Complex Forms of Structural Variation in the Human Genome: Haplotypes, Evolution, and Relationship to Disease

Complex forms of structural variation in the human genome: haplotypes, evolution, and relationship to disease

A dissertation presented

by

Linda M. Boettger

to

The Division of Medical Science

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

In the subject of

Genetics and Genomics

Harvard University

Cambridge, Massachusetts

November 2014

Dissertation Advisor: Professor Steven A. McCarroll   Linda M. Boettger

Complex forms of structural variation in the human genome: haplotypes, evolution, and relationship to disease

Abstract

Genomic mutations arise in many forms, varying from single base pair substitutions to complicated sets of overlapping copy number variants (CNVs). While each type of variation contributes to phenotype, complex structural variation, which contains multiple mutations, is difficult to type across many individuals and is largely omitted from genomic studies. This thesis presents methods to type complex structural variation, understand how it evolves, and integrate these complex variants into association studies to phenotypes.

We focused on four structurally complex regions in the human genome. The 17q21.31 region contains an inversion, previously uncharacterized overlapping copy number variants, and SNPs that associate to the female meiotic recombination rate and female fertility[1]. The haptoglobin (*HP*) gene at chromosome 16q22.2 contains a 1.7 kb tandem duplication[2], previously uncharacterized paralogous gene conversion, and nearby SNPs that associate to cholesterol levels[3]. The haptoglobin related gene (*HPR*) at chromosome 16q22.2, segregates as a multi-allelic copy number variant (mCNV) specifically in African populations. Lastly, complement component 4 (*C4*) at chromosome 6p21.3, contains a length polymorphism, paralogous sequence variation, and copy number variation segregating in humans and non-human primates[4].

We developed methods to characterize the complex structural variation in each of these four regions, type the variation at the population level and integrate it into association studies. Briefly, we determined the breakpoints of each

individual structural variant, typed each variant in a population cohort, and learned which variants segregate together through trio inheritance patterns. Once these structural haplotypes were defined, we phased them with surrounding SNP haplotypes and used this data as a reference panel for imputation into disease cohorts, and to better understand their evolutionary history.

We found that two overlapping duplications in the 17q21.31 region rose rapidly and independently to high frequency within European populations, and may account for the regional association to female fertility and the female meiotic recombination rate. We also found that a recurrent deletion in the *HP* gene associates to total cholesterol and LDL cholesterol levels. The methods developed in this thesis enable the integration of structurally complex variation into future association studies so that we can begin to understand their effects on phenotypes.

Table of Contents

Acknowledgements

There are many people who contributed to this thesis either through scientific or moral support whom I wish to acknowledge. I am thankful for their guidance, constructive criticism, and encouragement that made this work possible.

First I would like to thank my thesis advisor, Steve McCarroll, for his leadership, support, and attention to detail, all of which have helped me become a better scientist over the past five years. He has a great ability to select compelling research questions and assist his trainees in developing new approaches to answer these questions. He is also one of the very best at teaching scientific story telling. Assembling a study into a compelling story makes a significant difference in size of the audience it will reach and how well it will be understood. Starting to develop this skill in the McCarroll Lab will be extremely valuable to me over the course of my career.

The members of my Dissertation Advisory Committee - David Reich, Joel Hirschhorn and Monica Colaiacovo - were particularly helpful in providing insightful comments and new directions for my research.

I would also like to thank other members of the McCarroll and Reich Labs who have provided instrumental advice. My office mates Aswin Sekar, Avery Davis, Christina Usher, Matthew Baum and I discuss our projects with one another on a daily basis and their insightful comments have improved my research on many occasions. Priya Moorjani and David Reich have been very helpful in answering my questions related to evolutionary genomics. Nadin

Rohland and Tom Mullen provided great advice and assistance with sequence capture methods. Bob Handsaker has provided extremely helpful discussions that altered my thinking about several projects. Mike Zody was a wonderful collaborator for the 17q21.31 project and our thorough analysis would not have been possible without him.

The administrative support in the McCarroll and Reich Labs has been particularly helpful and allowed my time as a doctoral student to run much more smoothly. Elizabeth Fels always went above and beyond to schedule what I thought was impossible and give kind words of support during stressful times.

Lastly I would like to thank my family for their support and advice. My parents Mark Boettger and Carol Manning, and my brother, Dave Boettger, care a great deal about my research and have edited my scientific manuscripts on several occasions. My fiancé, Daniel Goodman, is a fellow member of the Harvard Medical School Department of Genetics, and he has been extremely loving and supportive during our graduate school careers. He never shies away from a detailed scientific discussion and has given valuable advice countless times.

Chapter 1
Introduction to complex structural genomics

Overview

Genomic variation spans from single nucleotides to megabases and from simple biallelic differences to complex sets of nested events. While each type of genomic variation has the potential to contribute to phenotypes, certain types of variation are more easily studied than others, and this greatly influences their representation in literature. Single nucleotide polymorphisms (SNPs) are the most commonly studied class of variation and are often typed using genotyping arrays. Arrays are readily scalable and the data can be easily interpreted, allowing for widespread use on a mass scale[5]. More recently, simple deletion and duplication polymorphisms have been discovered and typed with array comparative genomic hybridization (arrayCGH)[6] and high density genotyping arrays[7,8], allowing them to be studied alongside nucleotide variation on a large scale.

However, regions with more complex structure – those that have been affected by multiple cumulative structural mutations in human ancestors – remain an untapped reservoir of variation. Multiple structural changes at the same locus create multi-allelic variation with multiple structural forms segregating in the same population. These regions are difficult to study because they can contain overlapping structural mutations that are challenging to tease apart, and high copy number, which is difficult to measure accurately. The complex structure for a specific individual can be determined through the sequencing of clones[1]; however, this method is not easily scalable to large numbers of individuals. As a result, there is little understanding of how complex structural variation evolves, and this class of variation is largely omitted from association studies.

2

In this thesis we design strategies for studying complex structural variation and apply these methods to gain an in depth understanding of four structurally complex regions in the human genome.

1. We aim to design and apply methods for typing structurally complex haplotypes at the population level (Chapter 3).

2. We aim to understand how complex structure evolves within populations, across populations, and across species (Chapter 4).

3. We aim to develop and implement methods to incorporate complex genome structures into association studies (Chapter 5).

Genome Evolution: understanding different classes of variation

The evolution of nucleotide variation

The genome evolution of SNPs is commonly studied. The human nucleotide mutation rate is estimated to be 2.5×10^{-8} per nucleotide, per generation[9], but this rate varies throughout the genome. The rate is increased near recombination hotspots[10], high GC content[11], and insertions and deletions[12]. However, the overall clocklike consistency of single base mutations[13] under selective neutrality[14] has allowed both understanding of the relationships between organisms and the dating of species divergence times[15].

A nucleotide mutation arises on a specific haplotype background and continues to segregate with that haplotype. While recombination and recurrent mutation disrupt haplotypes, most SNPs have high linkage disequilibrium (LD) with other nearby nucleotides. A SNP in high LD with another SNP that is not

directly typed in a study can be used as a tag to infer the state of the unknown SNP.

Additionally, monitoring the length of SNP haplotypes can be informative for understanding how selection has affected specific regions. Haplotypes with low nucleotide diversity segregating at high frequency can be interpreted as regions evolving by positive selection[16,17]. Such regions in which a beneficial mutation has occurred rise to high frequency quickly without accumulating many nucleotide mutations or recombinations.

Furthermore, the rate of accumulation of nucleotide mutations in specific regions of the genome is used to gain understanding of the type of selection acting. For example, highly conserved regions of the genome with very few nucleotide differences between humans and mice[18], or even humans and yeast[19], have evolved by purifying selection. Under this scenario, most mutations are deleterious and are eliminated from the population. Additionally, the nucleotide diversity of species and populations can be informative for determining demographic events such as bottlenecks or population expansions[20,21].

The evolution of simple structural variation

The evolution of genome structural variation is less well understood. While many of the insights into genome evolution that were observed using SNPs may also apply to structural variation, there are also likely to be striking differences. Recently, geneticists have begun to explore the evolution of simple structural

4

variants, and CNVs have been discovered and typed in a diverse array of mammals including humans[7,8], great apes[22], rodents[23,24], dogs[25,26] and cows[27].

For a given species of great ape, the diversity of CNVs in the genome is correlated to that of nucleotides. While the rate of fixed deletions in great apes is relatively clocklike, on the chimpanzee lineage it has increased twofold[22].

Differences in the rate of structural and nucleotide evolution for a given species may be impacted by selection, but other forces also play a role. While the nucleotide mutation rate varies somewhat across different regions of the genome, the structural mutation rate likely varies more widely and depends more strongly on regional genetic architecture. For example, specific regions of the genome can be predisposed to structural mutations due to adjacent homology[28]. Additionally, CNV breakpoints are enriched for indels and microsatellites, indicating a relationship between different types of structural variants[8].

The majority of simple and common copy number variants in humans are well-captured by tag SNPs[7,8], suggesting that most of these arose from a single mutational event and continue to segregate on a specific haplotype. In the CEU population (Utah residents with Northern and Western European ancestry), 77% of simple and common CNVs are tagged by at least one SNP with $r^2 = 0.8$ (See Supplementary Table 1 for population identifiers). Interestingly, deletions appear to be tagged more often than duplications. This could be due to reversion mutations (deletions of one copy of a duplication resulting in the ancestral state) or recurrent mutation[8].

The evolution of complex structural variation

The evolution of complex structural variation is relatively unknown. In addition to the difficulty of typing this variation in large cohorts, complex structure cannot be inferred from a single surrounding SNP as multi-allelic variants are not in high LD with any single biallelic SNP. In this thesis we develop methods to type and analyze such regions in large cohorts, across populations and across species. One of the goals of the research presented in here is to thoroughly understand previously cryptic structurally complex loci and gain insight into their evolution.

Associating genomic variants to phenotypes

Association studies through direct typing of candidate genes

Natural genomic variation can provide insight into gene function and disease-causing variation. Until relatively recently, genotype-phenotype association studies were performed solely with candidate genes and candidate variants. While this method has successfully found phenotypic associations to single nucleotide[29-31] and structural[32-34] polymorphisms alike, there are multiple drawbacks. For example, direct typing of candidate variants severely restricts study sample size because each variant must be directly typed without the use of high-throughput genotyping technology. This strategy leaves variants of small effect undetectable, and risking false-positive associations. Additionally, typing only one specific variant in an LD block can result in incorrect assumptions about

6

the true causal variant. Finally, corrections for ancestry can be difficult to make without extensive genome-wide information.

Association studies using genotyping arrays

Genome-wide association studies (GWAS) can overcome many of the difficulties of the candidate gene approach. Because of the scalability of genotyping arrays, sample sizes have increased drastically, and are often in the tens or hundreds of thousands[35]. In addition to increased power, multiple SNPs are usually assayed within each haplotype, allowing for a more precise understanding of the most associated variants within an LD block, and ancestry correction can be made based on genome-wide SNPs.

Progress has been made recently to incorporate simple structural variation (deletions, duplications) into association studies. Genotyping arrays such as the Affymetrix SNP 6.0 are hybrid arrays on which both SNPs and CNVs can be ascertained[7]. While most common, biallelic, CNVs are in high LD with a SNP, arrays which include CNVs allow some duplications and deletions which are not captured by SNPs to be incorporated into association studies. Using a genome-wide approach, numerous copy number polymorphisms have been associated to phenotypes[8], including Crohn's disease[36] and body mass index[37].

Association studies of complex structural variation

Complex structural variation is largely omitted from association studies. This type of genomic variation is affected by multiple structural mutations and segregates as an allelic series, meaning that no single biallelic SNP can serve as a tag. In order to incorporate complex structural variation into association studies

and understand contribution to phenotype, we must be able to type this class of variation with high-throughput methods. This research seeks to develop methods to type complex structural variation in large cohorts and include it in association studies to human phenotypes.

The potential to leverage previous GWAS data to study complex structural variation through imputation

There is currently a wealth of data collected for GWAS that contains both SNP genotype and human phenotype information for thousands of individuals. While complex structural variants are not tagged by any individual biallelic SNP, a series of SNPs on a haplotype may be sufficient to infer complex structural variation through imputation. Imputation is a statistical method by which unobserved variants can be inferred through their association with a particular SNP haplotype. If complex structural variation were imputable using surrounding SNP haplotypes, previously collected GWAS data could be leveraged to study complex structural variation.

Chapter 2

Background information for four structurally complex loci

Introduction

In order to gain a broad perspective on the evolution of complex structural variation, we selected an interesting and diverse panel of structurally polymorphic regions for in-depth study. While structural polymorphisms have previously been documented in each of these regions, knowledge of precise structural forms and how they vary in different populations and species was previously unknown. It is also not known how these structural alleles relate to phenotypes. While the selected regions share common features, they are each affected by different types of structural polymorphism, have different evolutionary histories, and can teach us different principles about genome structural evolution (Table 2.1). Each of the chosen regions is described below.

Table 2.1. Structurally complex regions classified by different categories. This table compares four structurally complex loci by various criteria.

Locus	Thought to vary in all human populations	Thought to have multiple classes of structural variation	Structural polymorphism likely derived in humans	Part of a gene cluster	Thought to contain high copy number (>3)
17q21.31	X	X	X		X
HP 16q22.2	X		X	X	
HPR 16q22.2			X	X	X
C4 6p21.3	X	X			

17q21.31

The 17q21.31 region is associated with multiple phenotypes and has a long and interesting genetic history. There are two haplotypes at this locus that are inverted with respect to each other. The inversion prevented these haplotypes from recombining, and this lack of recombination has caused two haplotypes to become isolated and highly diverged[1]. The common allele of the inversion polymorphism is called H1, while the inverted form is called H2 (Figure 2.1).

The 17q21.31 region contains many genes but is perhaps most famous for the microtubule-associated protein (MAPT), which is thought to be involved in progressive supranuclear palsy and Parkinson's Disease[38-40].
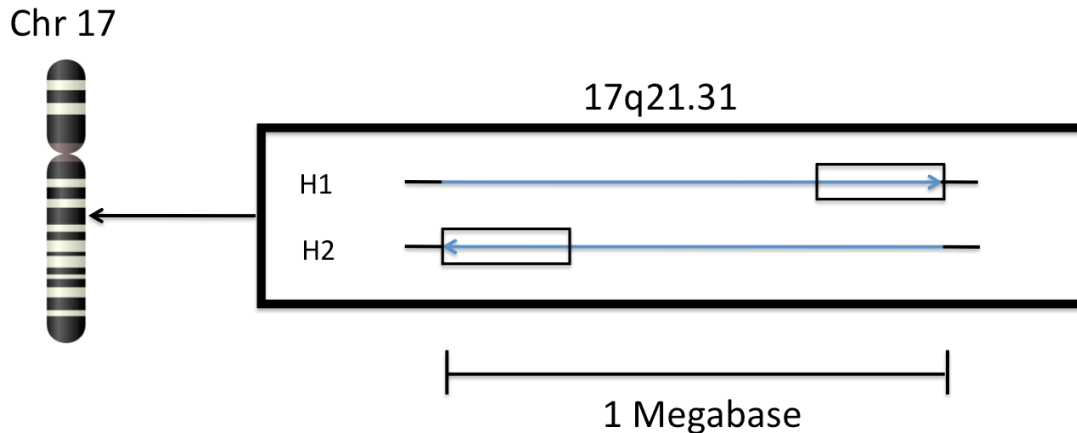
Chr 17

17q21.31

H1

H2

1 Megabase

Figure 2.1. The 17q21.31 region contains a megabase long inversion as well as regions of known copy number variability, indicated with small black boxes. The standard form of the inversion polymorphism is called H1, while the inverted form is called H2.

It is estimated that the inversion event took place 2.3 million years ago[41], which is an exceptionally long period of time for a population to maintain both haplotypes, as genetic drift should favor the fixation of one allele[42]. This population-genetic oddity led some researchers to hypothesize that the inversion was transmitted into humans from Neanderthal gene flow[43], but the publication of the Neanderthal genome demonstrated that the inversion is unlikely to be of Neanderthal origin[44]. Other possible explanations for the maintenance of these two haplotypes include frequency dependent selection, heterozygote advantage, and founder effects[1,41].

In addition to the noteworthy age of this inversion, it also has an unusual population distribution. The common H1 allele is predominant in most human populations, while H2 is present at a frequency of approximately 20% throughout most of Europe, 30% in Southern Europe, and <5% in Africa[45]. This distribution is surprising for a variant estimated to be much older than the date that modern humans are thought to have left Africa (100,000 years ago)[46], and could indicate that the H2 variant has been selected specifically in European populations.

An association study in an Icelandic population also supports the idea of positive selection. In their analysis, Stefansson et al. observed that women who are carriers of H2 have more children than those who are homozygous for H1[1]. In addition to this increase in fertility, women who carry H2 have a greater number of recombinations in their gametes ($p = 0.0002$)[1]. These two traits are correlated with each other, and it is thought that an increased rate of recombination makes women more fertile by decreasing the odds of a non-disjunction event in female germ cells[47].

12

The 17q21.31 region also harbors SNPs that associate with various phenotypes in GWAS. The region has reported genome-wide significant associations with Parkinson's disease[38], autism[48], Crohn's disease[49] and others. Interestingly, a recent GWAS examining variation in meiotic recombination also found an association with SNPs inside the 17q21.31 inversion region ($p=2.4\times10^{-6}$)[50]; however, none of the recombination-associated SNPs reached genome-wide significance (5.5×10^{-8})[51].

While the 17q21.31 region contains several interesting phenotypic associations, the variants responsible for these associations have yet to be identified. Additionally, this region is known to contain multiple multi-allelic copy number variants (mCNVs) that have not been characterized. In order to more thoroughly understand phenotypic associations and structural history of the 17q21.31 region, in the following chapters we will describe the accurate mapping and typing of its structure in large numbers of individuals.

Haptoglobin (*HP*)

Haptoglobin (*HP*) is a particularly important gene, which codes for both the haptoglobin protein (an antioxidant which binds free hemoglobin and cholesterol complexes)[52,53], and the zonulin protein (a key regulator of intercellular tight-junctions)[54]. The HP protein is proteolytically cleaved and held together by disulfide bonds[55], while zonulin is the pre-processed product of the *HP* gene[54].

HP1S

HP1F

HP2FS

"Fast" amino
acid changes

Sequence
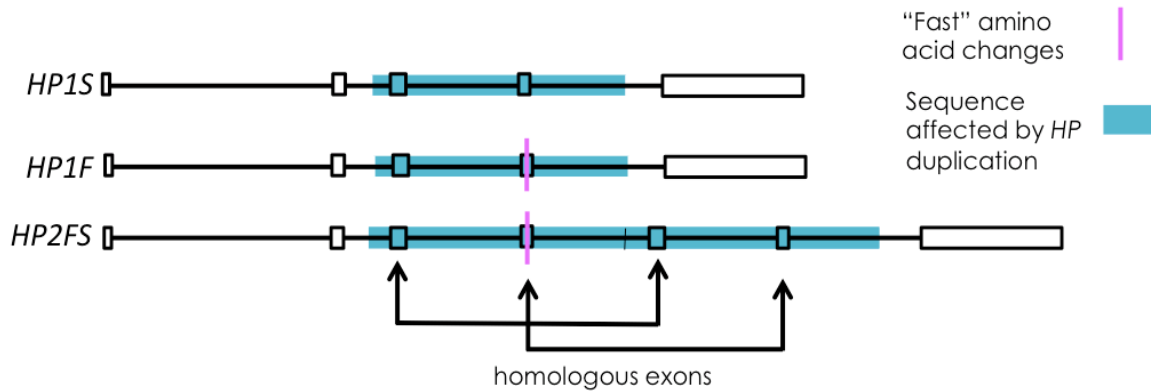affected by *HP*
duplication

homologous exons

Figure 2.2. A 1.7 kb duplication affects two exons of the *HP* gene. The sequence
affected by this duplication is indicated in blue, while the amino acid changes that
cause the protein to run faster on a gel are indicated with pink lines. The
common haplotypes of *HP* are called *HP1S*, *HP1F* and *HP2FS*. Black arrows
indicate homologous exons included in the duplication. The black boxes indicate
the locations of exons.

The haptoglobin (*HP*) gene contains a seemingly simple 1.7 kb tandem
duplication polymorphism, which was the second polymorphism to be discovered
in man[56], and continues to be widely studied. This CNV overlaps two exons of
the *HP* gene and is polymorphic in every human population yet examined[57]. The
duplicated form of haptoglobin is called *HP2*, while the unduplicated is known as
HP1. *HP* haplotypes are also classified by the "F" and "S" alleles, named for their
bands running "fast" or "slow" on a protein gel, and caused by two directly
adjacent charge-changing amino acids in exon four (Figure 2.2). The common
HP haplotypes are called *HP1F*, *HP1S* and *HP2FS*, as the *HP2* haplotype
usually has a copy of the fast and the slow form (Figure 2.2)[56,58]. These charge-

changing amino acids do not appear on SNP genotyping arrays, likely due to copy number variability, and their LD to other SNPs is not known.

The *HP2* allele is responsible for two striking molecular phenotypes: (1) it is required for the creation of the zonulin protein[54], and (2) it causes the HP protein, which is standardly a dimer, to form trimers or tetramers[56]. Interestingly, it has recently been documented that the duplication is not in high LD with any single SNP (max $r^2=0.16$)[59], and therefore this important structural variant is currently inaccessible to SNP-based association studies. We also note that this CNV is not observed with high-throughput CNV detection methods including probe intensity and read depth[8,60].

The *HP* CNV is one of the most highly studied variants in the human genome and can be directly typed by PCR or Western Blot[61]. Thus far in 2014 alone, more than fifteen studies have directly typed this variant and performed association studies to various phenotypes[62-77]. However, the necessity of directly typing *HP* structural variation severely restricts sample size. While individual studies have concluded that the *HP* duplication variant associates to a variety of phenotypes, including susceptibility to malaria, Crohn's disease[78,79], heart disease[62,80,81], leukemia[82-84], and HIV progression[85,86], we are not aware of any highly significant and reproducible associations (Table 2.2). Furthermore, association studies of *HP* structure that do not integrate nearby SNPs have led to controversy about the true causal variant[87].

15

Table 2.2. Association studies of the *HP* CNV to phenotypes have not been highly significant and reproducible. The phenotype, population risk allele, sample size, and p-value are listed for each study.

Phenotype	Study	Population	Risk allele	Sample size	P-value
Crohn's disease susceptibility	Papp 2007	Hungary	*HP1*	68 cases, 384 controls	0.03
	Maza 2008	Israel	*HP2*	382 cases, 3243 controls	0.05
High total cholesterol levels	Guthrie 2012	United Kingdom	no association	2,779	0.112
Heart disease susceptibility in diabetic individuals	Levy 2004	Framingham, USA	*HP1*	7,600	0.05
	Pechlaner 2014	various	no association	806	0.092
HIV progression	Delanghe 1998	caucasians	*HP2*	653	0.003
	Zaccariotto 2006	Brazil	no association	387 cases, 142 controls	0.85
Leukemia susceptibility	Nevo 1986	various	*HP1*	211 cases, 261 controls	0.001
	Campregher 2004	Brazil	no association	188 caes, 197 controls	best 0.061
	Atkinson 2007	Kenya	*HP1*	312	0.008
	Elagib 1998	Sudan	*HP1*	273 cases, 208 controls	0.001
Malaria susceptibility	Beiguelman 2003	Brazil	no association	182	0.764
	Bienzl 2005	Ghana	no association	290 cases, 580 controls	not reported
	Quaye 2000	Ghana	*HP1*	113 cases, 42 controls	0.04
	Aucan 2001	Gambia	no association	1,183	0.46

Promising functional data supports molecular differences between *HP* structural forms in the HP protein's antioxidant capacity[88], macrophage binding efficiency[89], and the ability to make the zonulin protein, which alters intestinal permeability[54]. However, because large cohort studies for relevant phenotypes have used SNP genotyping arrays, which do not effectively capture *HP* variation, and smaller direct-typing studies are likely underpowered, the true phenotypic contribution(s) of this highly studied and molecularly functional duplication remain uncertain.

The evolutionary history of *HP* structural variation is also not well understood. It is thought that the duplication is derived in humans because it has not been observed in other primates[90]. Maeda and Smithies et al. published a structural evolution model in which an *HP1F* haplotype and an *HP1S* haplotype underwent non-homologous recombination, thereby fusing to form *HP2FS*[2]. However, the lack of a good tag SNP for this common duplication indicates that a more complicated structural history is possible.

It has been proposed that balancing selection acts on the duplication[82,91], but this idea is difficult to demonstrate, given the uncertain age of the variant (estimates range from 100 thousand to 2 million years)[59,92]. In order to establish a role for balancing selection through allele age, a variant must be identical by descent and polymorphic in both humans and chimpanzees[93,94].

It will be necessary to further investigate the structural history of the *HP* region in order to understand why *HP2* is not in high LD with any SNP and not observed in genome-wide scans for structural variation. Additionally, in order to demonstrate robust phenotypic associations of *HP* structural variation to phenotype, it will be necessary to develop high-throughput typing methods.

The haptoglobin related gene (*HPR*)

The haptoglobin related gene *(HPR)* is another member of the haptoglobin gene cluster, lying 2.2 kb downstream of *HP* and sharing 90% sequence identity with it. *HPR* was created by an ancient primate triplication of *HP*, and the *HPR* gene is itself copy number variable in African populations (Figure 2.3)[95]. *HPR* is a key component of the Trypanosome Lytic Factor complex 1 (TLF1)[96]. TLF1

provides humans with immunity to most species of trypanosome, the cause of

African Sleeping Sickness[97]. An mCNV containing the *HPR* gene was first

observed in 1986 in African Americans[95], but this variation has yet to be fully

characterized for extent of copy number variation and distribution within African

populations. We hypothesized that this variant affects susceptibility to human

trypanosomiasis.
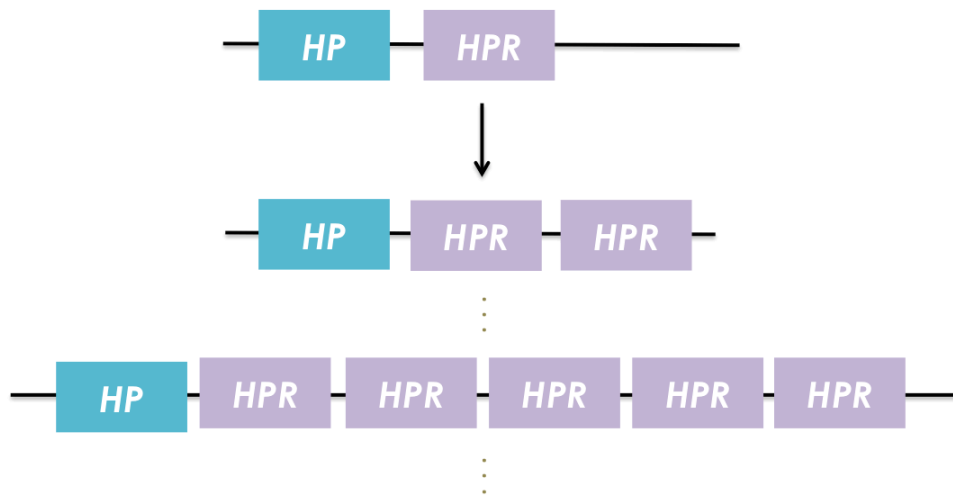
Figure 2.3. The *HPR* gene segregates as a multi-allelic CNV in African and

African American populations. The *HPR* gene is depicted as a lavender box,

while the *HP* gene (another member of the *HP* gene cluster) is shown as a blue

box. Copy number of up to five has been observed segregating on a single

haplotype.

Complement component 4 (*C4*)

Complement component 4 (*C4*) is a gene within the human leukocyte

antigen (HLA*)* region that contains complex structural variation segregating in

both humans and non-human primates. *C4* is a critical member of the complement pathway, which has roles in both the immune[98] and nervous systems[99,100]. Multiple forms of structural variation segregate at *C4*. There are two paralogs of this gene called *C4A* and *C4B*, which are defined by five base pair differences within a seventeen base pair region and result in functional sequence differences. *C4A* binds to proteins, while *C4B* binds the hydroxyl groups of carbohydrates[101]. Additionally, the insertion of a Human Endogenous Retroviral (HERV) generated a length polymorphism distinguished with the names *C4L* (long) and *C4S* (short). Finally, the copy number of each of these forms varies[102] (Figure 2.4).
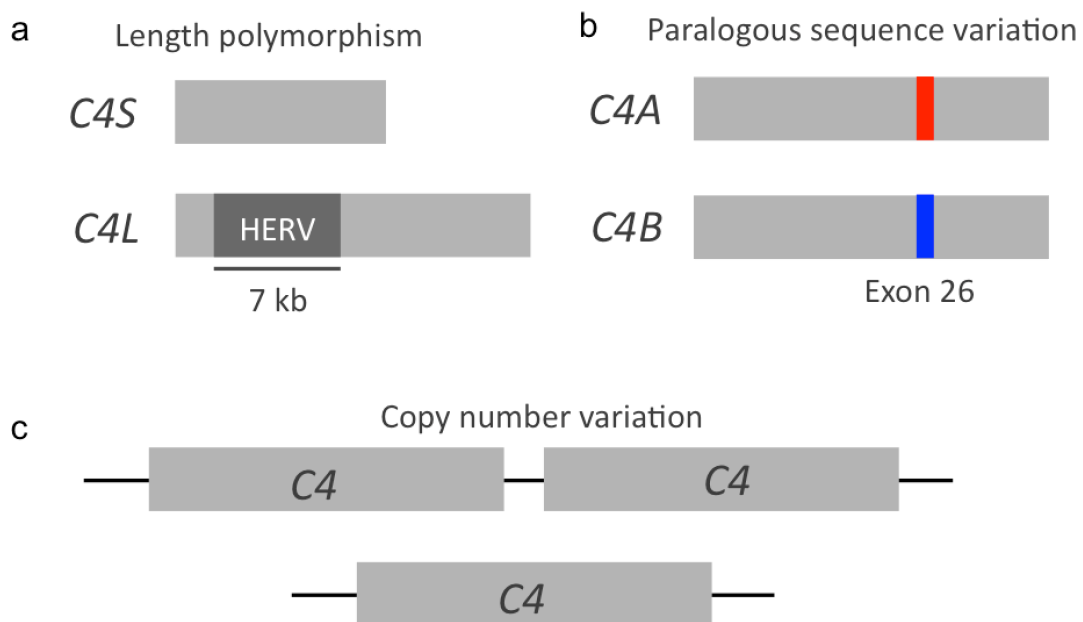

Figure 2.4. The *C4* gene is polymorphic for three types of structural variation: (a) A length polymorphism created through the introduction of a HERV. (b) Paralogous sequence variation in exon 26 that yields functional differences. (c) The entire *C4* gene is copy number variable.

Variation in *C4* and the HLA in general has deep coalescence. A few small studies have shown that several great apes share the L/S and A/B polymorphisms and are also copy number variable at this locus[4,103,104]. While extensive sequencing and genotyping of the *C4* locus in humans has led to a thorough understanding of the structural haplotypes segregating in human populations (Figure 2.5)[105], only a few great apes have been typed at *C4*.
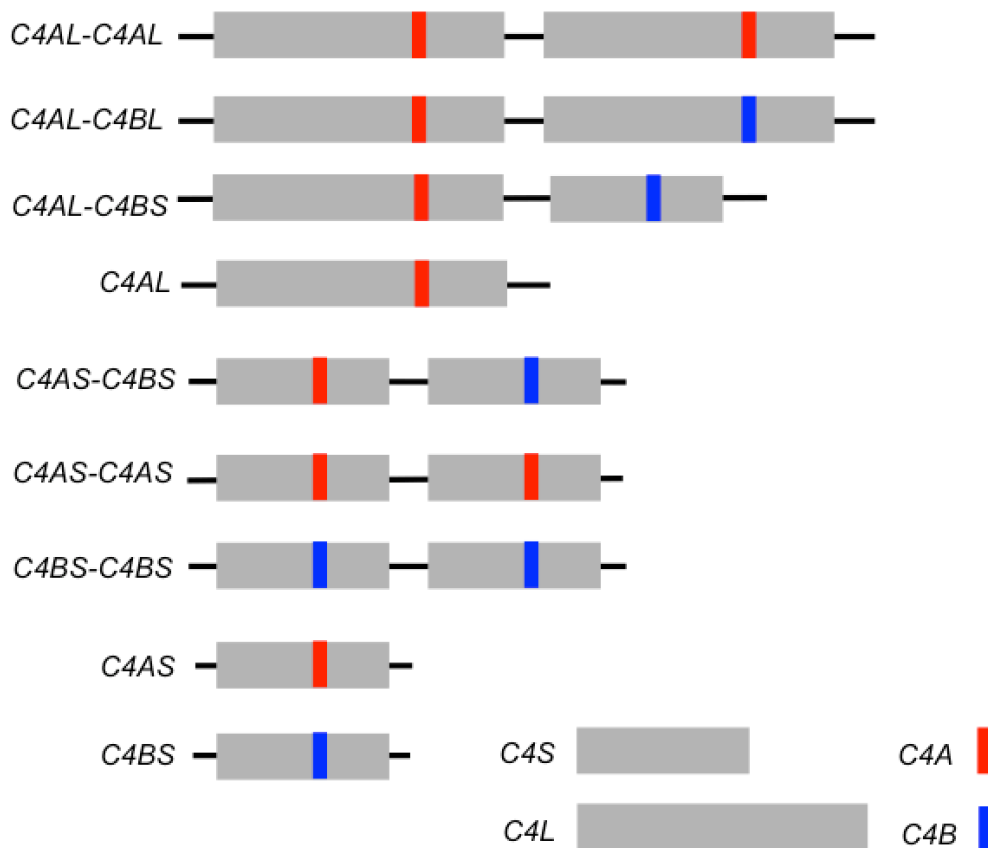

Figure 2.5. The *C4* gene is highly structurally polymorphic in humans. The length polymorphism in the gene is indicated with a short (*C4S*) or long (*C4L*) grey bar. The nucleotide differences for the *C4A* haplotype are shown in red and the nucleotides for the *C4B* allele are shown in blue.

20

While variation in the length polymorphism is present in most Old World monkeys and the great apes, the A/B paralog differences are present in apes and both Old World and New World monkeys[4] (Figure 2.6). The extent to which *C4* copy number varies in the great apes is unknown. In the following chapters we will explore how the frequencies of *C4* structural haplotypes vary in different non-human primate species, observing structural evolution on a macroevolutionary scale.

	A	B	L	S
Human	+	+	+	+
Chimpanzee	+	+	-	+
Gorilla	+	+	-	+
Orangutan	+	+	+	+
African Green Monkey	+	+	+	+
Cotton-top Tamarin	-	+	-	+

Old World: Human, Chimpanzee, Gorilla, Orangutan, African Green Monkey

New World: Cotton-top Tamarin

Figure 2.6. Great ape and monkey species are displayed with information about polymorphisms in the *C4* gene. A "+" indicates that the species has a certain variant, while a "-" indicates that the species appears to lack the variant.

In addition to gaining insight into how structurally complex regions evolve across species, more extensive *C4* typing in primates is relevant to the association of *C4* to human phenotypes. Work by Aswin Sekar and others in the McCarroll Laboratory has shown that structural variants in the *C4* gene are associated with lupus erythematosus (SLE) and schizophrenia (In preparation). Specifically, increased *C4A* copy number leads to an increase in *C4A* expression in lymphoblastoid cell lines, and both the copy number and expression level of *C4A* are associated with decreased risk of SLE. Conversely, increased expression of *C4A* is associated with increased risk of schizophrenia. In the brain, increased copy number of both *C4A* and *C4L* increase the expression of *C4A,* and elevated copy number of both of these variants associates to increased risk of the disease.

The *C4* haplotype most highly associated with SLE is *C4BS* and the haplotype most commonly associated with schizophrenia is *C4AL-C4AL*. In the following chapters we will investigate the prevalence of these haplotypes in non-human primates.

Chapter 3

Designing and applying methods for typing structurally complex haplotypes

Current high-throughput methods for identifying structural variants cannot define structural haplotypes with multiple alleles

Current methods used to identify and type copy number variation such as qPCR, array CGH, and fluorescent in situ hybridization (FISH) are inadequate for studying complex structural variation. Modern genotyping arrays often type both SNPs and CNVs genome-wide[7]. While simple duplications and simple deletions can be readily interpreted from this data, other regions are more complex and the precise series of mutations contributing to the variation are not obvious (Figure 3.1).

Because multiple and often overlapping structural variants segregate in structurally complex regions, individual polymorphisms can be difficult to tease apart. Due to the difficulty of typing structurally complex loci, they are largely omitted from genomics literature.
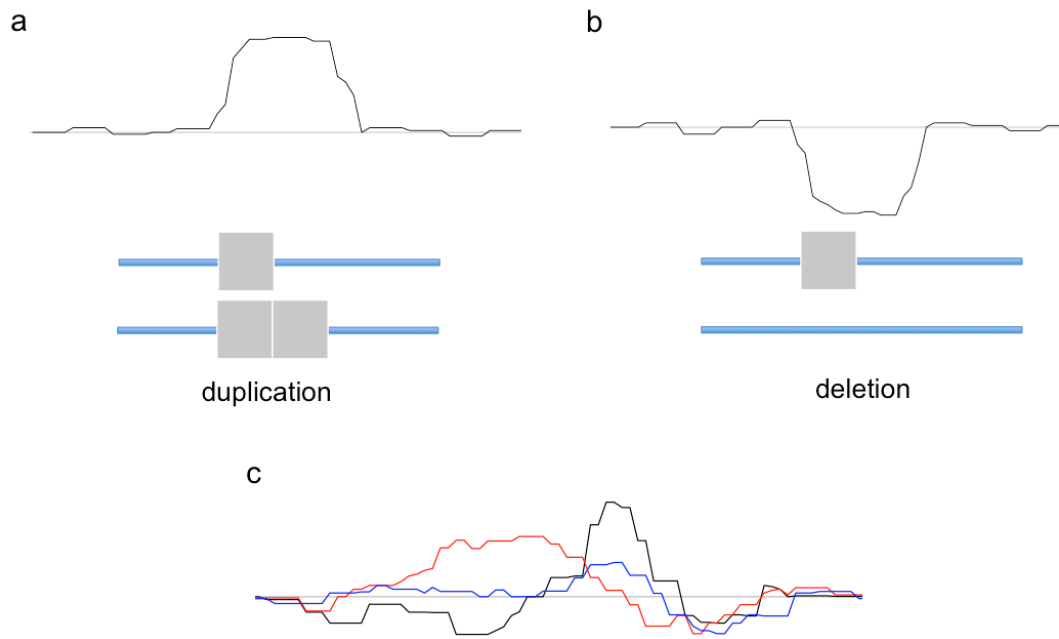
Figure 3.1. Structural polymorphisms depicted by normalized array probe intensity. (a) A simple duplication can be inferred from increased array probe intensity. (b) A simple deletion can be interpreted from decreased probe intensity. (c) Probe intensity differences can have a series of overlapping boundaries from which alleles are difficult to determine.

High copy number is common in structurally complex regions and is difficult to type

Another challenging aspect of structurally complex regions is that they often contain high copy number. Initially, we attempted to measure copy number of discrete regions within the 17q21.31 region using qPCR (Figure 3.2) and two-dimensional summarized array probe intensity (Figure 3.3); however, neither of these methods was precise enough for high-confidence genotyping at high copy number. Additionally, both these methods only provide relative copy number information such that many individuals need to be run at once.

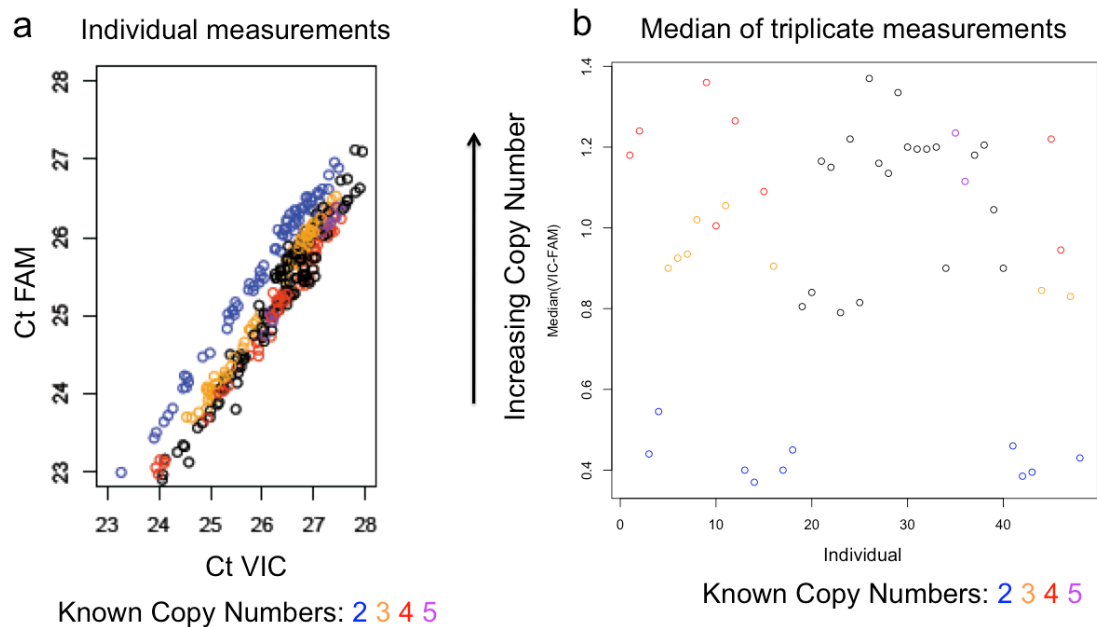Figure 3.2. Copy number genotyping with qPCR. Known copy numbers are plotted in a specific color while unknown values are plotted in black. (a) The threshold cycle (Ct) for two fluorophores (FAM and VIC) is shown for each individual. Copy number values of three and higher are difficult to determine. (b) The median value of VIC Ct minus the FAM Ct is plotted. Higher copy numbers cannot be determined with high confidence.
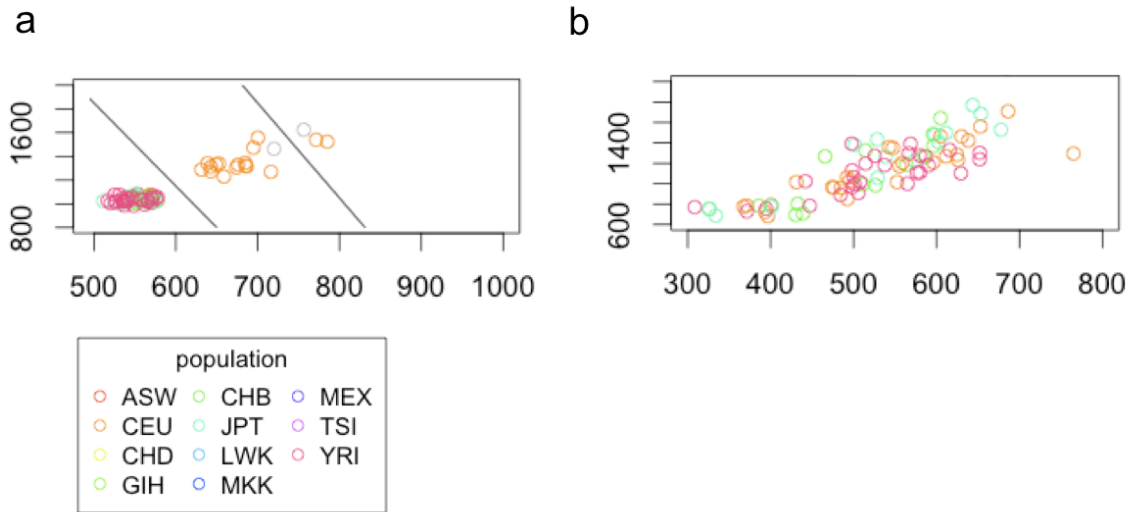
Figure 3.3. Copy number measurements from the 17q21.31 region using two-dimensional summarized array probe intensity. Each plot contains a set of individuals typed in the same genotyping run. The X axis displays probe intensities from the Illumina Human 1M array, while the Y axis shows intensities from the Affymetrix 6.0 array. The colors indicate which population each individual is from. (a) Summarized array probe intensity measurements for copy number genotypes 1-3. For copy numbers 1-3, most individuals localize to a specific cluster, allowing copy number determination. Grey circles indicate that the copy number could not be determined with high confidence (b) Summarized array probe intensity measurements for a region with high copy numbers. While there are clearly multiple copy number classes, a minority of copy number genotypes could be determined with high confidence. Population identifiers are defined in Supplementary Table 1.

A new method for understanding the variation in structurally complex regions

      In order to better understand the precise structural alleles segregating at structurally complex loci, we designed and published a set of methods that outlines a step-by-step approach[106]. Our approach involves

(i)   mapping the breakpoints for each individual structural variant using sequence and array data

(ii)   measuring copy number of each variant either by applying the Genome STRiP algorithm to whole genome sequence (WGS) read depth[107] or measuring copy number with a droplet-based digital PCR platform (ddPCR)[108]

(iii)   using patterns of inheritance in trios and statistical phasing in populations to determine which structural features segregate together on a chromosome.

These methods are readily scalable, enabling us to examine large panels of individuals.

Breakpoint mapping

      The strategy for mapping common breakpoints for structural variants depends on the available data and complexity of the locus. The general region for a structural breakpoint can be found through next generation sequencing and probe intensity by comparing individuals of high copy number to individuals of low copy number. Precise breakpoints can be found in non-aligning reads that are split-reads, containing the actual structural breakpoint. More complex regions

may require cloning and long-read sequencing in multiple individuals. The methods used to define structural variants for each of the four regions examined in this thesis are discussed below.

Copy number genotyping

Copy number genotyping with droplet-digital PCR

While qPCR and array probe intensity measurements were unsuccessful at typing high copy numbers, we became early adopters of droplet-digital PCR (ddPCR) technology, which allows for more precise copy number measurements.

Briefly, each experiment involves simultaneous interrogation of the CNV locus and an invariant two-copy control locus (Figure 3.4). For each locus, we designed a pair of PCR primers and a dual-labeled fluorescence/FRET oligonucleotide probes. Twenty microliter PCR reactions are compartmentalized through emulsion into approximately 20,000 droplets such that each droplet contains zero, one, or very few template molecules from each locus. PCR was performed on these emulsions and analyzed to count the droplets that were positive and negative for each fluorophore. By comparing this molecular count between locus X and the control, two-copy locus Y, the integer number of copies of locus X in each diploid genome was evaluated.
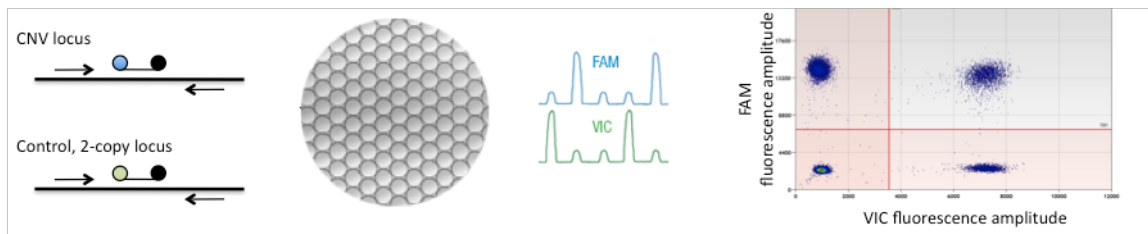
Figure 3.4. Copy number genotyping using the droplet-digital PCR method. The left panel shows that two assays are designed: the first to a CNV locus and the second to a two-copy control locus. These reactions are placed inside droplets and PCR amplification is performed. Each droplet is read for fluorescence in the FAM and VIC channels. A given droplet could be positive for FAM only, positive for VIC only, positive for both, or negative for both types of fluorescence.

Copy number genotyping with WGS read depth (Genome STRiP)

As a second method for determining the integer copy number of CNV segments in populations, we generalized the Genome STRiP genotyping method[107] to analyze duplications in low coverage sequencing data from the 1000 Genomes Project, Phase 1[60].

Briefly explained, for each CNV segment, the number of observed sequenced reads falling within the CNV segment was counted for each sample, and compared to the expected number of fragments per copy at the locus (Figure 3.5). The expected number of fragments was estimated based on the genome-wide sequencing coverage, correcting for the alignability of the segment and for sequencing bias based on the GC content of the segment.
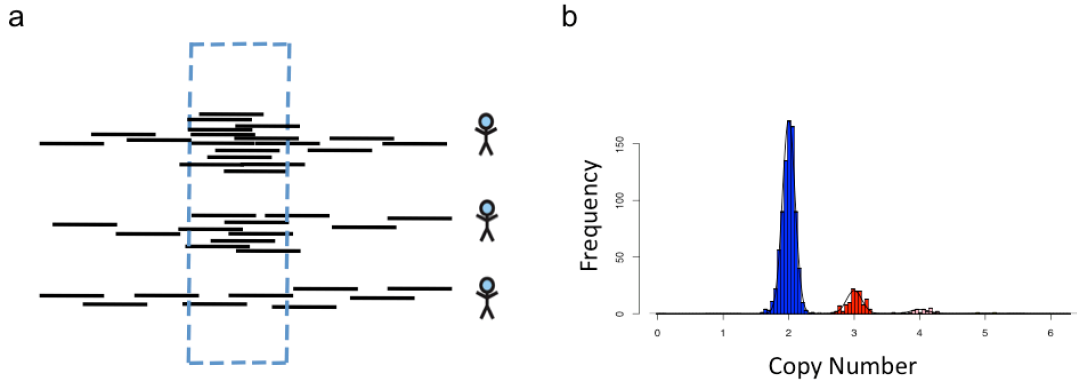
Figure 3.5. Genome STRiP copy number genotyping utilizes WGS read depth.
(a) Genome STRiP counts the number of observed reads for CNV locus and compares this to expected number based on normalized genome sequence coverage. (b) A Gaussian mixture model is used to account for statistical sampling and to determine copy number.

Using trio inheritance to determine allelic copy number

We addressed the difficulty of inferring haplotypic contributions to diploid copy number using inheritance patterns in trios (Figure 3.6) and employing a joint maximum-likelihood analysis of genotypes and allele frequencies when trio data was not available or informative. We considered all possible combinations of integer copy number (paternal transmitted, paternal untransmitted, maternal transmitted, maternal untransmitted) that were consistent with the diploid copy-number measurements from all three trio members from ddPCR or WGS read depth.

31
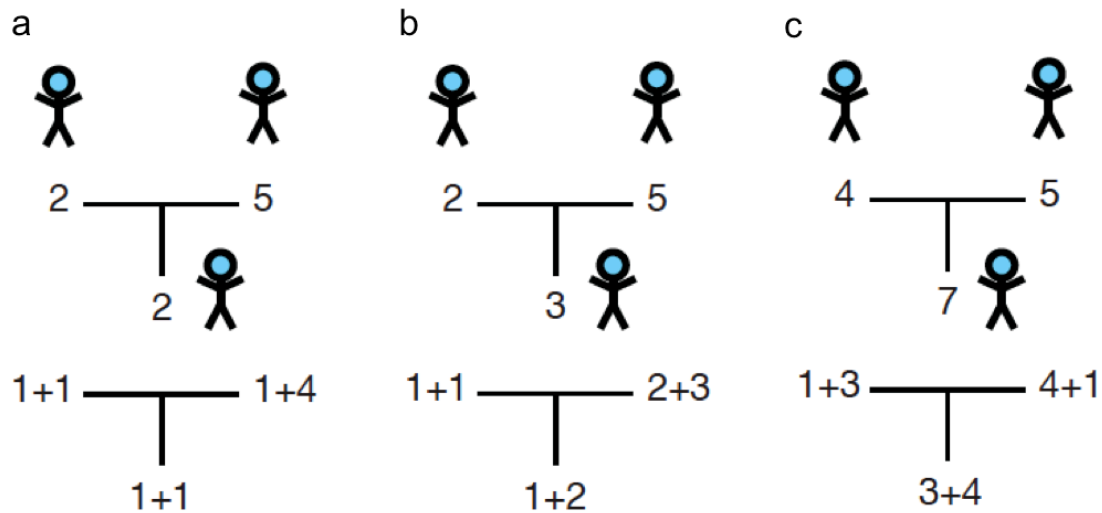
Figure 3.6. Examples of trios informing on haploid contribution to diploid copy number. Trios composed of father, mother and one offspring were used to determine the allelic copy numbers segregating in the population. (a) The father's total copy number is two and the mother's total copy number is five. Their child has a diploid copy number of two indicating that he inherited an allele of copy number one from each parent. (b) In this case a copy number one allele was inherited from the father and a copy number two allele was inherited from the mother. (c) In this trio a copy number three allele was inherited from the father and a copy number four allele was inherited from the mother.

In summary, our strategy for typing variation in structurally complex regions involves (1) determining breakpoints for discrete structural variants using array or sequence data (2) typing these discrete segments with ddPCR or WGS read depth and (3) determining which structural features segregate together through inheritance patterns in trios. We used variations on these three basic

steps to type the complex structural variation in the 17q21.31, *HP*, *HPR* and *C4* loci.

Typing 17q21.31 structural variation

There is a well known megabase-long inversion polymorphism in the 17q21.31 region[41], but there are also multiple uncharacterized copy number variants. To understand the structural alleles segregating in this region, we sought to identify and type all structural variants and determine how they relate to one another.

To identify the genomic span of each CNV in the 17q21.31 region, we used a combination of array and sequence data to pinpoint the breakpoints of each CNV. As a first step, we identified (at kilobase resolution) the estimated span of CNV segments using array-based data, and these were further refined by comparing read-depth profiles. This analysis used low-coverage sequence data from the 1000 Genomes Project, pooled across individuals with shared high or low copy number and then compared between these two groups. Searching the 1000 Genomes data for split reads identified precise breakpoints of these rearrangements.

We used several specific SNPs that serve as tags for orientation to determine the inversion state (H1 or H2). Through plotting copy number dosage for CNV Regions 1 and 2 in the context of the inversion state, we were able to discern two distinct yet overlapping duplications, which we named α and β. (Figures 3.7, 3.8). CNV Region 3 contains a non-overlapping section of a third duplication, which we named γ. (Figure 3.8).

33
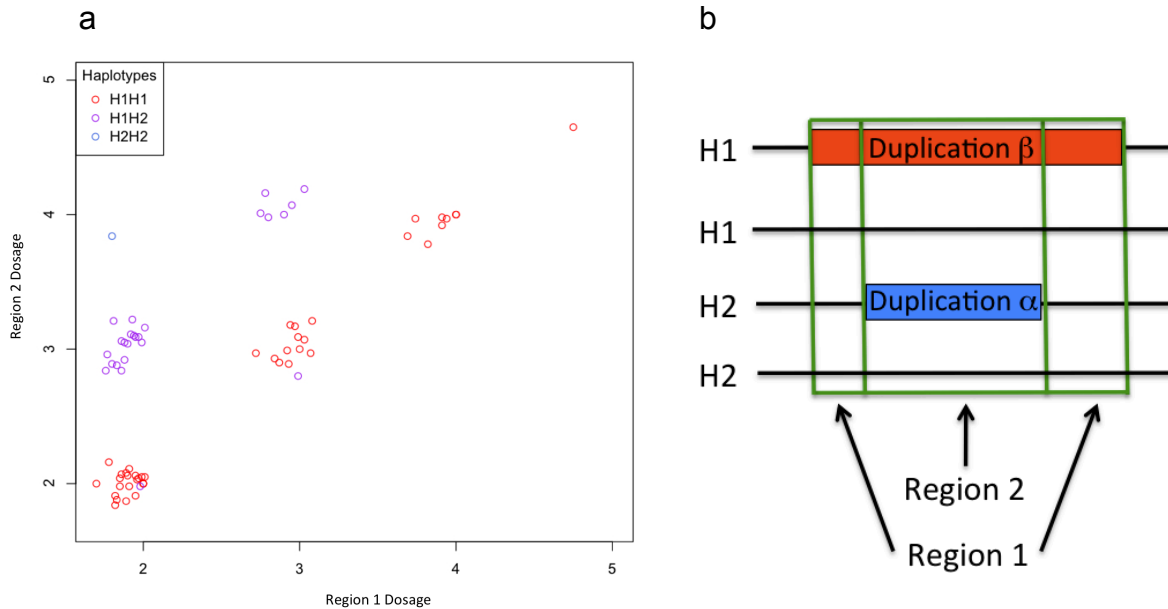
Figure 3.7. Relating the inversion to duplications. Comparison of copy number in Region 1 and Region 2 in the context of H1/H2 haplotype orientation makes it clear that copy number is influenced by two separate duplication polymorphisms segregating with different inversion states. Three main observations underlie this determination. (1) Copy number of Region 1 was always equal to copy number of Region 2 in H1 homozygotes. (2) Copy number in Region 2 was (in all but two cases, described below) equal to the sum of (i) copy number of Region 1 and (ii) the number of H2 haplotypes that an individual carries. This indicated that a duplication of Region 2 is present on most H2 haplotypes. (3) Copy number in Region 1 was only greater than two in individuals with at least one H1. These data indicated the existence of a long duplication which overlaps Region 1 and Region 2 segregating on the H1 background, and a shorter duplication overlapping Region 2 segregating on the H2 background.

Figure 3.8. (a) Assays for Regions 1, 2, and 3 were designed to target copy

number of α β and γ. Copy number of three copy number–variable segments of

17q21.31 was measured in populations using two approaches: analysis of read

depth in WGS libraries available for 942 individuals from the 1000 Genomes

Project Phase 1, which we applied to measure copy number of Region 1, Region

2 and Region 3, and a ddPCR approach, which we applied to analyze parent-

offspring trios from HapMap at specific sites within Region 1, Region 2, and

Region 3. Copy number of three copy number–variable segments of 17q21.31

(part a) was measured in populations using two approaches: analysis of read

depth in WGS libraries available for 942 individuals from the 1000 Genomes

Project Phase 1, which we applied to measure copy number of Region 1 (b),

Region 2 (c) and Region 3 (d), and a ddPCR approach, which we applied to

analyze parent-offspring trios from HapMap at specific sites within Region 1 (e),

Region 2 (f) and Region 3 (g). (Note that the frequencies of these copy-number

classes are not identical in b–d and e–g, as their frequencies stratify by

population, and the samples surveyed only partially overlap.) (h–j)

Determinations of copy number were concordant for genomes analyzed by both

methods in Region 1 (h), Region 2 (i) and Region 3 (j).
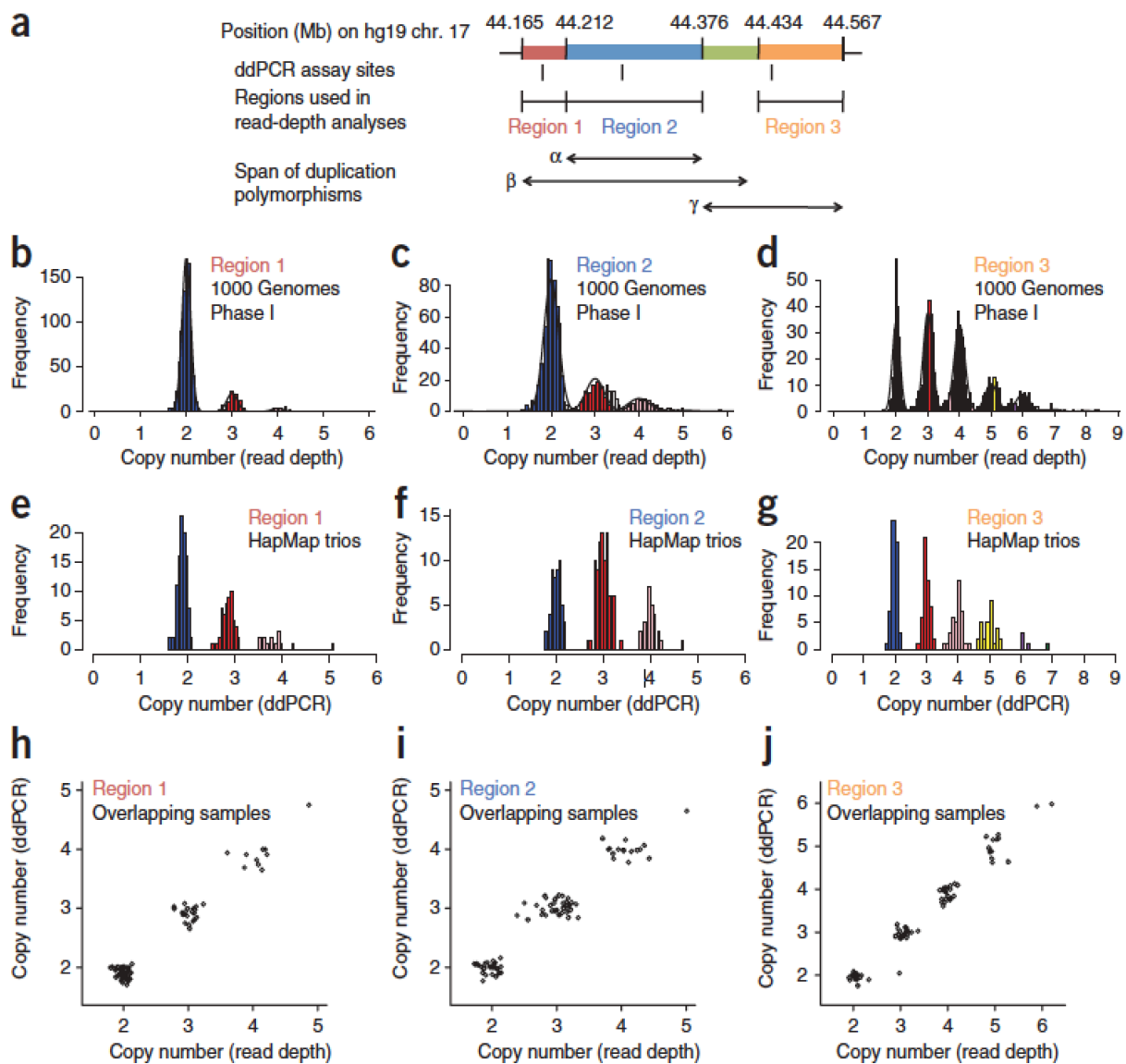
Figure 3.8 (continued)

In order to determine which structural features segregate together on a haplotype, we used inferences from inheritance patterns in trios and maximum likelihood analysis. While most haplotypes were decipherable using trios alone, others were estimated with a joint maximum-likelihood analysis of genotypes, allele frequencies and inheritance patterns in trios. Each population was analyzed separately. We considered all possible combinations of integer copy number (on each of the four haplotypes in a trio: paternal transmitted, paternal untransmitted, maternal transmitted, maternal untransmitted) that were consistent with the diploid copy-number measurements from all three trio members from ddPCR.

Using the copy number calls for each region, inheritance in trios to determine alleles, and inversion defining SNPs, we determined the haplotypes segregating in three populations (Figure 3.9). We learned that the numerous structural features of 17q21.31 segregate in a limited number of combinations, or nine structural haplotypes.

We found that the 17q21.31 inversion region contains a structural form with sequence in the standard orientation (*HP1*) and two segregating tandem duplications (β and γ). We also determined that the inverted form (*HP2*) has two duplications segregating (α and γ); however, these duplications are dispersed on the *HP2* form (Figure 3.9).

	YRI	CHB + CHS	CEU
H1.β1.γ4	0.01	0.08	0.01
H1.β1.γ3	0.07	0.31	0.17
H1.β1.γ2	0.49	0.18	0.10
H1.β1.γ1	0.43	0.42	0.26
H1.β2.γ1			**0.25**
H1.β3.γ1			0.01
H2.α1.γ2			0.01
H2.α2.γ2			**0.18**
H2.α2.γ1			0.01
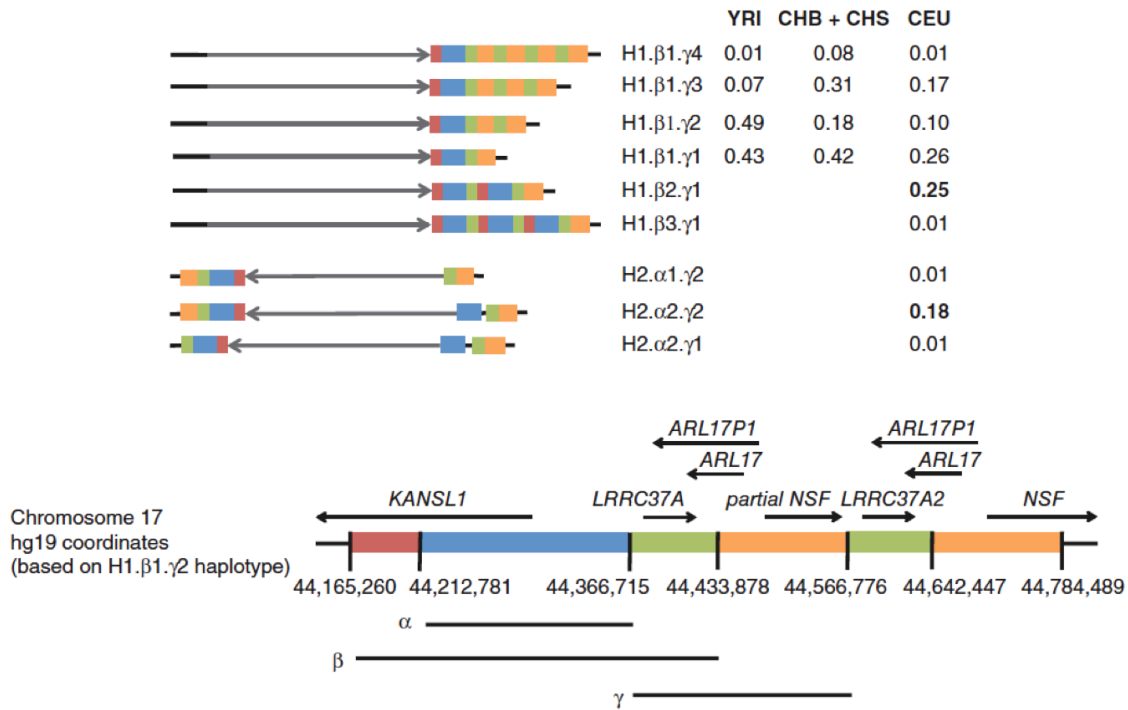
Figure 3.9. Structural forms of the human 17q21.31 locus and their population frequencies. Each haplotype is represented in a simplified form to highlight major structural differences. The schematic (bottom) indicates which genomic segment is represented by each color. The gray arrows indicate the orientation of the unique inverted region within 17q21.31. The α, β, and γ structural polymorphisms segregate as the nine common haplotypes shown. The table (right) lists allele frequencies for the nine structural haplotypes in different populations. YRI, Yoruba in Ibadan, Nigeria; CHB, Han Chinese in Beijing; CHS, Han Chinese South; CEU, Utah residents of Northern and Western European ancestry.

The H2 structural form and the α and β duplications were absent from both the African and Chinese populations. Interestingly, the haplotypes with the α

and β duplications (H1. β 2.γ1 and H2. α 2.γ2) are common in the CEU population, but absent from the YRI, CHB, and CHS populations.

Typing *HP* structural variation

The structure of haptoglobin has been studied since the 1950s[56,58]; however, current strategies for typing *HP* structural variation are low throughput and often incomplete, omitting differences between the F and S forms. We aimed to develop methods to type all structural polymorphisms in the *HP* gene and understand their frequency distributions in different populations.

While the breakpoint for the *HP2* duplication was previously determined[2], we also sought to also type the F/S variant for each haplotype. We began by comparing the left (*HP-Left*) and right (*HP-Right*) copies of the *HP* duplication (as they differ in the F/S variant) and noticed that the copies of the duplication are highly diverged from one another: *HP2-Left* is more closely related to the chimpanzee *HP* sequence than it is to *HP2-Right* (Figure 3.10a). This is an abnormal pattern for a human-specific duplication. Further inspection revealed that the *HP2-Left* and *HP2-Right* divergence is confined to a 300 base pair region and has high homology to *HPR*. A sequence comparison of *HP-Left*, *HP-Right*, and *HPR* confirms that a majority of this segment of *HP2-Left* that overlaps exon 4 is nearly identical to *HPR* (Figure 3.10b).
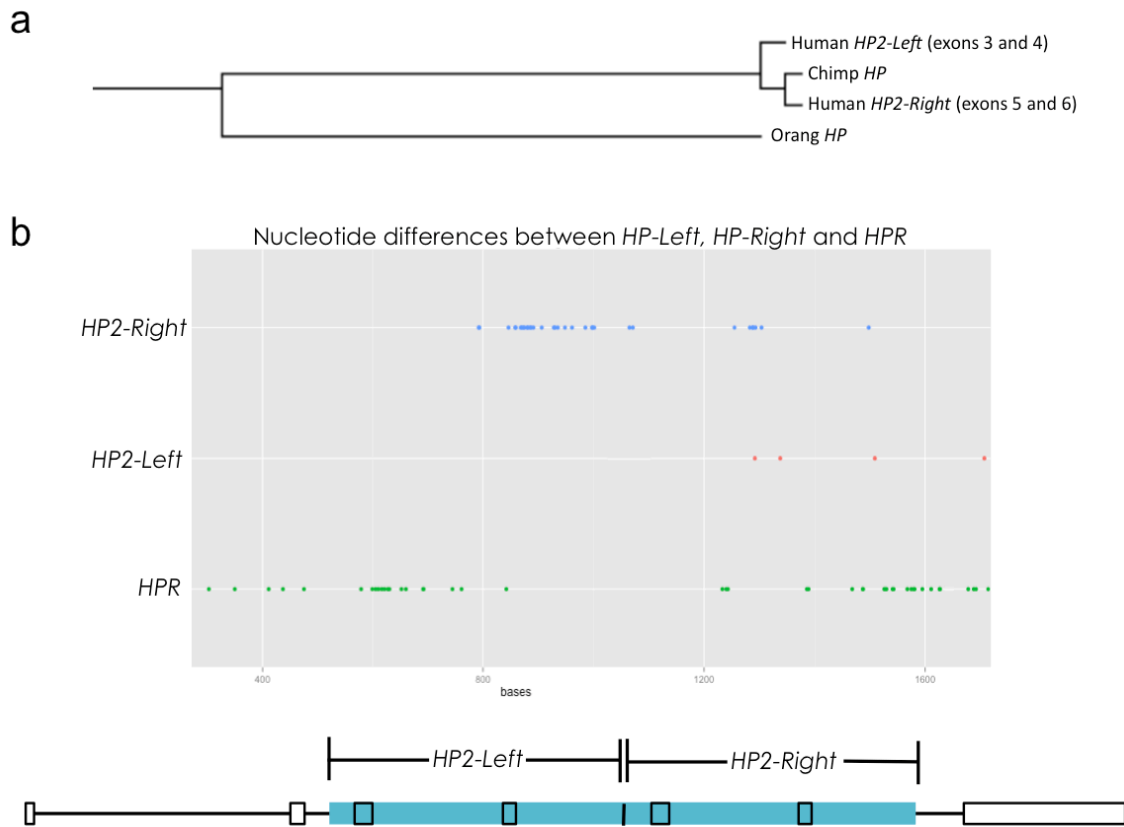
Figure 3.10. *HP2-Left* and *HP2-Right* are highly diverged. (a) A phylogeny constructed from regions homologous to the *HP* duplication in various species and paralogs in respective reference sequences. This phylogenetic tree shows that *HP2-Right* is more closely related to chimpanzee *HP* than it is to *HP2-Left.* (b) Representation of an alignment that identifies the non-matching base. Dots on this plot identify the outlier sequence that does not match the other two. The blue dots indicate a region in which *HP2-Left* and *HPR* match each other, while *HP2-Right* is different. The red dots indicate locations of bases where *HPR* and *HP2-Right* have the same base, while *HP2-Left* is different. The green dots indicate the bases at which *HP2-Left* and *HP2-Right* have the same base and *HPR* is different. The bottom diagram demonstrates the location of *HP2-Left* and *HP2-Right* within the *HP* gene.

40

We examined the ancestral state of *HP* in this region by sequencing the *HP* gene of several great apes (Figure 3.11, Supplementary Figure 1) and interpreted that a large segment of the differences between *HP2-Left* and *HP2-Right* were due to derived paralogous gene conversion from the nearby *HPR* gene, while another segment was highly diverged between *HP2-Left* and *HP2-Right*, but does not result from a recent paralogous gene conversion event as each haplotype contains a mix of derived and ancestral variants (Figure 3.11, Appendix Figure 1).

Next, we sequenced the *HP* structurally variant region of twenty-eight haplotypes (Supplementary Figure 2) from diverse SNP haplotype backgrounds determined by SNPs from the 1000 Genomes Project low coverage sequence. We found that the *HPR* paralogous gene conversion (first observed on *HP2FS*) is also present on *HP1F* (Figure 3.11) and responsible for the "fast" amino acid changes on both haplotypes. Additionally, we also found an *HP2* form lacking the *HPR* gene conversion, which we call *HP2SS,* because it has two of the slow forms (3.11).
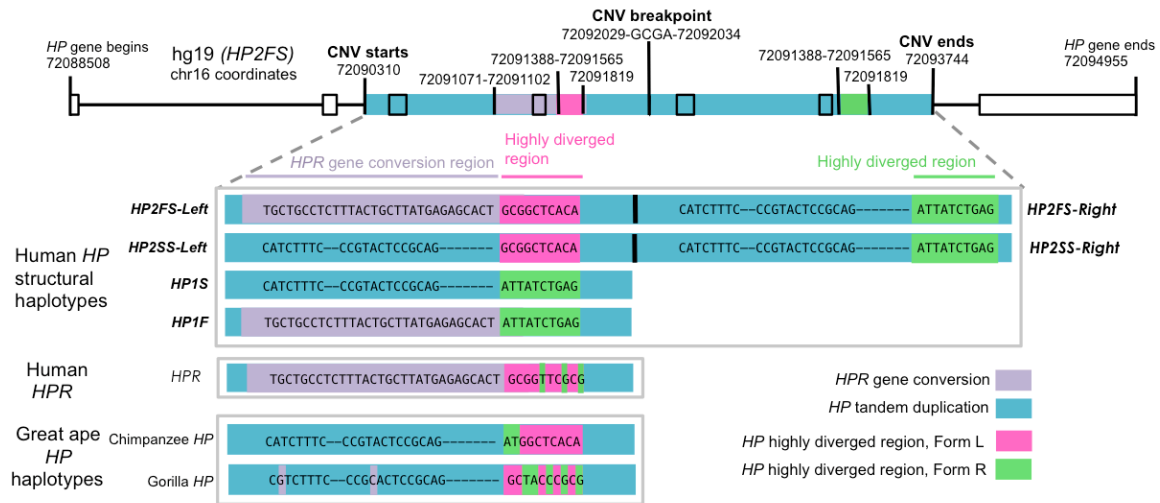
Figure 3.11. *HP* haplotypes contain paralogous gene conversion from *HPR* and a highly diverged region. (a) The polymorphic bases which define each region are shown. Haplotypes of *HP* in humans, human *HPR* and great ape *HP* are compared and the coordinates of each region are shown for hg19, which corresponds to *HP2FS*. The region of the duplication is shown in blue. The *HPR* gene conversion region in shown in lavender, while the non-gene converted form remains shown in blue. Form L of the highly diverged region is shown in pink, while Form R of the highly diverged region is shown in green. The black boxes indicate the locations of exons. *HPR* gene conversion is derived in human *HP2FS* and *HP1F*. The ancestry of the highly diverged region is unclear. See Supplementary Figures 1 and 2 for complete alignments.

In order to type these polymorphisms at the population scale, we designed ddPCR and PCR assays to each regional breakpoint including the *HPR* gene conversion and the highly diverged region (Figure 3.12a). We typed these polymorphisms in 589 individuals from four populations (Figure 3.12b), and

42

verified that the haplotypes transmit faithfully in trios. All but three of the typed

individuals were homozygous for one of the four structural alleles shown above.

The exceptions were a triplication variant, and deviations from the common

breakpoints for the *HPR* gene conversion and the highly diverged region.
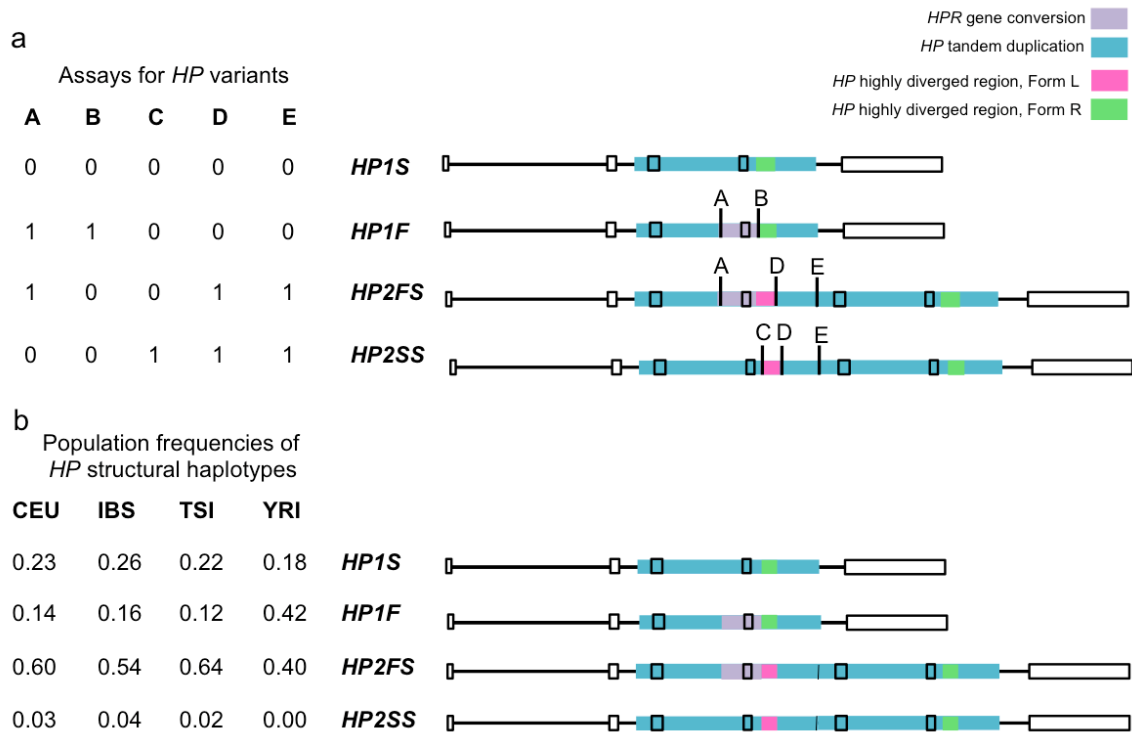
Figure 3.12. Typing breakpoints for common *HP* variation. The blue region represents the sequence that is duplicated. The lavender region is affected by paralogous gene conversion from *HPR.* The pink and green regions represent segments that are highly diverged between *HP2-Left* and *HP2-Right*. The black boxes indicate the locations of exons. (a) DdPCR and PCR assays (A-E) designed to the breakpoints of different *HP* variants. The table to the left lists the copy number result for each assay on each haplotype. For example, all five assays have a copy number of zero on the *HP1S* haplotype. For haplotype *HP1F*, the A and B assays yield copy number one, while the other assays yield copy number zero. (b) The frequency of each haplotype is indicated for each population. YRI, Yoruba from Ibadan, Nigeria; CEU, Utah residents of Northern and Western European ancestry; IBS, Iberians from Spain; TSI, Tuscans from Italy.

44

The frequency of the four major *HP* structural haplotypes is different between European and African populations. In fact, Europe and Africa have different major haplotypes (*HP2FS* in Europe and *HP1F* in Africa). Additionally, the *HP2SS* haplotype was not observed in Africa.

Typing *HPR* structural variation

It was documented in the 1980s that the *HPR* gene has segregating high copy number haplotypes in African Americans. Additionally, the breakpoints of this duplication had previously been observed through Sanger sequencing[95]. However, the extent to which this tandem duplication has expanded and the duplication's distribution throughout Africa was not well understood.

We determined the *HPR* copy number using both ddPCR and WGS read depth in the YRI and LWK populations. We found that our copy number calls were 92% concordant overall and 68% concordant for high copy numbers between these two methods. Disagreements between the two methods were resolved with triplicate ddPCR measurements.

We then used trios to determine haploid copy number for each individual. Previously, copy numbers of up to five on a single haplotype had been published, but our results show haplotypes with up to eight tandemly arranged copies of *HPR* (Figure 3.13a). Interestingly, we observe that the distribution of *HPR* copy numbers is bimodal: the most frequent copy number class is one followed by five (Figure 3.13b).

45

a   **Frequency in YRI**

0.85	*HPR.1*	
0.2	*HPR.2*	
0.3	*HPR.4*	
0.07	*HPR.5*	
0.02	*HPR.6*	
0.01	*HPR.7*	
<0.01	*HPR.8*	

HPR gene

b
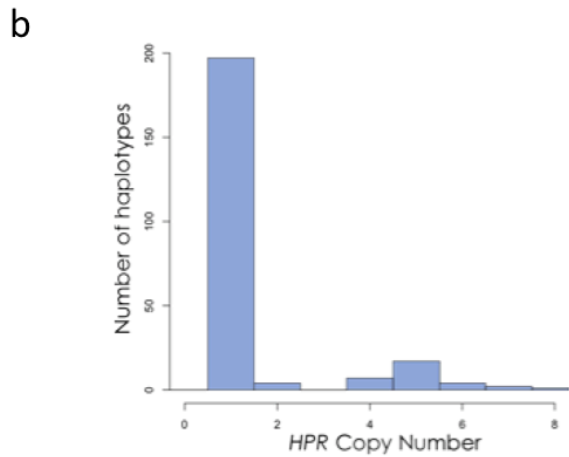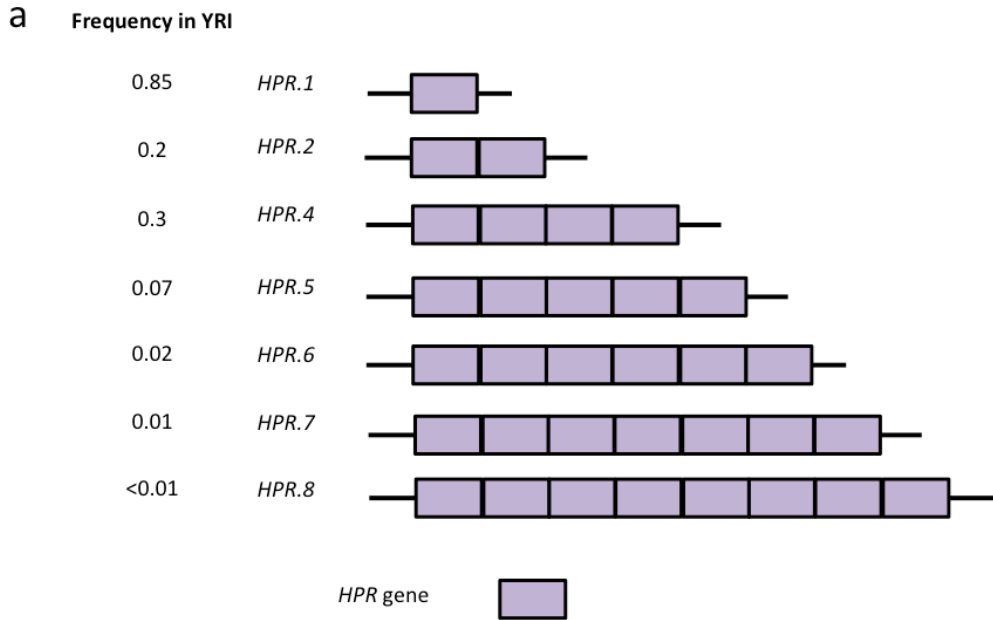
Figure 3.13. (a) The structural haplotypes of *HPR* are shown along with the frequencies determined. Tandemly arranged *HPR* genes are shown in purple. (b) The frequency of specific *HPR* copy numbers is plotted. The most common haplotype has copy number one and the second most common has copy number five. The distribution of copy numbers is bimodal.

46

Typing *C4* structural variation

      Complement component 4 (*C4*) is a gene within the *HLA* region that contains complex structural variation in both humans and non-human primates. Sequence variants define two paralogous forms of this gene (*C4A* and *C4B*), which bind either proteins (*C4A*) or carbohydrates (*C4B*)[101]. Additionally, there is a length polymorphism distinguished with the names *C4L* (long) and *C4S* (short).

      Previous research in a small number of individuals has shown that these structural polymorphisms also vary in other primates[103]. To understand how this structure has evolved in different species of great ape, we applied our methods for typing structurally complex loci to *C4* in several populations of humans and chimpanzees, as well as to other primates: bonobos, gorillas, orangutans. While five fixed nucleotide variants differentiate *C4A* from *C4B* in humans, these can differ among primates. We chose to define the *C4A/C4B* difference by a single base that governs function (protein or carbohydrate binding). Sanger sequencing was performed to identify the sequence differences between *C4A* and *C4B* defining bases in each species of great ape (Figure 3.14).

Human: *C4A*: GGC-TCG-TTC-CAG-GAC-C**C**C-T**G**T-CCA-GTG-**TT**A-**G**AC-AGG-AGC-ATG
C4B: GGC-TCG-TTC-CAG-GAC-C**T**C-T**C**T-CCA-GTG-**AT**A-**C**A**T**-AGG-AGC-ATG

Chimp: *C4A*: GGC-TCG-TTC-CAG-GAC-C**T**C-T**G**T-CCA-GTG-**TT**A-**G**AC-AGG-AGC-ATG
C4B: GGC-TCG-TTC-CAG-GAC-C**T**C-T**C**T-CCA-GTG-**AT**A-**C**A**T**-AGG-**G**GC-ATG

Bonobo: *C4A*: GGC-TCG-TTC-CAG-GAC-C**T**C-T**G**T-CCA-GTG-**TT**A-**G**AC-AGG-AGC-ATG
C4B: GGC-TCG-TTC-CAG-GAC-C**T**C-T**C**T-CCA-GTG-**AT**A-**C**A**T**-AGG-**G**GC-ATG

Gorilla: *C4A*: GGC-TCG-TTC-CAG-GAC-C**C**C-T**G**T-CCA-GTG-**TT**A-**G**AC-AGG-AGC-ATG
C4B: GGC-TCG-TTC-CAG-GAC-C**T**C-T**C**T-CCA-GTG-**AT**A-**C**A**C**-AGG-**G**G**A**-ATG

Orangutan: *C4A*: GGC-TCG-TTC-CAG-GAC-C**T**C-T**C**T-CCA-GTG-**TT**A-**G**AC-AGG-AGC-ATG
C4B: GGC-TCG-TTC-CAG-GAC-C**T**C-T**C**T-CCA-GTG-**AT**A-**C**A**C**-AGG-AGC-ATG
C4B': GGC-TCG-TTC-CAG-GAC-C**T**C-T**C**T-CCA-GTG-**T**TA-**C**A**C**-AGG-AGC-ATG

↑

This base is thought to cause the functional differences between *C4A* and *C4B*

Figure 3.14. Defining sequences for *C4A/C4B* in great apes. Nucleotides that define the difference between *C4A* and *C4B* in humans are colored in either red (*C4A*) or blue (*C4B*). Bases that differ between non-human primates and humans are printed in bold and underlined. *C4B'* was defined for the orangutan due to variation among its *C4B* sequences.

DdPCR assays were designed to *C4A* and *C4B* and *C4S* for several great ape species. Because genomic variation is not well known among great apes, it was difficult to develop a control assay in which we had confidence that there was neither copy number variation nor nucleotide variation. We ultimately designed an assay to a mammalian ultra-conserved element (UCE) identified in Derti et al. 2006[109]. Both the nucleotide sequence and copy number of ultra-conserved elements are highly conserved between species and these sequences play a role in chromosomal pairing during meiosis. We chose a specific region of

48

a UCE (chr12:106976559-106976654, hg19) that is identical and single copy in the reference sequences of humans, the great apes, dogs, horses and other mammals.

Using these assays we typed twenty western chimpanzees (*Pan troglodytes verus*), four central chimpanzees (*Pan troglodytes troglodytes*), eight bonobos (*Pan paniscus),* and three gorillas *(Gorilla gorilla*) (Figure 3.15).
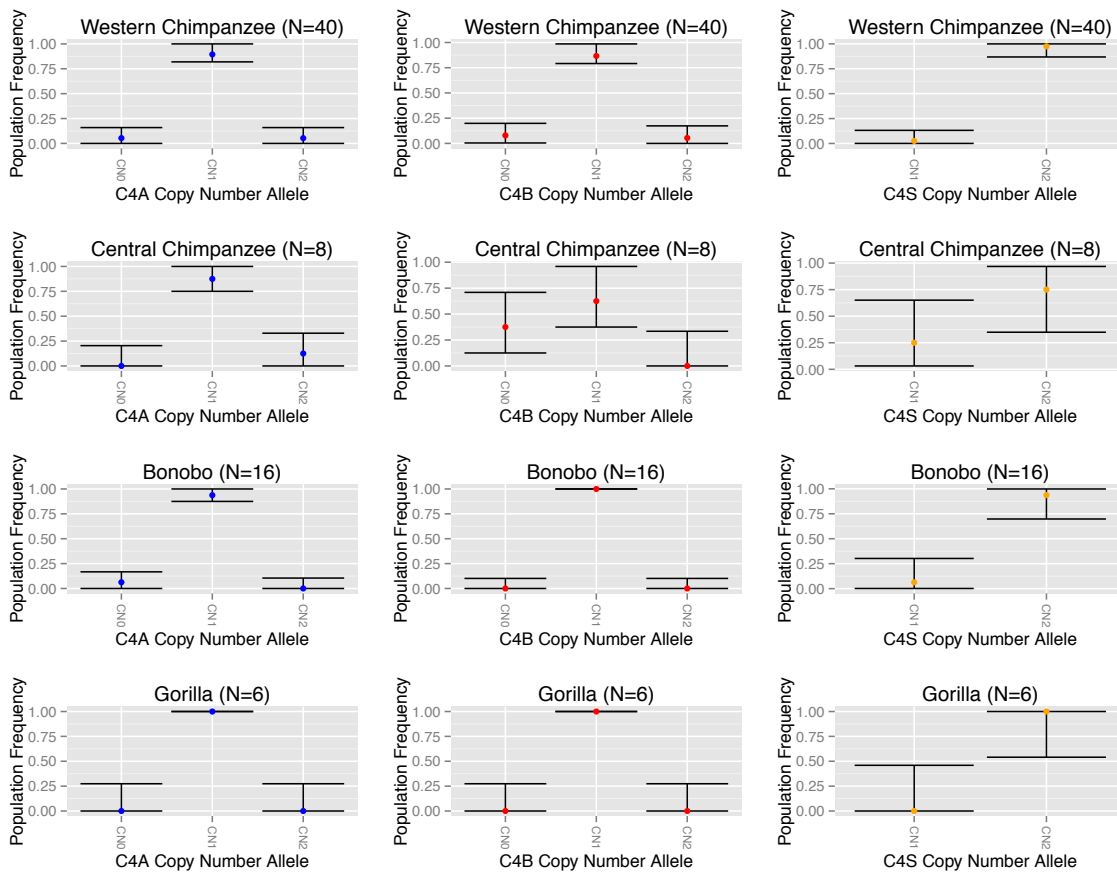
Figure 3.15. The frequencies of copy numbers for *C4A* (left column), *C4B* (center column) and *C4S* (right column) are plotted for different great ape populations and species: top row (western chimpanzee), second row (central chimpanzee), third row (bonobo), fourth row (gorilla).

While trios were not available for these great apes, we discerned the likely haplotypic confirmation of each structure based on frequency (Figure 3.16).

Haplotype model	Western Chimp	Central Chimp	Bonobo	Gorilla
C4A — C4B	0.875	**.625**	0.94	1.00
C4A — C4A	0.05	**0.125**	0	0
C4B — C4B	0.05	0	0	0
C4A	0.025	**0.25**	0	0
C4B	0	0	0.06	0

Figure 3.16. The frequency of inferred *C4* haplotypes for great ape species and populations. The most common haplotype in all species is *C4A-C4B*. Of the species and populations sequenced here, the Central Chimpanzees have the greatest amount of diversity at this locus.

Summary of typing structurally complex regions

The 17q21.31 region, *HP*, *HPR* and *C4* contain very different forms of complex structural variation, yet all were able to be typed (Figure 3.17) on a large scale using three basic steps: (1) mapping breakpoints for each discrete structural variant, (2) typing the copy number of each variant using ddPCR or WGS read depth, and (3) determining which variants segregate together on a haplotype using trio inheritance and population frequency information.

50
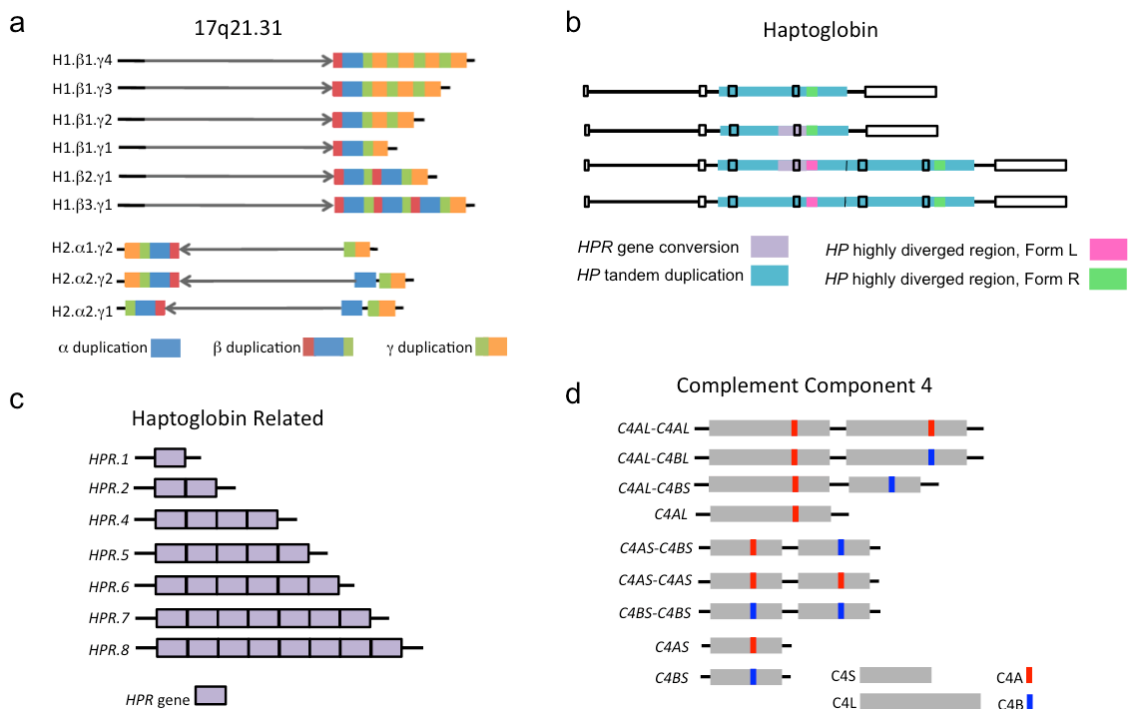
Figure 3.17. Structural haplotypes at four structurally complex regions. Structures are identified in legends below each figure. (a) Haplotypes at 17q21.31. The orientation of the arrow indicates the standard (H1) orientation, or the inverted (H2) orientation. The colored boxes indicate duplications. (b) The common structural haplotypes of *HP* are shown. Exons of *HP* are represented with by black boxes, while structural changes are indicated with color. (c) Seven structural haplotypes of *HPR* showing different numbers of tandemly arranged copies of the *HPR* gene. (d) The common human and great ape structural haplotypes of *C4* are shown. The long and short forms are distinguished by different length of grey rectangle, while the sequences differences (*C4A* vs. *C4B*) are indicated with colored boxes.

In the next chapter we will explore how these different forms of complex structural variation evolved by examining the SNP haplotype backgrounds on which each structural haplotype segregates.

Contributions

I designed the methods for typing complex structural variation under the guidance of Steve McCarroll. I designed all experiments and molecular assays, performed all molecular experiments, and did all data analysis in this chapter excluding the contributions of others mentioned below. The methods for typing complex structural variation were developed jointly with Aswin Sekar. Aswin Sekar provided advice for experiments involving the *C4* gene. Robert Handsaker performed the read depth analysis using Genome STRiP. Mike Zody analyzed clone sequence of the 17q21.31 to infer breakpoints and haplotypes of structural variation. David Reich provided some of the great ape DNA samples for the analysis of *C4*.

Chapter 4

Understanding how complex structure evolves within populations, across

populations, and across species

Introduction

Applying our methods for understanding structurally complexity across a diverse array of loci allows us to gain insight into the patterns of structural evolution within and between different genomic regions, populations, and species. While some structural changes are unique in our dataset such as the inversion on 17q21.31, other structural features are common to multiple loci such as multi-allelic CNVs as in *HPR* and 17q21.31.

In this chapter we will develop hypotheses about how the structure of each of these loci evolved. Specifically, we use information from the SNP haplotypes surrounding each structurally variant region to determine which structural haplotypes are more closely related.

17q21.31 structural evolution

While the frequencies for the different structural haplotypes in the 17q21.31 region vary widely between populations, we hypothesized that the structural diversity at 17q21.31 arose from a definable series of structural mutations; each mutation likely arose on a specific haplotype and may continue to segregate on that haplotype.

We analyzed the SNP haplotypes on which each 17q21.31 structural form segregates in European populations (Figure 4.1). The structural forms of 17q21.31 were strongly associated with SNP haplotypes on both sides of the distal end of the 17q21.31 inversion, where the polymorphic CNV copies reside (Figure 4.1).

Figure 4.1. Structural forms of 17q21.31 segregate on specific SNP haplotype backgrounds. The plot shows homozygosity and divergence (due to mutation and recombination) of the SNP haplotypes on which each structural form segregates in the European (CEU) trios analyzed in HapMap 3. The polymorphic CNV copies at the right end of the 17q21.31 inversion reside between the two origins of this plot (center). SNPs on the left half of the plot therefore reside within the unique inverted region of 17q21.31, whereas SNPs on the right half of the plot are distal to the 17q21.31 inversion. Branch points represent markers at which the depicted haplotypes diverge due to mutation and/or recombination with other haplotypes. In the plot, the structures are represented on the leaves in order to clarify their relationships to SNP haplotypes, but the variable parts of these CNVs actually reside (in genomic space) within the gap at the center between the two origins on the plot. The structural forms segregate on characteristic SNP haplotypes, both inside and outside the inversion region. Statistical imputation of structural alleles uses SNPs on both sides of the CNVs together with more distant markers not shown here.

Branches
(SNP haplotypes)
Allele frequency

1.0

0.5

0

Leaves
(structural isoforms)
H1.β1
H1.β2
H2.α1
H2.α2
γ = 1
γ = 2
γ = 3

Position (kb) relative to ends of CNVs at right end of 17q21.31 inversion
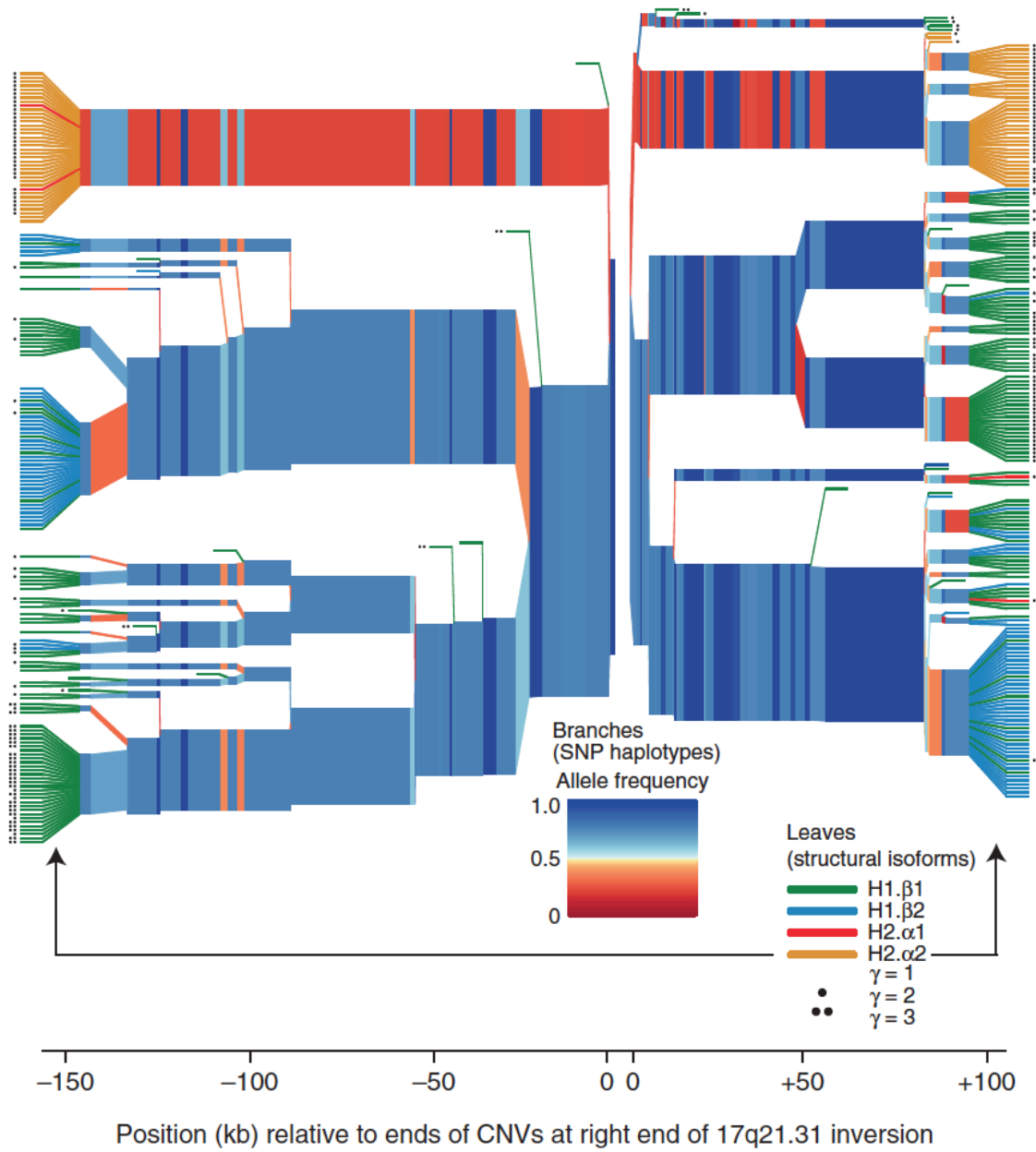
Figure 4.1 (continued)

Comparing 17q21.31 structural forms (Figure 3.9) and observing

similarities in SNP haplotype backgrounds for each structural haplotype (Figure

4.1) one can infer the likely mutational steps in the evolution of structure: a

duplication originates on a specific haplotype and then continues to expand on

that haplotype (Figure 4.2). We can see that the duplications existing in tandem

(on H1) have both expanded to become multi-allelic, while the dispersed

duplications (on H2) appear stable and remain biallelic (Figure 4.2). One

explanation for this phenomenon is that tandem duplications are predisposed to

further duplication events due to the proximity of homologous sequence enabling

non-allelic homologous recombination (NAHR)[110].

While most structural mutations likely are selectively neutral and may

expand or contract due to random chance, measures such as population

differentiation can be informative for inferring selection. For example, if a variant

is rare or absent in most populations, but at high frequency in a specific

population, this may indicate that the variant underwent a rapid rise in frequency,

which can be due to positive selection.

	YRI	CHB + CHS	CEU
H1.β1.γ4	0.01	0.08	0.01
H1.β1.γ3	0.07	0.31	0.17
H1.β1.γ2	0.49	0.18	0.10
H1.β1.γ1	0.43	0.42	0.26
H1.β2.γ1			0.25
H1.β3.γ1			0.01
H2.α1.γ2			0.01
H2.α2.γ2			0.18
H2.α2.γ1			0.01
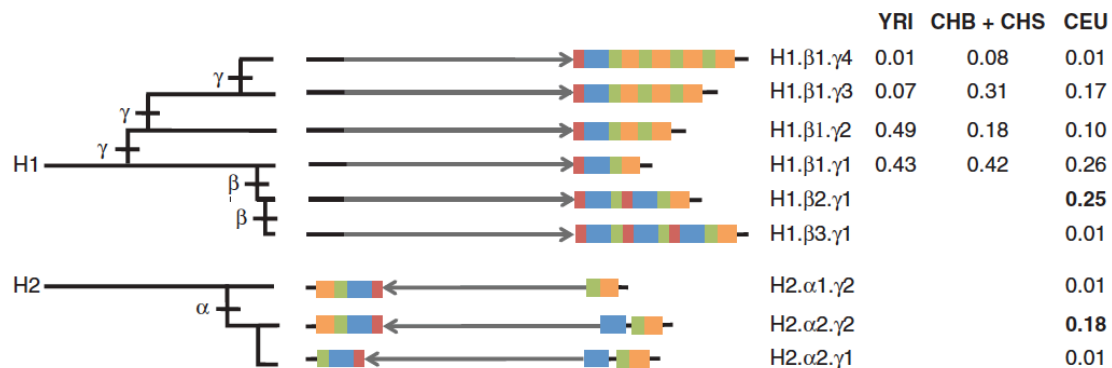
Figure 4.2. The evolution of 17q21.31 CNVs can be inferred as a series of

duplication expansions. The left panel shows a structural phylogeny, which

outlines a series of mutational steps that result in the present-day structural

haplotypes of 17q21.31. The frequency of each structural haplotype is displayed

for three populations.

While the H2 inversion is common in west Eurasians and rare in most other populations, which has been attributed to recent positive selection[1], we found that other structural variations at 17q21.31 show even greater population differentiation. Two distinct duplications (α and β), each affecting the 5′ coding exons of the *KANSL1* gene, have arisen independently on the H1 and H2 backgrounds and together comprise almost half of all forms of chromosome 17 in Europeans. In order to determine if the frequency increases of the α and β duplications are statistically significant, we calculated the fraction of SNPs that had a similarly high derived allele frequency in the European populations sampled (CEU, FIN, GBR, IBS, TSI); and that had a similarly low derived allele frequency across the non-European populations sampled (CHB, CHS, JPT, LWK, YRI) (Supplementary Figure 3).

We observed the α duplication segregating at an allele frequency of 19% in CEU and appearing only once among the 942 non-European chromosomes sampled by 1000 Genomes. For comparison, of 7,013 SNPs with allele frequency between 18% and 20% in Europe, only 37 SNPs (0.53%) were observed 0-1 times among the non-European population samples.

Evaluating the β duplication (allele frequency 26% in CEU, and not observed at all among the 942 CHB, CHS, JPT, LWK, and YRI chromosomes sampled in 1000 Genomes Phase 1), out of 6,127 SNPs with allele frequency between 25% and 27% in Europe, 4 SNPs (0.065%) were monomorphic in the non-European populations. These observations place both the α and β duplications among the human genome's most population-differentiated polymorphisms. The α and β duplications have reached these highly

differentiated allele frequencies in parallel at the same locus and in the same populations, in a pattern similar to that observed at other loci (such as the *LCT* and *APOL1* loci in African populations)[111,112] that have undergone recent positive selection.

We estimated two dates for each duplication: the time to coalescence of contemporary haplotypes and the age of the duplication events. The first can be estimated from the divergence of sequences flanking the duplications, the second by comparing the sequences of the duplication copies. To generate these data, we selectively captured and sequenced the 17q21.31 region in H1.β2 and H2.α2 homozygotes. We estimated the coalescence of the sampled β-duplicated H1 chromosomes at 12,000 years ago. Divergence of otherwise unique sequences within the β duplication suggests that the duplication itself occurred 20,000–27,000 years ago. For the α-duplicated H2 chromosomes, we estimate an average coalescence of 17,000 years ago, but the duplication itself seems to have occurred much earlier (>1 million years ago) than it rose to high frequency in west Eurasia, indicated by the large divergence of sequence within the α duplication.

HP structural evolution

A key mystery surrounding the evolution of *HP* structural variation is why no SNP is in high LD with the CNV. The currently accepted model of *HP* structural evolution proposed that *HP2* arose through the fusing of two diverged *HP1* alleles (*HP1F* and *HP1S*)[2], and is a textbook example of non-homologous recombination[113]. However, we hypothesized that the *HP* CNV's lack of LD to

surrounding SNPs indicates a more complex structural history. We began by phasing the *HP* structural alleles onto SNP haplotype backgrounds from the 1000 Genomes Project whole genome sequence using Beagle[114].

We first sought to understand why the *HP* CNV is not in high LD with any SNP. We considered that the forces which decrease linkage disequilibrium between proximal loci are (1) recombination and (2) recurrent mutation. We reasoned that if recombination were frequent in the *HP* structurally variant region, we would observe both low LD between SNPs and structure and low LD between SNPs on either side of the structural variants. Conversely, if *HP* structure were affected by recurrent mutation, we would observe low LD between SNPs and structure, but high LD between SNPs on either side of the structurally variant region (Figure 4.3).

We observed that SNP haplotypes almost always persist through the structurally variant region: common SNPs which reside on opposite sides of the CNV segregate with different structures and are routinely in high linkage disequilibrium with each other (r^2 >0.9) (Supplementary Figure 4). Likewise, a genome-wide scan of recombination rates has documented the lack of a recombination hotspot in this region[115]. These results are inconsistent with the hypothesis of high recombination and support recurrent structural mutation at *HP*.
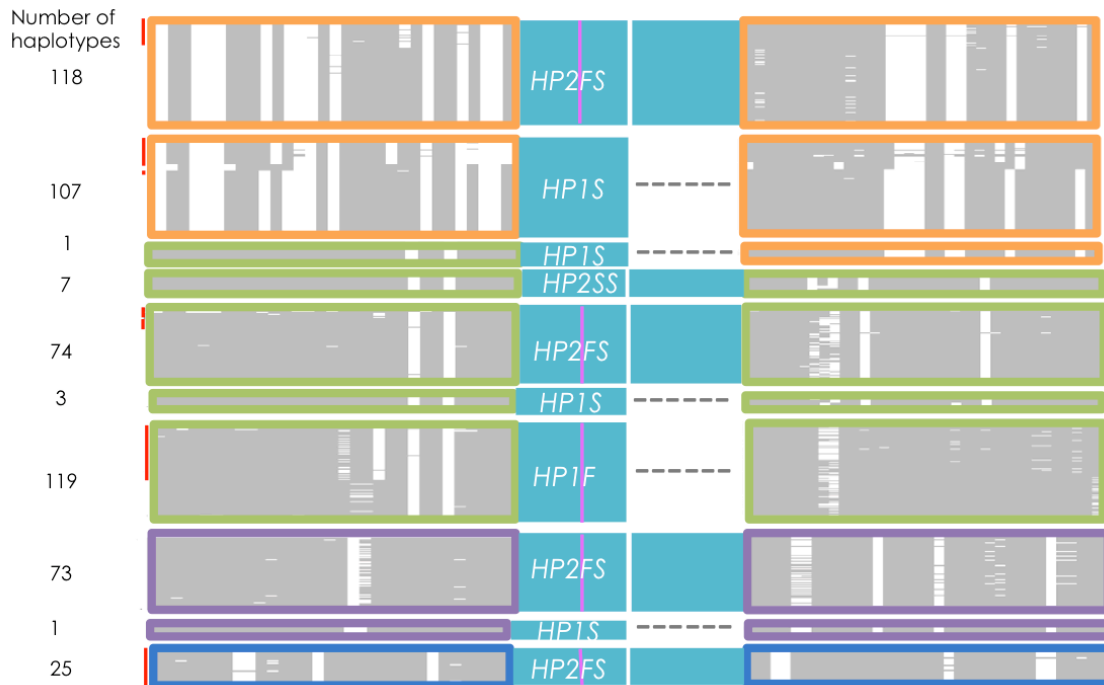
Figure 4.3. SNP haplotype background for *HP* structural haplotypes. This plot displays the SNP haplotypes (10 kb on each side) segregating with the *HP* structural variants. Each horizontal line represents an individual SNP haplotype. White represents the minor allele and grey indicates the major allele across the pooled populations (CEU, IBS, TSI, YRI). Population identifiers are defined in Supplementary Table 1. YRI individuals are indicated with red bars to the left of the plot. Haplotypes were first grouped by structure, then upstream SNP haplotypes were clustered by the k-means method, and downstream haplotypes were assigned the same order. Similar SNP haplotypes carrying different structures are indicated with colored outlines (orange, green, purple, blue).
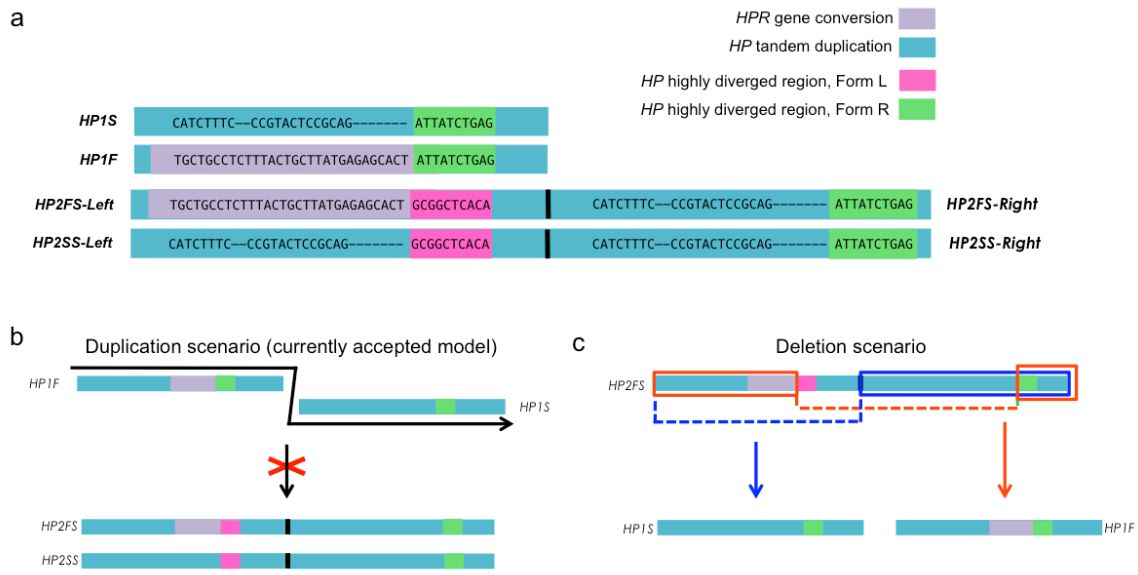
Figure 4.4 (a) A simplified alignment diagram shows base pair differences between *HP* structural forms inside the structurally variant region. Only the polymorphic bases are depicted. *HP1S* lacks the paralogous gene conversion, while *HP2SS-Left* contains only the right portion. Form L of the highly diverged region is only present on *HP2* alleles. (b) The current model of *HP* structural evolution is that and *HP1F* and an *HP1S* recombined to form *HP2FS*; however, *HP1F* and an *HP1S* allele do not have the necessary sequence to form either *HP2* allele as neither *HP1* has Form L of the highly diverge region. (c) *HP2FS* contains the necessary sequence to mutate into *HP1S* and *HP1F* through different deletion events. The deletion event which could create *HP1S* is depicted in blue, while the *HP1F* deletion event is shown in orange. In both cases the deleted sequence is underlined with a dashed line, while the preserved sequence is shown inside a box.

Next, we considered the directionality of the potential recurrent structural mutation. *HP1* could either recurrently duplicate to become *HP2,* or *HP2* could experience deletions to form new *HP1* alleles. We examined the structural forms to determine if each scenario is equally likely. We observed that while *HP1F* contains *HPR* gene conversion, as on *HP2FS-Left*, neither *HP1* allele has the Form L form of the highly diverged region (Figure 4.4).

Because Form L of the highly diverged region is not present in either of the *HP1* structural alleles (Figure 4.4a), a new *HP2* cannot be created by a simple tandem duplication*,* or by a fusion of *HP1F* and *HP1S* as is the current model[2] (Figure 4.4b). However, we observed that both *HP1F* and *HP1S* could be created by simple deletions in *HP2FS* (Figure 4.4c).

In fact, an *HP2* to *HP1* deletions have been observed at low frequency in the somatic and sperm cells of homozygous *HP2* individuals (mutation rate: 8×10^{-6} per cell)[116], demonstrating that the *HP* gene is predisposed to this structural mutation.

We next hypothesized that if *HP2* deletion events occur regularly, we may be able to observe recent deletions using long and specific SNP haplotypes. While the above short SNP haplotypes can be similar across many individuals, we reasoned that observing longer SNP haplotypes and considering all mutations and recombination events, would allow us to observe small groups of closely related haplotypes. We phased structural variation with a longer section of adjacent SNP haplotypes and created a tree diagram to group highly similar haplotypes (Figure 4.5).
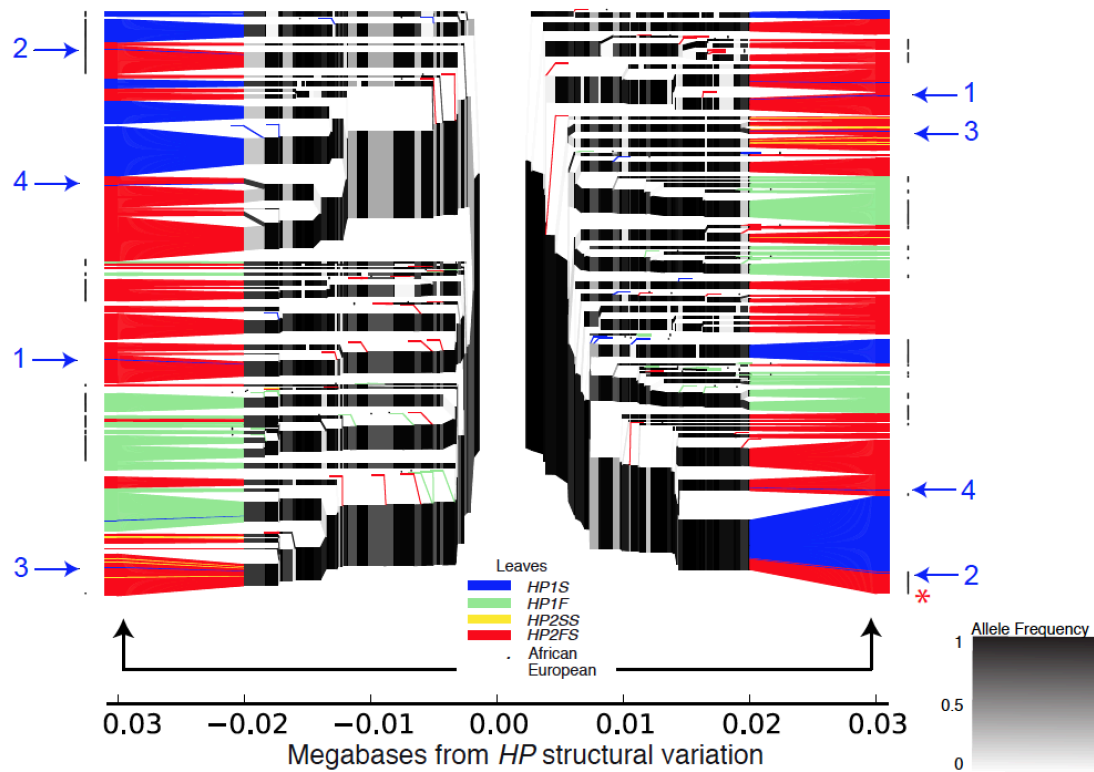
Figure 4.5. Lone *HP1S* alleles segregate with common *HP2FS* SNP haplotypes. The plot depicts homozygosity and divergence (due to mutation and recombination) of the SNP haplotypes on which each structural form segregates in European populations (CEU, IBS, TSI) and one African population (YRI). Branch points represent markers at which the depicted haplotypes diverge due to mutation and/or recombination with other haplotypes. We observe four *HP1S* alleles (indicated with blue arrows 1-4), segregating with *HP2FS* haplotypes for at least 20 kb surrounding the CNV on both sides. The African individuals are identified with a dot after the colored leaf. The red * indicates a SNP haplotype which segregates with *HP1S* in European populations and *HP2FS* in YRI. In the plot, the structures are represented on the leaves in order to clarify their relationships to SNP haplotypes, but the CNV actually resides (in genomic space) within the gap at the center between the two origins on the plot.

While most branches of the tree resolve into groupings of single structural haplotypes, several majority *HP2FS* branches contain single *HP1S* haplotypes. The structure of these *HP1S* alleles was phased with SNP haplotypes using the Drop-Phase method[117], in which DNA from physically linked alleles is partitioned into the same nanoliter-sized droplets more often than variants on separate chromosomes.

The observation that these *HP1S* structures segregate on standard *HP2FS* SNP haplotype backgrounds indicated to us that they may result from recent deletion events. We selected four putative deletion *HP1S* alleles, which each had a minimum of 20 kb perfectly matching *HP2FS* haplotypes on each side of the structurally variant region, for sequencing. We sequenced the structurally variant region in each putative recent deletion allele and the corresponding region in an *HP2* allele from the same SNP haplotype branch. In each case we found that the *HP1S* sequence more closely resembled the *HP2* from the same SNP haplotype than other *HP1S* alleles (Figure 4.6). While none of the sequenced *HP1F* alleles contained derived variation apart from the *HPR* gene conversion and highly diverged regions, the sequence was consistent with *HP2FS* sequences on similar SNP haplotypes (Supplementary Figure 2).
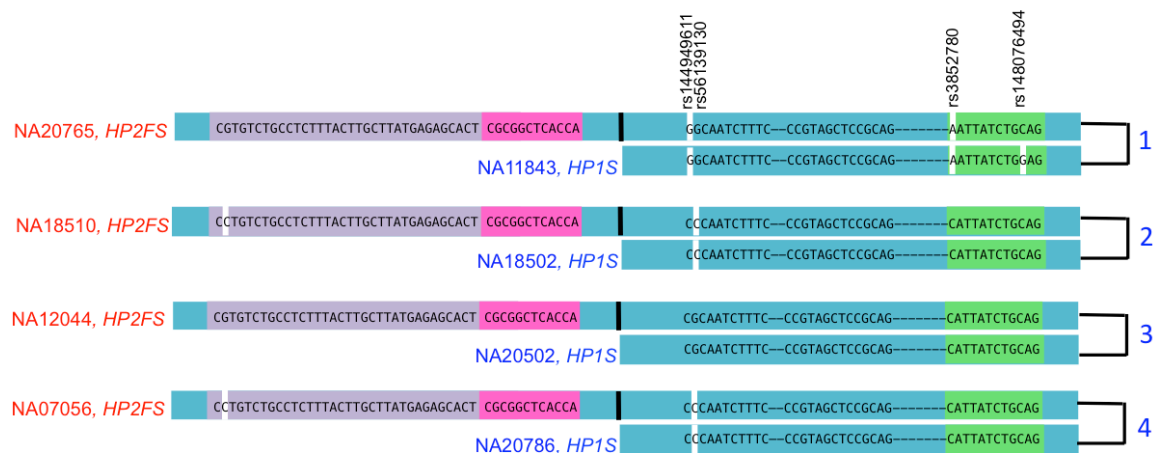
Figure 4.6. An alignment showing variation inside the structurally variant region from *HP2* and *HP1* structural alleles, which segregate with the same SNP haplotype background. The numbers 1-4 indicate the *HP1S* allele from Figure 4.5. Nucleotides that define the gene conversion and highly diverged region are shown along with other derived alleles (shown in white). *HP2-Right* and *HP1* deletion alleles on the same SNP haplotype background share most derived mutations inside the structurally variant region.

We next considered *HP* structural evolution at the population level. *HP2* is thought to be derived in humans since all other great apes are fixed for *HP1*, but we wondered if adjacent SNP haplotypes for *HP1* or *HP2* would appear less bottlenecked. Older variants segregate on more diverse SNP haplotypes because they have had more time to accumulate mutations. We found that SNPs proximal to *HP2* harbor a greater level of diversity than SNP haplotypes surrounding *HP1* in both African and European populations (Figure 4.7), indicating that modern *HP2* haplotypes pre-date most *HP1* haplotypes.
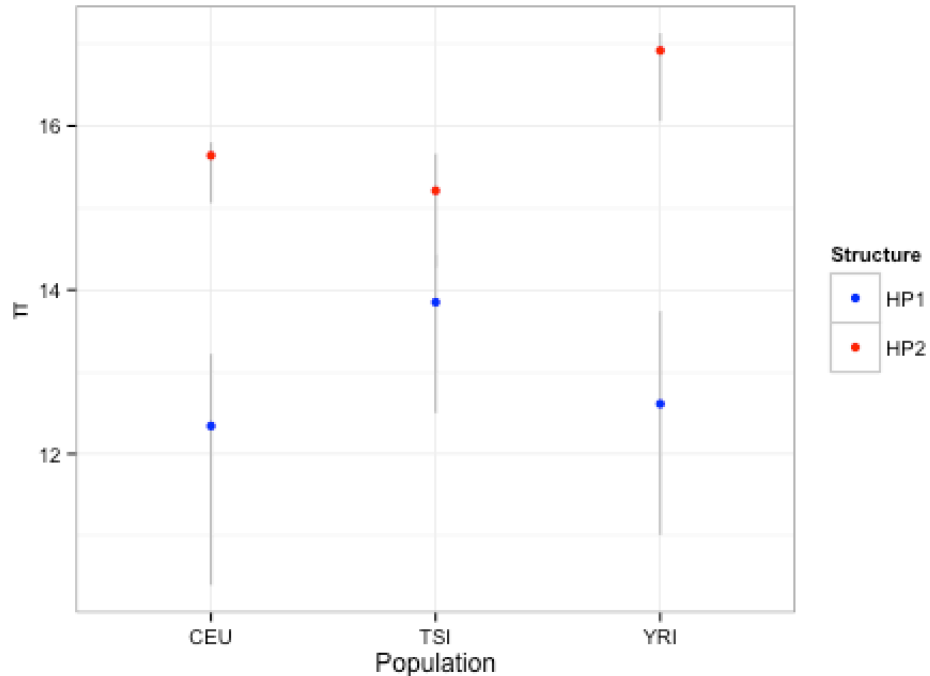
Figure 4.7. Measures of pairwise diversity (π) for *HP1* and *HP2* alleles within three populations. The 95% confidence intervals were obtained through 1000 bootstrap iterations.

Figure 4.3 shows that *HP1* SNP haplotypes are each closely related to a specific class of *HP2* SNP haplotype. We next quantified this distance in order to identify potential ancestral *HP1* alleles (Supplementary Figure 5). While the vast majority of *HP1* alleles appear closely related to specific class of *HP2*, we cannot rule out that a small number of segregating *HP1* alleles are ancestral and not derived deletions (Supplementary Figure 6).

We propose that the *HP* CNV is not in high LD with any SNP due to recurrent *HP2* to *HP1* deletion events, which have occurred on a variety of distinct SNP haplotype backgrounds.

67

The idea that the *HP2* allele is ancestral to *HP1* among common haplotypes led us to wonder if *HP2* and *HPR* paralogous gene conversion could be observed in our hominoid relatives. Previous estimates for the age of the *HP* CNV vary from 100 KYA to 2 MYA[59,92]. By examining high-coverage (~30x) Neanderthal[118] and Denisova[119] genomes, we observed that both species had alleles consistent with the human *HP2FS* haplotype (Supplementary Table 2). Because the *HP2FS* haplotype segregates in humans from every continent, including Africa, its presence in early hominins indicates that this haplotype was segregating in the ancestor of all three species and is not the result of introgression. The divergence time of these species is 400 to 600 KYA[120], which provides a lower-bound estimate for the age of both the *HP* CNV and the gene conversion from *HPR,* and both forms of the highly diverged region.

In order to confirm that the *HP* duplication is derived in humans, we assayed the breakpoint between the *HP-Left* and *HP-Right* segments in a panel of 64 chimpanzees and bonobos. None of the examined individuals produced an amplicon for this assay, indicating that they are all *HP1-HP1* homozygotes.

HPR structural evolution

The observation of 17q21.31 tandem duplication expansions led us to investigate the *HPR* region, which appears to have undergone a more extensive and rapid duplication expansion. While the γ duplication in 17q21.31 region was observed at copy numbers up to four on a single haplotype, we observed copy numbers of up to eight on a single haplotype for the *HPR* gene (Figure 3.13).

Interestingly, copy number at *HPR* is bimodal with the most common haplotypes being *HPR.1* and *HPR.5*.

Unlike the γ duplication in 17q21.31, which segregates in European, Asian and African populations, elevated copy number at *HPR* is perhaps more recent and is confined to African populations[65]. Supporting the idea that the *HPR* mCNV is relatively new, we estimated the coalescence of four sequenced *HPR* haplotypes with elevated copy number to be fairly recent (28 kya, 95% CI: 16-42 kya). Additionally, we can see from a haplotype plot of the SNPs surrounding *HPR* that all haplotypes with elevated copy number reside on a similar SNP haplotype background (Figure 4.9). This pattern suggests that present diversity is descended from a single duplication event. These lines of evidence support the model that *HPR* copy number underwent a rapid run-away expansion after one initial duplication.

Figure 4.9. Elevated *HPR* copy number alleles segregate on highly similar SNP haplotype backgrounds. The plot shows homozygosity and divergence (due to mutation and recombination) of the SNP haplotypes on which each structural form segregates in the three African (or African-derived) populations. SNPs on the left half of the plot exist upstream of the *HPR* duplication, whereas SNPs on the right half of the plot are downstream. Branch points represent markers at which the depicted haplotypes diverge due to mutation and/or recombination with other haplotypes. In the plot, the structures are represented on the leaves in order to clarify their relationships to SNP haplotypes, but the variable parts of these CNVs actually reside (in genomic space) within the gap at the center between the two origins on the plot. The group of haplotypes with elevated copy number resides on a similar SNP haplotype background; however, haplotypes with the same copy number do not cluster together within this group. Allelic copy number was ascertained in trios for the YRI population, and using a population-based maximum likelihood method for the ASW and LWK populations.

Leaves
(structural isoforms)

ASW duplicated
ASW unduplicated
YRI duplicated
YRI unduplicated
LWK duplicated
LWK unduplicated
● ● ● # of HPR duplications

Position (kb) relative to ends HPR duplication region
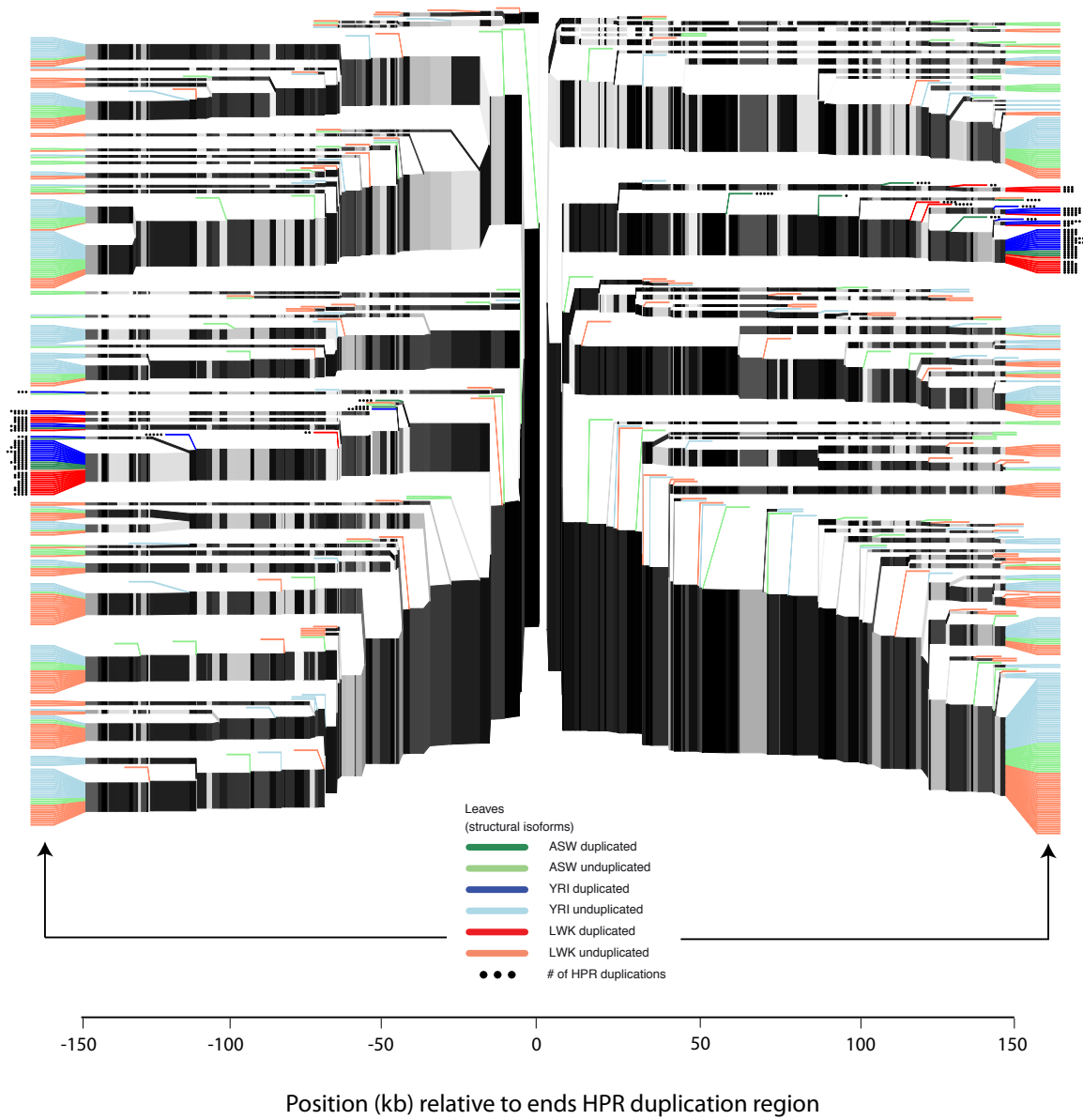
Figure 4.9 (continued)

In order to better understand the distribution of the CNV in African populations, we used phased CNV and SNP haplotypes as an imputation reference panel to impute the presence or absence of elevated copy number at *HPR* into other populations in and around Africa.

	Population	Number of haplotypes
1	Mozabite	58
2	Yoruba	252
3	Lemande	38
4	Bakola	58
5	San	10
6	Bulala	24
7	Biaka Pygmy	42
8	Mandenka	44
9	Mbuti Pygmy	26
10	Dinka	24
11	Bantu, ZA	16
12	Luo	64
13	Luya	150
14	Hadza	52
15	Bantu, KE	22
16	Amhara	52
17	Hadandawa	30
18	Baniamer	20
19	Palestinian	92
20	Druze	84
21	Bedouin	92
22	Hazara	44
23	Makrani	50
24	Balochi	48
25	Brahui	50

Figure 4.10. The frequency of elevated copy number at *HPR* in populations in and around Africa. The table lists each population as well as the number of imputed haplotypes used in the analysis. The left column indicates the number used to designate the population on the map. The number on each pie graph on the map indicates the identity of the population. The pink color indicates haplotypes imputed as having elevated copy number at *HPR* (2+), while the grey indicates haplotypes that were imputed as copy number 1.

72

We found the highest frequency of elevated copy number in the Dinka population and noted several populations in which we imputed no alleles with elevated copy number: San, Mbuti Pygmy, Hadandawa, Baniamer. Additionally, the *HPR* CNV segregates at low frequency in multiple populations in the Middle East (Figure 4.10).

Unlike the interpretable series of mutations that likely gave rise to 17q21.31 structural haplotypes, the *HPR* rapid expansion of a single element to high copy number leaves the evolutionary steps that created these haplotypes uncertain. One could imagine that duplications could accumulate in a stepwise fashion (one at a time), or make large jumps (e.g., four copies to seven copies). The observation that haplotypes carrying the same copy number do not always have the most similar SNP haplotype backgrounds also supports the model of a complicated mutational history (Figure 4.9).

C4 structural evolution

On a macroevolutionary scale, ancestral structural polymorphisms often become lost or fixed for different alleles in descendent species due to genetic drift, creating fixed differences between closely related organisms. However, regions like the HLA are known for deep coalescence[121], and the *C4* gene in this region is no exception. Our research confirms that both the length and sequence polymorphism vary in orangutans and humans, while the other great apes lack the *C4L* form. By comparing great ape species, we found that the sequence

differences between *C4A* from *C4B* can differ both between and within species (Figure 3.14).

In all Chimpanzee populations, Bonobos, and Gorillas in our study, the *C4AS-C4BS* haplotype is the most common (Table 4.1). While we see that Western Chimpanzees (*Pan troglodytes verus)* have at least four structural haplotypes, only the *C4AS-C4BS* is common.

Central Chimpanzees (*Pan troglodytes troglodytes)* appear to be more diverse, harboring several relatively common *C4* structural haplotypes. This observation is in line with the conclusion that Central Chimpanzees have more nucleotide diversity in their nuclear genomes than Western Chimpanzees[122].

Table 4.1. The observed frequency of *C4* structural haplotypes in great apes. The counts of observed *C4* structural haplotypes are indicated for each group. The frequency of each haplotype is shown in parentheses. Although the sample size of several populations is modest, we can begin to understand the level of *C4* diversity in different great apes. Interestingly, the *C4* types that are most highly associated with SLE (*C4BS*) and schizophrenia (*C4AL-C4AL*) are rare or absent in these great apes.

	Western Chimpanzee N haps=40	Central Chimpanzee N haps=8	Bonobo N haps=16	Gorilla N haps=6	Orangutan N haps=8	Human (CEU) N haps=240
C4AS-C4BS	35 (0.88)	5 (.63)	15 (0.94)	6 (1.00)		0
C4AS-C4AS	2 (0.05)	1 (0.13)	0	0		0
C4BS-C4BS	2 (0.05)	0	0	0		0
C4AS	1 (0.02)	2 (0.24)	0	0		0
C4BS	0	0	1 (0.06)	0		17 (0.07)
C4AL-C4AL	0	0	0	0		26 (0.11)
C4AL-C4BL	0	0	0	0	present	100 (0.42)
C4AL-C4BS	0	0	0	0		74 (0.31)
C4AL	0	0	0	0		5 (0.02)

Conclusion of the evolution of complex structural variation

The 17q21.31 region, *HP*, *HPR* and *C4*, all contain copy number variants, but the subsequent evolution of these CNVs differs greatly by region. In the 17q21.31 and *HPR* loci there was a multi-allelic expansion after an initial duplication. Regions with CNVs appear to be predisposed to creating multi-allelic CNVs, likely due to non-allelic homologous recombination. However, the *HP* gene commonly only segregates with two copies of the CNV, with an *HP*

75

triplication observed only once in this study. This pattern may be due to random chance, or the copy number of the *HP* CNV could be held at two by selection. As the CNV at copy number two allows *HP* to form tetramers, a triplication might allow larger quaternary structures of this protein, which might inhibit function.

Like tandem duplications, gene clusters (which evolve from ancient duplications) also contain adjacent homologous sequences that can spark structural mutation. We see an example of this phenomenon in the *HP* paralogous gene conversion from *HPR*. When double-stranded breaks occur, usually the other allele of the same gene provides the template; however in the case of the gene conversion in *HP*, a nearby paralog donated the sequence. This phenomenon has been observed for other tandemly arranged paralogs[123,124] and may be a common mode of sequence exchange between genes.

In addition to understanding how structurally complex regions evolve, it is also important to include this class of variation in association studies. In the next chapter we develop and implement methods to allow complex structural variants to be studied alongside nucleotide variation in association studies.

Contributions

I received guidance and input for work presented in this chapter from Steve McCarroll. I designed all experiments and molecular assays, performed all molecular experiments, and did all data analysis in this chapter excluding the contributions of others mentioned below. Aswin Sekar provided advice for experiments involving the *C4* gene. Figure 4.1 was produced by Robert

Handsaker. Sarah Tishkoff provided SNP genotype data for various populations provided by the HGDP[125] (Human Genome Diversity Project).

Chapter 5

Developing and implementing methods to incorporate complex genomic

structures into large-scale association studies

Developing methods to incorporate complex structural variation into association studies

Simple, common deletion and duplication polymorphisms have been typed in large cohorts and have been found to segregate on SNP haplotypes and have been associated with many human phenotypes via proxy SNPs[7,8]. Each of the structurally complex loci addressed in this thesis has greater than two structural forms, and we have shown that the *HP* structural polymorphism is recurrent; therefore, none of these polymorphisms is well tagged by an individual biallelic SNP and may be overlooked in GWAS.

Our observation that many structural haplotypes reside on specific and distinct SNP haplotype backgrounds (Figures 4.1, 4.5, 4.9), suggested the possibility that surrounding SNPs could be used to infer these structures. If this idea were achievable, it would allow complex structural haplotypes to be computationally determined from the wealth of SNP data already collected for many cohorts (for which phenotype information is available). This method would allow us to test if complex structural variation is responsible for regional associations to specific phenotypes. We use an imputation approach, which takes advantage of the r^2 correlation between an unknown variant and many surrounding SNPs to infer the state of the unknown polymorphism.

We constructed an imputation reference panel for each region composed of individuals for whom both structure and SNPs were determined. We used unrelated individuals from different populations in the 1000 Genomes Project[60] and HapMap[126] to create these reference panels. Using the imputation software Beagle[127], we then tested these reference panels using leave-one-out tests:

removing each individual from the reference panel, eliminating the individual's encoded structure, and attempting to impute the structure using the remaining reference panel. Comparing an individual's known structure to the imputed structure allows us to estimate the accuracy of the imputation. The results of these leave-one-out tests can be seen in Tables 5.1-5.3. From this study we observed that more stable structures at high frequency (17q21.31 inversion, *HP* duplication and gene conversion) were imputed with extremely high accuracy, but the more rapidly evolving tandem mCNVs proved more difficult. The least imputable structural variant was the exact copy number of the *HPR* mCNV (data not shown), followed by the multi-allelic γ duplication in 17q21.31 (Table 5.1). However, the presence or absence of the *HPR* elevated copy allele was imputable with high accuracy (Table 5.3).

Table 5.1. Correlation of 17q21.31 structural features with those inferred through imputation or biallelic tagging SNPs. Shown is the correlation (r^2) of experimental determinations of the state of each structural feature in each genome with either (i) imputed, probabilistic 'dosages' of each structural feature or (ii) the state of the single most-correlated proxy SNP (tag SNP) from each reference panel for the 17q21.31 region.

Structural feature	Tag SNP	Imputation
Copy number of α duplication	0.96	0.99
Copy number of β duplication	0.49	0.93
Copy number of γ duplication	0.27	0.84
Inversion state (H1 versus H2)	1.00	1.00

80

Table 5.2. Correlation of *HP* structural features with those inferred through imputation or biallelic tagging SNPs. Shown is the correlation (r^2) of experimental determinations of the state of each structural feature in each genome with either (i) imputed, probabilistic 'dosages' of each structural feature or (ii) the state of the single most-correlated proxy SNP (tag SNP) from the reference panel for the *HP* region.

Populations	African (YRI)		European (CEU,IBS, TSI)	
HP structural haplotype	Tag SNP	Imputation	Tag SNP	Imputation
HP1S	0.96	0.96	0.86	0.92
HP1F	0.82	0.94	0.83	0.98
HP2FS	0.76	0.92	0.4	0.92
HP2SS	NA	NA	0.58	0.74
CNV (*HP1* versus *HP2*)	0.76	0.92	0.44	0.94

Table 5.3. Correlation of *HPR* copy number with inferred copy number through imputation or biallelic tagging SNPs. Shown is the correlation (r^2) of experimental determination of the state of the structural feature in each genome with either (i) imputed, probabilistic 'dosages' of each structural feature or (ii) the state of the single most-correlated proxy SNP (tag SNP) from the reference panel for the *HPR* region.

Structural feature	Tag SNP	Imputation
Elevated copy number versus single copy	0.92	0.96

Phenotype hypothesis generation for 17q21.31 structural variation

Markers in the 17q21.31 region associate with female fertility and female meiotic recombination[1,50]; however, the genetic variants underlying these associations remain unknown. One of the key challenges to understanding association signals in the 17q21.31 region is the extensive linkage disequilibrium.

Recombination gradually disentangles variants that have arisen on a common haplotype, allowing these variants a degree of independence in an association study. The recombination rate varies throughout the genome[128], and the rate in the 17q21.31 region is particularly low due to the megabase-long inversion polymorphism. This structural variant has inhibited recombination to the degree that H1 and H2 have evolved with almost complete independence over the past 2.3 million years[41].

To overcome this challenge, we generated a hypothesis based on nucleotide association, overlapping duplications on different SNP haplotypes, and functional data.

17q21.31 and female miotic recombination rate and fertility

The 17q21.31 region contains SNPs that associate to the female meiotic recombination rate (which relates to fertility[47]), and women with at least one copy of H2 are more fertile than those who are homozygous for H1[1]. However, it is not known which alleles segregating with H1 or H2 contribute to this phenotype. The vast majority of H2 haplotypes segregate with the α duplication, while a minority of H1 haplotypes segregate with duplication β in the CEU population. Additionally, we have documented that both the the α and β duplications

underwent parallel increases in frequency in Europe over a short period of time. These ideas lead us to hypothesize that both the α and β duplications increase the female meiotic recombination rate and fertility. We sought to develop a hypothesis for which gene(s) in the duplications regions might play a role in the phenotype.

Our observation that the α and β duplications in the 17q21.31 region overlap the same gene (*KANSL1*) and underwent parallel increases in frequency in Europe, invited the hypothesis that they may be functional and influence a common phenotype. Based on the breakpoints of the duplications we hypothesized that they may cause novel truncated transcripts to be created.

Through reverse transcription PCR (RT-PCR) and sequencing we found novel fusion transcripts specific to individuals with the α and β duplications. These novel transcripts display the 5′ exons of *KANSL1* fused to cryptic exons that terminate the coding sequence (Figure 5.1). Notably, a similar truncation in the *Drosophila melanogaster* ortholog of *KANSL1*, *GC4699/E(nos)*, was identified in a mutagenesis screen for modifiers of *Nanos* and was found to enhance the effect of a *Nanos* hypomorph on age-dependent female fertility and germline stem cell differentiation[129]. The precise role of *KANSL1* in these processes is unknown, although the encoded protein is found within the MOF-MSL1v1 chromatin-modifying complex[130,131]. Proteins arising from these novel *KANSL1* transcripts would contain the coiled-coil domain.

The parallel increase in frequency of β and α in Europe, the functional affect that these duplications have on the *KANSL1* transcript (which affects female fertility in *Drosophila*) and the association of SNPs in the region to the

female meiotic recombination rate led us to hypothesize that these structural

polymorphisms affect the female meiotic recombination rate and fertility. The

female meiotic recombination rate and age-dependent fertility are related

phenotypes because more recombinations decreases the odds of a non-

disjunction in egg cells[47].

a

Fusion mRNA created from β duplication breakpoint

TGAGACGCAGGTCAGAATGGAAATGGGCTGCAGACCGGGCAGCTATTGTCAGCCG
CTGGAACTGGCTTCAGGCTCATGTTTCTGACTTGGAATATCGAATTCGTCAGCAAAC
AGACATTTACAAACAGATACGTGCTAATAAGGTTTCTGTGTGGAGACAGTAGAATAT
AAAAATAACACCTTCGCT

b

Fusion mRNA created from α duplication breakpoint

TACCCCTAGACGTGGGAACAACGCAAGTCCCACTTACAACACTTAAGAACATTCTC
ATGATGACCGTTGAACTGGAAAAACTTCCCAGCAGACCACAGGAGGTTGGCCCCA
GACTCACTGAGTGCCTGCAGCAGCCGTACAGACACAGCATCCTTGGCCACCTCAT
GCCCATCCCGGCCATCTAGGGTCAGCACAACCCAGATGAGGCCGCTGAAGGGCAC
CGGATGCCCAGGAATCACCACCTGGTACCAGAAGCGGTGCCAGCCAGCAGGTCCT
ATGCCCAAACACTTGGTGAGG

Figure 5.1. Fusion transcripts created from *KANSL1* duplications (α and β). (a)

KANSL1 (shown in blue) is fused *ARL17* (shown in red). This sequenced

breakpoint is also present in cDNA BC006271, likely a complete transcript of this

fusion mRNA. (b) *KANSL1* (shown in blue) is fused to *LRRC37A* (shown in

green), and yet another fusion occurs with a novel exon (shown in orange).

Phenotype hypothesis generation for *HP* structural variation

The *HP* structural variant was the second human polymorphism ever

discovered[58], and has been widely studied ever since. Haptoglobin codes for the

HP protein, an acute phase reactant that binds hemoglobin and cholesterol

molecules[53,132], and zonulin, a key regulator of intercellular tight junctions[54]. While the *HP* CNV affects the structure and function of both proteins[54,88,133] and is highly studied through the candidate gene approach, we are not aware of any highly significant and reproducible physiologic phenotypic associations, potentially due to lack of power. While some results from association studies are intriguing, there is a clear lack of consistency among the commonly studied phenotypes (Table 2.2).

Our imputation approach now allows for interrogation of structural genotype-phenotype associations to be tested in much larger sample sizes and could potentially allow for enough power to reach highly significant and reproducible associations.

HP and Malaria

Malaria is a disease transmitted by mosquitos that kills at least one million people each year[134]. Haptoglobin's potential involvement in an indivdual's predisposition to malaria is commonly studied[64,135,136] and cited, yet the results from association studies have been mixed (Table 2.2). It is important to examine the putative association between *HP* structural variants and malaria predisposition with a large cohort in order to resolve this contentious association.

HP and Inflammatory Bowel Disease (IBD) / Crohn's Disease

Inflammatory bowel disease is caused by an immune response which creates inflammation of the gastrointestinal tract and affects over 2.5 million people with European ancestry[137]. Crohn's disease and ulcerative colitis are the

most common types of IBD. While ulcerative colitis is confined to inflammation of the large bowel, Crohn's disease can affect any location in the gastrointestinal tract, and most commonly affects the the small intestine and colon[138]. There are two lines of evidence which suggest that the *HP* duplication may play a role in Crohn's disease.

First, SNPs in the region associate to Crohn's disease (best p=2.17x10^{-5}). Furthermore, the most associated SNP in the region is also the best tagging SNP for the CNV in European populations (rs217181, r^2 with CNV = 0.44) (Figure 5.2).
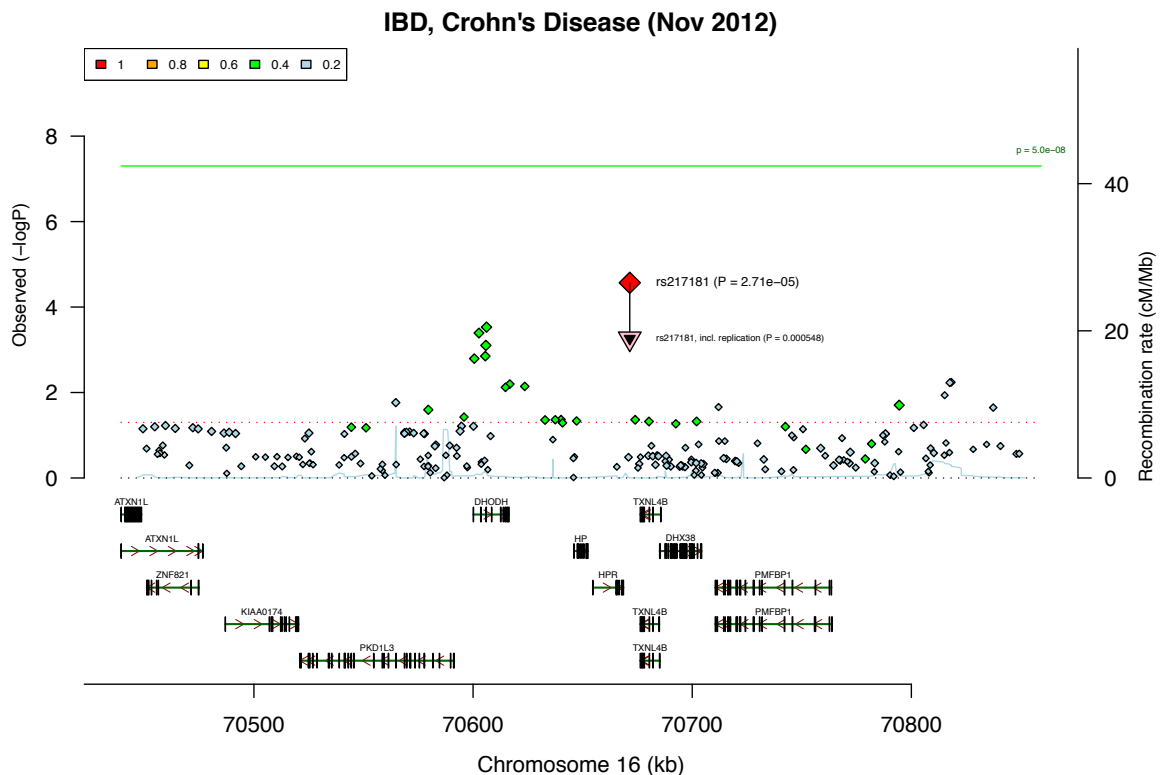

IBD, Crohn's Disease (Nov 2012)

Figure 5.2. SNPs in the *HP* region associate to Crohn's Disease. The best tag SNP for the *HP* CNV is the most associated SNP in the region but is below genome-wide significance. Potentially this association could be improved through imputation. This plot was made with Ricopili (http://www.broadinstitute.org/mpg/ricopili/).

Second, Crohn's disease is thought to relate to intestinal permeability[139] and zonulin is a key regulator of intercellular tight junctions (including cells of the intestines)[54]. While the haptoglobin protein is proteolytically cleaved and bound together by disulfide bonds, zonulin is uncleaved and thought to be the product of the *HP2* form only (though this has not been tested on a large scale)[54].

HP and Cholesterol Levels

Cholesterol levels measured in blood are highly predictive for the development of cardiovascular disease (CVD)[140], which is the leading cause of death for Americans[141]. While LDL cholesterol (LDL-C) can cause artery-clogging plaques, HDL cholesterol (HDL-C) clears cholesterol from the blood[142,143]. A high LDL-C to HDL-C ratio predisposes one to develop CVD[144]. Additinally, lowering LDL cholesterol levels using statins has been shown to reduce the incidence of CVD[145].

Evidence from GWAS suggests that *HP* may play a role in cholesterol levels. A GWAS for blood lipids found a highly significant association between total and LDL cholesterol levels and a SNP in the *HP* region (rs2000999, $p=3 \times 10^{-24}$)[3]. Additionally, this SNP is among the best tag SNPs for the structural variant ($r^2=0.13$), having a greater r^2 with *HP2* than 97% of SNPs in the region.

Protein-protein interactions suggest that HP may play a role in cholesterol levels. The ApoE protein is known to be involved in regulating cholesterol levels[146], and variants in or near this gene are highly associated with LDL and total cholesterol (TC) levels ($p=9 \times 10^{-147}$). Multiple studies have now verified that

the HP protein physically binds ApoE[53,147], and this invites the hypothesis that HP
may also affect cholesterol levels.

Phenotype hypothesis generation for *HPR* structural variation

HPR and trypansosomiasis

Trypanosomes cause trypanosomiasis or African sleeping sickness, which
is a devastating disease that affects humans and other mammals in rural Africa.
Between 50 and 70 thousand individuals are infected each year, and this disease
is often deadly if treatment is not received[148].

An interesting evolutionary arms race[149] has occurred between
trypanosomes and the human immune system. Host macrophages emit oxidative
bursts which have a general antimicrobial action[150]. Trypanosomes are able to
protect themselves against the oxidative burst with heme groups, which are
obtained through importing the host's haptoglobin-hemoglobin complex through
the trypanosome's haptoglobin-hemoglobin receptor[151]. However, HPR bound to
hemoglobin also binds this receptor and is imported by trypanosomes. HPR is a
key component of the Trypanosome Lytic Factor 1 complex (TLF1), and acts as
a trojan horse to import this complex into the trypanosome, where it kills it from
the inside[152]. While TLF1 protects humans from some species of trypanosome[153],
other species of trypanosome have evolved countermeasures[154,155].

Our population genetic study demonstrated that the *HPR* gene has
increased copy number in many individuals in Africa, the same continent in which

trypanosomiasis is most prevalent. We hypothesized that elevated copy number of *HPR* may be protective against Trypanosomiasis.

Imputation of complex structural variants and non-significant association results

We used these reference panels to impute structural variation into cohorts with relevant phenotypic information. For 17q21.31 we imputed the structural haplotypes into the AGRE[156] dataset, which contains parent and offspring genotypes, to test the association of the female meiotic recombination rate to structural variation. The *HP* structural haplotypes were imputed into the MalariaGEN[157] malaria case/control cohort and a Crohn's disease case/control cohort from the IBD Genetics Consortium[158]. We directly type the *HPR* CNV in a Ghanan cohort to examine the variant's association to trypanosomiasis because the efficacy of imputation was not tested in this population. The controls in the trypanosomiasis experiment were individuals who had malaria. In each of these experiments, the association study was either underpowered or suggested that no highly-significant association was present (Table 5.4).

The trypanosomiasis association to *HPR* was in the opposite direction from that which we had originally anticipated: the modest p-value (0.01) suggests that the single-copy allele of *HPR* is protective against trypanosomiasis. This study was likely underpowered and the association is not felt to be a true result.

Table 5.4. Association results for complex structural variants and various human phenotypes. Along with the structural variants and phenotype being examined the table lists the sample size and p-value for each association. * The controls for the association study between *HPR* copy number and trypanosomiasis were individuals who were treated at the same hospitals as those with trypanosomiasis and had malaria.

Structurally variant region	Phenotype	Number of individuals	Best p-value for structural variant
17q21.31	Female meiotic recombination rate	30 trios	0.24
HPR	Susceptibility to trypanosomiasis	64 cases 99 controls*	0.01
HP	Susceptibility to severe malaria	1,060 cases 1,500 controls	0.37
HP	Susceptibility to Crohn's disease	6,000 cases, 15,000 controls	9×10^{-4}

HP structure associates to Crohn's disease with a p-value of 9×10^{-4}; however, the top nucleotide association in the region was slightly more significant ($p = 2.71 \times 10^{-5}$). A larger study would need to be conducted in order to resolve the variant(s) underlying the putative association to Crohn's disease in the *HP* region.

While there is compelling evidence for the 17q21.31 region's involvement in the female meiotic recombination rate, we did not find an association to the

structural variants in this region. This association was done with an extremely limited sample size (30 trios) and should be examined in the future with a larger cohort.

_HP_ structural variation associates with cholesterol levels

We performed meta-analysis on 12,063 subjects of European descent from three cohorts with cholesterol information (ARIC[159], MESA[160], and CHS[161]) to examine the potential association between _HP_ structural variants and cholesterol levels. Individuals who were diabetic or taking cholesterol-lowering medication were omitted from the study. The first 10 principal components from a genome-wide principal components analysis (PCA) were used as covariates in the analysis to account for population substructure. The imputation was performed with a reference panel composed of the CEU, TSI and IBS populations and composed of SNPs from the Illumina OMNI, Affymetrix 6.0 and Illumina Human 1M arrays. The genomic region used included all assayed SNPs up to 1 megabase on both sides of the CNV.

Our analysis showed that the _HP2_ haplotype associates to total cholesterol levels (1.49×10^{-8})(Fig 5.3a) and LDL cholesterol levels (2.53×10^{-6}) (Fig 5.3b).

a

b
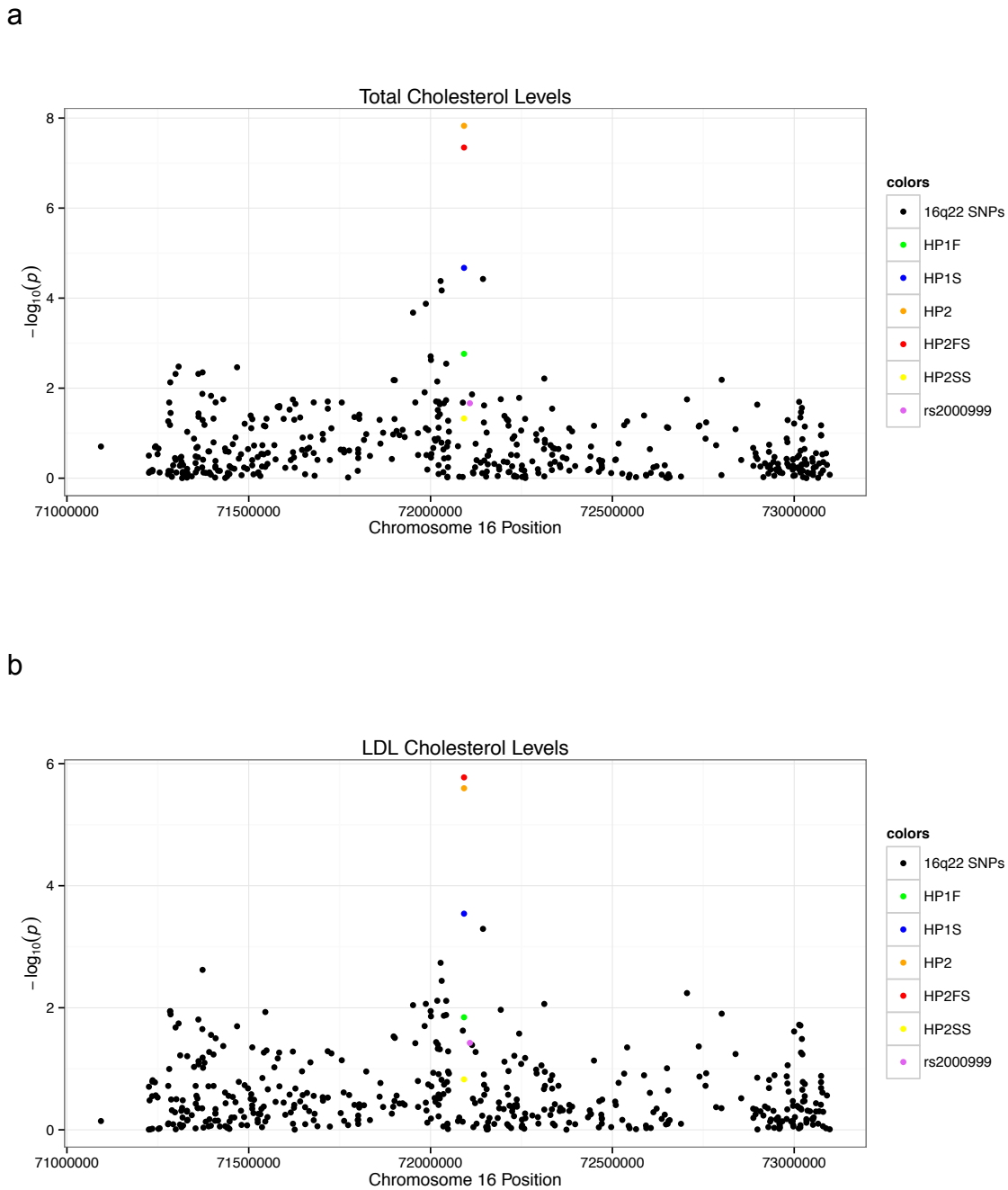
Figure 5.3. *HP2* associates to total and LDL cholesterol levels. Regional SNPs are shown in black, while the GWAS index SNP (rs2000999) is shown in pink. Other colors indicate *HP* structures (see legend). (a) *HP2* is most highly associated to total cholesterol levels. (b) *HP2FS* is the structural haplotype most associated with LDL cholesterol levels.

Additionally, we examined the effect size of each *HP* structural haplotype on both total cholesterol levels and LDL cholesterol levels. Both the *HP2FS* and the *HP2SS* haplotypes positively affected total and LDL cholesterol levels, while both *HP1S* and *HP1S* negatively affected total and LDL cholesterol levels (Figure 5.4).

a
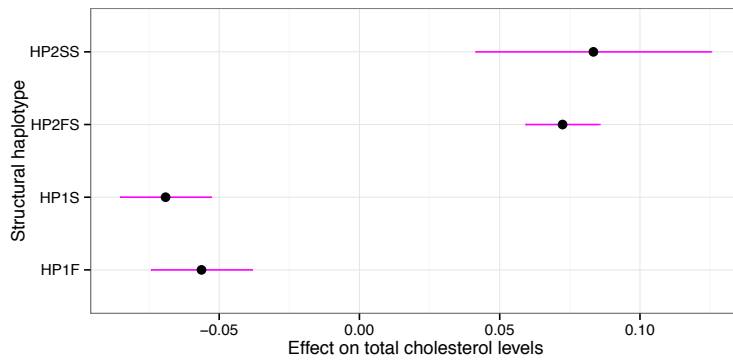
b
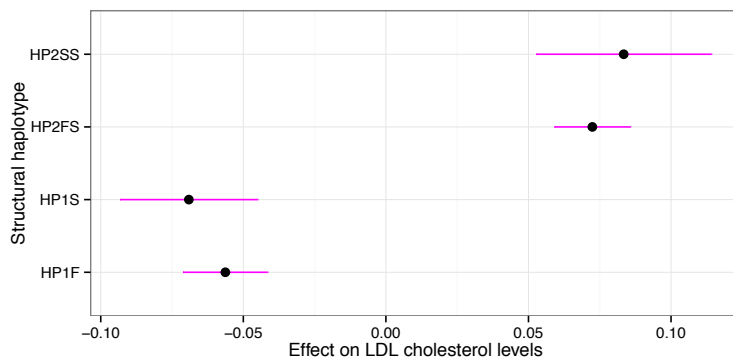
Figure 5.4. Effect size of *HP2* haplotypes on total cholesterol levels and LDL cholesterol levels. The black dots indicate the effect size (β) of each haplotype. The pink lines indicate the effect size +/- the standard error. (a) The effect of each *HP* structural haplotype on total cholesterol levels. (b) The effect of each *HP* structural haplotype on LDL cholesterol levels.

One potential mechanism by which *HP2* contributes to an increase in total cholesterol and LDL cholesterol levels is through its interaction with apolipoprotein E (ApoE). The ApoE protein is critical to maintaining low total cholesterol and LDL cholesterol levels[146], and the oxidation of ApoE is known to impair its function[162]. The HP protein directly binds ApoE[53,147] and serves as an ApoE antioxidant[53]; however, the HP2 form of the protein is a less efficient antioxidant[88]. We propose that *HP2* contributes to increased total and LDL cholesterol levels by providing insufficient antioxidant activity for ApoE.

Contributions

I received guidance and input for work presented in this chapter from Steve McCarroll. I designed all experiments and molecular assays, performed all molecular experiments, and did all data analysis in this chapter excluding the contributions of others mentioned below. Stephan Ripki performed the association study for *HP* structural haplotypes and Crohn's Disease. Rany Salem performed the association study for *HP* and cholesterol levels. The samples for the trypanosomiasis association study were provided by Martin Pollak. Genotypes for the malaria case-control dataset were provided by MalariaGEN Genomic Epidemiology Network[157].

Chapter 6

Discussion and future directions

Discussion

This thesis has discussed how we designed and applied methods for typing structurally complex haplotypes at the population level (Chapter 3), how we studied the evolution complex structure within populations, across populations, and across species (Chapter 4), and how we designed and implemented methods to incorporate complex genome structures into association studies (Chapter 5).

We observed that complex structural variation can occur on the megabase scale, affecting many genes in a region (17q21.31), or on the kilobase scale affecting a whole gene (*HPR*) or portions of a gene (*HP, C4*). Our comparisons of the structural evolution between and within regions showed that certain types of mutations can be stable: inversions, paralogous gene conversion, and dispersed duplications, while tandem duplications can be dynamic, rapidly accumulating additional structural mutations. Through a series of leave-one-out imputation trials we observed that the more stable structural mutations segregated on more uniform SNP haplotype backgrounds and imputed with greater accuracy.

Our constructed imputation reference panels for complex structural variation in several interesting regions will be extremely valuable for future association studies. They will make it possible to include multi-allelic structural variation in genomic studies while minimizing or avoiding costly and time-consuming direct testing of complex structure in large cohorts.

Future directions

Genome-wide studies of complex structural variation

This thesis examines a set of four structurally complex regions in the human genome; however, there are many more. A key extension of this work is to expand our methods of understanding complex structural variation to other complex loci. An important goal of the field of complex structural genomics is to ultimately develop genome-wide approaches to typing and incorporating complex structural haplotypes into association studies.

Extensive progress has recently been made to identify and type multi-allelic CNVs from genome-wide sequence data. Handsaker et al. have extended the GenomeSTRiP algorithm to scan whole genome sequence at the population level, map CNV breakpoints, generate precise genotypes at high copy number, and compute likelihoods for each copy number allele. They observed that simple deletion alleles can be imputed from surrounding SNPs, but the precise copy number of mCNVs often does not impute accurately[163] (as we observed for the *HPR* mCNV). Eventually, whole genome sequence may be used in association studies, in which case these mCNVs could be determined genome-wide using read depth. However, in the near term, targeted methods could be applied to mCNVs and other structural variants near significant GWAS variants.

Despite this recent advance in typing mCNVs, loci with overlapping CNVs and loci containing multiple classes of structural variation still require locus-specific strategies in order to be discovered and typed. Such highly complex

regions also currently need to be studied on an individual basis to be incorporated into association studies.

Haptoglobin's relationship to cholesterol levels

In this thesis we have demonstrated that the *HP2* structural allele associates to increased total cholesterol and LDL cholesterol levels. We hypothesize that this association relates to haptoglobin's physical interaction with the ApoE protein in which HP performs an antioxidant function. An important extension of this work is to test this hypothesis.

One method of examining whether the *HP2* allele affects cholesterol levels through its interaction with *APOE* is to test if alleles of these genes statistically interact. ApoE is itself an antioxidant. The E4 allele of *APOE* possesses the least antioxidant activity of the common alleles[164], and also is associated with increased LDL cholesterol[165]. We could test if individuals who have both *E4* at *APOE* and *HP2* at *HP* are more likely to have higher total and LDL cholesterol levels than expected from the two variants independently.

Appendix

Supplementary Tables and Figures

Supplementary Tables

Supplementary Table 1. Population identifiers

ASW African Ancestry in the Southwestern US
CEU Utah residents with Northern and Western European ancestry
CHB Han Chinese in Beijing, China
CHS Han Chinese South
CLM Colombian in Medellin, Colombia
FIN Finnish from Finland
GBR British from England and Scotland
JPT Japanese in Toyko, Japan
LWK Luhya in Webuye, Kenya
MEX Mexican Ancestry from Los Angeles USA
TSI Toscani in Italia
YRI Yoruba in Ibadan, Nigeria

Supplementary Table 2. Neanderthal and Denisova coverage of *HP2FS* haplotype.

| Chromosome	Position	Base	Denisova		Neanderthal		Variant	CNV copy
			Uniquely mapped reads	Mapped reads	Uniquely mapped reads	Mapped reads		
chr16	72091102	T	22	33	21	29	*HPR* gene conversion region	*HP2FS-Left*
chr16	72091103	G	23	35	18	26		
chr16	72091168	C	23	27	43	51		
chr16	72091169	T	22	27	42	51		
chr16	72091178	G	18	25	35	55		
chr16	72091180	C	17	25	34	57		
chr16	72091183	C	17	25	31	57		
chr16	72091185	T	16	24	29	57		
chr16	72091191	C	13	28	24	56		
chr16	72091192	T	12	27	24	57		
chr16	72091195	T	10	25	22	56		
chr16	72091201	T	7	22	18	59		
chr16	72091207	A	6	21	16	59		
chr16	72091216	C	4	22	12	61		
chr16	72091238	T	1	30	6	65		
chr16	72091245	G	1	36	4	64		
chr16	72091246	C	1	37	4	65		
chr16	72091258	T	0	40	3	68		
chr16	72091271	T	0	43	3	64		
chr16	72091295	A	0	35	2	61		
chr16	72091307	T	0	32	0	66		
chr16	72091308	G	0	32	0	63		
chr16	72091311	A	0	31	0	68		
chr16	72091382	G	0	19	0	40		
chr16	72091383	A	0	20	0	41		
chr16	72091384	G	0	20	0	39		
chr16	72091385	C	0	19	0	40		
chr16	72091386	A	0	21	0	43		
chr16	72091387	C	0	21	0	43		
chr16	72091388	T	0	21	0	47		
chr16	72091565	G	17	21	35	38	Highly diverged region Form L	
chr16	72091593	C	20	22	46	48		
chr16	72091598	G	21	22	47	48		
chr16	72091599	G	21	22	48	49		
chr16	72091602	C	21	21	48	49		
chr16	72091603	T	19	19	49	50		
chr16	72091614	C	21	21	50	51		
chr16	72091648	A	29	29	46	47		
chr16	72091808	C	34	34	49	50		
chr16	72091819	A	30	30	51	52		
chr16	72092826	C	23	23	46	46	Ancestral *HP* bases	*HP2FS-Right*
chr16	72092827	A	23	23	45	45		
chr16	72092892	T	25	25	48	48		
chr16	72092893	C	23	23	48	48		
chr16	72092902	T	25	25	48	48		
chr16	72092904	T	24	24	48	48		
chr16	72092907	T	24	24	50	50		
chr16	72092909	C	23	23	50	50		
chr16	72092915	T	26	26	52	52		
chr16	72092916	C	26	26	50	50		
chr16	72092919	C	24	24	47	47		
chr16	72092925	T	25	25	53	53		
chr16	72092931	T	25	25	53	53		
chr16	72092940	T	22	22	60	60		
chr16	72092962	G	22	22	69	69		
chr16	72092969	A	24	24	69	69		
chr16	72092970	G	23	23	68	68		
chr16	72092982	C	30	30	68	68		
chr16	72092995	T	30	30	66	66		
chr16	72093019	A	27	27	57	57		
chr16	72093031	A	37	37	61	61		
chr16	72093032	T	37	37	61	61		
chr16	72093035	G	36	36	59	60		
chr16	72093106	A	17	26	27	41		
chr16	72093107	G	17	27	26	40		
chr16	72093108	A	17	27	24	38		
chr16	72093109	G	16	27	24	39		
chr16	72093110	A	15	28	23	40		
chr16	72093111	G	15	29	23	41		
chr16	72093112	C	14	29	22	40		
chr16	72093289	A	32	33	42	42		
chr16	72093317	T	33	33	41	41	Highly diverged region Form R	
chr16	72093322	C	32	32	40	40		
chr16	72093323	A	30	30	39	39		
chr16	72093326	G	29	29	39	39		
chr16	72093327	A	29	29	39	39		
chr16	72093338	C	21	22	29	33		
chr16	72093372	G	12	24	20	32		
chr16	72093532	A	35	35	44	44		
chr16	72093543	C	28	30	36	44		

Supplementary Figures

Supplementary Figure 1. *HP* contains paralogous gene conversion from *HPR* and a highly diverged region. An alignment of the structurally variant region for human structural haplotypes and other primates shows distinct polymorphic sections. The paralog which provided the sequence is indicated, excluding the Marmoset and Squirrel Monkey sequences as these primates lack the triplication. Only the fixed differences between human *HP* structural forms are plotted. The blue and lavender region shows that human *HP1F* and *HP2FS-Left* have 33 derived mutations (with respect to the chimpanzee and bonobo *HP*) which match the human *HPR* gene. The series of mutation that gave rise to the highly diverged region is unclear, as the great apes are also highly polymorphic in this region, and multiple bases have potentially deep coalescence.

Supplementary Figure 2. Sequence differences between structural haplotypes within the structurally variant region of *HP*. The X axis lists the base pair position of each nucleotide variant (hg19). The Y axis indicates the structural haplotype and HapMap or 1000 Genomes Project individual who provided the sequence. The lavender sequence indicates bases which result from paralogous gene conversion from *HPR*, the blue bases indicate the standard *HP* base. The pink and green bases indicate alternate forms of the highly diverged region, while the grey and white colored bases indicate other sequence variation. The derived base (with respect to chimpanzee and bonobo) is indicated in white.

103
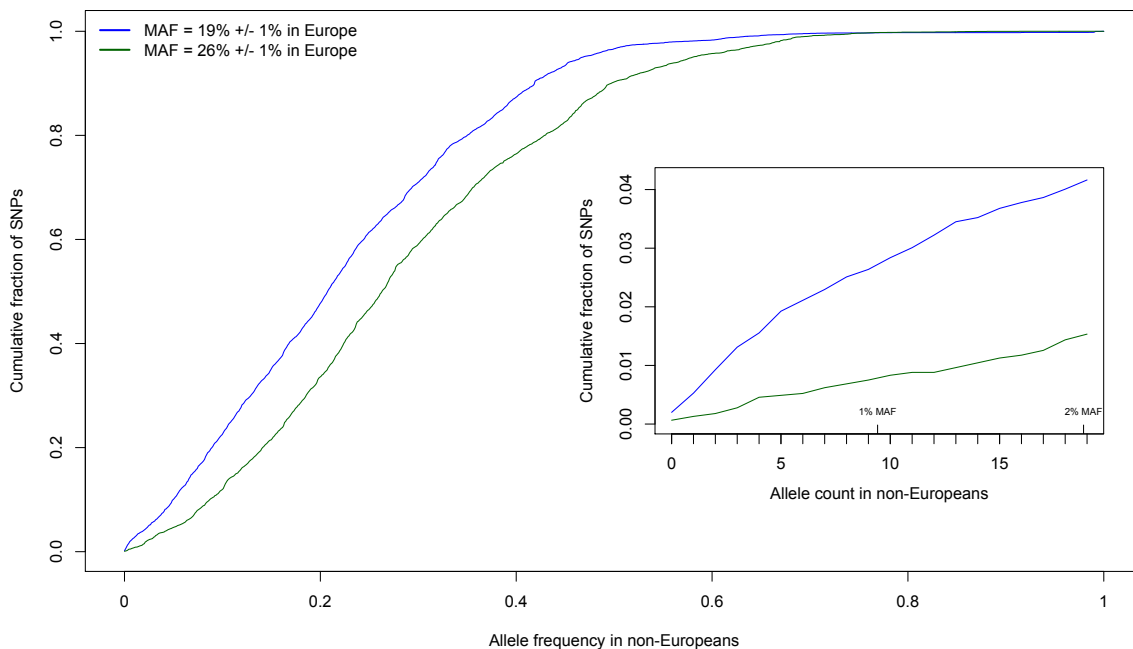
Supplementary Figure 3. SNPs with highly differentiated allele frequencies between European and non-European populations in 1000 Genomes phase 1. Minor allele frequency distribution of SNPs in non-Europeans (n=471, populations CHB, CHS, JPT, LWK and YRI) with MAF of 18% - 20% (blue) and 25% - 27% in Europeans (n=379, populations CEU, FIN, GBR, IBS and TSI), corresponding to the allele frequencies we observed for the alpha and beta duplications in CEU. Inset shows the low frequency portion of the same distribution by allele count in the non-European populations. The SNPs were ascertained and genotyped in Phase 1 of the 1000 Genomes Project on chromosome 17, excluding the region 17:43165000-45785000 (+/- 1Mb from the inversion region).
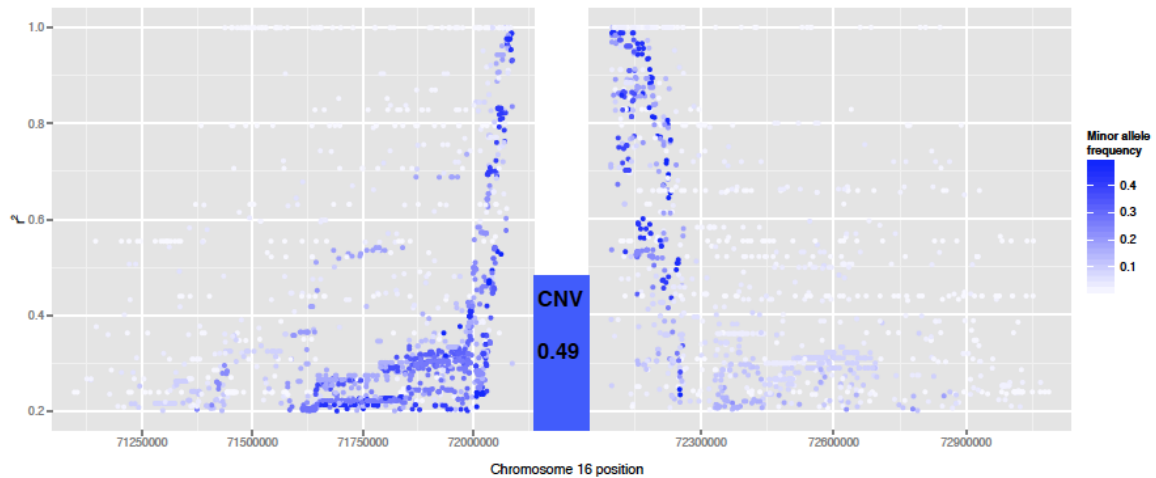
Supplementary Figure 4. Linkage disequilibrium between SNPs on opposite sides of the CNV and between SNPS and the CNV in European populations. The Y axis indicates position on chromosome 16, while the X axis indicates the r^2 for linkage disequilibrium. The linkage disequilibrium value for each SNP is the maximum LD shared with a SNP on the opposite side of the structural variation. As indicated, the maximum LD between the CNV and a SNP is 0.49. The minor allele frequency for each SNP is indicated by color.
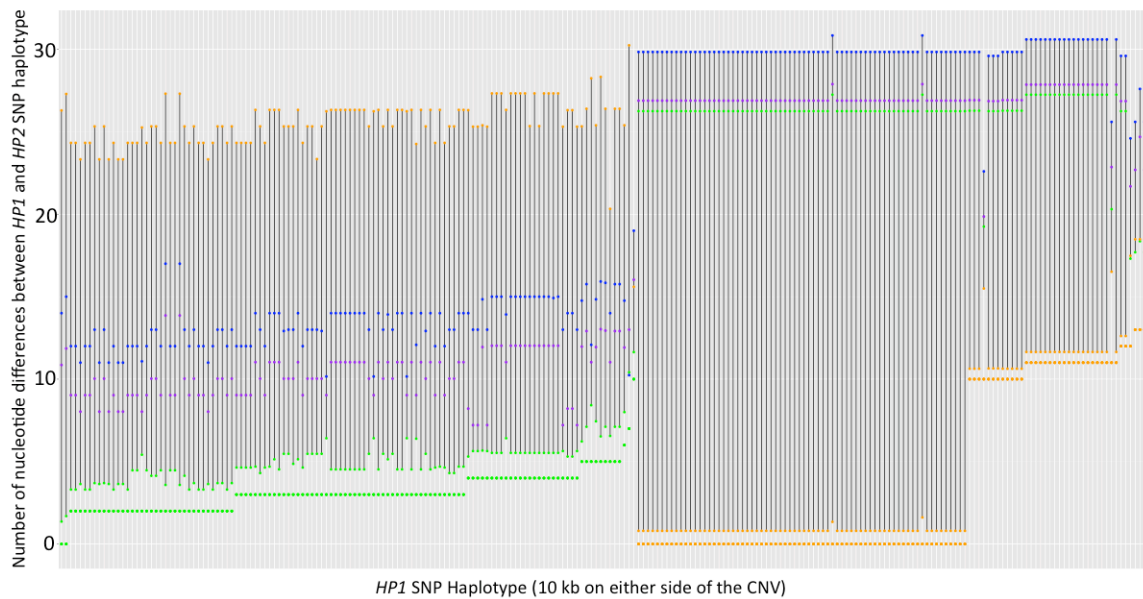
Supplementary Figure 5. The vast majority of *HP1* SNP haplotypes are closely related to a specific class of *HP2* SNP haplotype. The colored dots indicate one of four common SNP haplotypes in the region. These colors are the same as those in Figure 4.3. The each point on the X axis is a different *HP1* SNP haplotype. The colored dots on the Y axis indicate the number of nucleotide differences between the *HP1* SNP haplotype and each class of HP2 SNP haplotype (indicated by color). For the most closely related class, an additional dot is plotted, which shows the most closely related haplotype within that class. In most cases, each *HP1* haplotype is closely related to a specific type of *HP2* SNP haplotype. However, eight *HP1* haplotypes were of ambiguous origin (defined by the distance between the closest and the second closest SNP haplotype being fewer than nucleotides. These eight are examined below in Supplementary Figure 6.
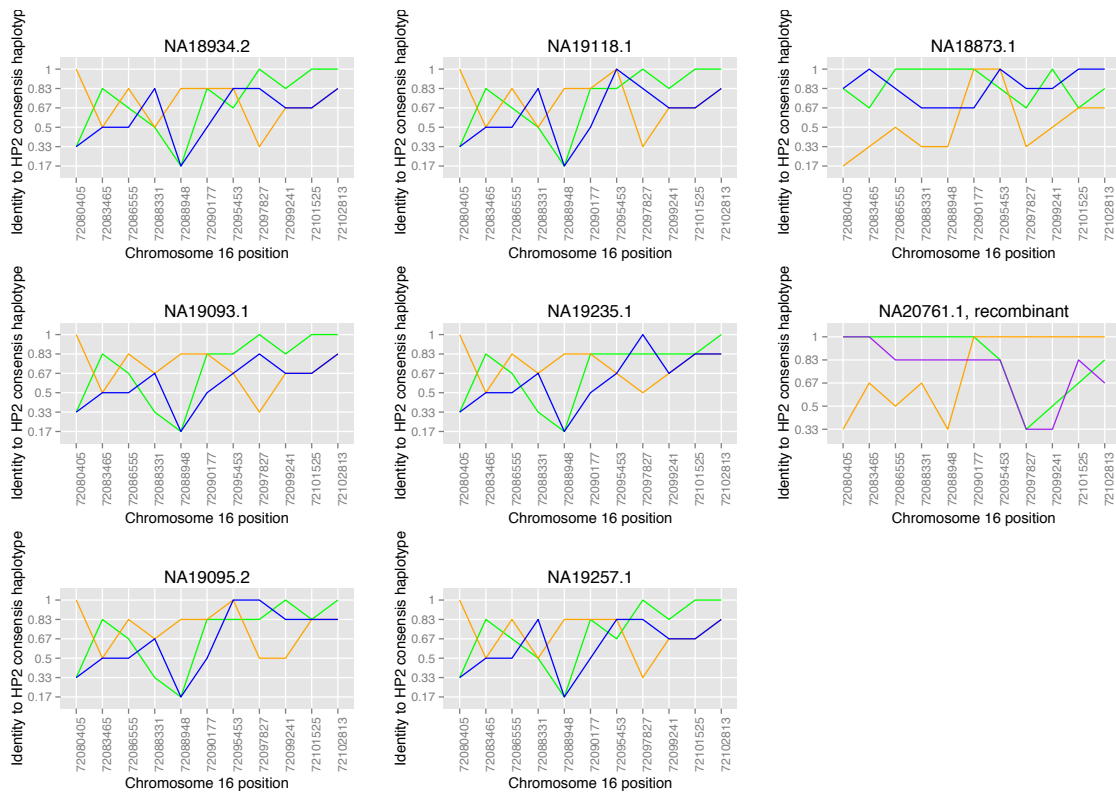
Supplementary Figure 6. Some *HP1* SNP haplotypes are not closely related to a specific class of *HP2* SNP haplotype. Each plot examines a specific *HP1* SNP that is not closely related to a specific, modern, *HP2* SNP haplotype. The X axis indicates the genomic coordinates of the SNPs and the Y axis indicates the percent identity to a specific *HP2* SNP haplotype (indicated by each color). The most closely related *HP2* SNP haplotype varies throughout the region. While NA20761 appears to be recombinant (the left half is closely related to the green haplotype while the right is closely related to the orange haplotype, the other *HP1* haploypes could be ancient deletions or ancestral *HP1s*, which are not derived from deletions in *HP2*. Distances between the haplotypes are measured with a sliding window of six SNPs.

References

1. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat Genet* **37,** 129–137 (2005).

2. Maeda, N., Yang, F., Barnett, D. R., Bowman, B. H. & Smithies, O. Duplication within the haptoglobin Hp2 gene. *Nature* **309,** 131–135 (1984).

3. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).

4. Dangel, A. W., Baker, B. J., Mendoza, A. R. & Yu, C. Y. Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. *Immunogenetics* **42,** 41–52 (1995).

5. Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G. & Chee, M. S. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* **37,** 549–554 (2005).

6. Goidts, V., Armengol, L., Schempp, W., Conroy, J. & Nowak, N. Identification of large-scale human-specific copy number differences by inter-species array comparative genomic hybridization. *Human genetics* **119**, 185-198 (2006).

7. Mccarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40,** 1166–1174 (2008).

8. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).

9. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297-304 (2000).

10. Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18,** 337–340 (2002).

11. Fryxell, K. J. & Moon, W.-J. CpG mutation rates in the human genome are highly dependent on local GC content. *Molecular Biology and Evolution* **22,** 650–658 (2005).

12. Tian, D. *et al.* Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455,** 105–108 (2008).

13. Zuckerkandl, E. & Pauling, L. Molecular disease, evolution and genetic heterogeneity. Horizons in Biochemistry, Academic Press, New York, 189–225 (1962).

14. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624 - 626 (1968).

15. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22,** 160–174 (1985).

16. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).

17. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449,** 913–918 (2007).

18. Ledbetter, J. A. *et al.* Evolutionary conservation of surface molecules that distinguish T lymphocyte helper/inducer and cytotoxic/suppressor subpopulations in mouse and man. *J. Exp. Med.* **153,** 310–323 (1981).

19. Smith, V. & Barrell, B. G. Cloning of a yeast U1 snRNP 70K protein homologue: functional conservation of an RNA-binding domain between humans and yeast. *The EMBO Journal* **10,** 2627–2634 (1991).

20. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).

21. Slatkin, M. & Hudson, R. R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555-562 (1991).

22. Sudmant, P. H. *et al.* Evolution and diversity of copy number variation in the great ape lineage. *Genome Res* **23,** 1373–1382 (2013).

23. Graubert, T. A. *et al.* A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet* **3,** e3 (2007).

24. Guryev, V. *et al.* Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet* **40,** 538–545 (2008).

25. Chen, W.-K., Swartz, J. D., Rush, L. J. & Alvarez, C. E. Mapping DNA structural variation in dogs. *Genome Res* **19,** 500–509 (2009).

26. Nicholas, T. J. *et al.* The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Res* **19,** 491–499 (2009).

27.	Liu, G. E. *et al.* Analysis of copy number variations among diverse cattle breeds. *Genome Res* **20,** 693–703 (2010).

28.	Koolen, D. A. *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat Genet* **38,** 999–1001 (2006).

29.	Widen, E., Lehto, M., Kanninen, T. & Walston, J. Association of a polymorphism in the β3-adrenergic–receptor gene with features of the insulin resistance syndrome in Finns. *N Engl J Med* **333**, 348-352 (1995).

30.	Shiang, R. *et al.* Association of transforming growth-factor alpha gene polymorphisms with nonsyndromic cleft palate only (CPO). *Am. J. Hum. Genet.* **53,** 836–843 (1993).

31.	Walston, J., Silver, K. & Bogardus, C. Time of Onset of Non-Insulin-Dependent Diabetes Mellitus and Genetic Variation in the β3-Adrenergic–Receptor Gene. *N Engl J Med* **333**, 343-347 (1995).

32.	Ruiz, J. *et al.* Insertion/deletion polymorphism of the angiotensin-converting enzyme gene is strongly associated with coronary heart disease in non-insulin-dependent diabetes mellitus. *Proc Natl Acad Sci USA* **91,** 3662–3665 (1994).

33.	Ohishi, M., Rakugi, H. & Ogihara, T. Association between a deletion polymorphism of the angiotensin-converting-enzyme gene and left ventricular hypertrophy. *N Engl J Med* **9**, 330 (1994).

34.	Rigat, B. *et al.* An insertion/deletion polymorphism in the angiotensin I-converting enzyme gene accounting for half the variance of serum enzyme levels. *J. Clin. Invest.* **86,** 1343–1346 (1990).

35.	Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).

36.	Mccarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* **40,** 1107–1112 (2008).

37.	Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* **41,** 25–34 (2009).

38.	The UK Parkinson's Disease Consortium and The Wellcome Trust Case Control Consortium 2 *et al.* Dissection of the genetics of Parkinson's

disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. *Hum Mol Genet* **20,** 345–353 (2010).

39.   Edwards, T., Scott, W. & Almonte, C. Genome‐Wide Association Study Confirms SNPs in SNCA and the MAPT Region as Common Risk Factors for Parkinson Disease. *Annals of human Genetics* **74**, 97-109 (2010).

40.   Höglinger, G. U. *et al.* Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat Genet* **43,** 699–705 (2011).

41.   Zody, M. C. *et al.* Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet* **40,** 1076–1083 (2008).

42.   Ridley, M. Evolution. 751 (2004).

43.   Hardy, J. *et al.* Evidence suggesting that Homo neanderthalensis contributed the H2 MAPT haplotype to Homo sapiens. *Biochem. Soc. Trans.* **33,** 582–585 (2005).

44.   Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).

45.   Donnelly, M. P. *et al.* The Distribution and Most Recent Common Ancestor of the 17q21 Inversion in Humans. *The American Journal of Human Genetics* **86,** 161–171 (2010).

46.   Templeton, A. Out of Africa again and again. *Nature* **416**, 45-51 (2002).

47.   Kong, A. *et al.* Recombination rate and reproductive success in humans. *Nat Genet* **36,** 1203–1206 (2004).

48.   Cantor, R. M. *et al.* Replication of Autism Linkage: Fine-Mapping Peak at 17q21. *The American Journal of Human Genetics* **76,** 1050–1056 (2005).

49.   Van Limbergen, J., Wilson, D. C. & Satsangi, J. The genetics of Crohn's disease. *Annu Rev Genomics Hum Genet* **10,** 89–116 (2009).

50.   Chowdhury, R., Bois, P. R. J., Feingold, E., Sherman, S. L. & Cheung, V. G. Genetic analysis of variation in human meiotic recombination. *PLoS Genet* **5,** e1000648 (2009).

51.   International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).

52.   Allison, A. C. & Rees, W. A. The binding of haemoglobin by plasma proteins (haptoglobins). *British medical journal* **2**, 1137–1143 (1957).

53.   Salvatore, A., Cigliano, L., Carlucci, A., Bucci, E. M. & Abrescia, P. Haptoglobin binds apolipoprotein E and influences cholesterol esterification in the cerebrospinal fluid. *J. Neurochem.* **110,** 255–263 (2009).

54.   Tripathi, A. *et al.* Identification of human zonulin, a physiological modulator of tight junctions, as prehaptoglobin-2. *Proceedings of the National Academy of Sciences* **106,** 16799–16804 (2009).

55.   Kurosky, A. *et al.* Covalent structure of human haptoglobin: a serine protease homolog. *Proc Natl Acad Sci USA* **77,** 3388–3392 (1980).

56.   Smithies, O. & Walker, N. F. Genetic control of some serum proteins in normal humans. *Nature* **176,** 1265–1266 (1955).

57.   Wobeto, V., Zaccariotto, T. R. & Sonati, M. F. Polymorphism of human haptoglobin and its clinical importance. *Genetics and Molecular Biology* **31**, 1415-4757 (2008).

58.   Smithies, O. Grouped variations in the occurrence of new protein components in normal human serum. *Nature* **175**, 307-308 (1955).

59.   Rodriguez, S. *et al.* Molecular and Population Analysis of Natural Selection on the Human Haptoglobin Duplication. *Annals of Human Genetics* **76,** 352–362 (2012).

60.   1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

61.   Koch W, Latz W, Eichinger M, Roguin A, Levy AP, Schömig A, Kastrati A. Genotyping of the Common Haptoglobin Hp 1/2 Polymorphism Based on PCR. Clin Chem. **48**, 1377-82 (2002).

62.   Pechlaner, R., Kiechl, S., Willeit, P. & Demetz, E. Haptoglobin 2-2 Genotype is Not Associated With Cardiovascular Risk in Subjects With Elevated Glycohemoglobin—Results From the Bruneck Study. *Circulation Research* **16**, e000732 (2014).

63.   Asleh, R. *et al.* Haptoglobin genotype-dependent differences in macrophage lysosomal oxidative injury. *Journal of Biological Chemistry* **289,** 16313–16325 (2014).

64.   Atkinson, S. H. *et al.* Epistasis between the haptoglobin common variant and α+thalassemia influences risk of severe malaria in Kenyan children. *Blood* **123,** 2008–2016 (2014).

65.   Hardwick, R. J. *et al.* Haptoglobin (HP) and Haptoglobin-related protein (HPR) copy number variation, natural selection, and trypanosomiasis. *Human genetics* (2013). doi:10.1007/s00439-013-1352-x

66.   Asleh, R., Ward, J., Levy, N. S. & Safuri, S. Accelerated atherosclerosis in individuals with the haptoglobin 2-2 genotype and diabetes is mediated by macrophage lysosomal injury and apoptosis. *Atherosclerosis* (2014).

67.   Soejima, M. *et al.* Genetic factors associated with serum haptoglobin level in a Japanese population. *Clin. Chim. Acta* **433,** 54–57 (2014).

68.   Lazalde, B., Huerta-Guerrero, H. M., Simental-Mendía, L. E., Rodríguez-Morán, M. & Guerrero-Romero, F. Haptoglobin 2-2 genotype is associated with TNF- α and IL-6 levels in subjects with obesity. *Dis. Markers* **2014,** 912756 (2014).

69.   Amor, A. J. *et al.* Haptoglobin genotype and risk of diabetic nephropathy in patients with type 1 diabetes mellitus: a study on a Spanish population. *Nefrologia* **34,** 212–215 (2014).

70.   Hamdy, G., Hendy, O. M., Mahmoud, H. & El-sebaey, A. Haptoglobin phenotypes as a risk factor for coronary artery disease in type 2 diabetes mellitus: An Egyptian study. *Egyptian Journal of Human Genetics* (2014).

71.   Khazaei, H. A. *et al.* Evaluation of haptoglobin phenotypes in association with clinical features of patients suffered from preterm labor disease. *Acta Med Iran* **52,** 106–110 (2014).

72.   Barbosa, L., Miranda-Vilela, A. L. & Hiragi, C. O. Haptoglobin and myeloperoxidase (− G463A) gene polymorphisms in Brazilian sickle cell patients with and without secondary iron overload. *Blood Cells* (2014).

73.   Moussa, A. *et al.* Association between haptoglobin 2-2 genotype and coronary artery disease and its severity in a tunisian population. *Biochem. Genet.* **52,** 269–282 (2014).

74.   Matos, A. *et al.* In women with previous pregnancy hypertension, levels of cardiovascular risk biomarkers may be modulated by haptoglobin polymorphism. *Obstet Gynecol Int* **2014,** 361727 (2014).

75.   Dzudzor, B., Nuwormegbe, S. & Asmah, R. H. Haptoglobin Genotypes And Longevity Among The Ghanaian Population. *IJSTR* (2014).

76.   Costacou, T., Secrest, A. M. & Ferrell, R. E. Haptoglobin genotype and cerebrovascular disease incidence in type 1 diabetes. *Diab Vasc Dis Res.* (2014).

77.   Costacou, T. *et al.* The Haptoglobin 1 allele correlates with white matter hyperintensities in middle-aged adults with type 1 diabetes. *Diabetes* (2014). doi:10.2337/db14-0723

78.   Papp, M. *et al.* Haptoglobin Polymorphisms Are Associated with Crohn's Disease, Disease Behavior, and Extraintestinal Manifestations in Hungarian Patients. *Dig Dis Sci* **52,** 1279–1284 (2007).

79.   Maza, I. *et al.* The association of Haptoglobin polymorphism with Crohn's disease in Israel. *J Crohns Colitis* **2,** 214–218 (2008).

80.   Levy, A. P. *et al.* Haptoglobin phenotype and prevalent coronary heart disease in the Framingham offspring cohort. *Atherosclerosis* **172,** 361–365 (2004).

81.   De Bacquer, D. *et al.* Haptoglobin polymorphism as a risk factor for coronary heart disease mortality. *Atherosclerosis* **157,** 161–166 (2001).

82.   Nevo, S. & Tatarsky, I. Serum haptoglobin types and leukemia. *Human genetics* **73,** 240–244 (1986).

83.   Campregher, P. V., Lorand-Metze, I., Grotto, H. Z. W. & Sonati, M. de F. Haptoglobin phenotypes in Brazilian patients with leukemia. *J. Bras. Patol. Med. Lab.* **40,** 307–309 (2004).

84.   Ping, C., Yuanzhong, C., Qilian, Z. & Xiuping, L. Relationship between Haptoglobin Genetic Polymorphisms and Acute Leukemia. *Journal of Fujian Medical University* **38,** 139–140 (2004).

85.   Delanghe, J. R. *et al.* Haptoglobin polymorphism, iron metabolism and mortality in HIV infection. *AIDS* **12,** 1027–1032 (1998).

86.   Zaccariotto, T. R. *et al.* Haptoglobin polymorphism in a HIV-1 seropositive Brazilian population. *J. Clin. Pathol.* **59,** 550–553 (2006).

87.   Cox, S. E. *et al.* Haplotype Association between Haptoglobin (Hp2) and Hp Promoter SNP (A-61C) May Explain Previous Controversy of Haptoglobin and Malaria Protection. *PLoS ONE* **2,** e362 (2007).

88.   Melamed-Frank, M. Structure-function analysis of the antioxidant properties of haptoglobin. *Blood* **98,** 3693–3698 (2001).

89.   Guetta, J., Strauss, M., Levy, N. S., Fahoum, L. & Levy, A. P. Haptoglobin genotype modulates the balance of Th1/Th2 cytokines produced by macrophages exposed to free hemoglobin. *Atherosclerosis* **191,** 48–53 (2007).

90. McEvoy, S. M. & Maeda, N. Complex events in the evolution of the haptoglobin gene cluster in primates. *J. Biol. Chem.* **263,** 15740–15747 (1988).

91. Atkinson, S. H. *et al.* The haptoglobin 2-2 genotype is associated with a reduced incidence of Plasmodium falciparum malaria in children on the coast of Kenya. *Clin. Infect. Dis.* **44,** 802–809 (2007).

92. Schultze, H. E. & Heremans, J. F. Molecular biology of human proteins: with special reference to plasma proteins. (1966).

93. Wiuf, C., Zhao, K., Innan, H. & Nordborg, M. The probability and chromosomal extent of trans-specific polymorphism. *Genetics* **168,** 2363–2372 (2004).

94. Leffler, E. M. *et al.* Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339,** 1578–1582 (2013).

95. Maeda, N., McEvoy, S. M., Harris, H. F., Huisman, T. H. & Smithies, O. Polymorphisms in the human haptoglobin gene cluster: chromosomes with multiple haptoglobin-related (Hpr) genes. *Proc Natl Acad Sci USA* **83,** 7395–7399 (1986).

96. Drain, J. Haptoglobin-related Protein Mediates Trypanosome Lytic Factor Binding to Trypanosomes. *Journal of Biological Chemistry* **276,** 30254–30260 (2001).

97. Hager, K. M. & Hajduk, S. L. Mechanism of resistance of African trypanosomes to cytotoxic human HDL. *Nature* **385,** 823–826 (1997).

98. Carroll, M. C. The complement system in regulation of adaptive immunity. *Nat Immunol* (2004).

99. Gasque, P. *et al.* Expression of the complement classical pathway by human glioma in culture. A model for complement expression by nerve cells. *J. Biol. Chem.* **268,** 25068–25074 (1993).

100. Terai, K., Walker, D. G., McGeer, E. G. & McGeer, P. L. Neurons express proteins of the classical complement pathway in Alzheimer disease. *Brain Res.* **769,** 385–390 (1997).

101. Carroll, M. C., Fathallah, D. M., Bergamaschini, L., Alicot, E. M. & Isenman, D. E. Substitution of a single amino acid (aspartic acid for histidine) converts the functional activity of human complement C4B to C4A. *Proc Natl Acad Sci USA* **87,** 6868–6872 (1990).

102. Yu, C. Y. *et al.* Dancing with complement C4 and the RP-C4-CYP21-TNX (RCCX) modules of the major histocompatibility complex. *Prog. Nucleic Acid Res. Mol. Biol.* **75,** 217–292 (2003).

103. Kawaguchi, H., Zaleska-Rutczynska, Z., Figueroa, F., O'hUigin, C. & Klein, J. C4 genes of the chimpanzee, gorilla, and orang-utan: evidence for extensive homogenization. *Immunogenetics* **35,** 16–23 (1992).

104. Paz-Artal, E., Corell, A., Alvarez, M., Varela, P. & Allende, L. C4 gene polymorphism in primates: evolution, generation, and Chido and Rodgers antigenicity. *Immunogenetics* (1994).

105. Wu, Y. L. *et al.* Sensitive and specific real-time polymerase chain reaction assays to accurately determine copy number variations (CNVs) of human complement C4A, C4B, C4-long, C4-short, and RCCX modules: elucidation of C4 CNVs in 50 consanguineous subjects with defined HLA genotypes. *J. Immunol.* **179,** 3012–3025 (2007).

106. Boettger, L. M., Handsaker, R. E., Zody, M. C. & Mccarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genet* **44,** 881–885 (2012).

107. Handsaker, R. E., Korn, J. M., Nemesh, J. & Mccarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43,** 269–276 (2011).

108. Hindson, B. J. *et al.* High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.* **83,** 8604–8610 (2011).

109. Derti, A., Roth, F. P., Church, G. M. & Wu, C.-T. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* **38,** 1216–1220 (2006).

110. Lupski, J. R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14,** 417–422 (1998).

111. Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* **39,** 31–40 (2007).

112. Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329,** 841–845 (2010).

113. Kothari *et al. Essentials Of Human Genetics Fifth Edition.* (Universities Press, 2009).

114.   Browning, S. R. Missing data imputation and haplotype phase inference for genome-wide association studies. *Human genetics* **124,** 439–450 (2008).

115.   Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310,** 321–324 (2005).

116.   Asakawa, J., Kodaira, M., Nakamura, N., Satoh, C. & Fujita, M. Chimerism in humans after intragenic recombination at the haptoglobin locus during early embryogenesis. *Proc Natl Acad Sci USA* **96,** 10314–10319 (1999).

117.   Regan, J.F., Kamitaki N., Legler, T., Cooper,S., Klitgord1, N., Karlin-Neumann, G., Wong, C., Hodges, S., Koehler, R., Tzonev, S., Steven A McCarroll. A rapid molecular approach for chromosomal phasing. *PLoS ONE.* (In Review).

118.   Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).

119.   Meyer, M., Kircher, M., Gansauge, M. T., Li, H. & Racimo, F. A high-coverage genome sequence from an archaic Denisovan individual. *Science* (2012).

120.   Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet* **13,** 745–753 (2012).

121.   Lawlor, D. A., Ward, F. E., Ennis, P. D., Jackson, A. P. & Parham, P. HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. *Nature* **335,** 268–271 (1988).

122.   Kaessmann, H., Wiebe, V. & Pääbo, S. Extensive nuclear DNA sequence diversity among chimpanzees. *Science* **286,** 1159–1162 (1999).

123.   Lazzaro, B. P. & Clark, A. G. Evidence for recurrent paralogous gene conversion and exceptional allelic divergence in the Attacin genes of Drosophila melanogaster. *Genetics* **159,** 659–671 (2001).

124.   Galtier, N. Gene conversion drives GC content evolution in mammalian histones. *Trends Genet* **19,** 65–68 (2003).

125.   Weiss, K. M., Kidd, K. K. & Kidd, J. R. Human genome diversity project. *Evolutionary Anthropology* **3**, 80-82 (1992).

126.   International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58

(2010).

127.   Browning, B. L. & Browning, S. R. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics* **84,** 210–223 (2009).

128.   Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327,** 836–840 (2010).

129.   Yu, L., Song, Y. & Wharton, R. P. E(nos)/CG4699 required for nanos function in the female germ line of Drosophila. *Genesis* **48,** 161–170 (2010).

130.   Smith, E. R. *et al.* A human protein complex homologous to the Drosophila MSL complex is responsible for the majority of histone H4 acetylation at lysine 16. *Mol Cell Biol* **25,** 9175–9188 (2005).

131.   Li, X., Wu, L., Corsa, C. A. S., Kunkel, S. & Dou, Y. Two mammalian MOF complexes regulate transcription activation by distinct mechanisms. *Molecular cell* **36,** 290–301 (2009).

132.   Spagnuolo, M. S. *et al.* Analysis of the haptoglobin binding region on the apolipoprotein A-I-derived P2a peptide. *J. Pept. Sci.* **19,** 220–226 (2013).

133.   Wejman, J. C., Hovsepian, D., Wall, J. S., Hainfeld, J. F. & Greer, J. Structure and assembly of haptoglobin polymers by electron microscopy. *J Mol Biol* **174,** 343–368 (1984).

134.   Malaria, R. B. World malaria report 2005. *World Health Organization and UNICEF* (2005).

135.   Elagib, A. A., Kider, A. O., Akerström, B. & Elbashir, M. I. Association of the haptoglobin phenotype (1-1) with falciparum malaria in Sudan. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **92,** 309–311 (1998).

136.   Quaye, I. K. *et al.* Haptoglobin 1-1 is associated with susceptibility to severe Plasmodium falciparum malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **94,** 216–219 (2000).

137.   Loftus, E. V. Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. *Gastroenterology* **126,** 1504–1517 (2004).

138.   Loftus, E. V. & Schoenfeld, P. The epidemiology and natural history of Crohn's disease in population-based patient cohorts from North America:

a systematic review. *Aliment Pharmacol Ther.* **16**, 51-60 (2002).

139. Wyatt, J. *et al.* Increased gastric and intestinal permeability in patients with Crohn's disease. *Am. J. Gastroenterol.* **92,** 1891–1896 (1997).

140. Sharrett, A. R., Ballantyne, C. M., Coady, S. A. & Heiss, G. Coronary heart disease prediction from lipoprotein cholesterol levels, triglycerides, lipoprotein (a), apolipoproteins AI and B, and HDL density subfractions. *Circulation* **104**, 1108-1113 (2001).

141. Virani, S. S., Wong, N. D., Woo, D., Turner, M. B. & Stroke, S. S. Heart disease and stroke statistics--2014 update: a report from the American Heart Association. *Circulation* (2014).

142. Libby, P., Schoenbeck, U., Mach, F., Selwyn, A. P. & Ganz, P. Current concepts in cardiovascular pathology: the role of LDL cholesterol in plaque rupture and stabilization. *Am. J. Med.* **104,** 14S–18S (1998).

143. Pieters, M. N., Schouten, D. & Van Berkel, T. J. In vitro and in vivo evidence for the role of HDL in reverse cholesterol transport. *Biochim Biophys Acta* **1225,** 125–134 (1994).

144. Kinosian, B., Glick, H. & Garland, G. Cholesterol and coronary heart disease: predicting risks by levels and ratios. *Ann. Intern. Med.* **121,** 641–647 (1994).

145. Law, M. R., Wald, N. J. & Rudnicka, A. R. Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis. *BMJ* **326,** 1423 (2003).

146. Ishibashi, S., Herz, J., Maeda, N., Goldstein, J. L. & Brown, M. S. The two-receptor model of lipoprotein clearance: tests of the hypothesis in 'knockout' mice lacking the low density lipoprotein receptor, apolipoprotein E, or both proteins. *Proc Natl Acad Sci USA* **91,** 4431–4435 (1994).

147. Cigliano, L., Pugliese, C. R., Spagnuolo, M. S., Palumbo, R. & Abrescia, P. Haptoglobin binds the antiatherogenic protein apolipoprotein E - impairment of apolipoprotein E stimulation of both lecithin:cholesterol acyltransferase activity and cholesterol uptake by hepatocytes. *FEBS J.* **276,** 6158–6171 (2009).

148. World Health Organization. Control and surveillance of African trypanosomiasis. Report of a WHO Expert Committee. WHO technical report series 881 (1998).

149. Dawkins, R. & Krebs, J. R. Arms races between and within species. *Proc.*

R. Soc. B **205**, 489-511 (1979).

150. Murray, H. W. & Cohn, Z. A. Macrophage oxygen-dependent antimicrobial activity. III. Enhanced oxidative metabolism as an expression of macrophage activation. *J. Exp. Med.* **152,** 1596–1609 (1980).

151. Tripodi, K. E. J., Menendez Bravo, S. M. & Cricco, J. A. Role of heme and heme-proteins in trypanosomatid essential metabolic pathways. *Enzyme Res* **2011,** 873230 (2011).

152. Smith, A. B., Esko, J. D. & Hajduk, S. L. Killing of trypanosomes by the human haptoglobin-related protein. *Science* **268,** 284–286 (1995).

153. Vanhollebeke, B. *et al.* A haptoglobin-hemoglobin receptor conveys innate immunity to Trypanosoma brucei in humans. *Science* **320,** 677–681 (2008).

154. De Greef, C. & Hamers, R. The serum resistance-associated (SRA) gene of Trypanosoma brucei rhodesiense encodes a variant surface glycoprotein-like protein. *Mol. Biochem. Parasitol.* **68,** 277–284 (1994).

155. DeJesus, E., Kieft, R., Albright, B., Stephens, N. A. & Hajduk, S. L. A single amino acid substitution in the group 1 Trypanosoma brucei gambiense haptoglobin-hemoglobin receptor abolishes TLF-1 binding. *PLoS Pathog* **9,** e1003317 (2013).

156. Geschwind, D. H., Sowinski, J. & Lord, C. The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *American Journal of human genetics* **69**, 463-6 (2001).

157. Achidi, E. A., Agbenyega, T., Allen, S., Amodu, O. & Bojang, K. A global network for investigating the genomic epidemiology of malaria. *Nature* **456**, 732-737 (2008).

158. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).

159. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am. J. Epidemiol.* **129,** 687–702 (1989).

160. Bild, D. E., Bluemke, D. A. & Burke, G. L. Multi-ethnic study of atherosclerosis: objectives and design. *Am J Epidemiol.* **156** (9), 871-881 (2002).

161. Fried, L. P., Borhani, N. O., Enright, P. & Furberg, C. D. The

cardiovascular health study: design and rationale. *Ann Epidemiol.* **3**, 263-76 (1991).

162.   Yang, Y., Cao, Z., Tian, L., Garvey, W. T. & Cheng, G. VPO1 mediates ApoE oxidation and impairs the clearance of plasma lipids. *PLoS ONE* **8,** e57571 (2013).

163.   Handsaker, R.E., Van Doren, V., Berman, J.R., Genovese, G., McCarroll, S.A. Alleles and haplotypes underlying large multi-allelic copy number variations in humans. *Nat Genet* (In review).

164.   Miyata, M. & Smith, J. D. Apolipoprotein E allele–specific antioxidant activity and effects on cytotoxicity by oxidative insults and β–amyloid peptides. *Nat Genet* **14**, 55-61 (1996).

165.   Xhignesse, M., Lussier-Cacan, S., Sing, C. F., Kessling, A. M. & Davignon, J. Influences of common variants of apolipoprotein E on measures of lipid metabolism in a sample selected for health. *Arterioscler. Thromb.* **11,** 1100–1110 (1991).