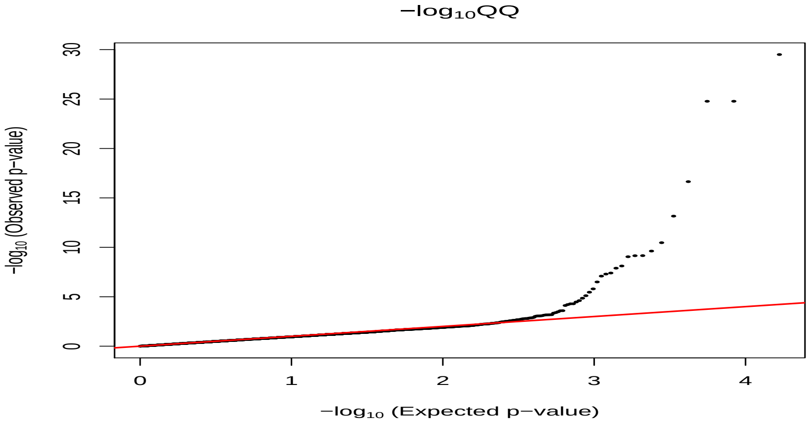
cnvHap model enables genotyping within different CN states



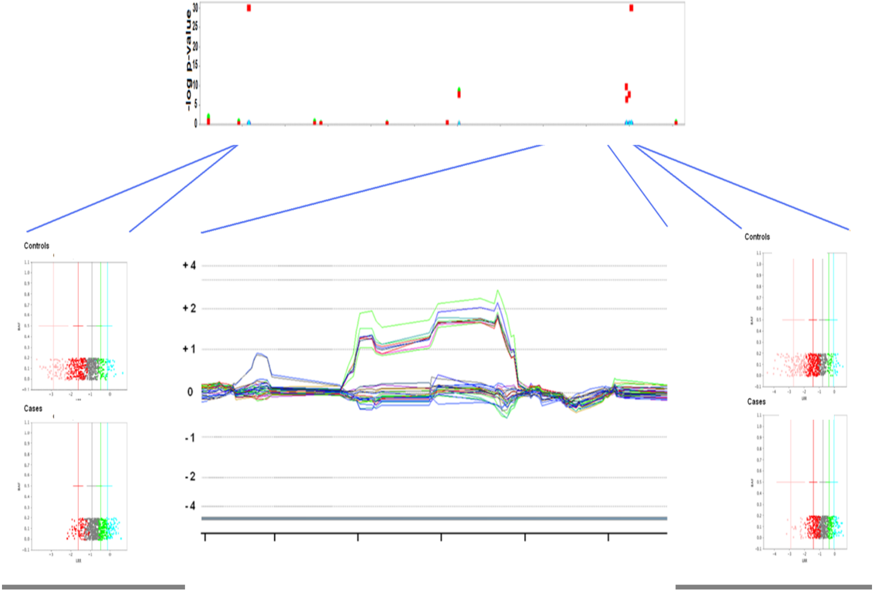
Strong CNV association on chromosome 1 for multiple metabolic traits



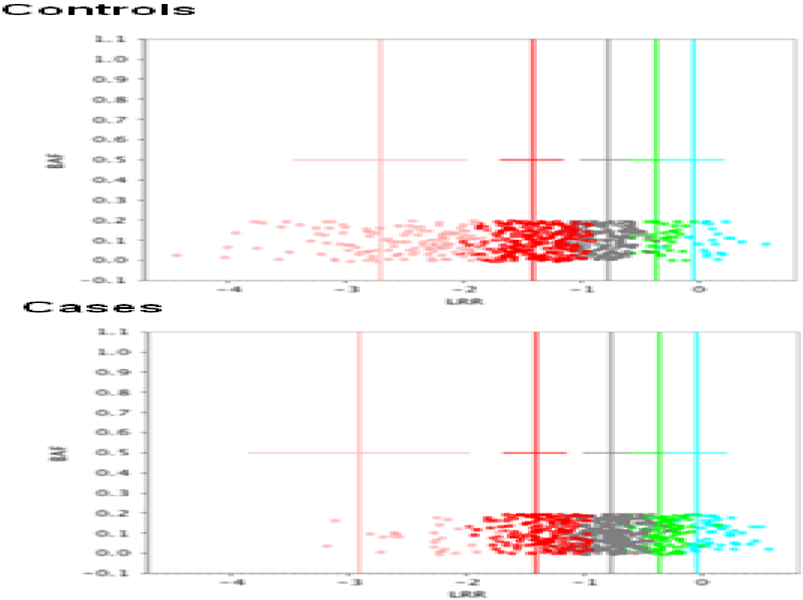
Detecting CNV associations



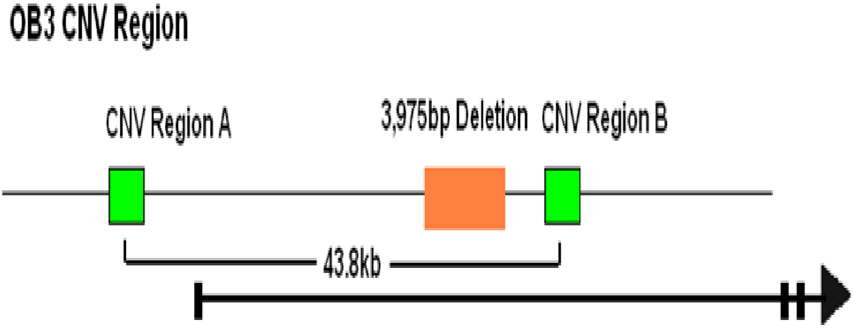
Discovery and validation of CNV association



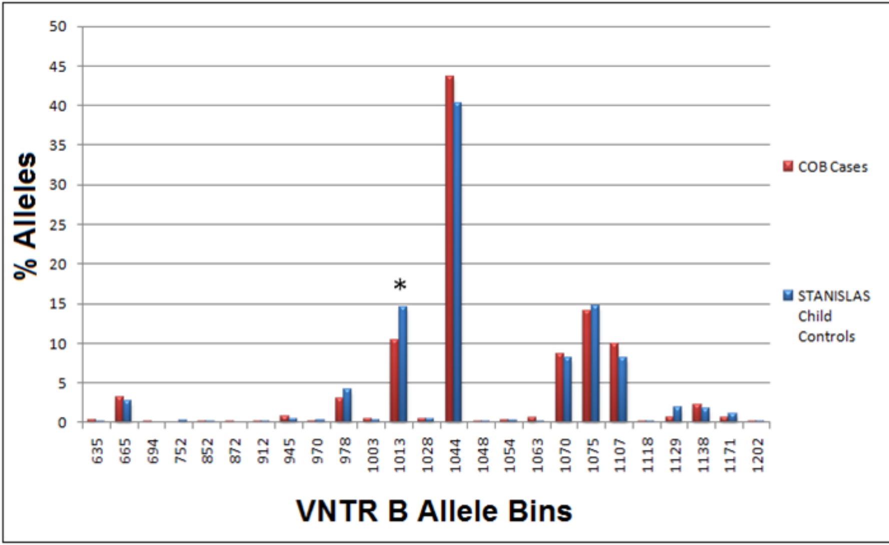
Discovery and validation of CNV association



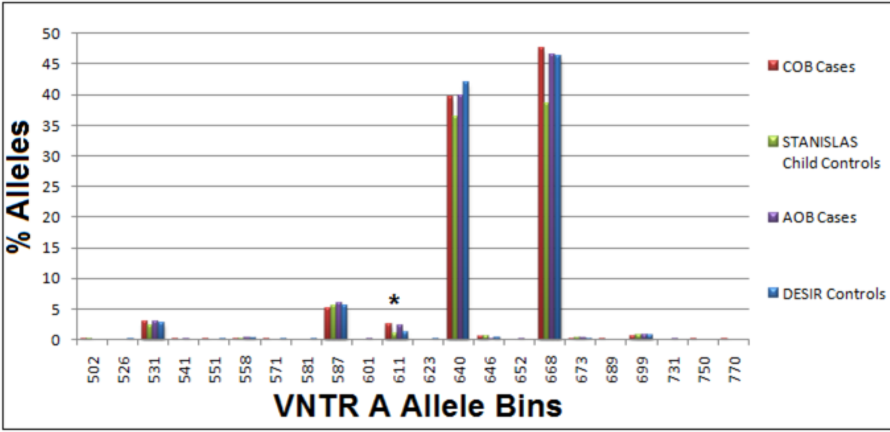
'OB3' CNV region



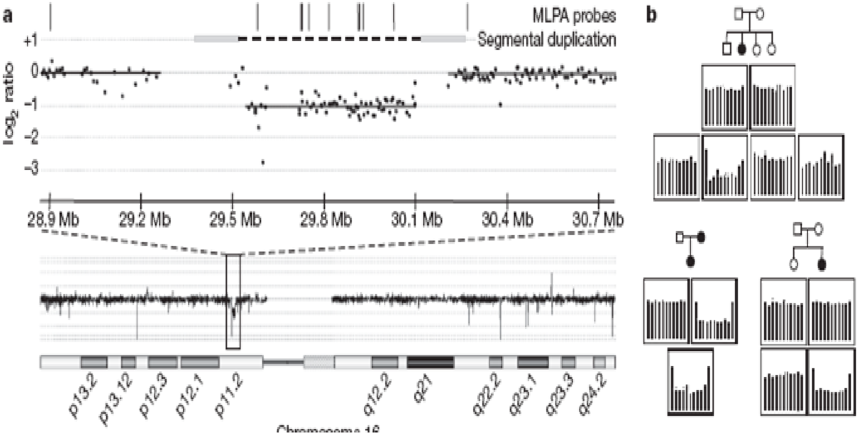
Discovery and validation of CNV association



Discovery and validation of CNV association



Identification and validation of deletions at 16p11.2



Walters et al,Nature 2010

Frequency of detected 16p11.2 deletions in multiple cohorts

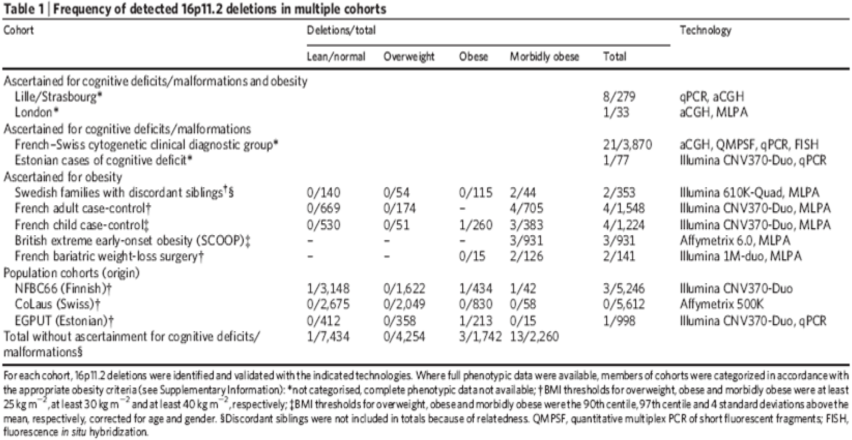


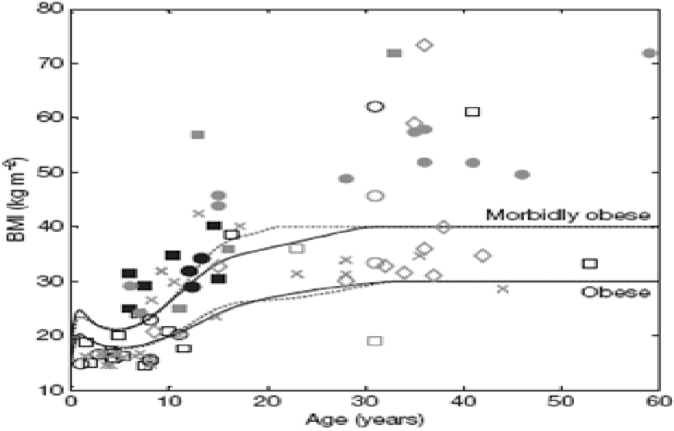
Table 1 | Frequency of detected 16p11.2 deletions in multiple cohorts

Cohort	Deletions/total					Technology
	Lean/normal	Overweight	Obese	Morbidly obese	Total	
Ascertained for cognitive deficits/malformations and obesity						
Lille/Strasbourg*					8/279	qPCR, aCGH
London*					1/33	aCGH, MLPA
Ascertained for cognitive deficits/malformations						
French–Swiss cytogenetic clinical diagnostic group*					21/3,870	aCGH, QMPSF, qPCR, FISH
Estonian cases of cognitive deficit*					1/77	Illumina CNV370-Duo, qPCR
Ascertained for obesity						
Swedish families with discordant siblings†§	0/140	0/54	0/115	2/44	2/353	Illumina 610K-Quad, MLPA
French adult case-control†	0/669	0/174	-	4/705	4/1,548	Illumina CNV370-Duo, MLPA
French child case-control‡	0/530	0/51	1/260	3/383	4/1,224	Illumina CNV370-Duo, MLPA
British extreme early-onset obesity (SCOOP)‡	-	-	-	3/931	3/931	Affymetrix 6.0, MLPA
French bariatric weight-loss surgery†	-	-	0/15	2/126	2/141	Illumina 1M-duo, MLPA
Population cohorts (origin)						
NFBC66 (Finnish)†	1/3,148	0/1,622	1/434	1/42	3/5,246	Illumina CNV370-Duo
CoLaus (Swiss)†	0/2,675	0/2,049	0/830	0/58	0/5,612	Affymetrix 500K
EGPUT (Estonian)†	0/412	0/358	1/213	0/15	1/998	Illumina CNV370-Duo, qPCR
Total without ascertainment for cognitive deficits/malformations§	1/7,434	0/4,254	3/1,742	13/2,260		

For each cohort, 16p11.2 deletions were identified and validated with the indicated technologies. Where full phenotypic data were available, members of cohorts were categorized in accordance with the appropriate obesity criteria (see Supplementary Information): *not categorised, complete phenotypic data not available; † BMI thresholds for overweight, obese and morbidly obese were at least 25 kg m^{-2}, at least 30 kg m^{-2} and at least 40 kg m^{-2}, respectively; ‡BMI thresholds for overweight, obese and morbidly obese were the 90th centile, 97th centile and 4 standard deviations above the mean, respectively, corrected for age and gender. §Discordant siblings were not included in totals because of relatedness. QMPSF, quantitative multiplex PCR of short fluorescent fragments; FISH, fluorescence *in situ* hybridization.

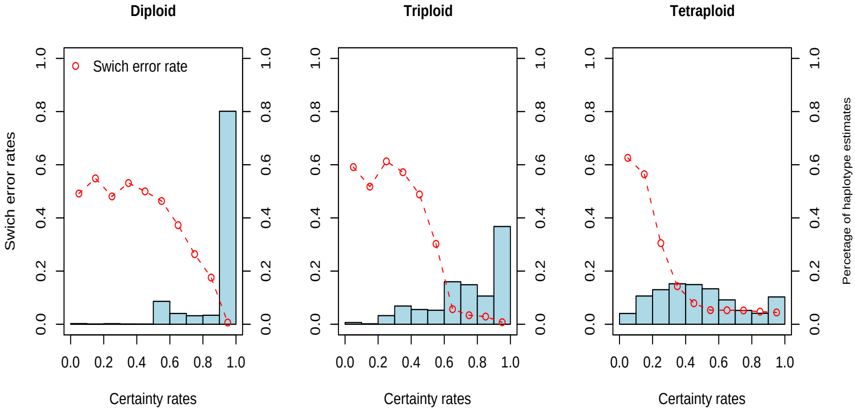
Walters et al,Nature 2010

Dependence of BMI on age in subjects having a deletion at 16p11.2



Lines denote the thresholds corrected for age and gender (solid, male; broken, female) for obesity and morbid obesity. Squares, male; circles, female; black, ascertained for developmental delay; grey, not ascertained for developmental delay; filled, ascertained for obesity; open, not ascertained for obesity; diamonds, first-degree relative

Phasing accuracy on paired male X chromosomes



Determining alleleic configuration of CNVS

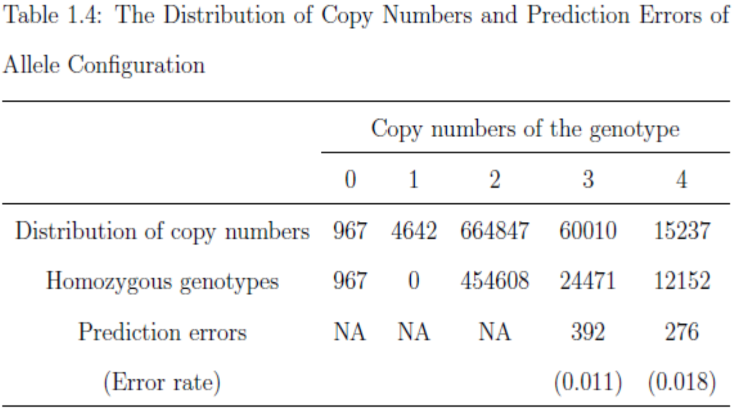
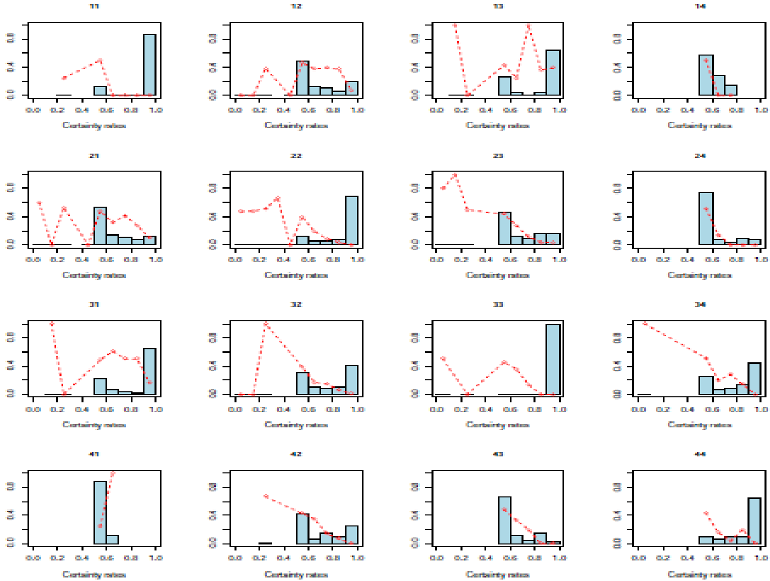


Table 1.4: The Distribution of Copy Numbers and Prediction Errors of Allele Configuration

	Copy numbers of the genotype				
	0	1	2	3	4
Distribution of copy numbers	967	4642	664847	60010	15237
Homozygous genotypes	967	0	454608	24471	12152
Prediction errors	NA	NA	NA	392	276
(Error rate)				(0.011)	(0.018)

Phasing CNVS



Comparison to CNVPhaser

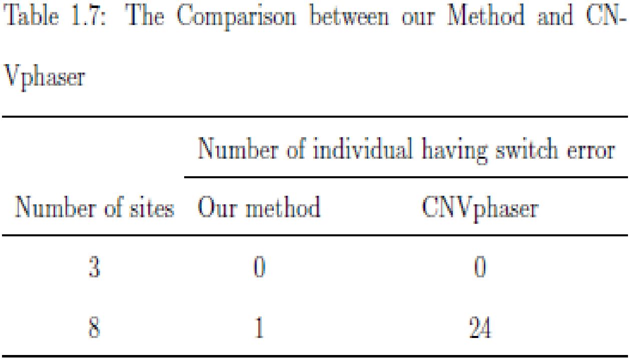


Table 1.7: The Comparison between our Method and CNVphaser

	Number of individual having switch error	
Number of sites	Our method	CNVphaser
3	0	0
8	1	24

Methods for detecting inversions

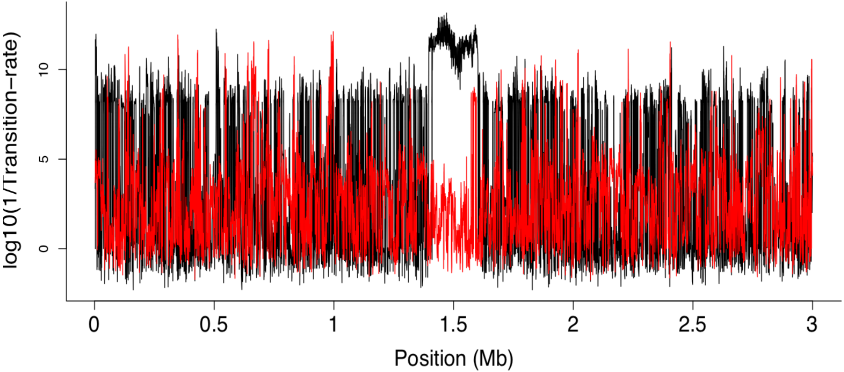
1. Sequencing — '1000 Genomes' Project (only certain pop's)
2. Aberant long range LD patterns — Bansal et al. (low power)
3. Suppression of recombination between inverted and non-inverted chromosomes.

Developed *invert*HMM to capture point 3:

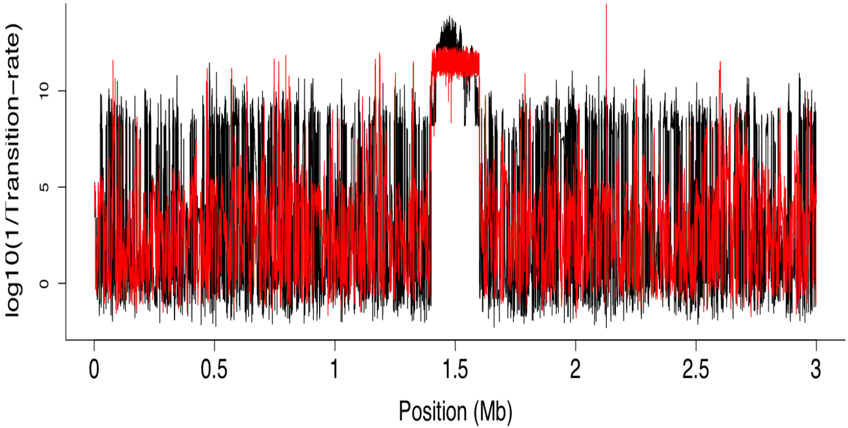
1. Use two hidden 'super' states to model inverted vs non-inverted haplotype ⇒ allows us to model recombination rate between inversion/ non-inversion
2. Use two hidden 'sub' states within each super-state to model underlying rate of recombination
3. Regions with low 'between' and 'within' super-state recombination are just regions of low recombination
4. Regions with low 'between' but normal 'within' are inversion candidates
5. We then predict from model which samples have inversion

*invert*HMM: Simulation with 200kb, 60% inversion

BLACK: Between-superstate transition rate (reciprocal, log-scale);
RED: Between-substate transition rate (reciprocal, log-scale)

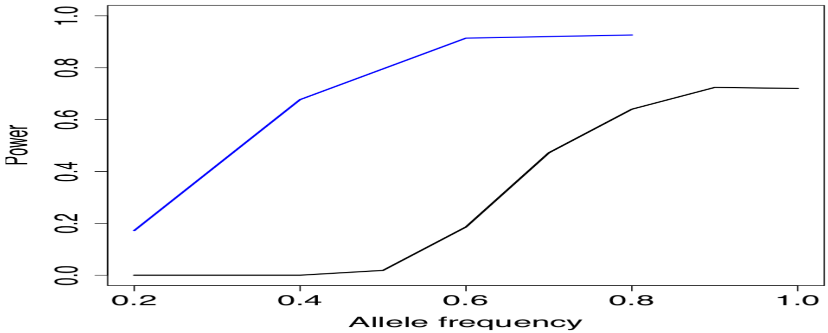


*invert*HMM: Simulation with 200kb of no recombination



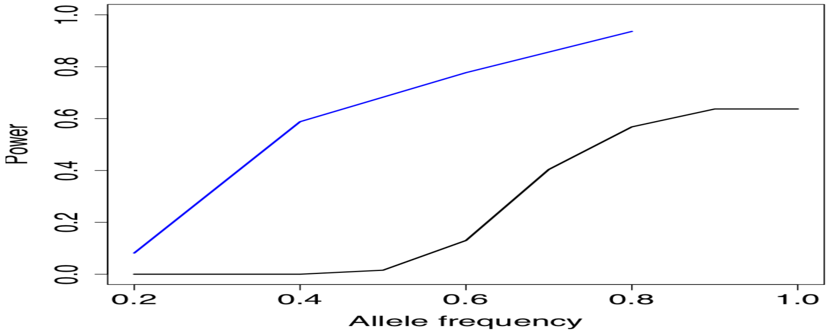
*invert*HMM: Power analysis (500kb inversion)

BLACK: LD method; BLUE: *invert*HMM



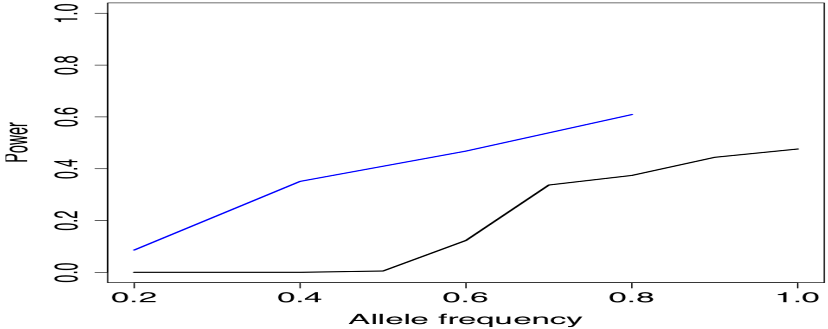
*invert*HMM: Power analysis (200kb inversion)

BLACK: LD method; BLUE: *invert*HMM



*invert*HMM: Power analysis (100kb inversion)

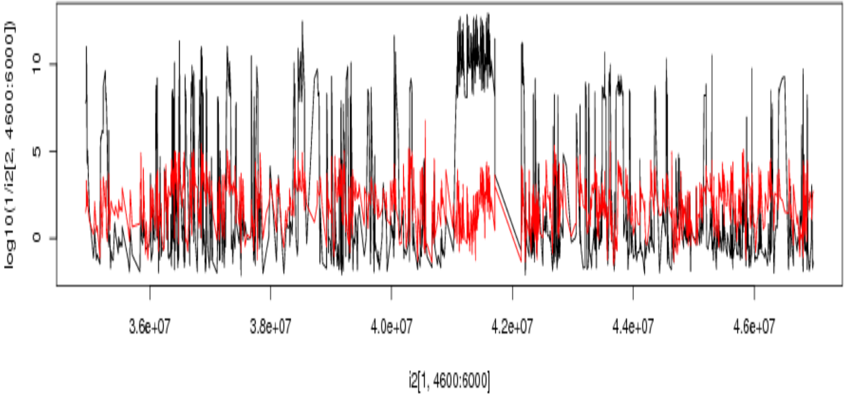
BLACK: LD method; BLUE: *invert*HMM



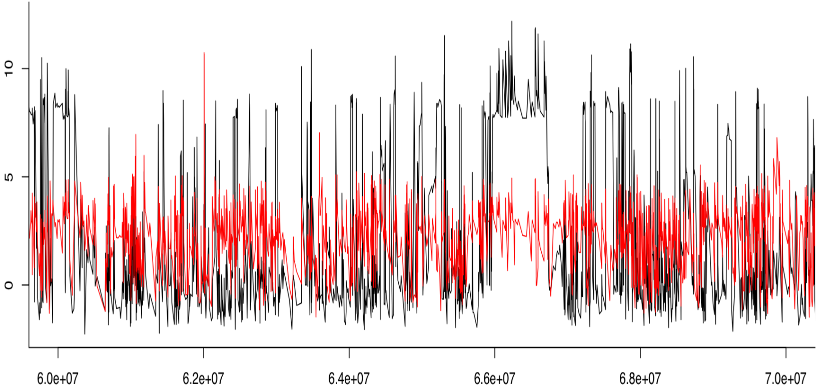
*invert*HMM: Applied to real data

- Scans over WTCCC & French data provided almost 400 candidates genome-wide
- Null distribution formed using a complex model incorporating demographic factors, and variation in recombination rate, calibrated to reflect real data (Schaffner et al. 2005)
- The method applied to the null data suggests just over half these candidates are real inversions (though results indicate null is too conservative here)

MAPT inversion, at ≈ 20% (chromosome 17)



Potential novel inversion



Acknowledgements

Imperial College

Paul O'Reilly
Shu-Yi Su
Clive Hoggart
Julian Asher
Penny Charoen
Harrieta Eleftherohorinou
Adam de Smith
Robin Walters
Alex Blakemore
David Balding
Phillipe Froguel
McGill University

Rob Sladek