Section 1: Supplementary Notes for "Origins and functional impact of copy number variation in the human genome"

Donald F. Conrad[1], Dalila Pinto[2], Richard Redon[1,3], Lars Feuk[2,4], Omer Gokcumen[5], Yujun Zhang[1], Jan Aerts[1], T. Daniel Andrews[1], Chris Barnes[1], Peter Campbell[1], Tomas Fitzgerald[1], Min Hu[1], Chun Hwa Ihm[5], Kati Kristiansson[1], Daniel MacArthur[1] Jeff MacDonald[2], Ifejinelo Onyiah[1], Andy Wing Chun Pang[2], Sam Robson[1], Armand Valsesia[1], Klaudia Walter[1], John Wei[2], Wellcome Trust Case Control Consortium, Chris Tyler-Smith[1], Nigel P. Carter[1], Charles Lee[5], Steve Scherer[2,6], Matthew E. Hurles[1]

[1] Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

[2] The Centre for Applied Genomics and Program in Genetics and Genomic Biology, The Hospital for Sick Children, MaRS CentreEast Tower, 101 College Street, Room 14-701, Toronto, Ontario M5G 1L7, Canada.

[3] Inserm UMR915, L'institut du thorax, Nantes, France

[4] Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden

[5] Department of Pathology, Brigham and Womens Hospital and Harvard Medical School, Boston, MA, USA

[6] Department of Molecular and Medical Genetics, Faculty of Medicine, University of Toronto M5S 1A8, Canada.

September 28, 2009

Contents

List of Figures

List of Tables

1 CNV Genotyping

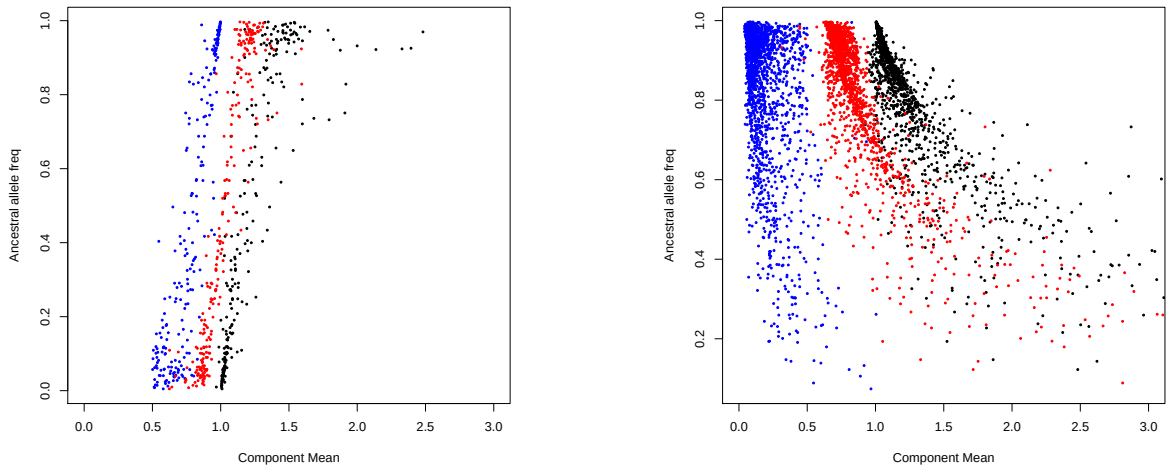
The tables and figures in this section provide supplementary results regarding CNV genotyping.

1.1 Absolute copy number / Ancestral state assignment

As described in Supplementary Methods, we fitted a likelihood-based model of absolute copy number using the intensity data and genotype calls derived from the 105k array, and then identified the maximum-likelihood absolute copy number for each CNV (Figure 1.1, Table 1.1). Independent of this, we assayed 8 chimpanzees with the same 105K genotyping array. Following divergence correction of this chimp data we genotyped the chimps using the identical genotyping models and calling parameters that were used for the human data. In the cases that a monomorphic chimp state could be assigned, it was almost always copy number "2" regardless of the alleles segregating in human, suggesting that most copy number variant alleles are the derived allele. Based on the reasonable agreement between the absolute copy number models and chimp genotypes we use the convention that the non 2-copy allele is the derived allele at each locus.

Model	Number of Loci	Chimp State
0, 1	14	3=0 10=1
1, 2	2221	76=1 1777=2
0, 1, 2	1853	32=0 173=1 1316=2
0, 1, 2,> 2	32	1=0 4=1 12=2 5=3
1, 2, > 2	277	33=1 91=2 13=3 4=4
2, 3	339	426=2 23=3
2, 3, > 3	664	69=2 18=3 59=4 3=5 1=8

Table 1.1: **Tabulation of absolute copy numbers observed at each locus.** "Model" indicates the absolute copy number of clusters parameterized by the model. "Number of Loci" refers to the number of CNV loci assigned to each model. "Chimp State" contains the breakdown of copy number seen at monomorphic chimp sites for each model type. The notation is "number of sites = copy number", so for model 2,3 there were 426 sites in chimp where the estimated ancestral copy number was 2 and 23 sites where it was 3.



(a) Duplications

(b) Deletions

Figure 1.1: **Observed cluster means for 3 component CNVs called as (a) duplications and (b) deletions.** The mean intensity ratio (Cy5/Cy3) for all three clusters at each locus (x-axis) are plotted against the frequency of the ancestral allele at that locus (y-axis). Because the reference DNA is a pool from 10 individuals, the cluster means for each absolute copy number is frequency dependent. These cluster means are colored by their ordered values (blue: lowest copy number component; red, middle; black, highest). These observed results agree quite well with the theoretically expected cluster positions (See Supplementary Methods).

2 Genomic Impact

The tables and figures in this section contain supplementary results pertaining to the genomic impact of CNVs. We further characterize the number and nature of functional elements predicted to be affected by CNVs in our map, and test for biases towards/away from various classes of genomic annotations and specific groupings of genes.

Validated genotyped deletions (N=4,136)	**# Loci (%)**	**# RefSeq Transcripts[1] (%)**	**# RefSeq genes[2] (%)**
Whole gene deletions	79 (1.91%)	272 (0.88%)	213 (1.06%)
Partial gene deletions	1354 (32.74%)	1658 (5.36%)	1159 (5.75%)
- Contained in intron	1142 (27.61%)	1324 (4.28%)	916 (4.54%)
- Overlapping exon(s)	191 (4.62%)	295 (0.95%)	212 (1.05%)
- Full exon (both breakpoints in introns), frameshift	35 (0.85%)	41 (0.13%)	34 (0.17%)
- Full exon (both breakpoints in introns), in frame	23 (0.56%)	38 (0.12%)	22 (0.11%)
- partial-exon (at least one breakpoint in exon), frameshift	8 (0.19%)	10 (0.03%)	8 (0.04%)
- partial-exon (at least one breakpoint in exon), in frame	11 (0.27%)	14 (0.05%)	12 (0.06%)
- Overlapping stop codon	54 (1.31%)	75 (0.24%)	65 (0.32%)
- Overlapping promoter	96 (2.32%)	163 (0.53%)	121 (0.60%)
Intergenic	2703 (65.35%)	-	-

Table 1.2: **Loss of function mutations.** [1] RefSeq transcripts, N=30,917, [2] RefSeq genes, N=20,174.

Dataset	All Valid CNVEs	All Genot CNVEs	Gen. Dels	Gen. Dups	Gen. Multi	Common CNVEs	Common Dels	Common Dups	Rare CNVEs	Rare Dels	Rare Dups
RefSeq genes	0.999	0.999	0.999	0.398	0.780	0.999	0.999	0.847	0.959	0.987	0.233
OMIM	0.999	0.996	0.997	0.669	0.897	0.756	0.802	0.474	0.998	0.999	0.505
AD genes	0.955	0.971	0.920	0.786	0.836	0.326	0.333	0.683	0.896	0.918	0.676
AR genes	0.996	0.924	0.887	0.743	0.961	0.999	0.995	0.999	0.860	0.955	0.362
Disease Dosage Sensitive Genes	0.784	0.599	0.400	0.938	0.726	0.730	0.603	0.999	0.760	0.630	0.894
CancerGenes	0.999	0.999	0.998	0.932	0.879	0.987	0.966	0.999	0.989	0.994	0.835
Pharmacogenetic	0.343	0.328	0.639	0.387	0.125	0.889	0.852	0.999	0.699	0.835	0.540
EnhancerElements	0.999	0.999	0.999	0.976	0.930	0.859	0.999	0.381	0.954	0.851	0.999
UltraConserved	0.999	0.999	0.988	0.999	0.999	0.999	0.999	0.999	0.820	0.624	0.999
miRNAs	0.718	0.956	0.804	0.964	0.874	0.999	0.999	0.999	0.882	0.703	0.999
Imprinting Genes	0.415	0.888	0.952	0.710	0.453	0.999	0.999	0.999	0.740	0.999	0.311
RecombHotspot	0.860	0.043	0.001	0.824	0.993	0.607	0.459	0.864	0.001	0.001	0.262
CpG Islands	0.000	0.507	0.999	0.000	0.014	0.933	0.986	0.236	0.005	0.200	0.000
DNAselhypersensitivity	0.132	0.745	0.999	0.000	0.007	0.963	0.995	0.038	0.006	0.030	0.013
Promoter.plusminus500	0.000	0.005	0.824	0.000	0.000	0.787	0.972	0.031	0.003	0.080	0.001
Stop Codons	0.000	0.102	0.999	0.000	0.000	0.823	0.956	0.183	0.225	0.909	0.004
WSSD (inferred read-depth)	0.000	0.000	0.966	0.000	0.000	0.218	0.999	0.000	0.001	0.311	0.000
GenomicSuperDups (sequence-alignment)	0.000	0.000	0.050	0.000	0.000	0.000	0.887	0.000	0.000	0.406	0.000

significant enrichment, after Bonferroni correction
significant enrichment, before Bonferroni correction
significant impoverishment, after Bonferroni correction
significant impoverishment, before Bonferroni correction

Figure 1.2: **Testing of enrichment/impoverishment of genomic features within CNVs.** Eleven CNV maps (columns) were constructed using subsets of the validated CNVs and each CNV map was intersected with a number of different genomic annotation datasets (rows). Cells contain the proportion of 1000 randomized CNV maps that result in greater annotation overlap than what was observed with the true CNV map. Despite a high level of interdependency between the individual tests, a Bonferroni correction was applied, taking into account that 18 different comparisons were performed with each set of simulated data (corrected p-value = 0.003). 'Gen.' = genotyped; 'Common' = variants $\geq$ 10% MAF; 'Rare' = variants $\leq$ 1% MAF.

2.1 Functional classification of genes intersected by CNV loci

Our analyses reveal that functional categories/biological processes such as cell adhesion activity (categories such as cell adhesion, cell-cell adhesion and homophilic cell adhesion, as well as cell recognition-notably enriched for proteins involved in gamete binding and fusion or fertilization), are found to be significantly enriched in all CNV loci (Figure 1.3). Other functional categories involved in defense and response to external stimulus were also found to be enriched especially in multi-allelic and duplication loci (response to biotic stimulus, response to bacterium, defense to bacterium, antigen processing and presentation of peptide antigen via MHC class II). Genes involved in nervous system development were found to be enriched especially in common loci.

Categories found to be significantly impoverished were all related to regulation of biological or cellular processes as well as primary metabolic processes (cellular metabolic process, macromolecule metabolic processes, primary metabolic processes, regulation of biological processes, RNA metabolic processes, and nucleobase, nucleoside and nucleotide and nucleic acid metabolic process).

Interestingly we found enrichment for genes in cell communication (cell-communication), as well as in signaling (phosphorus metabolic process and regulation of Ras protein signal transduction), opposed to previous results (Redon, et al. 06). These functional categories include members involved in cell-adhesion mediated signaling and G-protein mediated signaling, but also include tyrosine kinases, serine/threonine kinases, and phosphatases, that collectively participate in a variety of physiological and developmental processes such as cell growth, proliferation and differentiation [includes members of the Ras, FGF, EGF, PDGF signaling]. Given that many of these processes are known to be involved in medically relevant conditions, understanding the population genetics of associated-CNV loci is likely important to understand disease.

KEGG pathways found to significantly enriched in CNVs were metabolism of drugs, toxicity and xenobiotics (metabolism of xenobiotics by cytochrome P450, porphyrin and chrolophyll metabolism, and androgen and estrogen metabolism), and sugar metabolism (pentose and glucuronate interconversions and starch and sucrose metabolism). However, these observations are driven largely by one large gene cluster, the uridine diphosphate (UDP)-glucuronosyltransferase loci, where a few CNVs overlap multiple distinct genes that are functionally related.

Loci categories	Index	Term	Representation	List vs. background	p value	Adjusted p value (2-tail)[1]
Initial finding	**GO**	**biological process at level 3**				
Validated	0	cellular metabolic process (GO 0044237)	under	46.69% 53.31%	1.66E-06	**1.26E-04**
Validated	0	cell adhesion (GO 0007155)	over	61.7% 38.3%	6.99E-06	**2.66E-04**
Validated	0	macromolecule metabolic process (GO 0043170)	under	46.48% 53.52%	1.87E-05	**4.74E-04**
Validated	0	primary metabolic process (GO 0044238)	under	47.04% 52.96%	2.74E-05	**5.20E-04**
Validated	0	regulation of biological process (GO 0050789)	under	46.48% 53.52%	1.52E-03	**2.31E-02**
Validated	0	cell recognition (GO 0008037)	*over*	73.49% 26.51%	5.05E-03	*6.40E-02*
Validated	0	biosynthetic process (GO 0009058)	*under*	43.51% 56.49%	7.08E-03	*7.68E-02*
Validated	0	cell communication (GO 0007154)	*over*	52.81% 47.19%	9.77E-03	*9.28E-02*
Dups, Common	0	antigen processing and presentation (GO 0019882)	over	81.46% 18.54%	3.32E-04	**2.53E-02**
Multi	0	response to biotic stimulus (GO 0009607)	*over*	76.19% 23.81%	2.27E-03	*8.36E-02*
Multi	0	establishment of localization (GO 0051234)	*under*	29.19% 70.81%	3.30E-03	*8.36E-02*
Initial finding	**GO**	**biological process at level 4**				
Validated	1	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (GO 0006139)	under	42.78% 57.22%	7.15E-08	**1.54E-05**
Validated	1	cell-cell adhesion (GO 0016337)	over	65.98% 34.02%	1.20E-04	**1.18E-02**
Validated	1	regulation of metabolic process (GO 0019222)	under	43.88% 56.12%	1.64E-04	**1.18E-02**
Validated	1	regulation of cellular process (GO 0050794)	under	46.01% 53.99%	8.66E-04	**4.65E-02**
Dels	1	phosphorus metabolic process (GO 0006793)	*over*	60.34% 39.66%	6.92E-04	*7.44E-02*
Dups, Common	1	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II (GO 0002504)	*over*	87.03% 12.97%	1.32E-03	*9.47E-02*
Initial finding	**GO**	**biological process at level 5**				
Validated	2	homophilic cell adhesion (GO 0007156)	over	74.15% 25.85%	1.32E-06	**5.58E-04**
Validated	2	RNA metabolic process (GO 0016070)	under	43.68% 56.32%	8.19E-05	**1.73E-02**
Validated	2	regulation of cellular metabolic process (GO 0031323)	*under*	44.18% 55.82%	4.29E-04	*6.04E-02*
Multi	2	response to bacterium (GO 0009617)	over	90.69% 9.31%	2.51E-06	**1.06E-03**
Common	2	nervous system development (GO 0007399)	*over*	69.01% 30.99%	3.95E-04	*8.33E-02*
Initial finding	**GO**	**biological process at level 6**				
Multi	3	defense response to bacterium (GO 0042742)	over	90.69% 9.31%	1.06E-05	**6.66E-03**
Initial finding	**GO**	**biological process at level 8**				
Validated	5	regulation of Ras protein signal transduction (GO 0046578)	*over*	69.4% 30.6%	1.07E-04	*8.81E-02*
Initial finding	**KEGG**	**pathway**	**Representation**	**List vs. background**	**p value**	**Adjusted p value (2-tail)**
Validated	7	Pentose and glucuronate interconversions (hsa00040)	over	95.49% 4.51%	2.69E-07	**3.18E-05**
Validated	7	Starch and sucrose metabolism (hsa00500)	over	92% 8%	3.30E-07	**3.18E-05**
Validated	7	Porphyrin and chlorophyll metabolism (hsa00860)	over	91.37% 8.63%	6.55E-06	**4.21E-04**
Validated	7	Metabolism of xenobiotics by cytochrome P450 (hsa00980)	over	84.32% 15.68%	2.23E-05	**1.08E-03**
Validated	7	Androgen and estrogen metabolism (hsa00150)	over	86.7% 13.3%	8.81E-05	**3.40E-03**

[1] Statistically significant categories (FDR p <0.05) are highlighted in bold; categories with a 0.1< FDR p <0.05 are given in italic.

Figure 1.3: **Over- or under-representation of Gene Ontology categories (GO) and KEGG pathways.** This table lists all significant GO categories or KEGG pathways found to be enriched or impoverished for genes intersected by CNV (Supplementary Methods). 'Common' = variants $\geq$ 10% MAF.

Chr	Start	End	CNV	Left Gene (strand)	Right Gene (strand)
1	19472219	19487114	CNVR99.1	AKR7L-	AKR7A3-
1	103905687	103962929	CNVR266.8	AMY2B+	AMY2A+
1	110017689	110060631	CNVR299.5	GSTM2+	GSTM5+
1	146041404	146063219	CNVR345.1	NBPF11-	LOC728912-
1	146841681	146862201	CNVR348.4	NBPF15+	NBPF16+
1	151026773	151036902	CNVR360.1	LCE1E+	LCE1D+
1	153493683	153529298	CNVR370.1	SCAMP3-	PKLR-
1	159748409	159912837	CNVR383.1	FCGR2A+	FCGR2B+
1	195147464	195181215	CNVR459.2	CFHR4+	CFHR2+
2	86825725	87027632	CNVR859.2	RMND5A+	RGPD1+
3	49695755	49709630	CNVR1377.1	APEH+	RNF123+
5	732013	900681	CNVR2280.3	TPPP-	ZDHHC11-
5	140195897	140202238	CNVR2621.1	PCDHA1-PCDHA7+	PCDHA8+
5	140203353	140218974	CNVR2622.1	PCDHA1-PCDHA8+	PCDHA9+—PCDHA10+
5	140534809	140539116	CNVR2624.1	PCDHB7+	PCDHB8+
5	180309761	180362669	CNVR2719.1	BTNL8+	BTNL3+
6	32055886	32060381	CNVR2843.4	STK19+	C4B+
6	32056216	32114549	CNVR2843.6	STK19+	CYP21A2+
6	32058051	32114339	CNVR2843.6	C4B+	CYP21A2+
6	32066995	32093119	CNVR2843.3	C4B+	C4A+
7	74953470	74982353	CNVR3457.2	POM121C-	PMS2L3-
7	75898835	75982664	CNVR3462.1	ZP3+	UPK3B+
7	99647313	99775283	CNVR3508.2	STAG3+	PILRB+
10	46385159	46593846	CNVR4712.4	SYT15-	ANXA8-—ANXA8L1-
11	4924207	4932830	CNVR5040.1	OR51A4-	OR51A2-
11	71176703	71225602	CNVR5223.1	FAM86C+	DEFB108B+
12	10462673	10490066	CNVR5430.1	KLRC3-	KLRC1-
12	11396088	11435791	CNVR5435.2	PRB1-	PRB2-
12	132227275	132289498	CNVR5791.2	ZNF10+	ZNF268+
14	49171648	50012277	CNVR6152.2	C14orf104-	MAP4K5-
16	162918	167514	CNVR6569.1	HBA2+	HBA1+
16	2589819	2676116	CNVR6602.2	PDPK1+	KCTD5+
16	86447257	86517872	CNVR6851.1	SLC7A5-	CA5A-
17	24093092	24095700	CNVR7048.1	NEK8+	TRAF4+
17	33529260	33551942	CNVR7083.1	TBC1D3F-	TBC1D3E-—TBC1D3-
17	36636477	36649035	CNVR7095.1	KRTAP9-2+	KRTAP9-9+
17	36760531	36779150	CNVR7097.1	KRT33A-	KRT33B-
17	41521114	41789790	CNVR7114.3	KIAA1267-	ARL17-
17	41758316	42139813	CNVR7114.2	LRRC37A+	NSF+
17	41969226	42142846	CNVR7114.4	LRRC37A2+	NSF+
19	10350139	10366017	CNVR7534.1	TYK2-	CDC37-
19	48393880	48454467	CNVR7658.1	PSG4-	PSG9-
19	56823616	56840725	CNVR7702.1	SIGLEC5-	SIGLEC14-
19	59417076	59437785	CNVR7721.3	LILRB3-	LILRA6-
19	59940398	59949330	CNVR7725.4	KIR3DP1+	KIR2DL3+
19	60025725	60068860	CNVR7725.6	KIR3DL1+	KIR3DL2+
19	60046589	60064493	CNVR7725.5	KIR2DS4+	KIR3DL2+
21	10009191	10105624	CNVR7947.1	TPTE-	BAGE4-—BAGE-
22	37686073	37718472	CNVR8164.1	APOBEC3A+	APOBEC3B+
22	41232271	41284710	CNVR8171.2	SERHL+	SERHL2+
23	49030760	49129879	CNVR8331.1	PPP1R3F+	GAGE2A/B+—GAGE8+
23	49060021	49129879	CNVR8331.1	GAGE10+	GAGE2A/B+—GAGE8+
23	49180859	49256599	CNVR8332.1	GAGE8+GAGE2A/B	GAGE1+
23	134676299	134718802	CNVR8449.3	CT45-1+	CT45-4+—CT45-3+
23	134760888	134796563	CNVR8449.1	CT45-4-	CT45-6-
23	153953046	154098786	CNVR8495.1	BRCC3+	VBP1+

Table 1.3: **CNVs potentially producing fusion genes.**

Validation. Five candidate gene fusions were selected from the list of candidate gene fusions in the validated CNVs, prioritized by call frequency in the discovery data. These involved the genes *BTNL3/BTNL8*, *LCE1E/LCE1D*, *PCDHB7/8*, *AKR7L/AKR7A3*, *SIGLEC5/14*.

Breakpoints for four of the deletions were cloned by PCR:

- *AKR7L/AKR7A3* deletion: chr1:19486821-19472223, covering from the middle of the first intron of *AKR7L*, to the first intron of *AKR7A3*. Size: 14599bp
- *BTNL8* deletion: chr5:180307813-180363315, covering from fourth intron of *BTNL8*, to the fourth intron of *BTNL3*. Size: 55500bp.
- *LCE* deletion: chr1:151026967-151037452, covering from the 3 UTR of *LCE1E*, to the whole transcribed region of *LCE1D*. Size: 10486bp
- *SIGLEC5* deletion: chr19:56824371-56840789, covering from the third intron of *SIGLEC14*, to the third intron of *SIGLEC5*. Size: 16419bp.

Total RNAs from lymphoblastoid cells from samples predicted to have the fusion transcript (NA10851, NA11993, NA12006, NA18511, NA18523, NA18858) were first reverse transcribed into cDNA, then 3 RACE was used to check the possible fusion transcript. The gene *ACTB* was used as a control in each experiment, as well as a control RNA contained in the RACE kit.

Only RNA for the *BTNL8/BTNL3* fusion can be detected, while amplification for the control gene was detected in each case, suggesting that the experimental protocol should be fine. It appears that these genes are not normally expressed in lymphoblastoid cell lines or that the deletion had abolished the active transcription of the affected genes.

3 Mutation Mechanisms

In this section we present supplementary results pertaining to our characterization of CNV mutation mechanisms. Specifically, we describe detailed results of analyses of the relative contribution of different mutation processes, the effect of GC and ascertainment on the patterns of CNV around genes, analyses to detect dispersed duplications and analyses to discover DNA motifs enriched at CNV breakpoints. We also describe validation experiments for some VNTRs and putative dispersed duplications.

3.1 VNTRs

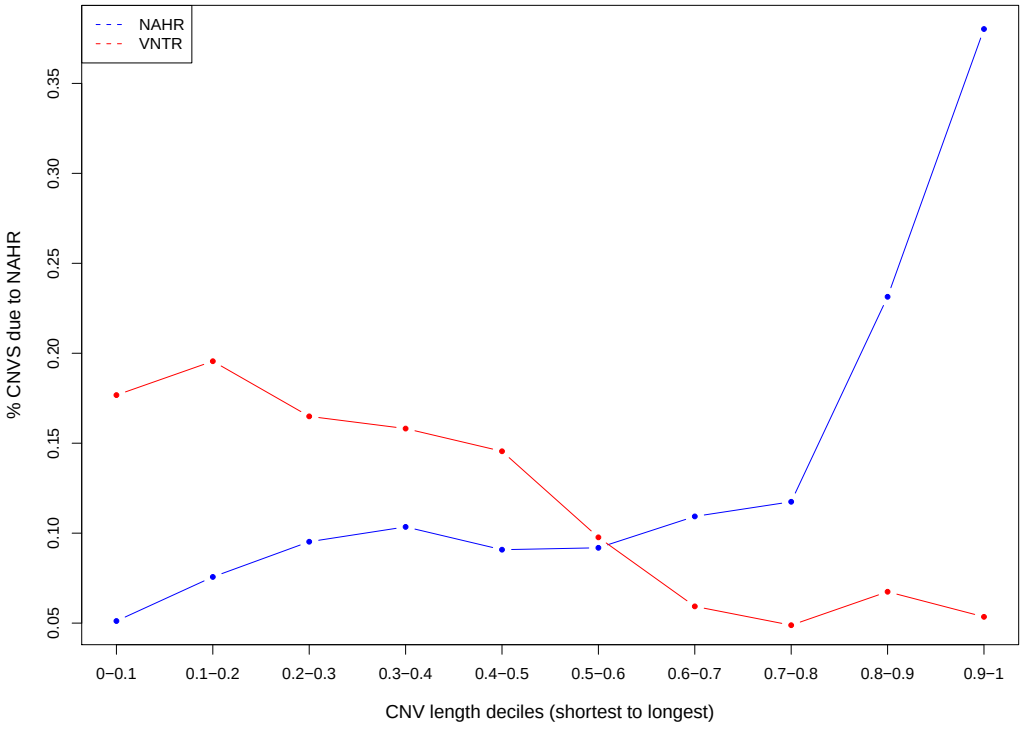


Figure 1.4: **Prevalence of mutation processes depend on size.** The proportion of events corresponding to NAHR and VNTR are plotted as a function of deciles of CNV length.

PCR Validation. We attempted PCR on 30 VNTRs shorter than 3kb, using a subset of 14 samples from the discovery panel. The experiments used a touch-down protocol to reduce non-specific amplification and 12 minute extensions. Normally, the size limits of bands should be about 10kb. PCR product was obtained for 29 of 30 reactions but interpretation at some loci was confounded by the presence of multiple bands. For 12 loci only 1 or 2 amplified bands per individual could be seen, and at 11 of these clear polymorphism was observed (primers listed in Table 1.4).

CNV	chr	start	end	length	orientation	primer Seq
CNVR3256.1	7	3969047	3970022	975	F	TCGATGGCTTCATCTACAACGAGAAAAATA
					R	GAAAGGGCAAACACAAAAGGTATAATTCCA
CNVR2399.1	5	23938693	23941114	2421	F	TCACAGAACCAATCCCGTATTGTGATTTAC
					R	AAGGCAAGCAATAAATGCAACTCTGAAAAT
CNVR4095.1	8	143061501	143064148	2647	F	AGGAGAGGCTAAAAATGCAGGTAAGAATCC
					R	AGGTTAAGTAACAGCCTCACTCCACAATCC
CNVR7103.1	17	37743451	37744141	690	F	AACTCCAGAGCAGGAACTTCTTAGAAACCA
					R	TATAAAACCCTCAAGAGTCAAGGAGGCAAG
CNVR6900.1	17	84932	85782	850	F	ACATTTTACAATCCACAACTGCACCTCATT
					R	CATTCAAGCCCAGAGGTAGGACATTAGAGT
CNVR3721.3	8	1824411	1825771	1360	F	TGCACTCTGAGAACTTACTGTAGCGATGAA
					R	AAGGTCAAACCTCTTTCCTCTGATATGTGC
CNVR8042.1	21	44653356	44655429	2073	F	TCTTCTTTCAAGAAATGTCTCTCCCTCTCC
					R	GCAACCAGTTTGGAAAGGAACAAGTAAAAC
CNVR6556.1	15	99488278	99488817	539	F	ACCACTGCTTACCCATTATGAGGGCTACTA
					R	AAACACTGGGGATTACAATTTGACATGAGA
CNVR1808.1	4	7970790	7971720	930	F	ACCTCTCACTACAACTCAACACCAGCTCTC
					R	GATAAGAACTCAATCGGCTCAGAAATGTCA
CNVR1230.1	2	242158351	242159398	1047	F	AAGCCTCTCTTTTCTCGCTCACGTATAAGA
					R	CCTCAATCTCTCGAGATGCTAAAGCTAAGG
CNVR8056.1	21	46142655	46143315	660	F	ACTTTCAGAGCATGTGTTTCTATCCTCTGG
					R	CAGGAAACTCTGGAAGAAATGTGTGAAAAA
CNVR3654.2	7	157569516	157571881	2365	F	ATCCTAAACCTAATTTTTCCCTCACGTGCT
					R	GATGATCCTGTTCATTCGCAGATTCTGTAT

Table 1.4: **PCR primers for VNTR assays.**

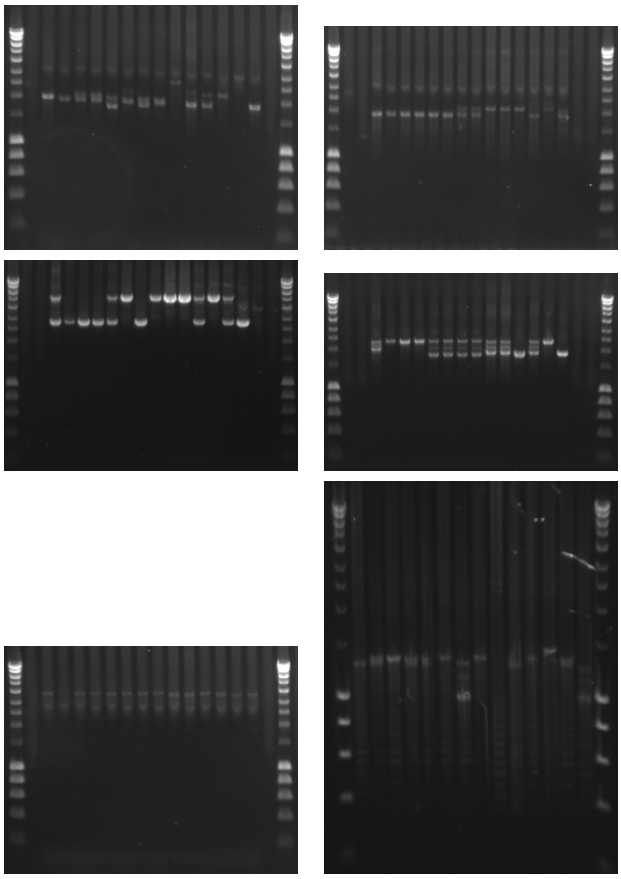


Figure 1.5: **Gels for PCR validaton of VNTRs.** Fourteen samples were used for each PCR, they are left-right: NA10851, NA12891, NA12892, NA12878, NA12044, NA11894, NA12287, NA12489, NA19240, NA18517, NA18858, NA19147, NA19190, NA18916. The loci assayed here are, going row-by-row and left to right, starting at the top of the image, Row1: CNVR1230.1 (Left) , CNVR1808.1 (Right); Row 2: CNVR2399.1 (Left), CNVR3256.1 (Right); Row 3: CNVR3654.2 (L), CNVR3721.3 (R); The marker is Bioline HyperLadder I. Bands were sequenced in pGEM-T easy vector.

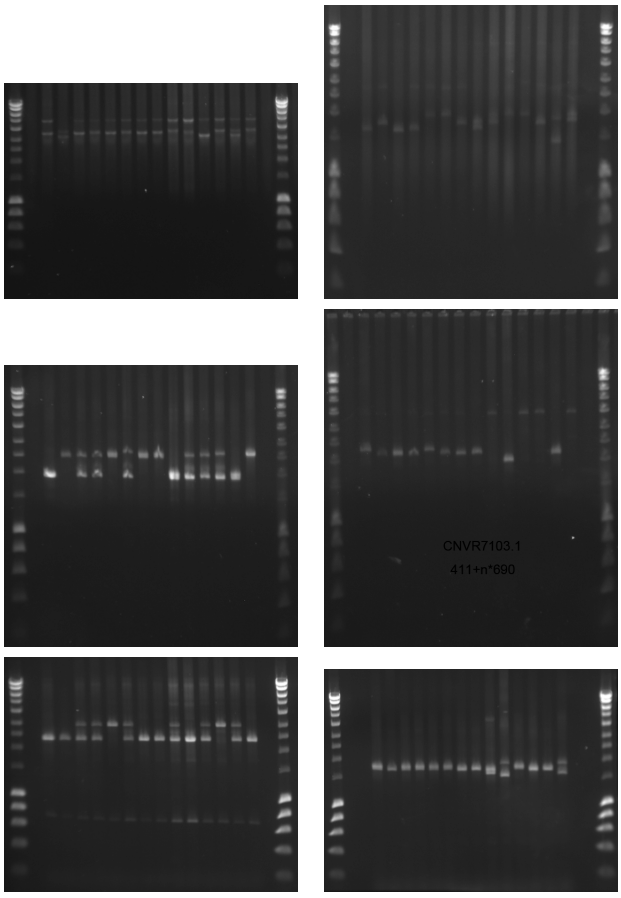


Figure 1.5: **(Continued). Gels for PCR validaton of VNTRs.** Fourteen samples were used for each PCR, they are left-right: NA10851, NA12891, NA12892, NA12878, NA12044, NA11894, NA12287, NA12489, NA19240, NA18517, NA18858, NA19147, NA19190, NA18916. The loci assayed here are, going row-by-row and left to right, starting at the top of the image, Row 1: CNVR4095.1 (L), CNVR6556.1 (R); Row 2: CNVR6900.1 (L), CNVR7103.1 (R); Row 3: CNVR8042.1 (L), CNVR8056.1 (R). The marker is Bioline HyperLadder I. Bands were sequenced in pGEM-T easy vector.

3.2 Sequence Context

Functional sequences	#CNVEs	%CNVEs	#Features	%Features	Total #Features
RefSeq genes	3340‡	37.7	2698	13.4	20174
Promoter region[1]	877*	10.2	1084	4.4	20174
miRNAs	32	0.4	18	4.7	685
CpG islands	775*	9	257	5.2	14867
Ultra-conserved elements	2	0	2	0.4	481
Enhancers	4	0	3	0.5	837
OMIM genes	374‡	4.3	247	8.1	2052
AR/AD genes	154‡	1.8	114	7.1	1613
Dosage sensitive genes[2]	50‡	0.6	32	22.1	145
Pharmaco-genes	45*	0.5	29	24.2	186
Imprinted genes	12	0.1	8	20.3	59
Cancer genes	54‡	0.6	43	14.1	384

[1] promoter region was defined as the region within 500 bp upstream and downstream from a given gene transcription start site (TSS). [2] Baylor's Agilent 105K array disease associated dosage sensitive genes. Statistical significance for the enrichment or paucity of functional sequences intersecting CNVEs was assessed by randomly permuting the genomic location of autosomal CNVEs (Supplementary Methods).
* Significant (P<0.05) enrichment
‡ Significant (P<0.05) paucity

Figure 1.6: **Functional sequences within validated CNVs.**

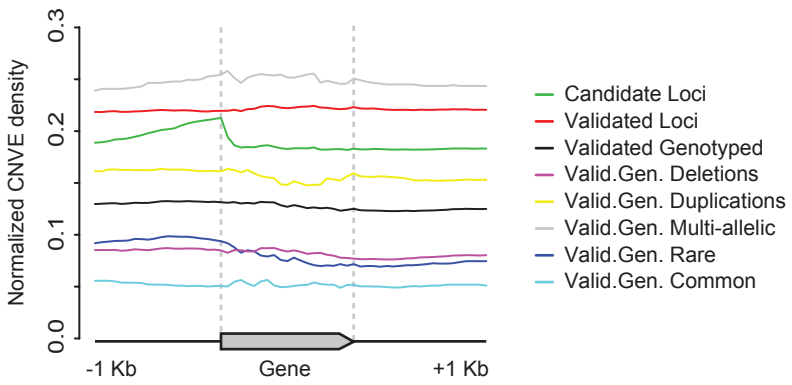


Figure 1.7: **Effect of GC and ascertainment on gene overlap analyses.** RefSeq genes and 1000 bp of 3' and 5' intergenic flanking regions were each divided into 20 bins, and the average number of CNVs overlapping each bin is shown for different CNV subsets. In the initial analysis ("Candidate loci") we found a modest enrichment in the density of CNVs at gene promoters. During validation we noticed that high-GC CNVs were enriched for false positives; as a result our validation criteria requires an average GC-content $< 65\%$. This filtering removes the 5' peak in CNV density near genes. Notes: Genes transcribed in opposite directions were first analyzed separately, and the density values estimated for each bin were then summed across all (plus and minus strand) genes and divided by the number of genes intersected by CNVs in each subset. Candidate Loci.= all variants detected in the discovery phase; Valid.= validated variants; Valid.Gen. = validated high-quality genotyped loci.

We used the program nestedMICA to identify the 20 most over-represented motifs of 6-13bp in size (See Supplementary Methods for full details). NestedMICA returns each motif as a postion weight matrix (PWM), and we plot these PWMs as Shannon Logos here (Figure 1.8). PWMs are ways of representing degeneracy in a motif; instead of representing a consensus sequence the PWM is a $4 \times n$ probability matrix where each column sums to one. Each column in the matrix correspond to a position in the motif, and is represented in a logo as a stack of the 4 DNA bases. At each position in the logo, each base is scaled by it probability of being observed at that position, and the total stack size is scaled by its information content. These logos were plotted with the R package *seqLogo*, written by Oliver Bembom.

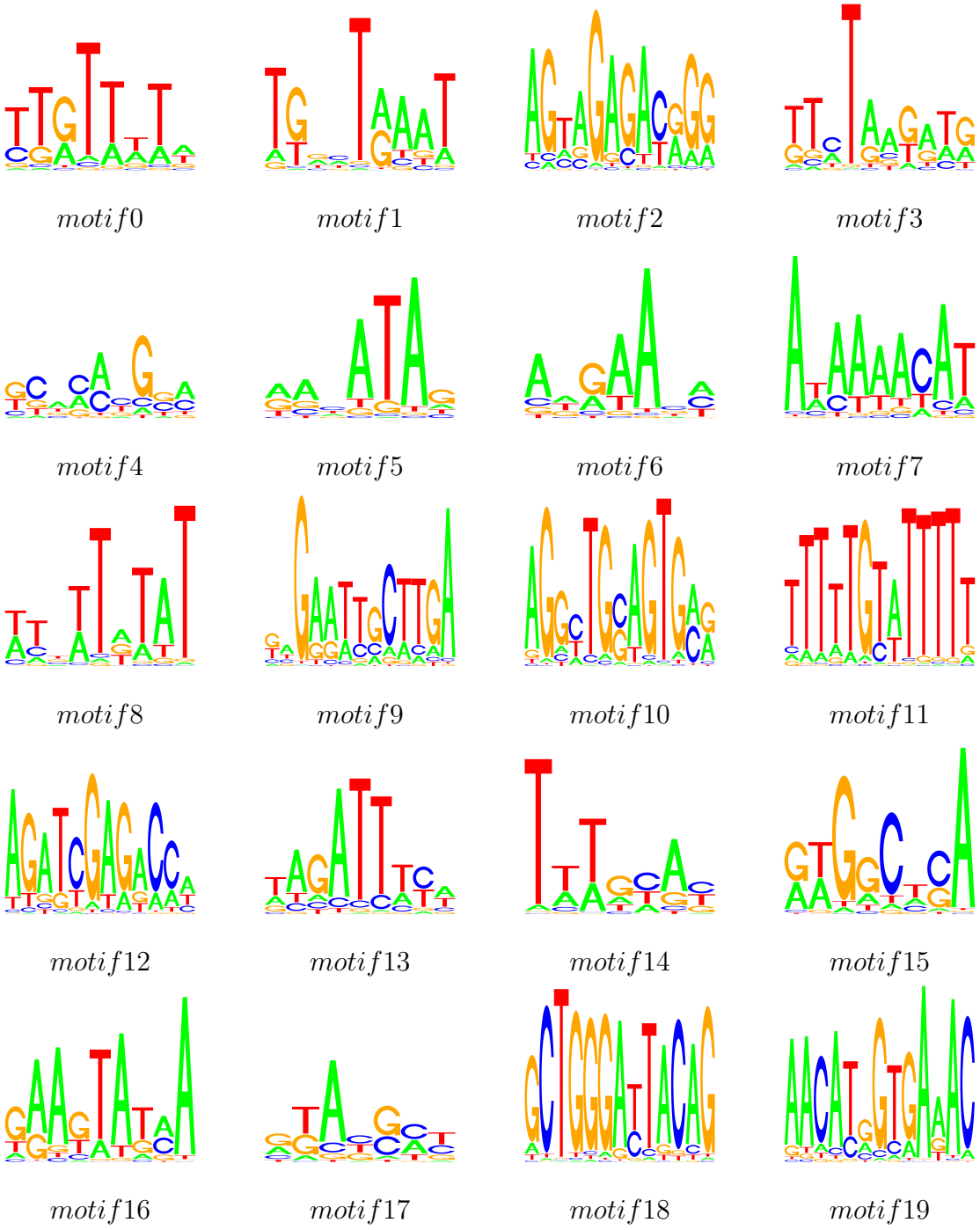


Figure 1.8: **Shannon Logos of 20 motifs identified by nestedMICA.**

3.3 Dispersed Duplications

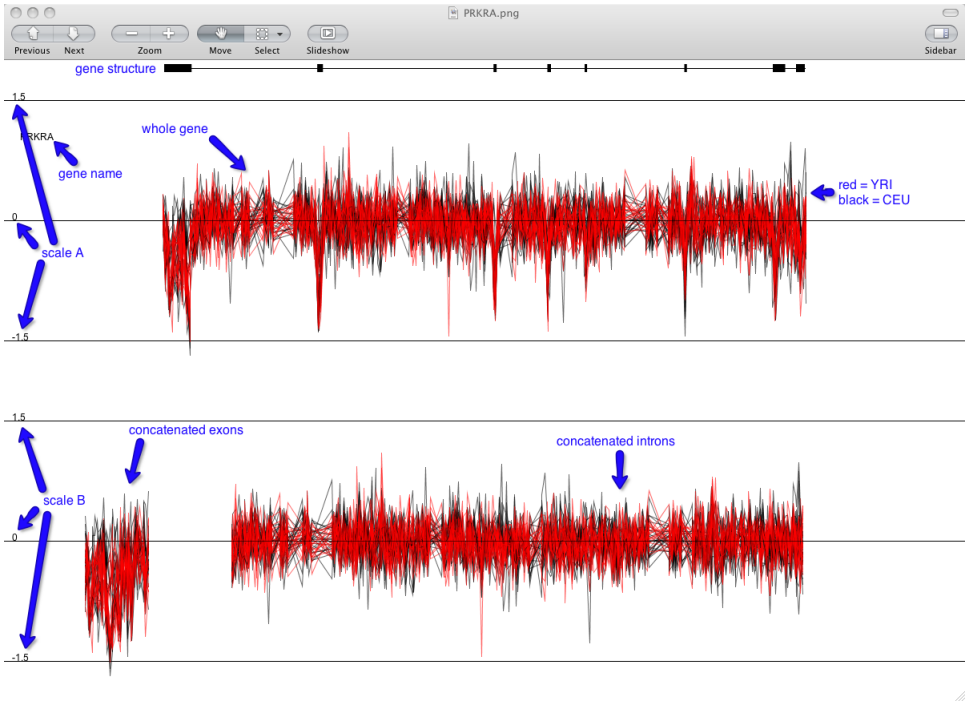


Figure 1.9: **Signal of polymorphic retroposition at the gene PRKRA from** *in silico* **splicing.** The figure is split into three sections: at the very top is the gene model for PRKRA; in the middle the 42M log2 ratios for all probes spanning the gene are plotted separately for each of the 40 discovery samples; at the bottom these same data are parsed into exonic probes (left) and intronic probes (right). In this example there appears to be a duplicative transposition of PRKRA in just the reference individual, which is apparent as a trend for log2 ratios < 0 at exonic but not intronic probes.

Name	Chr	Start	New Chr	New Start	Type	Genes	Code
Affy6_26	6	49539524	3	1364580	Dups	CENPQ:4	L
CNVR1.1	1	499	15	100263879		OR4F5:1	D
CNVR1040.1	2	161843446	4	49206896		0	D
CNVR1065.1	2	179004449	6	32544476	Dups	PRKRA:1	L
CNVR1328.1	3	26407277	1	189850097	Multi	0	L
CNVR1345.1	3	34224641	15	42839948	Dups	0	L
CNVR1372.1	3	46758249	20	16689497	Dups	TESSP5:1	L
CNVR1508.1	3	121780988	6	142289366		0	S
CNVR1532.3	3	129861994	2	168565491	Multi	0	L
CNVR1663.1	3	188066795	20	3961316		0	S
CNVR1770.1	4	3437802	11	62808519	Dups	DOK7:0	L
CNVR1778.1	4	4618969	5	130360762		0	S
CNVR1831.1	4	15215299	13	39522931	Dups	FBXL5:1	L
CNVR1832.1	4	15236048	13	39522931	Dups	FBXL5:1	L

Table 1.5: **CNVs with evidence of dispersed duplication.** Type: deletion/duplication/multiallelic, only coded for genotyped loci. Code:"L", interchromosomal LD; "D", overlapping inter-chromosomal segmental duplication; "C", validated germline event from Cancer Genome Project; "S", sequence underlying CNV shows poly-A tail and target site duplications. See Methods for full details.

Name	Chr	Start	New Chr	New Start	Type	Genes	Code
CNVR2.2	1	340394	15	100268630		OR4F16:1	D
CNVR2151.1	4	166377398	11	57882483	Dups	KLHL2:0	L
CNVR2479.1	5	61273372	18	6019122	Dups	0	L
CNVR2485.1	5	65063884	1	13047215		NLN:0	S
CNVR2526.1	5	90660057	6	136903092	Multi	0	L
CNVR2542.1	5	99541856	21	42761422	Multi	0	L
CNVR2562.1	5	110311433	6	136645520		0	S
CNVR2717.1	5	180108619	9	8307236	Dups	0	L
CNVR2862.1	6	36748154	4	17188213		0	S
CNVR3159.1	6	166919026	15	81450426		RPS6KA2:0	S
CNVR318.1	1	119943910	13	58670160	Multi	0	L
CNVR3398.3	7	55757645	14	69712132	Dups	0	L
CNVR3533.1	7	111019447	5	115579000	Multi	0	L
CNVR3598.2	7	143539578	6	114333506	Multi	0	L
CNVR3618.4	7	151532774	21	10105718	Multi	MLL3:2	L
CNVR3751.1	8	4014857	2	64381478	Multi	CSMD1:0	L
CNVR4280.1	9	41552631	10	47993093	Multi	ZNF658B:1	L
CNVR4339.2	9	74995705	8	39465563		0	S
CNVR4684.1	10	32903336	8	78655139	Dups	CCDC7:1/C10orf68:0	L
CNVR4685.1	10	33229325	19	14592686	Dups	ITGB1:1	LD
CNVR4841.2	10	89138473	1	105062330	Multi	0	L
CNVR4841.3	10	89139568	7	13621304	Multi	MINPP1:3	L
CNVR5131.1	11	34128668	6	1051881	Dups	ABTB2:1	L
CNVR5186.1	11	57923640	4	166377389		0	S
CNVR5545.1	12	54066882	2	178636416	Dups	0	L
CNVR5712.1	12	123062059	2	3898560	Dups	ZNF664:1	LC
CNVR5889.1	13	56701502	16	16584576	Multi	0	L
CNVR6073.1	14	19594433	20	51835473	Dups	OR4L1:1	L
CNVR6172.1	14	64084537	3	2709841	Dups	C14orf50:3	L
CNVR6233.1	14	92782265	1	9039330	Dups	BTBD7:1	L
CNVR6266.1	14	102020212	4	130514072	Multi	KIAA0329:0	L
CNVR6330.9	15	26710413	2	172121715	Multi	0	L
CNVR6366.1	15	41780690	6	40482266	Multi	0	L
CNVR6564.2	15	100209456	6	170775689		OR4F4:1	D
CNVR6608.1	16	3622891	6	47107965	Dups	0	L
CNVR6808.1	16	80395173	12	50843083	Dups	PLCG2:0	L
CNVR6855.1	16	86722175	13	102017780	Dups	0	L
CNVR6927.1	17	924032	12	124353045	Dups	ABR:0	L
CNVR7043.1	17	22560736	17	16871486		0	D
CNVR7070.1	17	29087983	1	88752190	Dups	0	LC
CNVR7131.1	17	47252967	18	2737840	Multi	CA10:0	L
CNVR7273.1	18	14789518	3	44405336	Multi	0	L
CNVR7280.1	18	22001856	5	79957900		PSMA8:0	S
CNVR73.8	1	13081767	20	12316097	Dups	0	L
CNVR7358.1	18	60186868	14	27933537	Dups	0	L
CNVR756.1	2	37811582	4	49314563	Multi	0	L
CNVR7588.2	19	24343380	17	31426031	Dups	0	L
CNVR7629.1	19	41449253	4	132863778		0	D
CNVR7658.2	19	48000706	13	98502630	Multi	PSG1:5/PSG6:6/PSG7:6/PSG11:5	L
CNVR7702.1	19	56823616	1	14998430	Multi	SIGLEC5:4/SIGLEC14:7	L
CNVR7821.2	20	25679048	7	57673452		0	D
CNVR7992.1	21	31354018	15	23168062		0	S
CNVR8072.6	22	14969740	11	76210801	Multi	0	L
CNVR8085.1	22	16871523	17	22289266	Dups	MICAL3:0	LD
CNVR8103.16	22	20783598	2	158609008	Multi	0	L
CNVR8108.1	22	22101503	7	1885258	Dups	0	L
CNVR8136.1	22	31257932	6	24780180	Dups	SYN3:0	LC
CNVR8221.1	22	49414530	4	22188023	Multi	0	L
CNVR861.3	2	87510594	18	42476587	Multi	0	L

Table 1.5: **CNVs with evidence of dispersed duplication.** Type: deletion/duplication/multiallelic, only coded for genotyped loci. Code:"L", interchromosomal LD; "D", overlapping inter-chromosomal segmental duplication; "C", validated germline event from Cancer Genome Project; "S", sequence underlying CNV shows poly-A tail and target site duplications. See Methods for full details.

Name	Chr	Start	New Chr	New Start	Type	Genes	Code
CNVR927.1	2	111724832	4	65475118	Dups	0	L
CNVR979.1	2	129958916	12	11577532	Multi	0	L

Table 1.5: **CNVs with evidence of dispersed duplication.** Type: deletion/duplication/multiallelic, only coded for genotyped loci. Code:"L", interchromosomal LD; "D", overlapping inter-chromosomal segmental duplication; "C", validated germline event from Cancer Genome Project; "S", sequence underlying CNV shows poly-A tail and target site duplications. See Methods for full details.

Gene	Chr	Start	End	CNV	T-tests	Nexon	Nintron
ABTB2	11	34129111	34335378	0.026	18	85	3020
AHNAK	11	61957592	62070908	0.000	8	351	1130
AP2B1	17	30938395	31077549	0.000	6	118	1899
API5	11	43290081	43322658	0.000	10	75	438
CCDC50	3	192529568	192599152	0.099	9	165	1030
DST	6	56430744	56615653	0.000	8	555	2456
EIF4B	12	51686329	51722260	0.000	6	77	371
EXPH5	11	107881368	107969584	0.000	7	178	1060
FAM3C	7	120776141	120823658	0.000	25	49	757
FAM45A	10	120853601	120887213	0.040	31	45	289
FAM45B	10	120853619	120886505	0.019	11	31	289
FAM91A1	8	124850063	124896871	0.000	9	111	671
MLL2	12	47699025	47735374	0.000	6	368	256
MTHFD1	14	63924512	63996478	0.000	10	75	862
MUC16	19	8820520	8953018	0.000	8	826	1207
MUC5AC	11	1132474	1245302	0.081	6	383	1261
MYST2	17	45221070	45261457	0.000	11	65	509
NAB1	2	191222093	191265737	0.000	8	86	657
NUS1	6	118103310	118138579	0.000	29	90	455
PAK2	3	197951125	198043915	0.076	13	112	1070
PCDHA10	5	140215818	140372113	0.021	16	99	2237
PCDHA9	5	140207541	140372113	0.070	43	155	2317
PCLO	7	82221257	82630133	0.000	24	415	6175
PRKCI	3	171422914	171506464	0.000	31	100	902
PRKRA	2	179004388	179024204	0.040	31	40	272
RAB6A	11	73064331	73149849	0.064	10	67	894
RBM39	20	33754945	33793607	0.000	6	60	441
SALL1	16	49727387	49742684	0.000	10	99	176
SRRM2	16	2742331	2761414	0.000	6	171	168
TDG	12	102883723	102906785	0.000	11	65	245
TERF1	8	74083651	74122541	0.000	11	55	533
TTC3	21	37367441	37497278	0.000	34	185	1737
TTN	2	179098964	179380395	0.037	8	2244	2646
TYRO3	15	39638524	39658818	0.000	31	76	235
USP10	16	83291056	83371028	0.034	21	65	1151
WASF2	1	27604713	27689256	0.000	27	78	919
ZFHX4	8	77756070	77942076	0.000	9	251	2950
ZNF219	14	20628045	20642703	0.000	6	68	194
ZNF664	12	123023623	123065922	0.091	40	91	636

Table 1.6: **Ancestral loci of potential polymorphic retrogenes identified from** *in silico* **splicing of CGH data.** T-tests: number of significant t-tests. CNV: proportion of gene spanned by 42M CNVs. Nexon, Nintron: number of exonic,intronic probes.

PCR validation of retroposed genes. PCR assays were designed for four putative retrotransposed sequences using a three primer design: each assay consists of a duplication-specific primer, a primer specific to the non-duplication allele, and a common primer that

works for both alleles (Table 1.7). In this way the genotype of each sample can be read directly off of the gel by considering band sizes and counts.

Locus	CNVR4685.1	CNVR7070.1	CNVR5712.1	CNVR8136.1
Ancestral Chromosome	10	17	12	22
Location	33224858-3324595100	29087983-29088776	123062509-123065927	31257893-31258560
Target Chromosome	19	1	2	6
Location	14593345-14595100	Around 88745413	Around 3909526	Around 24791967
Insertion-specific Primers	5'-GAGATTGCACC ACTGTATTCTAGC-3' 5'-GGAGTTTGCTAA ATTTGAAAAGGA-3'	5'-TGAACCAGGA GTAAAACAGGC-3' 5'-GGTTTAAATCTCAA TCTAGTTCACTGC-3'	5'-CCCTATGGTG TATGTGCAGC-3' 5'-TCGAGCTTTAAA GTCCATAATTG-3'	5'-TTGCCAGGT TGGTTCTGAT-3' 5'-CAAGCCAGGT GGACACAGTA-3'
Expected Size of Product	342bp	N/A	N/A	N/A
No-insertion-specific Primers	5'-GAGATTGCACCA CTGTATTCTAGC-3' 5'-TGGCTTTTTGAATA AATGACAGAA-3'	5'-TGAACCAGGAG TAAAACAGGC-3' 5'-GGCTTGTAATTT TTCCTGGTACAT-3'	5'-CCCTATGGTGTA TGTGCAGC-3' 5'-GGCTATGATATA CCAGCCTAAGACA-3'	5'-TTGCCAAGGT TGGTTCTGAT-3' 5'-TATGAAACCCC CTAAACGTCTTA-3'
Expected Size of Product	277bp	370bp	745bp	437bp

Table 1.7: **PCR primers for assaying four duplicative transpositions.**

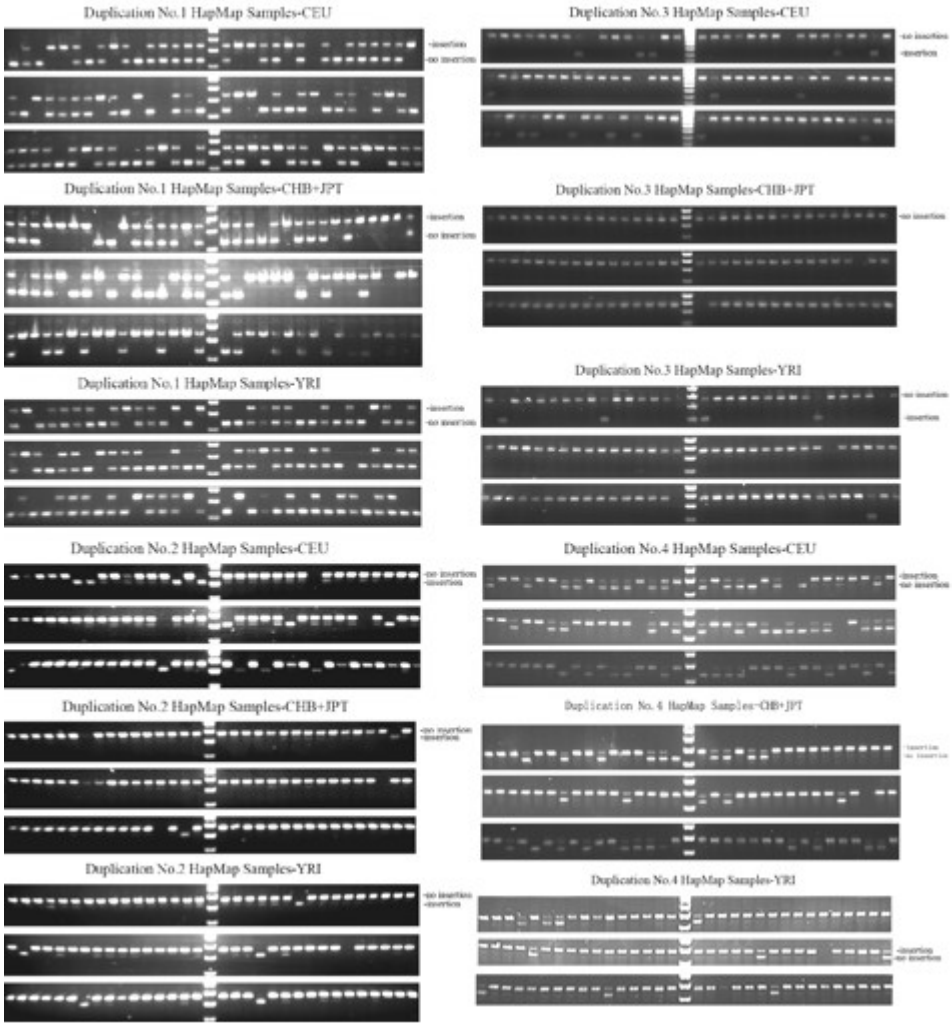


Figure 1.10: **PCR results for four dispersed duplications on HapMap samples.** Duplication 1: CNVR4685.1; Duplication 2: CNVR7070.1; Duplication 3: CNVR5712.1; Duplication 4: CNVR8136.1.

4 Population Genetics Analyses

4.1 Selection analyses

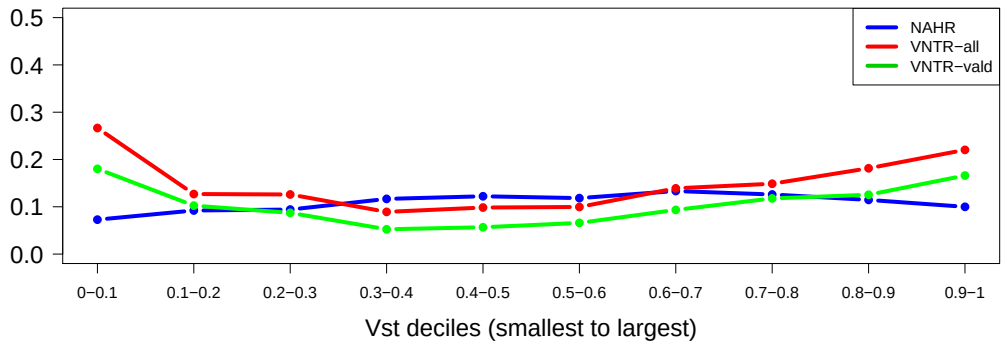


Figure 1.11: **Population differentiation as a function of mutation mechanism.** CNVs were subdivided by likely mutation process (NAHR and VNTR) and the proportion of events falling into each process was tabulated within each decile of V_{st}. We examined two groupings of VNTR: all VNTRs discovered to be polymorphic with the discovery array (VNTR-all), and the subset of those validated by 105k array (VNTR-vald). There was a significant difference in the proportion of CNVs formed by VNTR-all and NAHR in the top and bottom decile ($p < 10^{-5}$ by permutation).

4.2 LD analyses

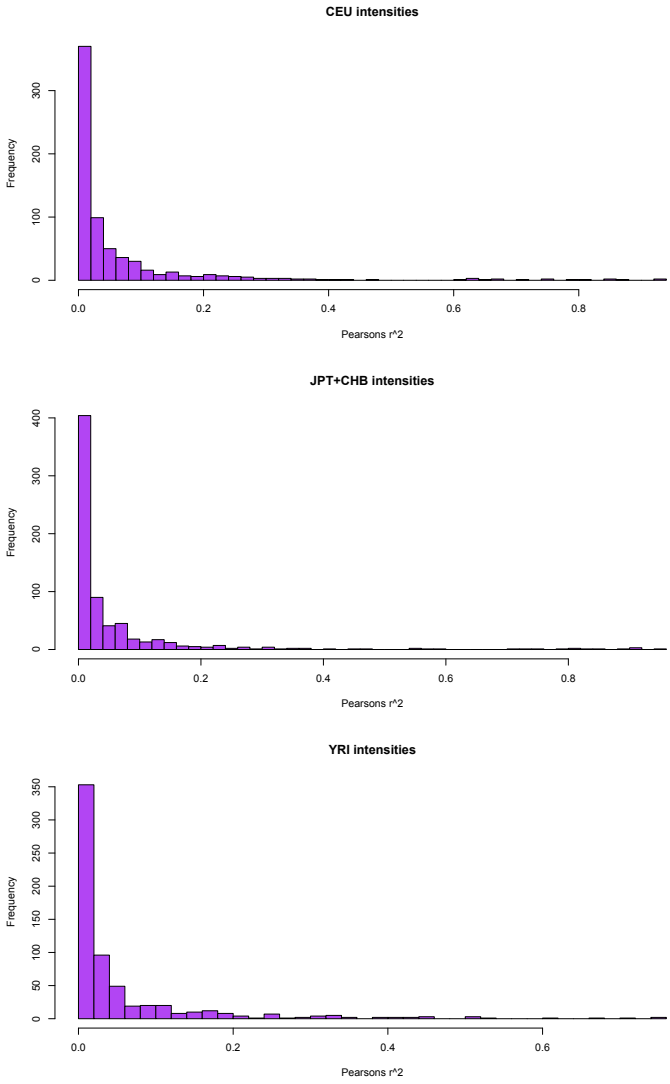


Figure 1.12: **Distribution of squared Pearson correlation between CNV intensities and GWAS hit SNP genotypes.** We employed two strategies for investigating the patterns of linkage disequilibrium between CNVs and GWAS hit-snps. Bi-allelic CNVs that could be reliably genotyped were phased into HapMap haplotypes and then subjected to conventional r^2 calculations. For all other validated CNVs we calculated pairwise Pearson correlations between raw (non-integer) copy number values and GWAS hit SNP genotypes. The distribution of these correlations is plotted here, separately for each HapMap population.

A WTCCC Authors

WTCCC Management Committee: Peter Donnelly (Chair)[1], Panos Deloukas[2], Audrey Duncanson[3], Matthew E Hurles[2], Dominic P Kwiatkowski[1,2], Mark I McCarthy[1,4], Willem H Ouwehand[5,6], Miles Parkes[7], Nazneen Rahman[8], Nilesh J Samani[9,10], John A Todd[11]

WTCCC CNV Committee: Nicholas Craddock[12] (co-chair), Matthew E Hurles[2] (co-chair), Chris Barnes[2], Niall Cardin[13], Donald F Conrad[2], Peter Donnelly[1], Eleni Giannoulatou[13], Chris Holmes[13], Jonathan L Marchini[13], Richard Pearson[1], Vincent Plagnol[11], Samuel Robson[2], Nilesh J Samani[9,10], Kathy Stirrups[2], Martin B Tobin[14], Damjan Vukcevic[1], Louise Wain[14], Chris Yau[13].

Autoimmune Thyroid Disease: Oliver J Brand[15], Jayne A Franklyn[15,16], Matthew J Simmonds[15], Stephen CL Gough[15,16].

Ankylosing Spondylitis: David M Evans[17], Millicent Stone[18,19], B Paul Wordsworth[20], Matthew A Brown[20,21].

Breast Cancer: Jaswinder Bull[8], Darshna Dudakia[8], Bernadette Ebbs[8], Diana Eccles[22], Anna Elliot[8], Polly Gibbs[8], Anita Hall[8], Sarah Hines[8], Debbie Hughes[8], David Pernet[8], Anthony Renwick[8], Richard Scott[8], Sheila Seal[8], Katarina Spanova[8], Clare Turnbull[8], Margaret Warren-Perry[8], Michael R Stratton[2,8], Nazneen Rahman[8].

Bipolar Depression: Gerome Breen[23,24], Sian Caesar[25], Anne Farmer[24], I Nicol Ferrier[26], Liz Forty[27], Katherine Gordon-Smith[25,27], Elaine Green[27], Detelina Grozeva[27], Ian R Jones[27], Lisa A Jones[25], George Kirov[27], Peter McGuffin[24], Michael C O'Donovan[27], Michael J Owen[27], Ellie Russell[27], David St Clair[23], Allan H Young[26,28], Nicholas Craddock[12,27].

Coronary Artery Disease: Stephen G Ball[29], Anthony J Balmforth[29], Peter S Braund[9], Paul R Burton[14], Suzanne Rafelt[9], John R Thompson[14], Alistair S Hall[29], Nilesh J Samani[9,10].

Crohns Disease: Tariq Ahmad[30], Katarzyna Blaszczyk[31], Francesca Bredin[7], Hazel E Drummond[32], Alistair Forbes[33], Derek P Jewell[34], Charlie Lees[32], James Lee[7], John Mansfield[35], Dunecan C O Massey[7], Alex Mentzer[36], Elaine R Nimmo[32], Natalie J Prescott[31], Jeremy D Sanderson[36], Jack Satsangi[32], Christopher G Mathew[31], Miles Parkes[7].

Hypertension: Morris J Brown[37], David G Clayton[11], John Connell[38], Anna Dominiczak[38], Martin Farrall[39], Philip Howard[40], Toby Johnson[40], G Mark Lathrop[41], Kate L Lee[40], Abiodun Onipinla[40], Nilesh J Samani[9,10], Sue Shaw-Hawkins[40], John Webster[42], Patricia B Munroe[40], Mark Caulfield[40]

Multiple Sclerosis: Alastair Compston[43], Stephen J Sawcer[43].

Rheumatoid Arthritis: Anne Barton[44], John Bowes[44], Ian N Bruce[44], Paul Emery[45], Steve Eyre[44], Edward Flynn[44], Paul Gilbert[44], Pile Harrison[46], Anne Hinks[44], Lynne Hocking[47], John D Isaacs[48], Paul Martin[44], Ann E Morgan[49], David M Reid[47], Deborah PM Symmons[44], Sophia Steer[50], Wendy Thomson[44], Anthony G Wilson[51], Paul Wordsworth[20,46], Jane Worthington[44]xs

Type 1 Diabetes: Oliver S Burren[11], Jason D Cooper[11], Kate Downes[11], Matt Hardy[11], Joanne MM Howson[11], Meeta Maisuria-Armer[11], Nigel R Ovington[11], Vincent Plagnol[11], Helen Schuilenburg[11], Debbie J Smyth[11], Helen E Stevens[11], Neil M Walker[11], Chris Wallace[11], Matthew Woodburn[11], David G Clayton[11], John A Todd[11].

Type 2 Diabetes: Amanda J Bennett[4], Rachel M Freathy[52], Chris J Groves[4], Neelam Hassanali[4], Graham A Hitman[53], Hana Lango-Allen[52], Cecilia M Lindgren[1], Andrew P Morris[1], Kirstie Parnell[52], John RB Perry[52], Inga Prokopenko[1], Nigel W Rayner[1], Neil Robertson[1], Mary Travers[1], Mark Walker[54], Michael N Weedon[52], Eleftheria Zeggini[1,2], Andrew T Hattersley[52,55], Timothy M Frayling[52], Mark I McCarthy[1,4].

1958 Birth Cohort Controls: Wendy L McArdle[56], Susan M Ring[56], David P Strachan[57].

UK Blood Service Controls: Anthony Attwood[2,5,6], Jennifer D Jolley[5,6], Jennifer G Sambrook[5,6], Jonathan Stephens[5,6], Nicholas A Watkins[5,6], Willem H Ouwehand[2,5,6].

Sample processing, genotyping and sequencing: Hazel Arbury[2], Sanjeev Bhaskar[2], John Burton[2], Chris M Clee[2], Alison J Coffey[2], Andrew Dunham[2], Sarah Edkins[2], Emma Gray[2], Rhian Gwilliam[2], Eleanor Howard[2], Sarah Hunt[2], Kirsten E McLay[2], Michael L Mimmack[2], Kimmo Palin[2], Michael A Quail[2], Carol E Scott[2], Elilan Somaskantharajah[2], Ins Barroso[2], Aarno Palotie[2], Panos Deloukas[2].

CNV Analysis: Jan Aerts[2], Chris Barnes[2], Niall Cardin[13], Donald F Conrad[2], Peter Donnelly[1], Eleni Giannoulatou[13], Naomi Hammond[2], Chris Holmes[12], Kevin Lewis[2], Jonathan L Marchini[13], Richard Pearson[1], Vincent Plagnol[11], Samuel Robson[2], Kathy Stirrups[2], Martin B Tobin[14], Damjan Vukcevic[1], Louise Wain[14], Chris Yau[13], Matthew E Hurles[2]

Fine-mapping and sequence analysis: Adam Auton[13], Jake Byrnes[1], Julian Maller[1], Andrew P Morris[1], Simon Myers[13], Gil McVean[13], Peter Donnelly[1]

WTCCC Principal Investigators: Matthew A Brown[20,21], Paul R Burton[14], Mark Caulfield[40], Alastair Compston[43], Nicholas Craddock[12], Panos Deloukas[2], Martin Farrall[39], Stephen CL Gough[15,16], Alistair S Hall[29], Andrew T Hattersley[52,55], Adrian VS Hill[1], Matthew E Hurles[2], Dominic P Kwiatkowski[1,2], Mark I McCarthy[1,4], Christopher G Mathew[31], Willem H Ouwehand[5,6], Miles Parkes[7], Marcus Pembrey[58], Nazneen Rahman[8], Nilesh J Samani[9,10], Jack Satsangi[32], Michael R Stratton[2], John A Todd[11], Jane Worthington[44], Peter Donnelly[1].

[1]The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK.
[2] The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA UK.
[3] The Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK.
[4] Oxford Centre for Diabetes, Endocrinology and Medicine, University of Oxford, Churchill Hospital, Oxford OX3 7LJ, UK.
[5] Department of Haematology, University of Cambridge, Long Road, Cambridge, CB2 0PT, UK
[6] National Health Service Blood and Transplant, Cambridge Centre, Long Road, Cambridge CB2 2PT, UK.
[7] IBD Genetics Research Group, Addenbrooke's Hospital, Cambridge, CB2 0QQ, UK
[8] Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton SM2 5NG, UK.
[9] Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Groby Road, Leicester LE3 9QP, UK.
[10] Leicester NIHR Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester, LE3 9QP, UK
[11] Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics,
Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge CB2 0XY, UK.
[12] Department of Psychological Medicine, Henry Wellcome Building, School of Medicine,
Cardiff University, Heath Park, Cardiff CF14 4XN, UK.
[13] Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG
[14] Department of Health Sciences and Genetics, University of Leicester, 217 Adrian Building, University Road, Leicester, LE1 7RH, UK
[15] Centre for Endocrinology, Diabetes and Metabolism, Institute of Biomedical Research,
University of Birmingham, Birmingham, B15 2TT, UK
[16] University Hospital Birmingham NHS Foundation Trust, Birmingham, B15 2TT, UK
[17] MRC Centre for Causal Analyses in Translational Epidemiology, Department of Social Medicine,
University of Bristol, Bristol, BS8 2BN, UK
[18] University of Toronto, St. Michael's Hospital, 30 Bond Street, Toronto Ontario Canada M5B 1W8
[19] University of Bath, Claverdon, Norwood House, Room 5.11a Bath Somerset BA2 7AY UK
[20] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Nuffield Orthopaedic Centre,
University of Oxford, Windmill Road, Headington, Oxford, OX3 7LD, UK
[21] Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital,
University of Queensland, Ipswich Road, Woolloongabba, Brisbane, Queensland, 4102, Australia
[22] Academic Unit of Genetic Medicine, University of Southampton, Southampton, UK
[23] University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen AB25 2ZD, UK
[24] SGDP, The Institute of Psychiatry, King's College London, De Crespigny Park, Denmark Hill, London SE5 8AF, UK
[25] Department of Psychiatry, University of Birmingham, National Centre for Mental Health, 25 Vincent Drive, Birmingham, B15 2FG, UK
[26] School of Neurology, Neurobiology and Psychiatry, Royal Victoria Infirmary, Queen Victoria Road, Newcastle upon Tyne, NE1 4LP, UK
[27] MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Heath Park, Cardiff, CF14 4XN, UK
[28] UBC Institute of Mental Health, 430-5950 University Boulevard Vancouver, British Columbia, Canada V6T 1Z3.
[29] Multidisciplinary Cardiovascular Research Centre (MCRC), Leeds Institute of Genetics,
Health and Therapeutics (LIGHT), University of Leeds, Leeds, LS2 9JT, UK
[30] Genetics of Complex Traits, Peninsula College of Medicine and Dentistry University of Exeter, EX1 2LU
[31] Department of Medical and Molecular Genetics, Kings College
London School of Medicine, 8th Floor Guys Tower, Guys Hospital, London, SE1 9RT, UK
[32] Gastrointestinal Unit, Division of Medical Sciences, School of Molecular and Clinical Medicine,
University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK
[33] Institute for Digestive Diseases, University College London Hospitals Trust, London NW1 2BU, UK
[34] Gastroenterology Unit, Radcliffe Infirmary, University of Oxford, Oxford, OX2 6HE, UK
[35] Department of Gastroenterology & Hepatology, University of Newcastle upon Tyne
Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK
[36] Division of Nutritional Sciences, King's College London School of Biomedical and Health Sciences, London SE1 9NH, UK
[37] Clinical Pharmacology Unit, University of Cambridge, Addenbrookes Hospital, Hills Road, Cambridge CB2 2QQ, UK;
[38] BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow, G12 8TA, UK;
[39] Cardiovascular Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK;
[40] Clinical Pharmacology and Barts and The London Genome Centre, William Harvey Research Institute, Barts and
The London School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK;
[41] Centre National de Genotypage, 2, Rue Gaston Cremieux, Evry, Paris 91057, France;
[42] Medicine and Therapeutics, Aberdeen Royal Infirmary, Foresterhill, Aberdeen, Grampian AB9 2ZB, UK;
[43] Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 2QQ, UK
[44] arc Epidemiology Unit, Stopford Building, University of Manchester, Oxford Road, Manchester, M13 9PT, UK
[45] Academic Unit of Musculoskeletal Disease, University of Leeds, Chapel Allerton Hospital, Leeds, West Yorkshire LS7 4SA, UK
[46] University of Oxford, Institute of Musculoskeletal Sciences, Botnar Research Centre, Oxford, OX3 7LD, UK
[47] Bone Research Group, Department of Medicine and Therapeutics, University of Aberdeen, Aberdeen, AB25 2ZD, UK
[48] Institute of Cellular Medicine, Musculoskeletal Research Group, 4th Floor, Catherine Cookson Building, The Medical School,
Framlington Place, Newcastle upon Tyne, NE2 4HH, UK
[49] NIHR-Leeds Musculoskeletal Biomedical Research Unit, University of Leeds, Chapel Allerton Hospital,
Leeds, West Yorkshire LS7 4SA, UK
[50] Clinical and Academic Rheumatology, Kings College Hosptal National health Service Foundation Trust,
Denmark Hill, London SE5 9RS, UK
[51] School of Medicine and Biomedical Sciences, University of Sheffield, Sheffield, S10 2JF, UK
[52] Genetics of Complex Traits, Peninsula College of Medicine and Dentistry, University of Exeter, Magdalen Road, Exeter, EX1 2LU, UK
[53] Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London, E1 1BB UK
[54] Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University,
Framlington Place, Newcastle upon Tyne NE2 4HH, UK
[55] Genetics of Diabetes, Peninsula College of Medicine and Dentistry, University of Exeter, Barrack Road, Exeter, EX2 5DW, UK
[56] ALSPAC Laboratory, Department of Social Medicine, University of Bristol, BS8 2BN, UK.
[57] Division of Community Health Sciences, St George's, University of London, London SW17 0RE, UK

[58] Avon Longitudinal Study of Parents and Children, University of Bristol, 24 Tyndall Avenue, Bristol BS8 1TQ, UK.