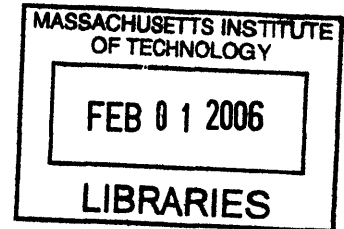


**Studies at the Hemochromatosis (HFE) Locus:
Gene conversions, Haplotypes, and an Association Analysis**

by

Junne Kamihara-Ting

B.A. Biochemical Sciences
Harvard University, 1997



Submitted to the Department of Biology
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN BIOLOGY

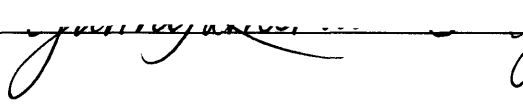
at the
Massachusetts Institute of Technology
December 2005

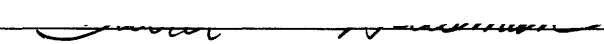
[February 2006]

ARCHIVES

© 2005 Massachusetts Institute of Technology. All rights reserved

The author hereby grants MIT permission to reproduce and to distribute publicly paper
and electronic copies of this thesis document in whole or in part.

Signature of Author:  Department of Biology
January 2, 2006

Certified by:  David E. Housman
Ludwig Professor of Biology
Thesis Supervisor

Accepted by:  Stephen P. Bell
Professor of Biology
Chair, Biology Graduate Committee

TABLE OF CONTENTS

Abstract.....	3
Acknowledgements.....	5
Chapter 1 Introduction.....	7
Chapter 2 Haplotype analysis of recombination events in the HFE locus.....	26
Chapter 3 Association analysis of the HFE locus with residual age of onset in Huntington’s disease.....	90
Chapter 4 Conclusions and Prospects for future work.....	129

Studies at the Hemochromatosis (HFE) Locus: Gene conversions, Haplotypes, and an Association Analysis

by

Junne Kamihara-Ting

Submitted to the Department of Biology
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Biology

Abstract

Haplotype-based association studies offer an exciting potential methodology for the identification of genes that contribute to complex traits. There is thus great interest in understanding the biological forces that shape haplotypes. We have studied a well-characterized genetic locus surrounding the gene responsible for hereditary hemochromatosis (HFE) to investigate the impact of meiotic recombination events upon haplotype structure in this region. First we identified crossover hotspots in order to define the boundaries of haplotype blocks in this locus. We then found that gene conversion events play a significant role in shaping haplotype structure within these haplotype blocks. These gene conversion events were not limited to recombination hotspots and occurred with a frequency as high as 1 in 10^4 per site per generation. Gene conversions lead to the creation of new haplotypes and we suggest that they are important for the spread of disease alleles in a population. In addition, we discuss how these events can be used as important tools in haplotype-based association studies.

We also present an association study in a large Venezuelan cohort to search for genes that contribute to residual age of onset in Huntington's disease. We demonstrate significant association between multiple alleles in a region on chromosome 6p21.3. We identify two candidate genes in this region, HFE and histone H1t and demonstrate significant association of this region with age of onset in a male-specific model.

Thesis supervisor: David E. Housman
Title: Ludwig Professor of Biology

*In memory of Thomas Halpin Jr. and for all those who like him
make our research at MIT possible.*

ACKNOWLEDGEMENTS

I would like to thank the many people who have been instrumental to the work contained in this thesis as well as to my own learning process over the last few years. Dianne Keen laid the foundation for my work in the HFE locus with her impressive work in the CFTR locus. Gorka Alkorta-Aranburu, a student in the laboratory, has provided generous technical as well as intellectual support throughout the years. Javier Gayan, our statistical collaborator at Oxford, ran numerous statistical tests for our association study in the HFE locus. Ray Chung, who provided us with many of the HFE samples, also offered helpful suggestions and discussion. Rhianna Cohen provided technical help with much of the genotyping as well as with mouse work as a technician in the lab. Neelesh Chudasama, a talented undergraduate, worked on the transferrin and IRP2 genotyping with me as well as on lymphoblast cell culture.

The entire Housman lab, past and present, have offered many moments of mentoring and advice: Julia Alberta, Laura Riba-Ramirez, Amanda Shearman, Ruth Bodner, Al Charest, Michelle Maxwell, Brenda Luciano, Shanie Coven, Kevin McMahon, Steve Kovach, Alex Kazantsev, and Amanda Thole. Patty, Maria, Maria, Russell, and Shelby have also made working nights much more enjoyable.

I would especially like to thank Jill Crittenden for her patient support and mentoring when I first started in the lab and J. Michael Andresen for many helpful discussions and for allowing me to investigate the iron metabolism candidates as part of his larger association study. Myra Coufal has been a great colleague with whom to share all the joys and challenges of being a Biology graduate student at MIT. Katie Rose

Boissonneault has provided me with constant encouragement, friendship, prayer and support as well as careful editing of this and all my written work during my time in the Housman lab.

My committee members, Frank Solomon, Sue Lindquist, Angelika Amon, Ann Graybiel, and Nancy Andrews, have each been especially supportive and amazing mentors to me. I have been privileged to receive their constant encouragement over the years. Finally, my Advisor, David Housman, has been a true mentor since he was my professor in HST 160, when I was introduced not only to genetics but to the first Huntington's patients I would ever meet. He has offered key insights, given me the courage and opportunities to ask my own scientific questions, and has inspired me by his passion for teaching and for bettering the human condition.

For my husband Dave, and our families, who with their constant care provide me with unspeakable love and unconditional support.

Chapter 1

Introduction

Complex traits, Association studies, and the International HapMap Project

Complex traits, including diabetes, schizophrenia, and cardiovascular disease, are traits that do not segregate in a clear Mendelian fashion. There is currently great interest in finding genes responsible for contributing to complex traits. These traits are considered complex because more than one gene contributes to overall phenotype, and there is often interaction between these genes as well as interactions between the genes and the environment.

One strategy for finding the genes that contribute to complex traits is based on linkage analyses. This approach has proven to be highly successful for finding genes that are inherited in a Mendelian fashion in which a single gene contributes to a single phenotype. For Mendelian traits, this approach carries with it a systematic certainty that there is a single site within the genome with a variant sequence responsible for the disorder. As additional family members are added to the study, the candidate interval within which the variant sequence is known to lie becomes progressively narrowed. The success of this approach for Mendelian disorders has led to interest in applying this strategy for complex trait analysis; however, studies based on linkage alone have significant limitations. While the localization of a relevant gene can be achieved through linkage analysis, our ability to define a candidate interval is usually restricted to a large genetic region of 20 cM or more since each gene has only a partial contribution to overall phenotype. Major increases in study size do not effectively reduce this interval.

Association studies offer a promising alternative or adjunct to linkage for the discovery of genes that contribute to complex traits. These studies compare populations, not families, to find alleles that are shared among an affected group and not among an

unaffected control group. Association studies, in effect utilize the same basic approach that genetic linkage studies use in families to narrow the candidate interval around an allele that contributes to the disease phenotype. In Figure 1, we have attempted to illustrate the commonality between these approaches.

Linkage analyses are performed in families in which a single disease-causing allele is passed through several generations in the family. This type of analysis relies on the identification of crossover events which surround the gene containing the disease-causing mutation and which occur over two or more generations in the family. In Figure 1a we can trace the transmission of a disease gene from a parent who is shown affected by a clinical phenotype (disease) inherited in an autosomal dominant Mendelian pattern with full penetrance. One of this parent's chromosomes (red) carries a mutation (indicated by the star) that causes his disease. During meiosis, this parent's chromosomes will undergo crossing over such that each of his children will receive some combination of his red and green chromosomes. The children receiving the red portion carrying the disease-causing variant will be affected. As more related individuals are collected, the interval containing the disease-causing allele narrows further and further. Finally, it may become possible to identify a single candidate region such as a transcription unit, within the interval.

The power of association studies for locating the gene of importance in causing a clinical phenotype in a particular chromosomal interval is based on analogous logic. Figure 1b illustrates this concept. In the association study, the individuals under study are treated as though they are members of an extended family based on the concept that a single ancestral mutation is assumed to be responsible for the disease phenotype in many

of the affected individuals in the study. Despite the absence of the many individuals who represent the family members who connect these individuals, the goal is to identify chromosomal regions which are identical by descent (i.e. derived from a common ancestor) among a significant number of affected individuals compared to control populations. Thus, in Figure 1b, the red chromosome with the disease-causing mutation is now representative of a common ancestral chromosome. After many generations and crossover events, descendants of this initial ancestor will carry smaller and smaller shared (red) chromosomal segments. Individuals who receive the portion of the red chromosome carrying the disease-causing variant will carry that particular genetic contribution to phenotype.

Association studies thus take advantage of markers that are inherited together with, or said to be in linkage disequilibrium (LD) with a disease causing allele. These markers are inherited together more frequently than expected by chance, thus disobeying Mendel's law of independent assortment. In Figure 1b, these markers would lie on the red portion of the chromosome that is inherited together with the disease-causing variant. Meiotic crossover events create the boundaries of these stretches of LD, and an important foundation to the success of association studies lies in the patterns of crossover recombinations that occur in the genome. These crossover events are known to occur not randomly, but limited to punctate regions called recombination hotspots (RHS) (Daly et al., 2001). The molecular characteristics of specific RHS have been described by sperm typing analyses using markers flanking the sites of crossing over (Jeffreys et al., 2001; Jeffreys et al., 2000). RHS flank LD blocks, which themselves, by comparison, are thought to be relatively cold regions for recombination (Daly et al., 2001) . Each stretch

of LD can also be described by the specific set of alleles inherited together there, known as a haplotype block. Human populations share common haplotype blocks, stemming from shared ancestry, and each block can usually be described by only a few distinct haplotypes (2-5) in any given population (Patil et al., 2001). Haplotype block lengths are thought to reflect population history: older African haplotypes which have had more chances for recombination events tend to be shorter, while younger, European populations tend to have longer haplotype blocks for example (Reich 2001).

Another critical component of association studies is the availability of appropriate markers. One of the consequences of the human genome sequencing project was the immediate identification of over 1.4 million single nucleotide polymorphisms, or SNPs, which serve as excellent candidate markers for association studies (Sachidanandam et al., 2001). SNPs are ancient sites of variation that are present on the level of single DNA bases and make up the majority of variations present between any two individuals (Shastry, 2002). SNPs have a low mutation rate and are found frequently throughout the genome (about 1 SNP per kb in both exonic and intronic regions) (Cargill et al., 1999; Shastry, 2002). The National Center for Biotechnology Information (NCBI) has a public database called “dbSNP” that contains SNP information that serves as a freely accessible resource for the scientific community (<http://www.ncbi.nlm.nih.gov>). By 2003, this database already contained 5.7 million SNPs with unique positions in the genome (Consortium, 2003).

One approach to association studies that holds great promise is to look for haplotypes defined by SNP markers that are enriched in a disease-carrying population and not in an unaffected group (Cardon and Abecasis, 2003). Comparing blocks of

haplotype, rather than entire genome-wide sequence, reduces the complexity of comparing two populations (in this case disease vs. non-disease). The rationale is that markers on these shared haplotype blocks, if not themselves the causal mutations, will be in LD with the causal variant. This strategy was used to uncover a “risk haplotype” enriched in a population with Crohn’s disease, one of two common types of inflammatory bowel diseases (Rioux et al., 2001).

The International HapMap Project, commonly known as the “sequel” to the Human Genome Project, was designed to provide a resource of information about SNP genotypes, frequencies, and measurements of LD across the genome (Consortium, 2003). Phase I of the project released over 1 million SNPs genotyped in 269 individuals with African, Asian, and European descent (30 trios-2 parents and child from Utah population in United States, 30 trios from Yoruba people in Ibadan, Nigeria, 45 unrelated individuals from the Han Chinese in Beijing, China, and 44 unrelated Japanese people from Tokyo, Japan) (Altshuler et al., 2005). Phase II of the project will include an additional 4.6 million SNPs. “Common” SNPs, with minor allele frequencies of greater than 5% were specifically selected for this project. An additional 10 regions of the genome each 500 kb in length were sequenced in 48 individuals to discover both common as well as rare SNPs. These SNPs were then genotyped in the entire panel of 269 individuals. The block-like structure of LD across the genome as well as the limited haplotype diversity inferred from smaller studies was confirmed by analyses of initial HapMap data releases (Altshuler et al., 2005).

Recombination events that shape haplotypes: crossovers and gene conversions

The exciting potential of haplotype-based association studies for the discovery of disease genes contributing to complex traits makes the understanding of biological mechanisms that shape haplotypes of great interest to the scientific community. There are still many gaps, however, in our understanding of the nature of these forces and their impact on haplotype structure. The two major biological mechanisms known to shape haplotypes are mutation and recombination. Mutations can change a given haplotype block by altering a single nucleotide or a stretch of sequence. The impact of this mechanism on haplotype is closely related to the rate at which such an event can occur. Hotspots for mutation, known as CpG dinucleotides, have extremely high mutation rates. The CpG single base mutation in the FGFR3 gene, for example, associated with achondroplasia (a dominantly-inherited condition with short stature), has one of the highest recorded mutation rates of about 1×10^{-5} per generation (Crow, 2000). Mutations at non-CpG sites, however, estimated by comparing pseudogene sequence divergence between chimpanzees and humans, occurs much less frequently, on the order of $1-2.7 \times 10^{-8}$ per nucleotide per generation (Nachman and Crowell, 2000).

The second important force governing haplotype structure is meiotic recombination. Meiotic recombination, the exchange of genetic material between homologous chromosomes (homologs), ensures the maintenance of diversity from generation to generation. Although the term “recombination” is commonly used to refer specifically to crossover events alone, it is more classically defined to encompass both crossover events as well as gene conversion events. While these two processes can both arise during a meiotic event, their products are distinctly different. Crossover events

result in the reciprocal exchange of chromosomal segments between homologs, while gene conversion events result in the non-reciprocal exchange of genetic material that can arise after heteroduplex formation and mismatch repair of a single homolog. Much of our understanding of meiotic recombination events derives from studies in yeast in which all four products of a single meiotic event can be recovered. Recent advances in SNP resequencing and sperm typing (Arnheim et al., 2003; Carrington and Cullen, 2004; Kauppi et al., 2004), have improved our ability to study these events in mammalian systems.

Meiotic crossover events, as mentioned earlier, take place at RHS and result in the formation of haplotype block boundaries. Crossovers occur once per chromosome or once per chromosome arm per generation and are thought to be essential for the successful completion of meiosis. Studies in yeast indicate that crossovers are initiated by double strand breaks catalyzed by the enzyme Spo11 (Lichten and Goldman, 1995). While definitive shared sequence motifs driving RHS have not been identified, several observations have been made in yeast regarding RHS location. Correlation has been seen, for example, with sites of transcription factor binding (referred to as α -hotspots), sites of nuclease sensitivity (β -hotspots), and GC-rich region (γ -hotspots). These are all thought to be associated with the modification of histones that leads to the availability of the chromosome to the recombination machinery (Petes, 2001).

In mammals, RHS are determined in several ways. First, they can be inferred by the local breakdown of LD between markers. This can be performed as haplotype analysis over a set of markers demonstrating the location of block boundaries and the “swapping” of chromosomal segments between haplotypes. This methodology is useful

in showing where historical crossovers have taken place. RHS can also be inferred by pairwise measurements of LD. The most commonly used description of LD between two markers is D' . If two loci, A and B have alleles A,a and B,b with frequency (p), then $p_{AB}=p_A * p_B$ if complete linkage equilibrium (independent assortment) is observed. Any deviation from this is measured by a value D, where $D=p_{AB}*p_{ab}-p_{Ab}*p_{aB}$. This measurement of LD, however, depends highly on allele frequencies, so the measurement D' is used more frequently, in which $D'=(p_{AB}-p_A p_B)/D_{max}$, where D_{max} equals the maximum absolute value of the numerator (Strachan and Read, 2004). D' measurements are limited by the fact that they are limited to pairwise comparisons between markers. This method also shows evidence for historical sites of crossover recombination between blocks of conserved LD (Gabriel et al., 2002).

To observe RHS over a single generation, pedigree mapping, in which crossovers can be identified using parental information, can be used. Only limited information can be provided by such analyses, however, given the low frequency with which crossovers occur over each generation. Another more robust method is to directly observe crossover events in single sperm, called sperm typing. Using allele specific PCR in single sperm, Jeffreys et. al. characterized several hotspots in the MHC II region of chromosome 6 (Jeffreys et al., 2001; Jeffreys et al., 2000). These studies reported hotspots of 1-2 kb in width, at sites where RHS were previously identified using pedigree recombination mapping. In addition, the hotspots identified by sperm typing were also shown to correlate with the location of hotspots that could be identified using LD breakdown measured by the D' statistic. Most crossovers were simple, with only a few showing evidence of accompanying gene conversions. Interestingly, hotspots were also reported

in clusters of 2-3 sites of elevated recombination that were each separated by 1-7 kb (Jeffreys et al., 2001).

More recently, coalescent models have also been developed to estimate RHS from population-wide variation data. Coalescent methods use currently existing polymorphism information to derive inferences about most recent common ancestry. McVean et. al. used a coalescent-based method to identify RHS and showed that this method could predictably localize RHS in regions where hotspots were previously identified by fine-scale sperm mapping or pedigree mapping (McVean et al., 2004). This group applied this methodology to predicting RHS and recombination rates across the genome using HapMap data (Myers et al., 2005). They report over 25000 RHS occurring on average about once every 50 kb. While they did not find any single sequence motifs common to all hotspots, several sequences were enriched in hotspots more than others. Among these, was the short motif (CCTCCT) that was found associated with hotspots when in the context of two retrovirus-like retrotransposons: THE1A and THE1B (Myers et al., 2005).

Another important product of meiotic recombination is gene conversion events that results in the unidirectional transfer of genetic information. These recombination events are becoming recognized as important factors in fine-scale haplotype evolution. While crossover events produce the boundaries of haplotype blocks, gene conversion events result in the decay of LD between markers within a block over a short distance. The importance of this mechanism in fine-scale haplotype mapping has emerged from studies that model LD decay which demonstrate that crossover events alone are insufficient to explain the lack of LD that is often observed between markers that are otherwise expected to be tightly linked . Inclusion of gene conversion into these models

allows for a better fit with existing LD data suggesting the important role of gene conversion events in shaping haplotype (Ardlie et al., 2001; Frisse et al., 2001). Padhukasahasram, et. al. (2004) modeled fine-scale LD using markers along chromosome 21 and proposed that gene conversion events occur at a ratio of 1.6-9.4 times the frequency of crossover events (Padhukasahasram et al., 2004).

Our molecular understanding of gene conversions derives largely from studies in yeast. In yeast, gene conversions can accompany two major products of recombination called crossover products and noncrossover products. While both of these pathways follow a double strand break, only the crossover pathway gives rise to chromosomes with reciprocal exchange. The mechanisms that produce crossover and noncrossover products were once thought to derive from the relative resolution of a single intermediate containing two junctions known as Holliday junctions (Szostak et al., 1983). Recently, however Borner, et. al. (2004) showed evidence supporting an “Early Crossover Decision” model, in which the decision between these two pathways is made early in prophase, before the formation of a stable common intermediate can be observed (Bishop and Zickler, 2004; Borner et al., 2004). By extrapolation, this model gives rise to the possibility that gene conversion events associated with noncrossover products are not limited to regions where crossover events take place.

In mammals, many observations of intra-allelic gene conversions have been reported especially for duplicated regions of the genome. Studies of interallelic meiotic gene conversions have also recently emerged. Single sperm analysis has been used to examine these events at sites of crossover hotspots. In the mouse, several groups have looked at gene conversion events occurring at previously established RHS where

crossovers were also known to occur (Guillon and de Massy, 2002; Yauk et al., 2003). Guillon and de Massy (2002) examined a RHS in the Proteasome subunit b type 9 (Psmb9) locus and reported 16 gene conversions in 6000 molecules of sperm DNA, with conversion tract lengths of less than 540 bp. In humans, Jeffreys and May (2004) similarly examined three loci with previously defined crossover hotspots. Using single sperm typing, they found gene conversion events in each of these hotspots with mean tract lengths ranging from 55-290 bp. They reported gene conversion: crossover frequency ratios ranging from 4:1 to 15:1. They also demonstrated that within each hotspot a gradient of recombination activity could be observed, in which the location of peak crossover activity corresponded with the location of peak gene conversion activity (Jeffreys and May, 2004). However, they do not address the possibility of conversion events outside of these defined crossover hotspots.

The HFE region and this present work

In the present work, we address the limitations in our understanding of the impact of recombination on haplotype structure by investigating the behavior of haplotypes in a region where disease-causing mutations are already characterized. The objective was to study haplotypes in a region where mutations were already known to gain insight that could impact future association studies in the search for genes whose role in disease are not yet known. The hemochromatosis (HFE) gene, associated with autosomal recessive hemochromatosis Type 1 in humans, was an ideal candidate region for this study. Hereditary hemochromatosis is a disease of iron overload common to individuals of Northern European decent. Phlebotomy can successfully reduce iron overload, but if left

untreated, clinical sequelae can include serious complications such as liver cirrhosis, diabetes, and heart failure. The HFE gene, located on chromosome 6p21.3, is an HLA Class I-like gene, with two major mutations that can lead to hemochromatosis: C282Y (G to A transition at nucleotide 845), and H63D (C to G transversion at nucleotide 187) (Feder et al., 1996). C282Y is thought to interfere with the ability of HFE to reach the cell surface where it can regulate the interaction between the transferrin receptor with transferrin (Feder et al., 1998; Feder et al., 1997). While 80-90% of patients who present with clinical symptoms of hemochromatosis in Northern Europe are C282Y homozygotes, the clinical penetrance is not complete (Feder et al., 1996; Waalen et al., 2002). The significance of the H63D mutation is also complex and is thought to be most clinically significant when found in a compound heterozygote accompanying a C282Y mutation. The prevalence of these mutations is very high in individuals with European ancestry (Merryweather-Clarke et al., 1997). 1 in 8 individuals in Northern Europe carry the C282Y mutation and 1 in 200 are homozygotes, while 25% of individuals throughout Europe are thought to carry the H63D allele (Distante et al., 2004). A third mutation, S65C, is present in 3% of individuals from Northern Europe (Distante et al., 2004). Like H63D, S65C is thought to become clinically significant when associated with a C282Y mutation although again with incomplete penetrance. The C282Y mutation lies on an extended haplotype suggesting a relatively recent origin. The HLA A3/B7 alleles have long been associated with C282Y, and with the exception of a report from a population in Sri Lanka (Rochette et al., 1999), this mutation is thought to have only a single origin. H63D, on the other hand, lies on a haplotype extending only about 700 kb. There have been several reports placing H63D on unique haplotypes, including an investigation that

placed it on 3 unique haplotypes defined by three polymorphisms spanning 10.6 kb in the HFE gene (Rochette et al., 1999).

Although defined by a Mendelian recessive inheritance pattern, the penetrance of the HFE mutations and the fact that more than one mutation can interact to produce a phenotype make the HFE locus an ideal candidate for a study to address issues that will impact future association studies to find genes involved with complex traits.

We demonstrate the location of a local cluster of recombination hotspot activity in the HFE gene using haplotype analysis in the immediate region surrounding the HFE mutations. We also report gene conversions in the human locus using C282Y homozygotes and present the first direct evidence for gene conversions arising from female meiosis in the mouse. Based on the high frequency with which we observe conversion events and our analysis of H63D haplotypes, we also suggest that gene conversion may have been a mechanism which led to the spread of this disease allele. As an additional study we also present an association analysis to investigate the impact of alleles in the HFE locus on age of onset in a large Huntington's disease cohort. We show that a specific haplotype exists in the population containing alleles that lead to a later age of onset of the disease.

REFERENCES

- Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S., Daly, M. J., and Donnelly, P. (2005). A haplotype map of the human genome. *Nature* 437, 1299-1320.
- Ardlie, K., Liu-Cordero, S. N., Eberle, M. A., Daly, M., Barrett, J., Winchester, E., Lander, E. S., and Kruglyak, L. (2001). Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am J Hum Genet* 69, 582-589.
- Arnheim, N., Calabrese, P., and Nordborg, M. (2003). Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *Am J Hum Genet* 73, 5-16.
- Bishop, D. K., and Zickler, D. (2004). Early decision; meiotic crossover interference prior to stable strand exchange and synapsis. *Cell* 117, 9-15.
- Borner, G. V., Kleckner, N., and Hunter, N. (2004). Crossover/noncrossover differentiation, synaptonemal complex formation, and regulatory surveillance at the leptotene/zygotene transition of meiosis. *Cell* 117, 29-45.
- Cardon, L. R., and Abecasis, G. R. (2003). Using haplotype blocks to map human complex trait loci. *Trends Genet* 19, 135-140.
- Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., *et al.* (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 22, 231-238.
- Carrington, M., and Cullen, M. (2004). Justified chauvinism: advances in defining meiotic recombination through sperm typing. *TRENDS in Genetics* 20, 196-205.
- Consortium, T. I. H. (2003). The International HapMap Project. *Nature* 426, 789-796.
- Crow, J. F. (2000). The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet* 1, 40-47.
- Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat Genet* 29, 229-232.
- Distante, S., Robson, K. J., Graham-Campbell, J., Arnaiz-Villena, A., Brissot, P., and Worwood, M. (2004). The origin and spread of the HFE-C282Y haemochromatosis mutation. *Hum Genet* 115, 269-279.
- Feder, J. N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D. A., Basava, A., Dormishian, F., Domingo, R., Jr., Ellis, M. C., Fullan, A., *et al.* (1996). A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet* 13, 399-408.

- Feder, J. N., Penny, D. M., Irrinki, A., Lee, V. K., Lebron, J. A., Watson, N., Tsuchihashi, Z., Sigal, E., Bjorkman, P. J., and Schatzman, R. C. (1998). The hemochromatosis gene product complexes with the transferrin receptor and lowers its affinity for ligand binding. *Proc Natl Acad Sci U S A* 95, 1472-1477.
- Feder, J. N., Tsuchihashi, Z., Irrinki, A., Lee, V. K., Mapa, F. A., Morikang, E., Prass, C. E., Starnes, S. M., Wolff, R. K., Parkkila, S., *et al.* (1997). The hemochromatosis founder mutation in HLA-H disrupts beta2-microglobulin interaction and cell surface expression. *J Biol Chem* 272, 14025-14028.
- Frisse, L., Hudson, R. R., Bartoszewicz, A., Wall, J. D., Donfack, J., and Di Rienzo, A. (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am J Hum Genet* 69, 831-843.
- Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., *et al.* (2002). The structure of haplotype blocks in the human genome. *Science* 296, 2225-2229.
- Guillon, H., and de Massy, B. (2002). An initiation site for meiotic crossing-over and gene conversion in the mouse. *Nature Genetics* 32, 296-299.
- Jeffreys, A., and May, C. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genetics* 36, 151-156.
- Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 29, 217-222.
- Jeffreys, A. J., Ritchie, A., and Neumann, R. (2000). High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum Mol Genet* 9, 725-733.
- Kauppi, L., Jeffreys, A., and Keeney, S. (2004). Where the crossovers are: recombination distributions in mammals. *Nature Review Genetics* 6, 413-424.
- Lichten, M., and Goldman, A. S. (1995). Meiotic recombination hotspots. *Annu Rev Genet* 29, 423-444.
- McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* 304, 581-584.
- Merryweather-Clarke, A. T., Pointon, J. J., Shearman, J. D., and Robson, K. J. (1997). Global prevalence of putative haemochromatosis mutations. *J Med Genet* 34, 275-278.
- Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310, 321-324.

Figure 1: Schematic to illustrate the basic concepts underlying linkage and association studies for the discovery of a disease-causing mutation.

a: Linkage analyses

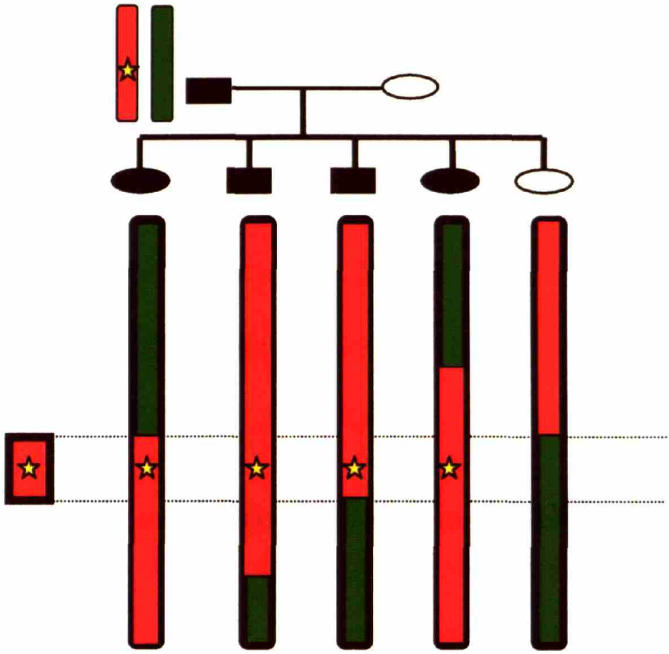
Linkage analyses are performed in families. A hypothetical parent affected by a disease passes on a disease-causing allele (star) to each of his children receiving the portion of his red chromosome containing the mutation. Over generations, each successive crossover recombination event narrows the interval further.

b: Association studies

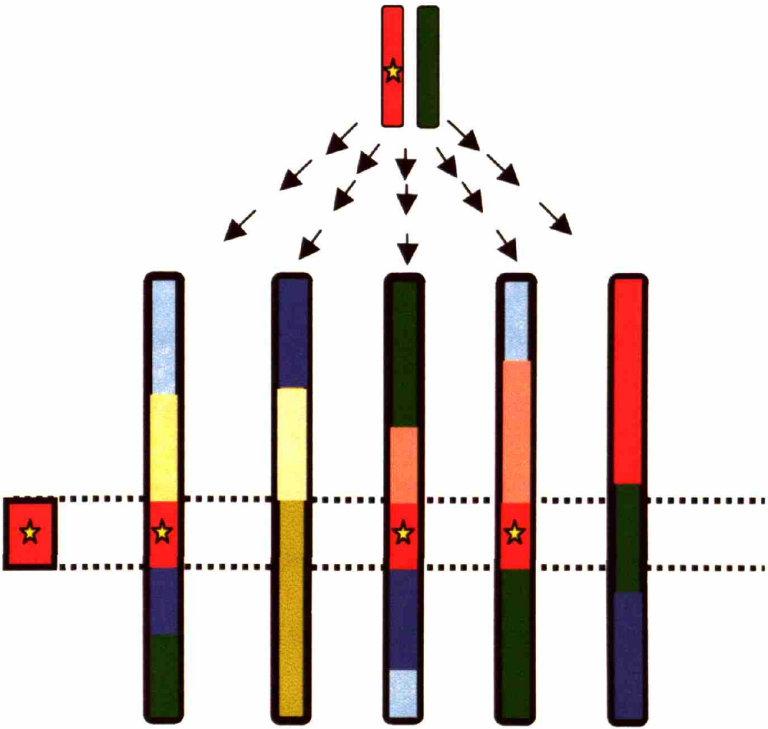
Association studies are performed in populations, not families, but use a similar logic. Here, the red chromosome with the disease-causing mutation represents an ancestral chromosome. Each descendent who receives the portion of the red chromosome with the mutation will be affected by the disease. Many generations will result in many crossover recombination events that lead to smaller and smaller shared segments from the red chromosome.

Figure 1

a



b



Chapter 2

Haplotype analysis of recombination events in the HFE locus

ABSTRACT

Haplotype structure in human populations is influenced by three biological processes: *de novo* mutation, crossing over, and gene conversion. In order to gain a better understanding of the relative contributions of these processes to haplotype evolution at a specific locus in the human genome, we have chosen to analyze the human hemochromatosis locus (HFE). Using panels of chromosomes from different populations, we have identified sites of increased crossing over within this region. One relatively recent mutation in the HFE gene, C282Y, has allowed us to identify chromosomes with a shared common ancestry. By comparing chromosomes marked by the C282Y mutation, we identified a site of gene conversion which is not found within a recombination hotspot. We have also assessed haplotype structure around the more widespread H63D mutation, which suggests that gene conversion was a likely mechanism for the movement of the H63D mutation onto different haplotype backgrounds.

The presence of gene conversions in human chromosomes outside of recombination hotspots led us to systematically search for gene conversion events in real experimental time. We therefore investigated recombination events taking place in the area surrounding the HFE locus in two fully informative mouse backcrosses. These studies allowed us to identify two clearly documented gene conversion events in female meiosis in a survey of 23,573 potential sites in the mouse. The occurrence of gene conversion at detectable frequencies in both mice and humans in the context of these studies suggests that gene conversion is a very significant contributor to haplotype evolution in mammals and that haplotypes which occur as a consequence of gene

conversion may have great potential utility in identifying the locations of genetic variants which contribute to human phenotypes with complex inheritance patterns.

INTRODUCTION

Haplotype-based association studies offer great potential for the identification of loci contributing to human complex trait phenotypes. The mechanisms that contribute to the evolution of new haplotypes in the human genome are thus of interest and importance in understanding the most effective use of this powerful methodology. Biological mechanisms that shape haplotypes include mutation and meiotic recombination. The term recombination encompasses two types of events that both result in the exchange of chromosomal segments. These include crossover events, in which chromosomal segments undergo reciprocal exchange, and gene conversion events, involving the unidirectional transfer of genetic information from a donor chromosome that itself remains unchanged.

Sites of crossover are referred to as recombination hotspots (RHS). These RHS are restricted to punctate regions along the genome which form the boundaries of haplotype blocks. In addition to crossover events, gene conversion events are also known to occur at these hotspots in mammals as well as in lower eukaryotic organisms.

Our ability to investigate the impact of recombination on haplotype evolution is often restricted by our inability to observe these events in real time and by the difficulty in specifically identifying how groups of chromosomes are related through generations of human history. We have taken two approaches, the first in humans and the second in mice that have allowed us to compare chromosomal blocks with a defined point of common ancestry. We used these approaches to study recombination events and show that both crossover and gene conversion events have significant impact on haplotype evolution.

To study these events first in human chromosomes, we collected chromosomes that are identical by descent because they carry a unique disease-causing mutation and are thus related to each other over finite historical periods in the regions immediately surrounding the mutation. We have taken this approach to investigate haplotype evolution in the region of chromosome 6p21.3 surrounding the HFE gene. HFE contains a group of disease-causing mutations that include two major mutations, C282Y and H63D, which are associated with hemochromatosis, a disease of iron-overload. The C282Y mutation is thought to have a single origin, and thus chromosomes marked by this mutation share a single common ancestor. Given the relatively young age of the mutation and its single origin, we have analyzed haplotypes in C282Y-carrying chromosomes in order to identify gene conversion events. In addition, we analyzed haplotypes surrounding the older H63D mutation to demonstrate the possibility for the spread of this disease allele via gene conversion.

Our ability to observe recombination events in humans is limited by our inability to directly observe these events in real experimental time. To address this issue and to investigate the possibility of gene conversion events not limited to recombination hotspots, we looked at the inheritance of haplotype in DNA isolated from backcrossed progeny resulting from two crosses from genetically divergent strains of mice. Here we were able to directly track the inheritance of markers over a single generation to monitor recombination events in the area we examined. In both these systems, we examine the two types of genetic recombination events and their contribution to haplotype evolution.

MATERIALS AND METHODS

DNA samples. Human: Diversity panel of 31 ethnically diverse human DNA samples were obtained from Coriell Cell Repositories (Camden, NJ, <http://locus.umdj.edu/ccr/>). NA00522: Kikuyu, NA00726: Korean, NA01850: African American, NA02064: Ghana, NA02347: Swedish, NA02430: Italian, NA02476: Zulu, NA02743L Greek, NA02783: Iranian, NA03043: !Kung, NA03579: Cuban, NA03580: Greek, NA03721: African American, NA03780: Spanish, NA04428: Mexican American, NA05052: African American, NA03735: African American, NA10418: Finnish, NA10810: Japanese, NA10923: German, NA10965: American Indian, NA11321: Chinese, NA11322: Chinese, NA11323: Chinese, NA11324: Chinese, NA11373: Cambodian, NA11589: Japanese, NA12556: French, NA12558: French, NA14611: East Indian. Pygmy DNA also obtained from Coriell included NA10494, NA10471, NA10492, NA10496, NNA10469, NA10470. Primate DNA included: NA03448, NG03612, NG03657, NG03610: Pan troglodytes, NG05251: Gorilla gorilla, NG05253: Pan paniscus, NG06209: Pongo pygmaeus.

The following HFE samples were obtained from Coriell Cell Repositories: NA13591, NA14180, NA14620, NA14621, NA14628, NA1463, NA14640, NA14646, NA14650, NA14651, NA14652, NA14654, NA14655, NA14656, NA14657, NA14685, NA14686, NA14688, NA14689, NA14690, NA14691, NA14702, NA14703, NA14712, NA14715, NA14857, NA16000. All other HFE samples were provided by N. Andrews (Children's Hospital, Boston), and R. Chung (Massachusetts General Hospital, Boston). Population panel DNA was provided by: Kosrae: M. Karayiorgou (Rockefeller, New York), Basque and Spanish: M. Ramos (U. of Basque Country, Spain), Dutch: F. Baas

(Academic Medical Center, Amsterdam), African American: A. Menon (University of Cincinnati, Cincinnati), Vietnamese: E. Schurr (McGill University, Montreal).

Venezuela DNA and kindred information (Wexler et al., 2004) was provided by the Hereditary Disease Foundation (Rockefeller, NY). When applicable, DNA from whole blood was isolated using a phenol/chloroform extraction protocol. Genome-wide amplification using GenomiPhi™ was performed to increase DNA yield when necessary.

Mouse: Mouse DNA was provided by: R. Swank (Roswell Park, Buffalo).

Control DNA for mouse was obtained from Jackson Laboratories (Bar Harbor, Maine) including PWK/PhJ (003715), C57BL/6J (00664), and Spret/EiJ (001146).

SNP Discovery. Human: SNP discovery in the human HFE locus was performed by sequencing three samples derived from a C282Y homozygote, a C282Y heterozygote, and a H63D homozygote. Overlapping primers pairs flanking 400-600 bp regions were designed to cover a total of ~50 kb surrounding the HFE gene. Two regions (of approximately 500 bp each) were eliminated due to the presence of pseudogene-like sequences that would make unambiguous genotyping difficult. SNPs significant for either mutation-carrying chromosome were selected for analysis.

Chimpanzee (*pan troglodyte*) sequencing was also performed or chimp genome sequence (recently available) was used to designate the “ancestral” reference allele at each SNP.

Mouse: SNP discovery in the mouse HFE locus was made by sequencing ~3 kb using overlapping primer pairs flanking 400-600 bp of non-repeat sequences in regular intervals across 1 MB surrounding the HFE locus. Sequencing was performed in DNA from PWK/Ph and Spret/Ei strains. C57BL/6J sequence was also obtained or the database sequence was referenced.

Genotyping. Human: PCR amplification of a 500-1000 bp region surrounding each SNP was performed in 96-well format using a thermocycler with a final reaction volume of 12-50 μ l. PCR products were then denatured and spotted onto Hybond N+ membranes (two identical membranes per PCR plate). Allelic discrimination was performed by using allele-specific oligo (ASO) hybridizations. ASOs were designed as 17-mers with the allele of interest in the 9th position. Each membrane was then probed using an ASO labeled with $\gamma^{33}\text{P}$. Membranes were hybridized for 1.5 hr-overnight at 54^o C, washed, and exposed to phosphor screens for subsequent visualization. Images were acquired using a Storm Phosphoimager [®] (GE) after 24 hours of exposure and analyzed visually or with ImageQuant [®] software (GE). **Mouse:** Mouse genotyping was performed by KBioscience (Hoddesdon Herts, UK).

Resolution of haplotypes. Haplotypes were resolved when pedigree information was available to allow the unambiguous assignment of genotypes to each of the two chromosomes. This was possible in the samples from the large Venezuelan cohort using pedigree information for extended kindreds. Haplotypes were determined manually by assigning each allele to one of two parental chromosomes using genotype information for each parent, child, or extended family member. Each parent's contribution to the child's haplotypes was determined. Whenever possible, this was done by identifying homozygous sites in an individual. For example, in the simplest scenario, a parent may have two identical haplotypes, homozygous at every site examined. In such a case, this parent must have contributed one of these two homozygous haplotypes to the offspring. By subtracting this haplotypes from the summed haplotypes (represented by genotype information for each SNP marker) of the offspring, the remaining haplotype would be

that haplotype contributed by the other parent. This parent's second haplotype would then be determined by subtracting the haplotype donated to the child from the parent's summed haplotypes.

HapMap data deriving from the CEPH (Centre d'étude du polymorphisme humain) population collected for human genetic mapping studies of Utah residents with northern and western European ancestry (Altshuler et al., 2005) was also performed. Each haplotype was manually subtracted by comparison of triads (mother, father, child). Phased haplotypes were determined using genotype data, and no call was made whenever resolution could not be unambiguously made.

RESULTS

Identification of a local recombination hotspot in the human HFE locus

In an effort to develop a high-resolution picture of haplotypes found in the region surrounding the HFE locus, 41 SNP markers were genotyped in this region. These SNPs span a genomic region of 45.8 kb on 6p21.3 that includes two histone genes, histone 1H4c and histone 1H1t, which are downstream of HFE. The arrangement of these SNP markers in relation to the genomic locus is shown schematically in Figure 1. These SNP markers were selected from an original group of 44 SNP markers that were found by sequencing three individuals, a homozygote and a heterozygote for the C282Y mutation, and one homozygote for the H63D mutation. Information about each marker is shown in Table 1. The name of each SNP used in this study along with corresponding reference SNP ID numbers (rs#) assigned by NCBI (National Center for Biotechnology

Information) in the SNP public database (dbSNP) is shown. The SNPs used in the analysis of each population in this study is also indicated in Table 1.

We genotyped the 41 SNPs indicated in a collection of 32 individuals from diverse ethnic and racial backgrounds (this collection of samples will henceforth be referred to as the diversity panel). The genotype patterns for these individuals are shown in Figure 2. Each column represents one individual, whose two chromosomes are shown together in a single column. The genotype at each diallelic SNP marker is represented by a colored box. Green was used to represent homozygosity for the reference allele at a SNP, while red was used to represent homozygosity for the alternative allele at a SNP. A blank box indicates that no genotype was available for that site. The “ancestral” allele of each SNP, as determined by chimpanzee (*pan troglodyte*) sequence, was designated the reference allele whenever this information was available. (In the rare case in which this information was not available, the reference allele is arbitrarily assigned to correspond to the Genbank sequence). Blue was used to represent heterozygosity for both alleles at that particular marker.

Two major haplotypes (designated A and B in Figure 2) in the region from SNP 487 to SNP 525 (shown between brackets) became evident by grouping individuals with similar genotype patterns together. These chromosomes were arranged according to their genotypes within this block, to illustrate the presence of two major homozygote groups above SNP 525. These groups carry two chromosomes with the same haplotypes (marked as A/A and B/B in Figure 2). A third group of chromosomes with many heterozygote genotypes presumably results from individuals with an A haplotype on one of their chromosomes and a B haplotype on their other chromosome (designated A/B in

Figure 2). The two homozygote groups above SNP 525 are associated with more than one haplotype below. In addition, several individuals with the heterozygote block above SNP 525 have a homozygote block below this SNP. This suggests the presence of a local site of crossing over, or a recombination hotspot (RHS), somewhere below SNP 525. Historical crossovers at this hotspot would explain the apparent swapping of modular chromosomal blocks that would lead to these results.

In an effort to further locate sites of crossing over positioned within the region surrounding the HFE locus, we looked for chromosomes with clear evidence for historical recombination within a cohort of kindreds from Venezuela. 30 SNPs, shown schematically in Figure 3, were genotyped and used in this analysis. Recombinant chromosomes were defined specifically as those in which a single homozygous block is preceded or followed by a single heterozygote block in the region we genotyped. Our rationale was that this type of pattern would most likely arise from historical crossover events that occurred at the boundary between these two blocks. Information from recombinant chromosomes was isolated and each recombinant was classified according to the position of the apparent site of crossover. We totaled the number of unique recombinant types for crossovers occurring at each site observed. This is shown graphically in Figure 3, in which the height of each peak indicates the number of recombinant types of chromosome at each site. We reasoned that totaling each type of recombinant would offer insight into the number of historical crossover events that occurred. As shown in Figure 3, the largest number of recombinant patterns (9 types) was found between SNP 525 and SNP 532-(H63D). The second largest number (6 types) was found close to this site between SNPs 532b-3 and 532-2 (see Figure 3). The third

largest number (3 types) was observed between SNP 536 and SNP 538. These findings support the localization of a local site for crossing over below SNP 525.

To corroborate our findings with data derived from other populations and to confirm that our analysis is valid over an extended chromosomal region, we analyzed blocks of genotyping data recently made available from the HapMap project (public release #19, 10/24/05, <http://hapmap.org>) (Altshuler et al., 2005) from the same region of chromosome 6. Figure 4 shows a selection of genotype data from Japanese chromosomes arranged with SNPs running 5' to 3'. Using this data we illustrate the presence of four recombinant chromosomes (each indicated by a * in Figure 4) defined by the presence of a single heterozygote block and a single homozygote block. These recombinant chromosomes are shown alongside non-recombinant chromosomes to demonstrate that the boundaries of apparent crossover for two of these chromosomes are in the same location as we observed in our data (see large arrow in Figure 4). Two additional recombinant chromosomes with sites for crossing over outside of the region we examined are also shown for comparison. On Figure 4, the region corresponding to the location of our own analysis is shown by a bracket. We also include SNP markers extending out a total of 140 kb (distance from topmost SNP to bottommost SNP).

HapMap reports of local recombination rates modeled using genotype data (Myers et al., 2005) also confirms our findings of an elevated region of recombination in the region we examined. Table 2 lists the recombination rates (from hapmap.org) reported between markers as shown (SNPs that are common to both our analysis and the HapMap analysis are designated with both SNP names). Recombination rates are summarized in the last two columns and given as a rate (cM/MB) and genetic distance

(cM). These values are given between the “starting” SNP (shown in the left group of columns) and the “ending” SNP (shown in the middle group of columns). Elevated recombination rates (highlighted yellow in Table 2) above SNP 532-1(H63D) confirms our finding. Interestingly, HapMap also reports a second elevated region of recombination between SNPs 532-2 and SNP 542. This region encompasses an area where we also note a historical recombination activity, as shown by the third highest peak in Figure 3 between SNPs 536 and SNP 538. These data together suggest a cluster of local recombination activity in this region.

To place these locally elevated recombination rates in context of genome-wide recombination activity, we also searched HapMap estimates of genome-wide recombination hotspot reports (Myers et al., 2005). These data do not report a recombination hotspot in this area. The nearest reported hotspots, in fact, lie over 400 kb upstream and over 40 kb downstream from the HFE gene. This indicates that either recombination was not detectable using their methodology and/or that the local regions of elevated recombination (henceforth referred to as local RHS) are relatively cold areas for crossing over in relation to a genome-wide measurement of recombination.

Gene conversion events lead to the creation of new haplotypes

The next step in this project was to investigate the occurrence of gene conversion events in the HFE locus by comparing chromosomes that are identical by descent because they carry a shared disease-causing mutation. Previous work in our group has shown that this strategy is an especially effective way of demonstrating recombination events in human chromosomes. This work examined a region of chromosomes 7q31 surrounding

the $\Delta F508$ mutation of the CFTR locus responsible for cystic fibrosis and found evidence for gene conversion events not limited to local sites of crossovers (Keen Kim, 2002). In order to apply this methodology to the HFE locus, we analyzed haplotypes on chromosomes marked by the C282Y mutation in the HFE gene. We genotyped 42 SNP markers (shown in Table 1) spanning the same 45.8 kb region in DNA from 39 patients homozygous for the C282Y mutation to identify shared haplotypes in this region. These genotypes are presented in Figure 5a. The SNPs used in this analysis are shown schematically in Figure 5b.

C282Y is thought to have a single occurrence in history (Feder et al., 1996) and thus each C282Y-carrying chromosome is expected to be identical by descent on the haplotype that immediately flanks this mutation. While this was the case for the majority of chromosomes we genotyped, we also found a new haplotype created by nucleotide sequence changes (arrows labeled GC, Figure 5a and c). These sequence changes were noted at SNP 563 and SNP 565-2. The region between these SNPs spans approximately 1.5 kb and is located between Hist1H4c (about 1.7 kb upstream) and Hist1H1t (about 1.3 kb downstream). The concordant change of two SNPs in a region of less than 2 kb is consistent with the properties of a gene conversion event. Notably, these events were not found at the local RHS for crossing over described above.

We considered alternative explanations to the gene conversion scenario observed at this site. *De novo* mutation appears to be an unlikely explanation for these results. The sequence changes seen in our C282Y haplotypes are not found at CpG hotspots for mutation. The genome-wide mutation rate for non-CpG sites in the human genome has been estimated to be approximately $1-2.7 \times 10^{-8}$ per nucleotide per generation (Nachman

and Crowell, 2000). The history of the C282Y mutation (previous studies have estimated that the C282Y mutation arose between 62 and 250 generations ago based on extended haplotype analysis (Distante et al., 2004)) suggests that chromosomes which carried the C282Y mutation would not have been transmitted through enough generations to be likely to have accumulated one, let alone two *de novo* mutations.

A single crossover at the local RHS described above found beneath C282Y (between SNP 536(C282Y) and SNP 538) could be used to explain our finding if we could identify a haplotype identical to the C282Y haplotype at all sites except for the mutation and the stretch of observed sequence change. The closest such haplotype is shown in Figure 6 on the chromosome labeled 3721 Afric Amer. One finding argues strongly against a crossover at this site. A single SNP, SNP 547, found through sequencing of a C282Y homozygote, was seen only in C282Y chromosomes. The allele frequency of this SNP in the diversity panel in chromosomes not carrying C282Y was 0 (see Figure 6). This SNP is therefore likely to be a mutation that occurred close in time to the C282Y mutation on an ancestral chromosome carrying C282Y. The absence of this SNP in the candidate African American chromosome makes it unlikely that a crossover occurred below at the hotspot between SNP 536(C282Y) and SNP 538 producing the observed result. A crossover occurring below this site is even less likely given the recombination data presented above.

The remaining mechanism consistent with the novel haplotype, gene conversion, involves the unidirectional transfer of genetic information from one haplotype to another. In order for the novel haplotype to have arisen via gene conversion-, the appropriate donor sequences must be present on a non-C282Y haplotype Figure 6 shows the

genotypes from a subset of individuals from the diversity panel, with SNP markers common to those genotyped in the C282Y homozygotes shown. 30 chromosomes are shown that carry the appropriate SNP alleles between SNP 563 and SNP 565-2 that could serve as donor alleles for the event noted at these SNPs (arrow, right panel). This indicates that chromosomes that carry the appropriate reservoir sequence which can be transferred onto the C282Y haplotype through gene conversion are present in these populations.

Although we had observed only a single gene conversion event in the C282Y chromosomes we had studied, we wished to make a quantitative estimate of the sample size surveyed to identify this event, which in turn would suggest an estimate of the frequency with which a gene conversion event outside a recombination hotspot might occur. The selection of chromosomes specifically marked by the C282Y mutation suggests that the stretch of haplotype immediately surrounding the mutation derives from a single common ancestor. The data in Figure 6 support that assumption. To estimate the number of generations we have surveyed, we make the assumption that the chromosomes we have surveyed are related to each other by a star shaped genealogy in which each chromosome is related to the ancestral chromosome by an independent path. This assumption clearly contributes to an overestimate of the number of generations surveyed, since the chromosomes we are sampling may well have more recent common ancestors. The estimated age of the C282Y mutation, ranging from 62 to 250 generations (Distante et al., 2004), gives a common denominator with which we can make an estimate regarding the frequency of gene conversion events given the number of generations which have passed since the C282Y mutation occurred. The size of the sample in which

we observed the gene conversion event was a survey of 42 SNP sites in 39 homozygotes for the C282Y mutation. We thus estimate that we surveyed a total of 3276 sites total (given 2 chromosomes per individual). Our observation of gene conversion would thus be consistent with a frequency of gene conversion at non recombinant hotspots of 1.2-4.9 x 10⁻⁶ per site per generation or higher (given the limitations of the assumption of a star shaped genealogy). This result suggested that gene conversions not limited to recombination hotspots could occur with a high enough frequency to lead to the creation of new haplotypes within haplotype blocks.

To test this theory we used genotype data from the HapMap project and manually resolved haplotypes from 30 trios (father, mother, child). Thus far we have inferred haplotype structure from genotype information derived from the two chromosomes of each individual. Parental genotypes allow us to determine the phased haplotypes on each of these two chromosomes at most sites. We used data from 180 SNP markers spanning 141.2 kb in a population residing in Utah with northern and western European ancestry originally collected by the Centre d'etude du polymorphisme humain (CEPH) for human genetic mapping studies. 40 of these haplotypes are shown in Figure 7. Putative historical gene conversion events are shown (solid arrows) that change a haplotype block at the site indicated. Interestingly, two of the sites shown involve coordinate changes of less than 1 kb (834 bp and 291 bp), highly suggestive of a conversion tract. These findings are not likely to be due to genotyping error since these data were generated by comparing genotypes of three individuals in a parent offspring trio. When any disagreement of genotypes caused by possible genotyping error was found, the trio with these disagreements was eliminated in the analysis for that marker.

Gene conversion events directly observed over a single generation in progeny of two large mouse backcrosses

The significance of gene conversions in shaping haplotypes over time depends on the frequency with which these events occur over a single generation. When observing human chromosomes for recombination events, we are limited by the fact that we are observing a snapshot of chromosomes that have evolved over generations of time. Our crude estimation of gene conversion frequency for example, was dependent upon age estimates of the C282Y mutation and assumptions regarding the genealogical relationships among the chromosomes surveyed. For chromosomes with even less well defined relationships, assessing the dynamics of gene conversion are even more challenging.

We therefore sought to develop a system in which we could directly observe the occurrence and transmission of gene conversion events in a mammalian model system in real experimental time. To accomplish this goal we have followed the inheritance of a series of SNP markers tightly linked to the mouse HFE locus over a single generation using progeny from two large interspecific mouse backcrosses. 658 N2 samples from ((PWK x C57BL/6J-*gm/gm*) x C57BL/6J-*gm/gm*) backcrosses and 570 N2 samples from ((*M. spretus* x C57BL/6J-*gm/gm*) x C57BL/6J-*gm/gm*) backcrosses were collected. The breeding scheme and summary of samples used is shown in Figure 8a and b. These samples were genotyped for 20 SNPs spanning 1 MB of chromosome 13 containing the mouse HFE locus. These SNPs are shown schematically in Figure 9a and listed in Table 3. SNPs were identified by sequencing a PWK, *M. Spretus*, and C57BL/6J homozygote, respectively. All 20 SNPs had one allele in the C57 strain and an alternate allele that was

shared by both the *M. Spretus* and PWK strains. 12827 successful genotypes were obtained from progeny samples deriving from the (PWK x C57BL/6J-*gm/gm*) x C57BL/6J-*gm/gm* backcross, while 10926 successful genotypes were obtained from progeny samples deriving from the (*M. spretus* x C57BL/6J-*gm/gm*) x C57BL/6J-*gm/gm* backcross.

Gene conversion candidates were readily detected using our strategy (see Figure 8a, II). In order to examine whether these events were limited to recombination hotspots, crossover events were also noted by the continuous change of marker alleles (see Figure 8a, III). These crossovers are shown in Figure 9a and were detected between SNPs JK_29 and JK_32 (8 such events occurring at this site were detected in progeny from the PWK x C57BL/6J backcross), and one event was detected between JK_06 and JK_08 (in *M. Spretus* x C57BL/6J backcross). Other events were apparent between JK_32 and JK_36 (4 events in *M. Spretus* x C57BL/6J backcross), but these seen in the terminal SNP genotyped, so further downstream genotyping needs to be performed to rule out possible gene conversion.

Of the 23753 total genotypes obtained, two gene conversion candidates were observed and genotyping error was ruled out by direct sequencing. Both conversion events did not occur at the sites of crossover detected in other samples (above). These conversion events are shown schematically in Figure 9b. The first conversion event was observed at SNP JK_08, and seen in one progeny of the ((PWK x C57BL/6J-*gm/gm*) x C57BL/6J-*gm/gm*) backcross. JK_08 is located between two genes (*Abt1* and *Btn1a1*) on chromosome 13. Further sequencing of the region located immediately adjacent to this conversion event showed at least one other SNP marker located 21 bp upstream from

marker JK_08 that was also included in the gene conversion tract. The second conversion event was observed at SNP JK_19 in one progeny of the ((*M. spretus* x C57BL/6J-*gm/gm*) x C57BL/6J-*gm/gm*) backcross. JK_19 is located between two exons in the mouse HFE gene. Further sequencing around this SNP change was accompanied by at least 10 coordinate changes observed in SNPs spanning a tract of 604 bp. These data are consistent with gene conversion tract lengths of less than 1-2 kb reported in other gene conversion studies (Guillon and de Massy, 2002; Jeffreys and May, 2004). Further sequencing will allow us to determine the outer boundaries of each conversion tract.

H63D is found on multiple haplotype backgrounds, suggesting transfer via a gene conversion-like mechanism

The high frequency with which gene conversion events were observed lead us to support the hypothesis that these events could serve as a viable mechanism for the spread of disease alleles in a population. Currently, when a disease-causing mutation is found on more than one haplotype background, a common explanation offered is that these mutations occurred independently at the same site. Given the low frequency of mutations at non-CpG nucleotides, however, this model becomes less likely as more haplotypes carrying the mutation are found. Another mutation in the HFE locus, H63D, is one such mutation that is found with high frequency throughout the world (Distante et al., 2004) and reports have placed it on different haplotype backgrounds (Rochette et al., 1999). In the present study we examined several different populations using a high-resolution scale to determine whether we can detect H63D on different haplotypes immediately surrounding the mutation.

We genotyped 464 individuals (928 chromosomes) from different populations as summarized in Table 4. We found 7 H63D homozygotes and 58 H63D heterozygotes among these samples. Using these samples, we defined each major haplotype by the coordinate change of 2 or more SNPs in a block above or below the H63D mutation. Since these chromosomes were derived from unrelated individuals, we could only resolve haplotypes from genotype data unambiguously in H63D homozygotes or in those chromosomes carrying homozygote blocks. Nevertheless, as shown in Figure 10, these chromosomes were sufficient to confirm that H63D can be found on multiple haplotypes within a short-range distance from the mutation. Figure 10 shows just three samples which together illustrate H63D on multiple haplotypes. The majority of H63D chromosomes we examined were associated with one of the two major haplotypes shown in this figure (leftmost and rightmost, top). An additional major haplotype below the mutation can be deduced from the H63D homozygote sample (Sp1815) shown with heterozygous sites below the mutation. In total, we observed at least two major haplotypes within 10 kb above the mutation, and two major haplotypes within 17 kb below the mutation.

We then turned to chromosomes derived from related individuals in order to unambiguously assign haplotypes surrounding the H63D mutation. In a large Venezuelan pedigree, we analyzed 89 individuals (out of a total of 755 examined) including 15 homozygotes and 74 heterozygotes for the H63D mutation. Using extended family information, we resolved haplotypes in these chromosomes and isolated those carrying the H63D mutation. Among these, we found one major haplotype block below the H63D mutation and two major haplotypes, labeled A and B in Figure 11 above the

H63D. A third haplotype block B', (far right in Figure 11) likely results from a gene conversion at SNP 500-2 on haplotype B. Of the 104 haplotypes carrying H63D that we examined, we observed 20 A haplotypes, 81 B haplotypes, and 3 B' haplotypes.

We also selected H63D-carrying chromosomes from the HapMap data set (Altshuler et al., 2005) and performed a similar analysis using data from trios of the CEPH population of Utah. These resulting haplotypes are shown in Figure 12. Using genotype information from 101 SNP markers covering a region of 141.2 kb, we found two major haplotype blocks (labeled A and B in Figure 12) above H63D that correspond to the two haplotypes seen in Venezuela. In the CEPH population we also observed a third major variation of haplotype A (far left column). We also found at least three major haplotypes below H63D and a fourth minor haplotype shown to the far right on Figure 12. Each major haplotype above H63D (A and B) was associated with more than one major haplotype below H63D. If we include other minor haplotypes (likely produced by gene conversion events) with variations at one or more nucleotide changes, there are as many as 5 haplotypes above H63D within 5 kb (with 7 within 20 kb) and as many as 5 haplotypes below H63D within 11 kb of the mutation.

Given the large number of haplotypes on which we found H63D, it is unlikely that this finding is due to recurrent mutation. A second possibility is that crossovers carried H63D from one framework haplotype to another. H63D is found right at the boundary of two regions with elevated recombination fractions, see above and Figure 12. As shown schematically in Figure 13, a single occurrence of the H63D mutation in history followed by crossovers above and below H63D could account for the four haplotypes seen in the CEU chromosomes (haplotypes: Ax, Bx, Ay, By). For H63D to

move from haplotype A-x to B-y by crossover alone, a minimum of two sequential crossovers would be necessary to give these results (from A-x to B-x, then from B-x to B-y). The probability that such a double crossover would occur can be represented by the product of the recombination frequencies above H63D and below H63D. To compute these frequencies, we summed the genetic distance in cM from the first informative marker (distinguishing haplotype A from B) above H63D (at marker rs 2794719, 2289 bp above the mutation) to H63D, and the first informative marker below H63D (rs 6918586, 6205 bp below the mutation) to H63D. Recombination rates from HapMap, as shown in Table 2 were used for these computations. A rate for marker rs 6918586 was not available, so the closest available marker, rs 1150660, located 10.2 kb from H63D was used. The recombination distances, given as the sum of distances of intervening markers, was 0.004 cM above H63D and 0.008 cM below H63D. Therefore, the crossover frequencies at these sites are given by 4×10^{-5} crossovers per generation above H63D and 8×10^{-5} crossovers per generation below H63D. To produce three of the four haplotypes shown by crossover events alone, two sequential crossovers (once above H63D followed by once below H63D) would happen at a frequency of approximately 3.2×10^{-9} per generation.

Instead, given the high frequency with which we observed gene conversion events in our mouse experiment, this mechanism, either alone or accompanying a single crossover event, becomes a much more attractive explanation for the spread of the H63D mutation onto different haplotype backgrounds.

DISCUSSION

Gene conversions have been well studied in yeast and lower eukaryotes in which all four products of a single meiotic event can be directly recovered. Studies of meiotic gene conversions in mammals, on the other hand, is exceptionally challenging precisely because of our inability to isolate the products of a single meiosis. We have taken two approaches that have allowed us to better observe these events in mammalian systems. First, we chose to examine chromosomes marked by a disease-causing allele, C282Y of the HFE locus. This mutation has been reported to lie on an extended haplotype framework and most studies estimate that the mutation arose between 62-77 generations ago (the full range of age estimates of C282Y extend from 62-250 generations) (Distante et al., 2004). We have examined markers within 25 kb upstream and downstream of the C282Y mutation, where infrequent events such as mutations, have not had significant time to adequately erode LD between surrounding markers. We find, instead, evidence of gene conversion events, shown by the coordinate change of a short stretch of sequence approximately 1.5 kb in length. We found 1 event in two chromosomes in 42 SNPs sampled across 78 chromosomes, leading us to estimate these events on the order of $1.2-4.9 \times 10^{-6}$ per site per generation. A previous study in our group in the CFTR locus using chromosomes homozygote for the $\Delta F508$ mutation, estimated gene conversion events occurring within the bounds of two hotspots at a rate of 8.3×10^{-7} events per site per generation (Keen Kim, 2002). This estimate, subject to the same caveats regarding genealogical history and age of the $\Delta F508$ mutation, nevertheless is of the same order of magnitude as our estimate for the HFE locus of 1 gene conversion event in 10^6 per site per generation.

To illustrate that these events shape haplotypes within a haplotype block, we resolved haplotypes in trios from the CEPH population of Utah. Using a selection of samples with similar haplotypes, we illustrate that two haplotypes can differ by short, (<1 kb) tracts of allelic change, highly suggestive of gene conversion events (Figure 7). These haplotype differences can lead to the eventual creation of new haplotypes by these mechanisms as the LD between markers decays over time and these new haplotypes are propagated in the population.

To directly measure gene conversion events over a single generation, we tracked the inheritance of markers in progeny from two large mouse backcrosses using markers spanning 1 MB of the region surrounding the mouse HFE locus. Using this system, we were able to differentiate gene conversions from double crossovers. Crossovers are thought to occur only once per chromosome or once per chromosome arm and to exhibit the phenomenon of interference, in which a single crossover event will deter a second event from happening nearby. Since we were observing these events occurring over a single generation, we could rule out double crossovers and clearly identify gene conversion events.

We observed 2 gene conversion events in 23,753 genotypes surveyed from this mouse study. From this we make the first estimate of gene conversion occurring in a single generation in female meioses to be as high as 1 in 10^4 , which is as much as four orders of magnitude higher than mutation rates for non-CpG sites. The estimation of female meiotic gene conversion events we made in this region is lower than those reported for events detected in other regions by sperm typing analysis of male meioses in mouse and humans. Sex differences may account for the different frequencies we

observed. In addition to sex differences, however, another major difference between our study and others reporting gene conversion frequencies is that we report gene conversion events that are not located at apparent hotspots for crossing over. This finding correlates with both our human data as well as with recent yeast data supporting the notion that gene conversions are not limited to crossover hotspots. Thus far in mammalian systems, however, gene conversions have only been reported at crossover hotspots and the estimations of frequency of conversion events have also been limited to sites that are also crossover hotspots. This raises the possibility that gene conversion events occur more frequently at recombination hotspots, although as we demonstrate, they are not limited to these sites.

Given the high frequency with which gene conversions may occur at recombination hotspots, we propose that gene conversion events that happen at the local hotspot we identified may have led to the propagation of the H63D mutation onto many different haplotypes. The high frequency with which we observe gene conversion events when compared to the low genome-wide mutation rate at non-CpG nucleotides supports the proposal that gene conversion is a likely mechanism for this observation. It is also likely that selective pressures may have acted to amplify this effect, such that conversion-containing chromosomes would persist more than expected in populations. Various proposals have been made regarding candidate selective forces. One suggestion has been that the transition from a hunter-gatherer to a farming society made mutations that could protect from iron-deficiency anemia beneficial in light of an iron-poor diet (Distante et al., 2004). Another more intriguing suggestion has been that the HFE protein, expressed at cell surfaces, may serve as a receptor for an infectious agent (Rochette et al., 1999).

The H63D mutation might disrupt this interaction, and lead to a selective advantage. Such possibilities make it plausible that the H63D-containing chromosome was propagated by both gene conversion and selective pressures.

This study demonstrates the presence of gene conversion events in the HFE locus that are not limited to sites with elevated crossover activity. Using estimates from human chromosomes marked by a disease-causing mutations and a direct measurement of these events in mouse backcross progeny, we demonstrate that gene conversions at these sites can occur at least as frequently as $1.2-4.9 \times 10^{-6}$ per site per generation. Given this high frequency, we propose that gene conversion events, perhaps in the context of selective pressures, can lead to the propagation of disease alleles, exemplified by H63D, in populations.

REFERENCES

- Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S., Daly, M. J., and Donnelly, P. (2005). A haplotype map of the human genome. *Nature* 437, 1299-1320.
- Distante, S., Robson, K. J., Graham-Campbell, J., Arnaiz-Villena, A., Brissot, P., and Worwood, M. (2004). The origin and spread of the HFE-C282Y haemochromatosis mutation. *Hum Genet* 115, 269-279.
- Feder, J. N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D. A., Basava, A., Dormishian, F., Domingo, R., Jr., Ellis, M. C., Fullan, A., *et al.* (1996). A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet* 13, 399-408.
- Guillon, H., and de Massy, B. (2002). An initiation site for meiotic crossing-over and gene conversion in the mouse. *Nat Genet* 32, 296-299.
- Jeffreys, A. J., and May, C. A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet* 36, 151-156.
- Keen Kim, J. D. (2002) High-resolution Linkage Disequilibrium in the Cystic Fibrosis Transmembrane Conductance Regulator Gene: Implications for Association Mapping, Massachusetts Institute of Technology, Cambridge.
- Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310, 321-324.
- Nachman, M. W., and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297-304.
- Rochette, J., Pointon, J. J., Fisher, C. A., Perera, G., Arambepola, M., Arichchi, D. S., De Silva, S., Vandwalle, J. L., Monti, J. P., Old, J. M., *et al.* (1999). Multicentric origin of hemochromatosis gene (HFE) mutations. *Am J Hum Genet* 64, 1056-1062.
- Wexler, N. S., Lorimer, J., Porter, J., Gomez, F., Moskowitz, C., Shackell, E., Marder, K., Penchaszadeh, G., Roberts, S. A., Gayan, J., *et al.* (2004). Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc Natl Acad Sci U S A* 101, 3498-3503.

Table 1: Summary of SNP markers used in each analysis.

The SNP markers used in each analysis is shown. The name of each SNP used in our study is shown alongside the reference SNP ID (rs#) assigned by the NCBI database. Columns 3-5 indicate the SNPs used in the analyses of DNA from different populations as shown.

Table 1

JK SNP name	NCBI name: rs #	Genotyped in Diversity panel	Genotyped in Venezuela	Genotyped in C282Y +/-	Reference (chimp) allele	Alternate allele	Chr 6 position	Dist from prev SNP (bp)
485	rs1935235	*		*	G	C	26175760	-
487	rs9358904	*	*	*	G	A	26176544	784
489-1	rs9295683	*	*	*	T	C	26177473	929
489-2	rs9295684	*	*	*	C	T	26177647	174
493	rs9295685	*	*	*	G	C	26179703	2056
494	rs9968910	*		*	A	C	26180073	370
495	rs9366634	*		*	G	C	26180374	301
495-2	rs6942196	*	*	*	A	G	26180782	408
500	rs807205	*	*	*	G	C	26183013	2231
500-2	rs1539183	*	*	*	G	C	26183029	16
501 [#]	rs9393684	*	*	*	C	G	26183509	480
505	rs9358905	*	*	*	A	T	26185817	2308
505-2	rs10946805	*		*	C	T	26185869	52
515	rs9295687	*	*	*	C	T	26190688	4819
516 [#]	rs4529296	*	*	*	G	C	26191113	425
517	rs9379825	*	*	*	C	A	26191849	736
521-2	rs1971508	*	*	*	C	A	26193785	1936
522	rs2006736	*	*	*	T	C	26193995	210
524	rs2794720	*	*	*	G	C	26195180	1185
525	rs2858993		*		T	A	26195834	654
532-1(H63D)	rs1799945	*	*	*	C	G	26199157	3323
532b-3(S65C)	rs1800730		*		A	T	26199163	6
532-2	rs2071303	*	*	*	T	C	26199314	151
534	rs807208		*		N/A :C	T	26200125	811
536(C282Y)	rs1800562	*	*	*	G	A	26201119	994
536-2	rs1800758	*		*	G	A	26201214	95
538	rs2858996	*	*	*	G	T	26202004	790
538-2	rs2071302	*		*	T	C	26202108	104
542	rs707889	*	*	*	G	A	26203909	1801
550	rs1150659	*	*	*	G	A	26208001	4092
553	rs1150660			*	C	A/-	26209418	1417
555	rs198857	*		*	C	G	26210395	977
556	rs1543680	*	*	*	G	A	26211155	760
557	rs198855	*	*	*	T	A	26211376	221
558	rs198854	*	*	*	T	C	26212035	659
563	rs198848	*		*	C	T	26214303	2268
565	rs198846	*	*	*	G	A	26215441	1138
565-2	rs198845	*	*	*	G	T	26215768	327
566	rs198844	*		*	G	C	26216260	492
571-2	rs707894	*		*	A	C	26218577	2317
575	rs198840	*	*	*	T	G	26220142	1565
575-2	rs198839	*		*	G	T	26220598	456
577	rs198838	*		*	T	C	26221318	720
577-2	rs198837	*		*	T	A	26221376	58
577-3	rs198836	*		*	A	T	26221594	218

[#]: in these SNPs, the non-chimp allele is designated as reference for the Venezuela panel only

Figure 1: Schematic representation of SNPs in the HFE locus genotyped in human chromosomes.

41 SNP markers identified through sequencing were used in this analysis. These SNPs span a 45.8 kb region of a locus on 6p21.3 as shown. Each SNP marker is represented by an open circle and shown from left to right. In addition to the HFE gene, this region also includes two histone genes: 1H4C and 1H1t, respectively, downstream from HFE as shown. (See also Table 1).

Figure 1

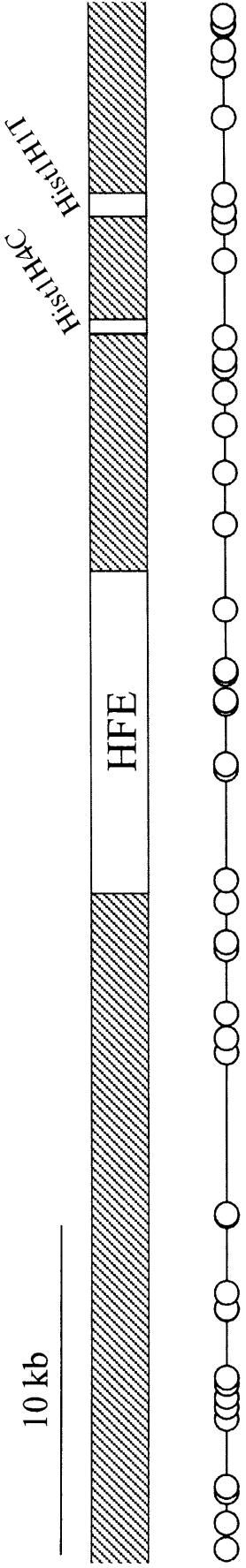


Figure 2: Genotypes of individuals in diversity panel to demonstrate location of local recombination hotspot.

The genotypes of the 32 individuals from the diversity panel are shown. Each column represents the two chromosomes of the individual whose ethnic identity is shown above. SNPs run from top to bottom (see Figure 1 and Table 1 for a summary of these SNP markers). The genotype at each SNP is represented by a color. Homozygosity for the reference allele (1/1) at each SNP is illustrated with a green box. Homozygosity for the alternate allele (2/2) is shown with a red box. A blank box indicates that no genotype is available for that site. Heterozygosity at a SNP (1/2) is represented by a blue box. Individuals with similar patterns are arranged together revealing two groups with largely homozygous blocks of SNPs in the area indicated by a bracket (between SNP 487 and SNP 525). These two groups are shown as A/A and B/B above. A third group, with many heterozygous sites, most likely consists of individuals carrying one A haplotype and one B haplotype (shown as A/B) above. Each block above SNP 525 is associated with more than one block below (shaded region), suggesting the presence of a local site where historical crossover events have occurred, shown by the arrow as a recombination hotspot (RHS).

Figure 2

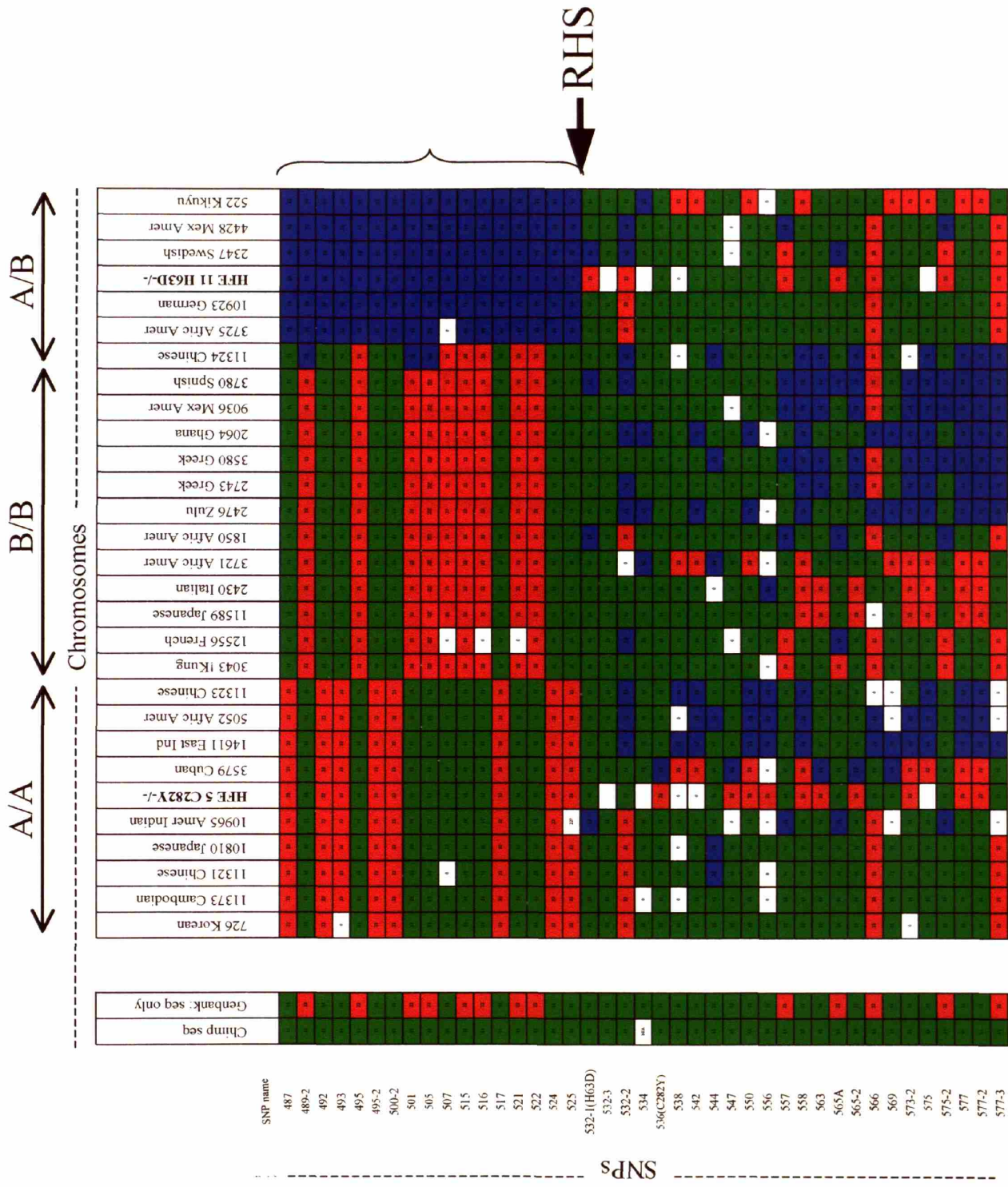


Figure 3: Location of apparent sites of crossing over determined by the number of recombinant chromosomal types observed in Venezuelan chromosomes

SNPs used in this analysis are shown schematically above in relation to the HFE locus and as tick marks below the graph. The graph indicates the location of apparent sites of crossing over determined by recombinant chromosomes in a large Venezuelan pedigree. Peaks were placed midway between the two SNPs between which recombination was observed. The height of each peak correlates to the number of recombinant chromosomal types observed at each location. The largest peak was observed between SNP 525 and SNP 532-1(H63D) with 9 recombinant types, and the second tallest peak was observed between SNP 532-1(H63D) and SNP 532b-3(S65C) with 6 recombinant types. [Note that there is a minor peak adjacent to the second tallest peak that falls between SNP 532-1(H63D) and SNP 532b-3(S65C)]. The third tallest peak was observed between SNP 536 and SNP 538 with 3 recombinant types.

Figure 3

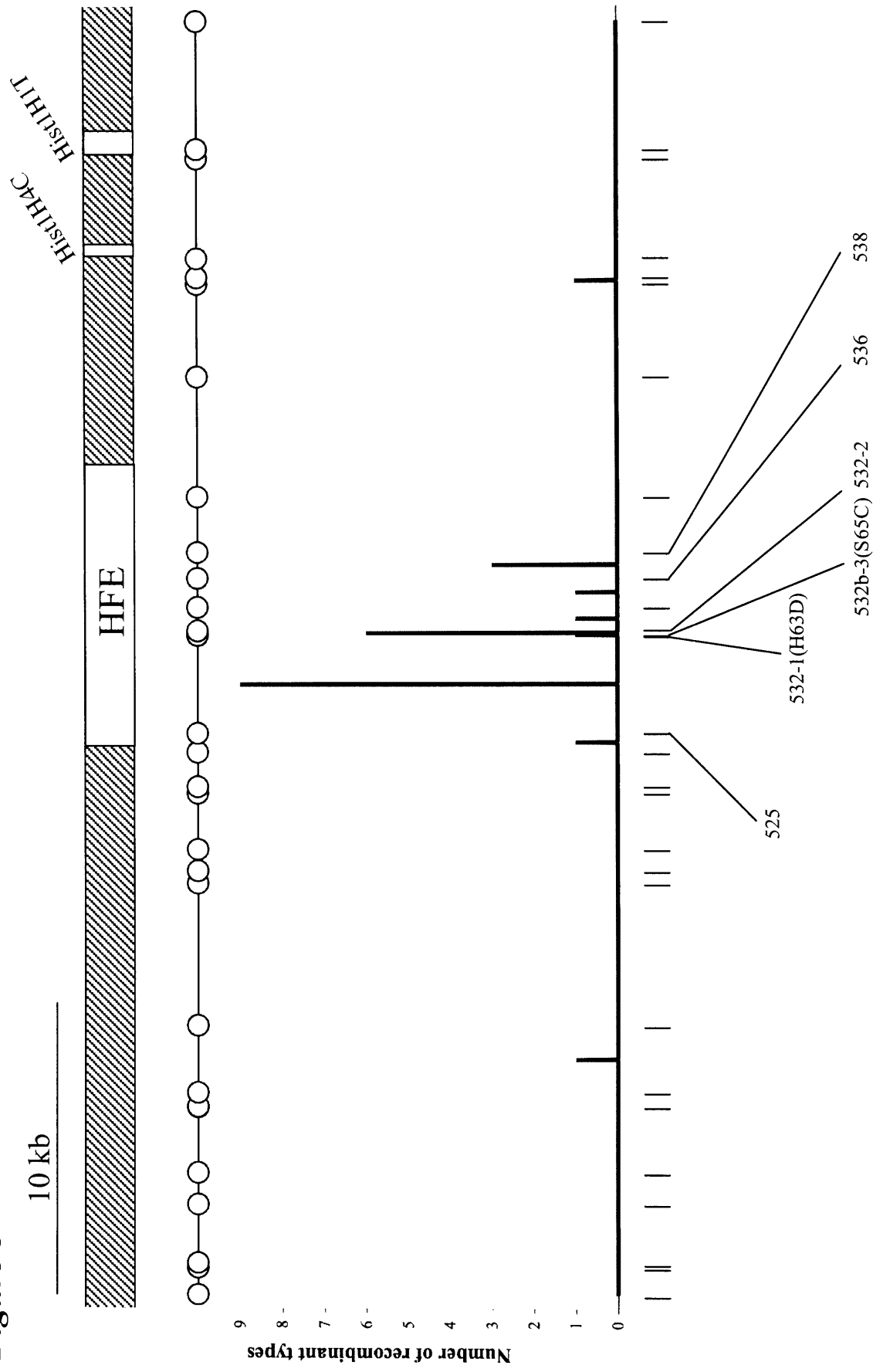


Figure 4: Genotypes from HapMap Japanese panel to demonstrate location of local recombination hotspot.

Genotypes from HapMap are shown in a selection of chromosomes from the Japanese panel with SNPs running from top to bottom and each individual's two chromosomes summed by a single column. Non-recombinant homozygous chromosomes are shown alongside four recombinant chromosomes (each indicated by a *), two of which demonstrate an apparent site of crossing over in the region we identified (shown as RHS, large arrow). The other two sites of crossing over are indicated by the open arrows. Brackets indicate the region studied in our analysis with corresponding SNP names shown to the right.

Figure 4

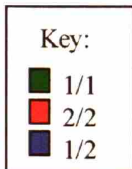
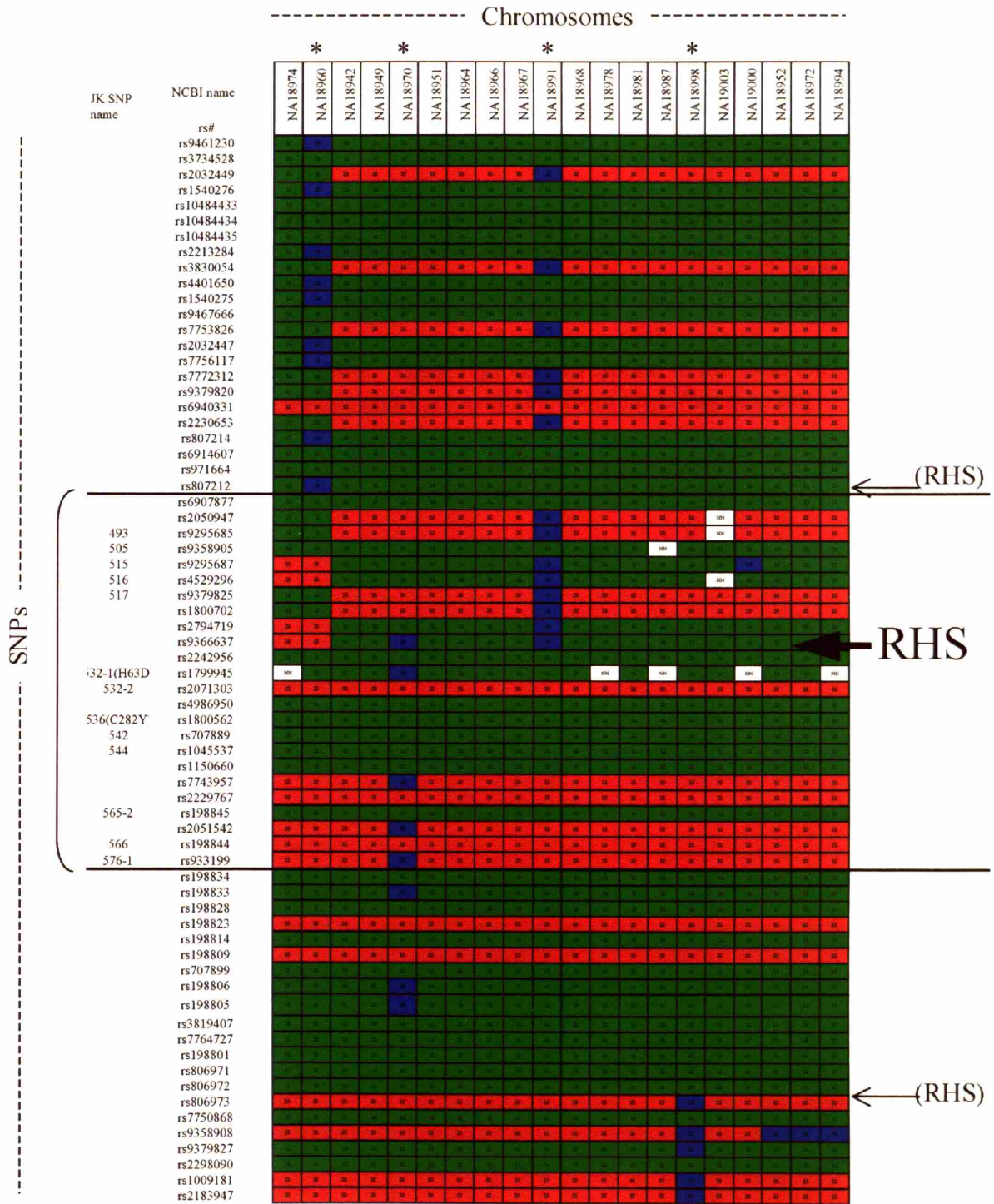


Table 2: Recombination rates in the region we examined from HapMap dataset.

Where available, the SNP names used in our analysis are shown next to the corresponding NCBI SNP designation (rs#). The recombination rate, given in cM/MB between the starting SNP and ending SNP is shown in column 7. The corresponding genetic distances, given in cM, is shown in column 8. Regions with elevated recombination are highlighted in yellow. These regions correspond to the local sites RHS we identified. The rates shown on this chart are available from <http://hapmap.org>.

Table 2

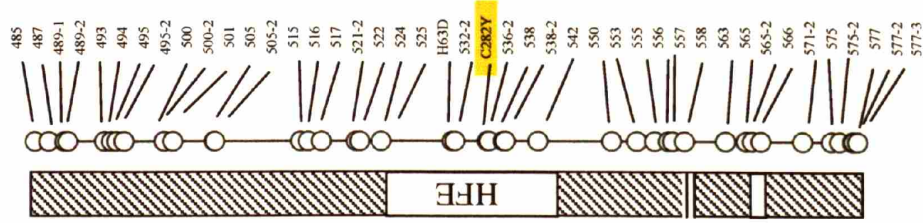
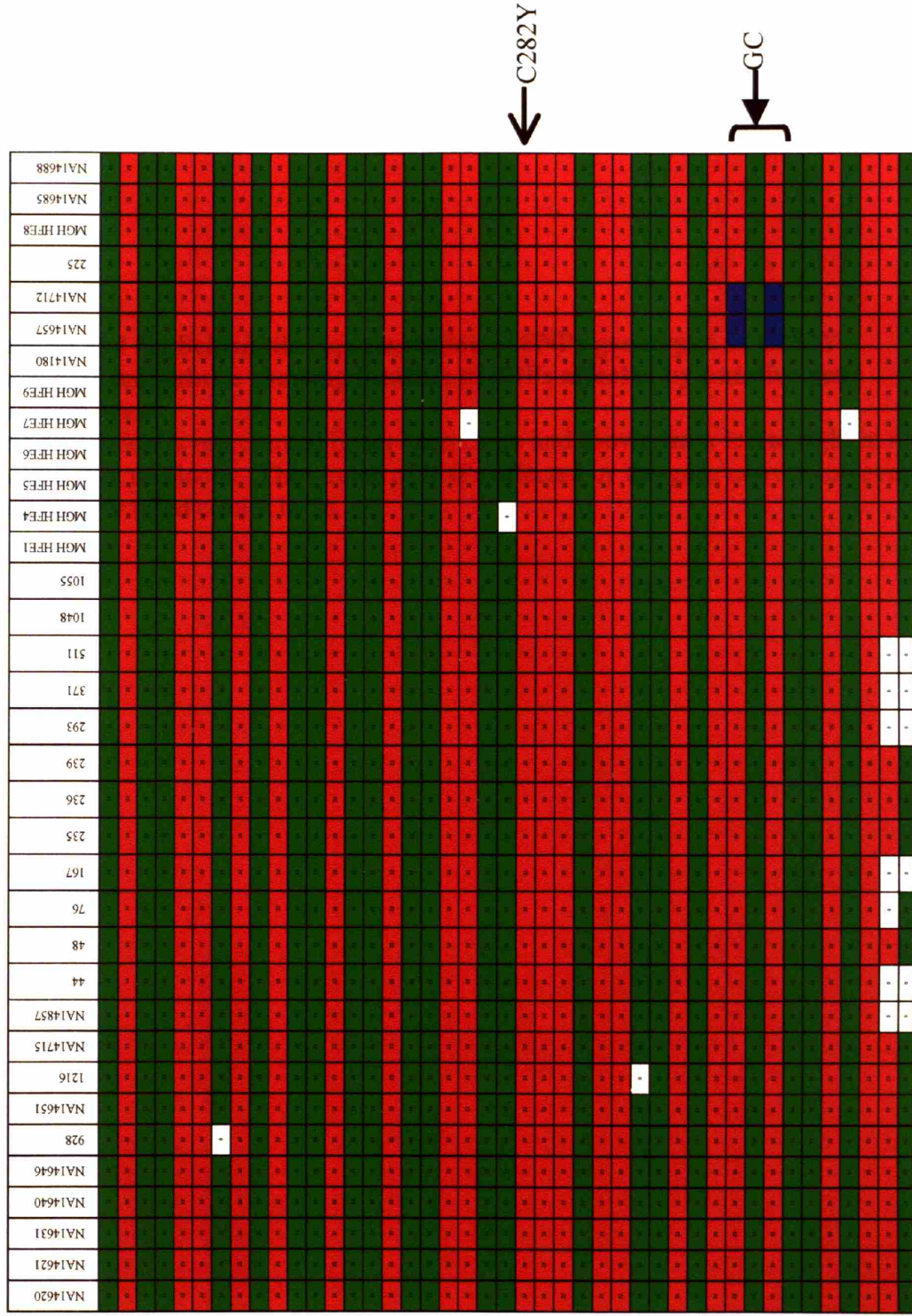
Starting SNP			Ending SNP			Recombination rate		
JK SNP name	rs#	Chr6 position	JK SNP name	rs#	Chr6 position	Rate	cM/Mb	Avg cM
	rs807212	26173600	489-2	rs9295684	26177647	0.07077		0.00029
489-2	rs9295684	26177648		rs2050947	26178057	0.07023		0.00003
	rs2050947	26178058	493-1	rs9295685	26179703	0.06970		0.00011
493-1	rs9295685	26179704		rs9295687	26190688	0.06966		0.00077
	rs9295687	26190689	515	rs4529296	26191113	0.06950		0.00003
515	rs4529296	26191114		rs9379825	26191849	0.06942		0.00005
	rs9379825	26191850	517	rs1800702	26194441	0.06930		0.00018
517	rs1800702	26194442		rs2858993	26195834	0.06946		0.00010
	rs2858993	26195835	525	rs2794719	26196868	0.06955		0.00007
525	rs2794719	26196869		rs9366637	26197076	1.27069		0.00026
	rs9366637	26197077	532-1(H63D)	rs1799945	26199157	1.79696		0.00374
532-1(H63D)	rs1799945	26199158		rs2071303	26199314	0.96281		0.00015
	rs2071303	26199315	532-2	rs1800562	26201119	0.64497		0.00116
532-2	rs1800562	26201120		rs707889	26203909	0.59368		0.00166
536-1(C282Y)	rs707889	26203910	542	rs1045537	26204726	0.24840		0.00020
542	rs1045537	26204727	(544-1)	rs1150660	26209418	0.18706		0.00088
(544-1)	rs1150660	26209419		rs7743957	26211976	0.21873		0.00056
	rs7743957	26211977		rs2229767	26212360	0.23237		0.00009
	rs2229767	26212361	565-2	rs198845	26215768	0.23238		0.00079
565-2	rs198845	26215769		rs2051542	26216146	0.20293		0.00008
	rs2051542	26216147	(566)	rs198844	26216260	0.15725		0.00002
(566)	rs198844	26216261		rs707892	26217211	0.11628		0.00011
	rs707892	26217212	(573-2)	rs198841	26219649	0.10454		0.00025
(573-2)	rs198841	26219650		rs933199	26220871	0.10473		0.00013
	rs933199	26220872	(577)	rs198838	26221318	0.10120		0.00005
(577)	rs198838	26221319		rs198834	26222350	0.10069		0.00010
	rs198834	26222351			26222486	0.09273		0.00001

Figure 5a: Genotypes of C282Y homozygote individuals showing nucleotide sequence changes suggestive of gene conversion.

45 SNP markers spanning a 45.8 kb region are arranged from top to bottom; the schematic of SNPs from Figure 5b is shown vertically to illustrate the position of each SNP relative to each other. The position of SNP 536(C282Y) is shown with an open arrow. Note that all haplotypes in the region are identical except for the short stretch of change resulting in heterozygosity at SNP 563 and SNP 565-2, indicated by the arrow.

Figure 5a

Chromosomes



Key:

- 1/1
- 2/2
- 1/2

Figure 5b and 5c: Schematic representation of SNPs genotyped in C282Y homozygote individuals showing location of SNPs affected by putative gene conversion event.

Each SNP marker is represented by an open circle and shown from left to right in relation to the genes in this locus. The stretch of SNPs affected by the gene conversion event is represented by filled circles in 5c.

Figure 5

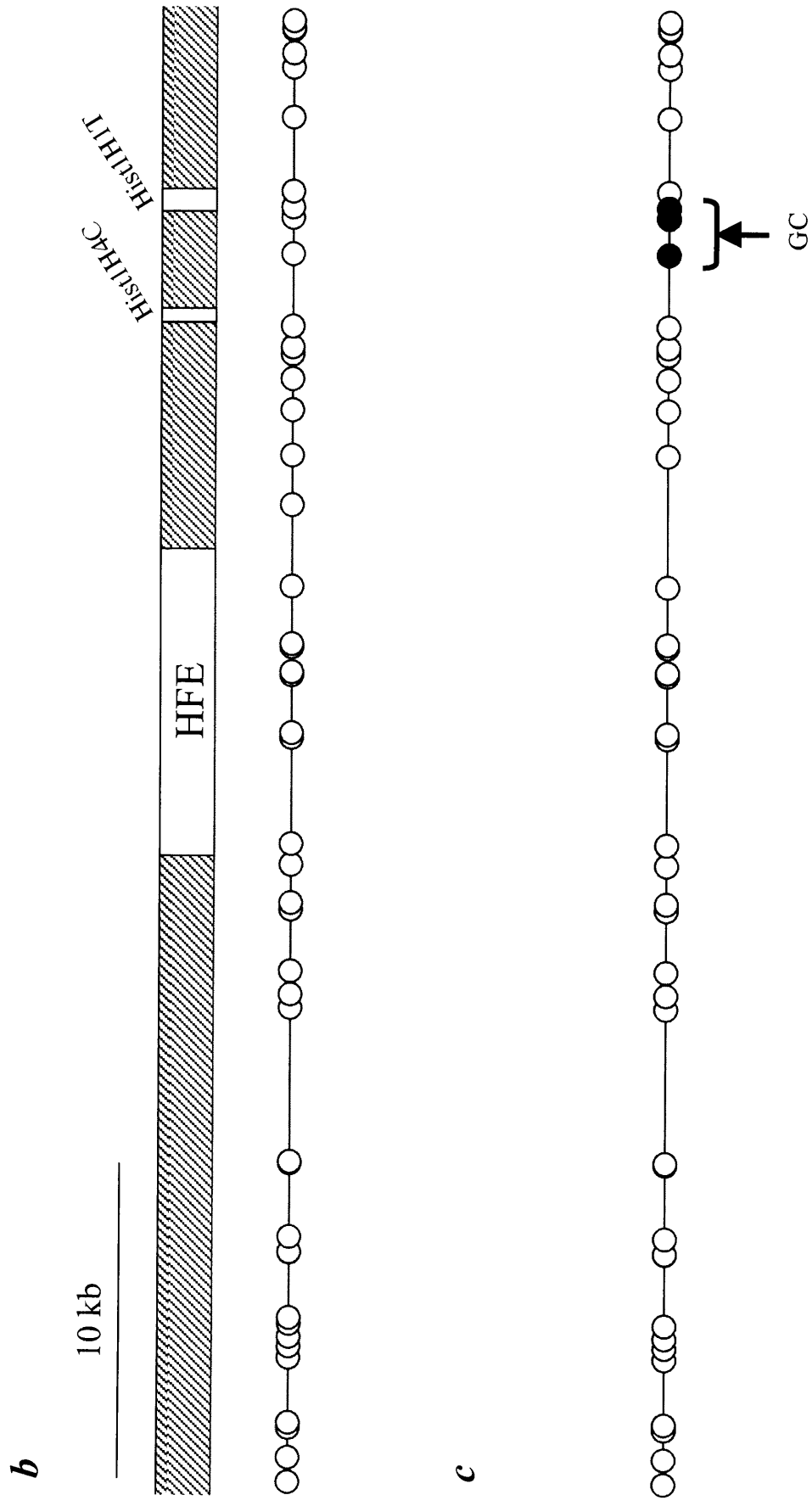


Figure 6: Candidate donor chromosomes from the diversity panel that carry the appropriate haplotype from SNP 563 to SNP 565-2 that could produce the resultant gene conversion event observed in C282Y homozygotes.

A C282Y homozygote is shown at the far left, and the C282Y homozygote with the gene conversion event is shown to its right. The gene conversion event spanning SNPs 563-565-2 would require a pattern of three green homozygote SNPs in the donor chromosome. Such candidate donor chromosomes are shown to the right with the appropriate SNPs circled (arrow).

Figure 6

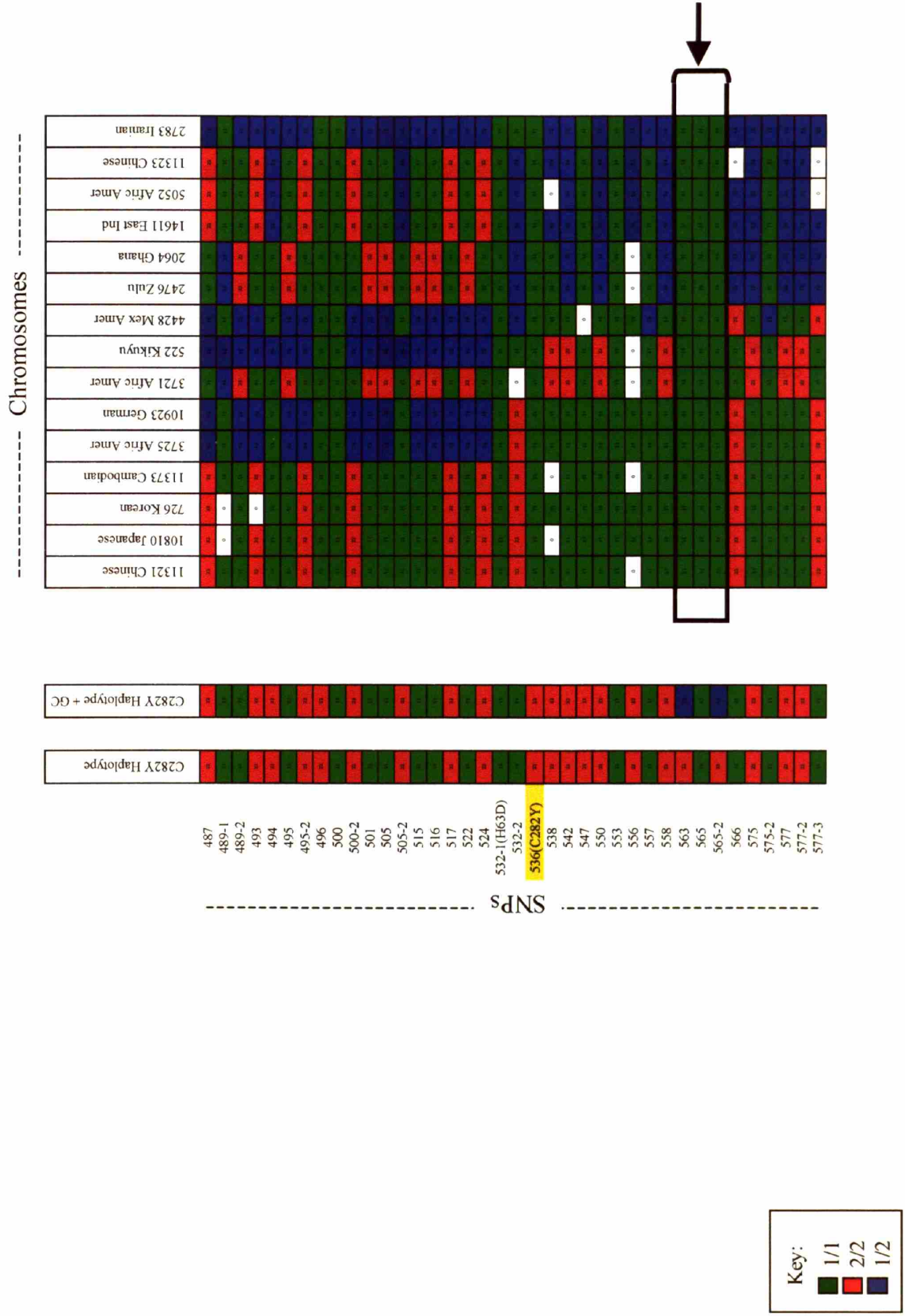
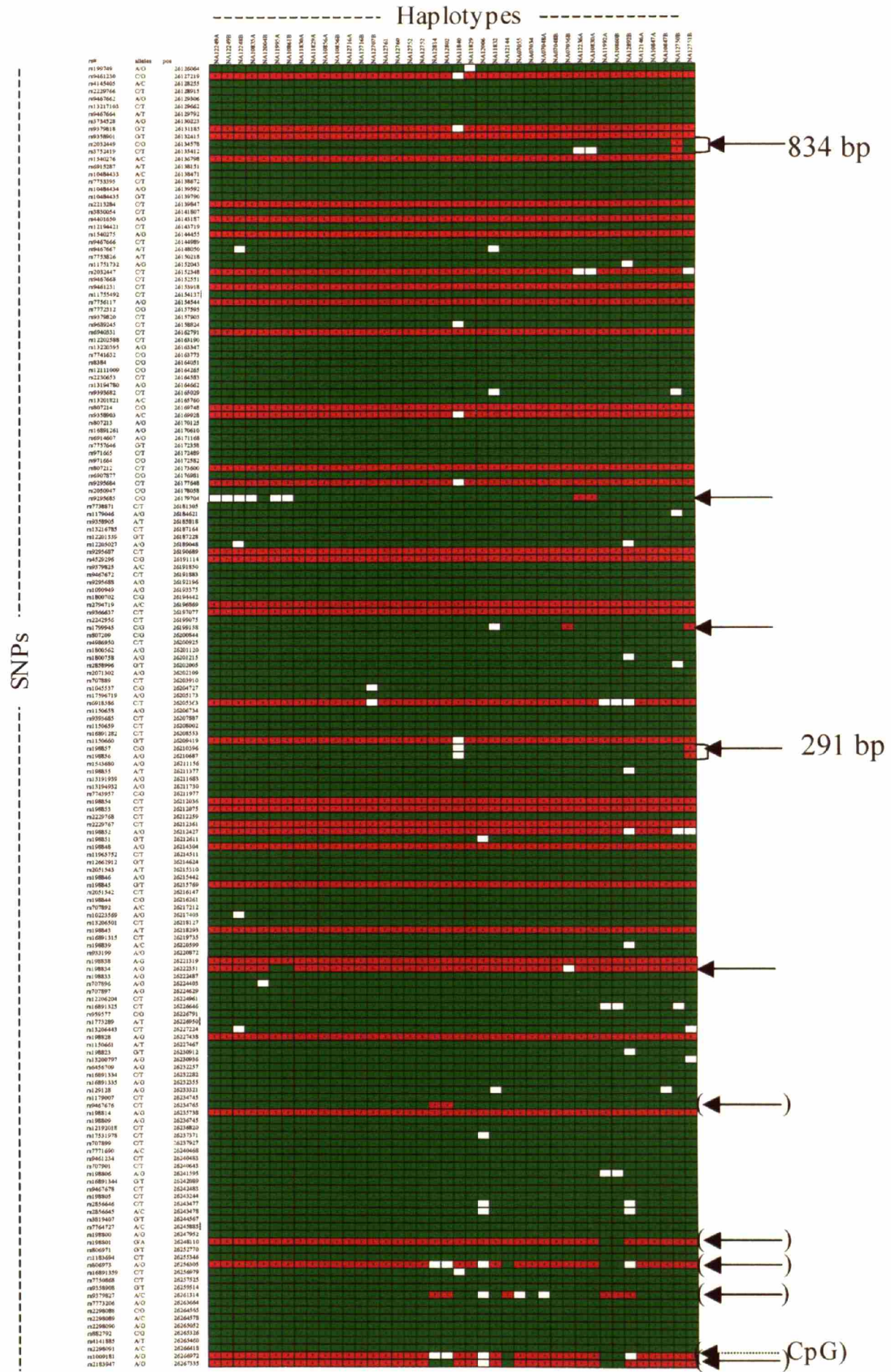


Figure 7: Haplotypes resolved from HapMap genotype information of CEPH UTAH panel of trios to illustrate putative gene conversion events that can lead to haplotype evolution.

Genotype data from HapMap was used to manually resolve haplotypes for 180 SNP makers spanning 141.2 kb centered on the HFE locus. Chromosomes were derived from the CEPH population in Utah (northern and western European ancestry). Arrows show putative gene conversion events indicated by punctate sequence changes. Two of these events involve the coordinate changes of several markers within a 1 kb interval highly suggestive of gene conversion as shown. Arrows in parentheses indicate possible sites of gene conversion events that may alternatively or additionally result from crossover events, suggested by the coordinate change of several markers in the same haplotype. One CpG site at a candidate SNP is indicated.

Figure 7



Key:

- 1
- 2

Figure 8: Strategy for identifying recombination events in two large backcrosses.

a: Schematic to illustrate strategy.

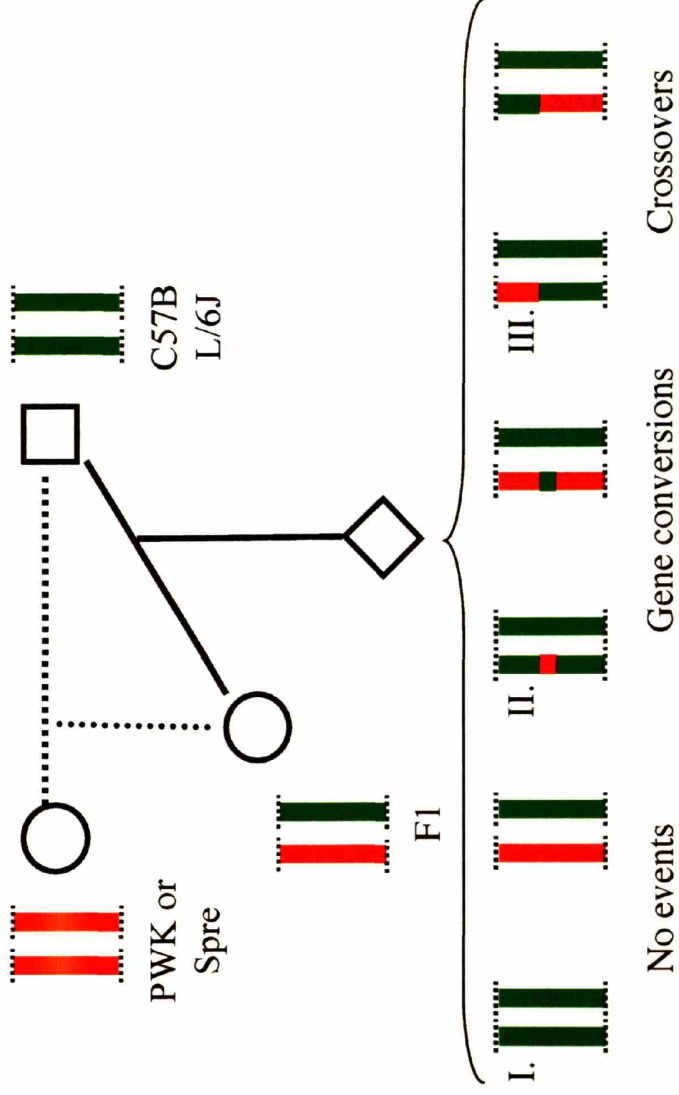
Samples were collected from a ((PWK x C57BL/6J-*gm/gm*) x C57BL/6J-*gm/gm*) backcross and from a ((*M. spretus* x C57BL/6J-*gm/gm*) x C57BL/6J-*gm/gm*) backcross, shown schematically. Recombination events were directly observed by genotyping 20 SNP markers covering 1 MB of chromosome 13q. This region is illustrated as stretches of red or green. The results expected from this strategy are shown below. I. No recombination events in the region; II. Gene conversion events in which only a short stretch of markers change from expected; III. Crossover events in which a contiguous portion of the region examined changes from expected.

b: Summary of genotyping analysis for each backcross.

20 markers were genotyped for each sample. The number of successful genotypes indicated represents those in which an unambiguous genotype could be attributed to the given SNP marker.

Figure 8

a



b

Number of backcrossed progeny DNA genotyped:
Successful genotypes

Backcross A	Backcross B	Total
658	570	1228
12827	10926	23753

Backcross A: (PWK x C57BL/6J-*gm/gm*) x C57BL/6J-*gm/gm*
Backcross B: (*M. Spetus* x C57BL/6J-*gm/gm*) x C57BL/6J-*gm/gm*

Figure 9: Schematic representation of SNPs genotyped in mouse chromosomes showing location of RHS and gene conversions identified.

a: SNPs genotyped in mouse chromosomes.

Each SNP marker is represented by an open circle and shown from left to right. 20 SNPs span a region of 1 MB on chromosome 13 (see also Table 3). The location of two sites of crossover with the

number of events observed at each site are shown (RHS).

b: Gene conversions identified.

The same SNP markers are indicated as above. Each filled circle represents a SNP marker where a gene conversion was observed. JK_19 is found in a region with a higher density of markers that cover the mouse HFE locus, these markers, along with the corresponding mouse HFE gene, are shown in the inset.

Figure 9

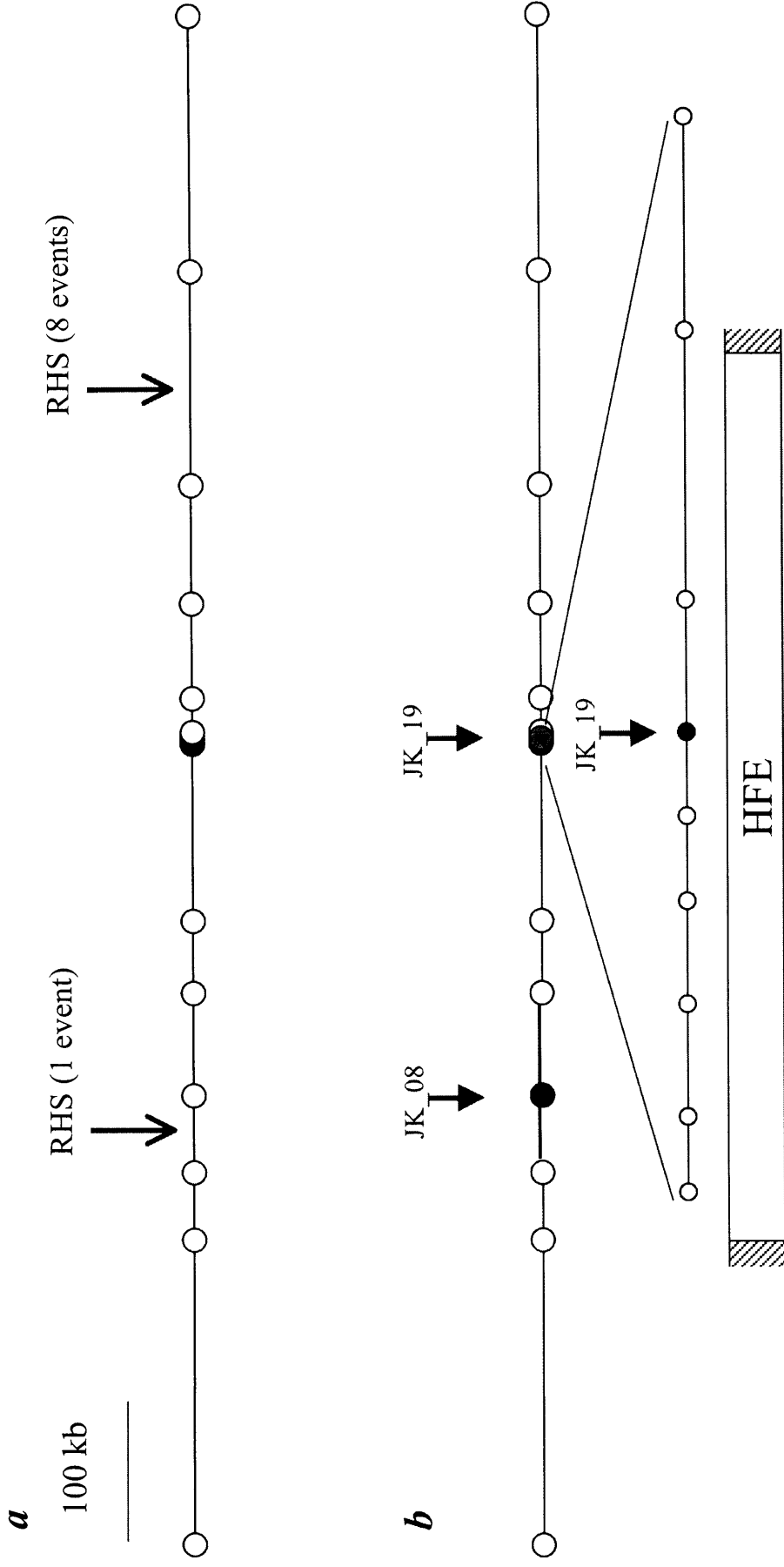


Table 3: Summary of SNP markers used in analysis on mouse chromosome 13.
20 SNP markers identified through sequencing were used in this analysis. These SNPs 1 MB region of a locus on the syntenic region of mouse chromosome 13 containing the HFE gene.

Table 3

SNP name	rs #	C57 allele	PWK/Spre allele	Chr 13 position	Dist from prev SNP (bp)
JK 01	N/A	G	T	22617275	-
JK 05	N/A	A	G	22834552	217277
JK 06	N/A	A	C	22882184	47632
JK 08	N/A	T	G	22936746	54562
JK 83	N/A	----	GAGT	23008473	71727
JK 10	N/A	G	T	23059003	50530
JK 14	N/A	T	G	23184071	125068
JK 15	rs8267033	T	C	23184665	594
JK 16	rs8267100	C	T	23185545	880
JK 17	rs8267020	T	C	23186354	809
JK 18	rs8267097	G	T	23187022	668
JK 19	rs8267015	G	A	23187679	657
JK 20	rs8267079	A	G	23188719	1040
JK 21	N/A	A	G	23190833	2114
JK 23	N/A	C	T	23192516	1683
JK 24	N/A	T	C	23216037	23521
JK 26	N/A	T	C	23282439	66402
JK 29	N/A	A	C	23366239	83800
JK 32	N/A	G	A	23516498	150259
JK 36	N/A	A	T	23697245	180747

Table 4: Summary of population samples genotyped to demonstrate that H63D can be found on multiple haplotype backgrounds over a short distance.

The number of individual samples for each population is shown. 7 H63D homozygotes and 58 H63D heterozygotes were identified.

Table 4

	Number of individuals sampled	
	Total number	H63D heterozygotes
Dutch	128	28
Spanish	37	7
African American	91	8
Kosrae	118	5
Basque	10	4
Viet	74	6
Pygmy	6	0
Primates	7	0
<hr/>		
Total Samples:	464	58
Total Chromosomes:	928	116*

*(of which 58 have H63D)

Figure 10: H63D-carrying chromosomes illustrate that the mutation can be found on multiple haplotypes.

Two major haplotypes above H63D were seen in the populations we sampled and shown here. The haplotype block below H63D as illustrated in the homozygous blocks in the samples to the far left and far right represents the one most frequently seen in the populations we sampled. A third haplotype can be inferred from the middle samples, an H63D homozygote.

Figure 10

----- Chromosomes -----

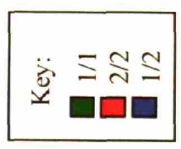
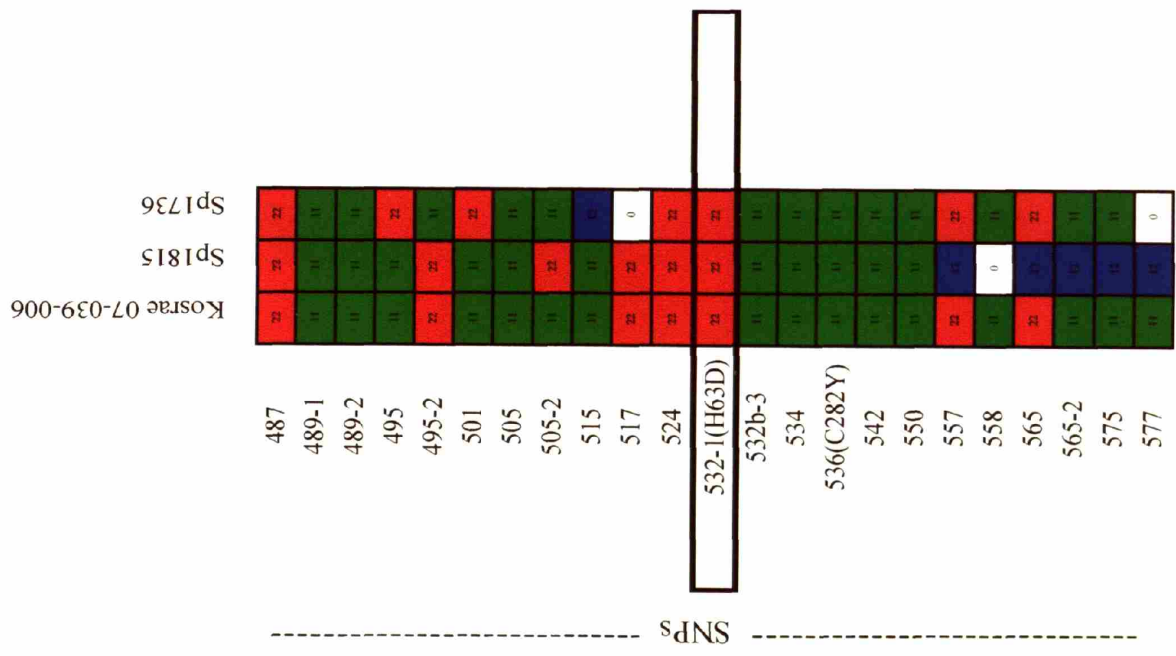


Figure 11: Haplotypes carrying H63D derived from Venezuelan panel.

These three haplotypes represent the three haplotypes carrying H63D in the Venezuela population. They share a common haplotype block below H63D. Two major haplotype blocks above H63D are also shown (A and B). A third minor haplotype, B' (far right), most likely results from a historical gene conversion event at SNP 500-2 that occurred on haplotype B. Of the 104 haplotypes carrying H63D that we studied, 20 A haplotypes, 81 B haplotypes, and 3 B' haplotypes were observed. Each haplotype was subtracted manually from chromosomes using genotype and pedigree information.

Figure 11

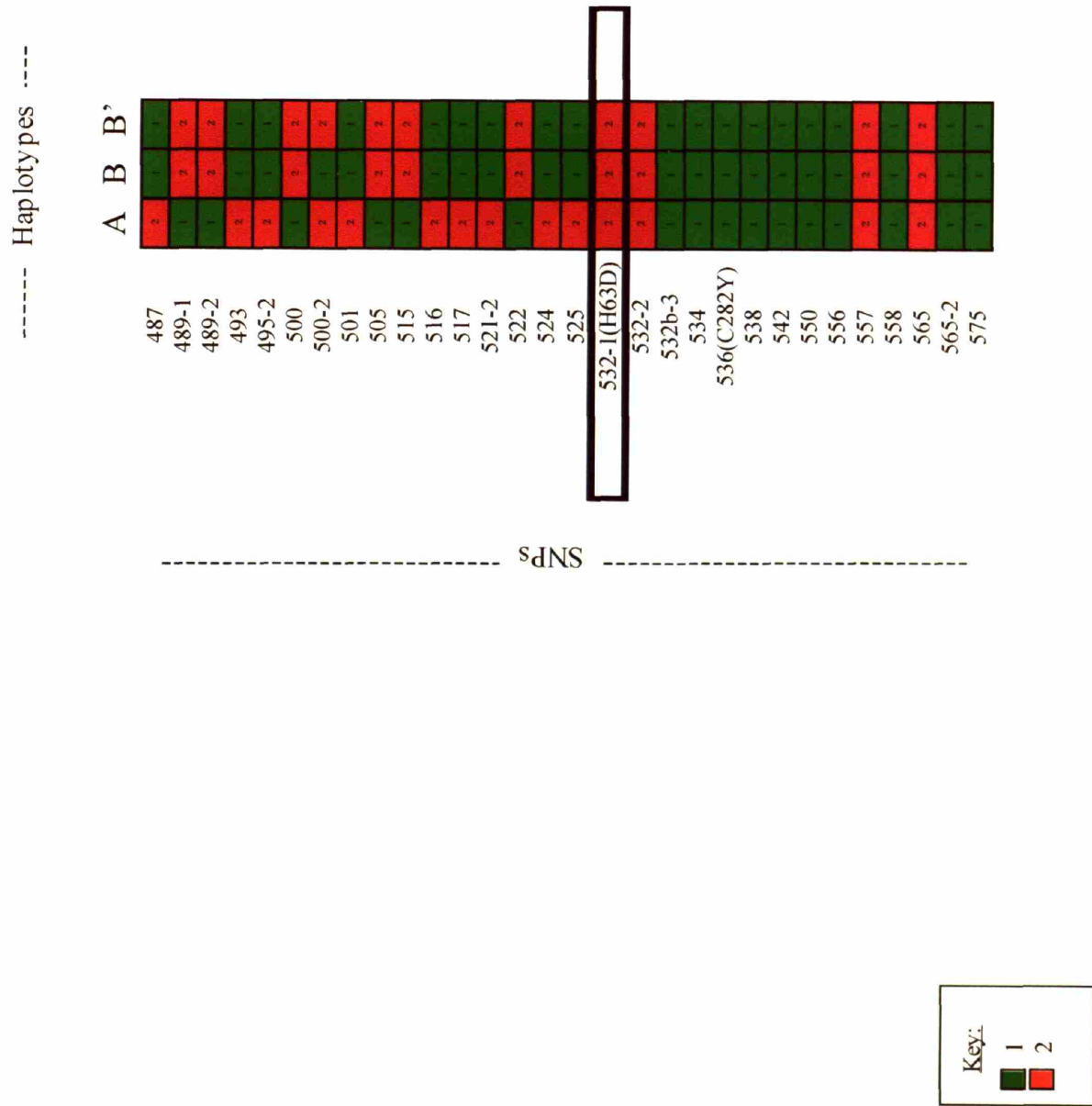


Figure 12: Haplotypes carrying H63D derived from CEPH triads HapMap data.

Genotype data was obtained from hapmap.org, and each haplotype was derived manually using information from each triad. 19 haplotypes are shown, with 101 SNP markers running from top to bottom covering 141 kb of chromosome 6p21 surrounding the HFE gene. All haplotypes carry the H63D mutation, as labeled. Above the H63D mutation, two major haplotypes A and B are present. Below the H63D mutation, two major haplotypes are shown as X (with two variations) and Y. Minor variations within each major haplotype group would also allow us to name further haplotypes on which H63D is found in this population. The area within the bracket corresponds to the region we studied in our populations, with corresponding SNP names shown to the right.

Figure 12

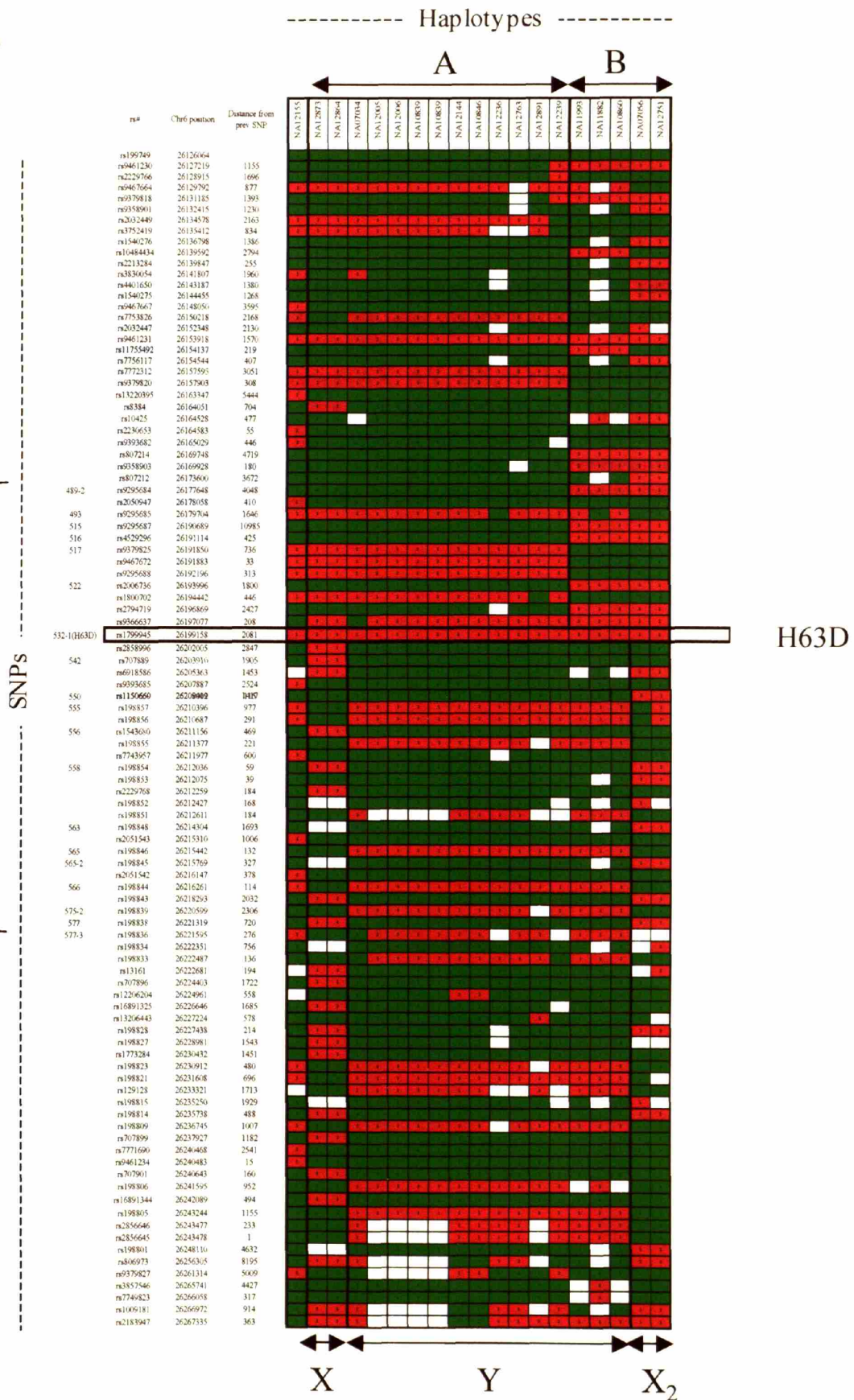
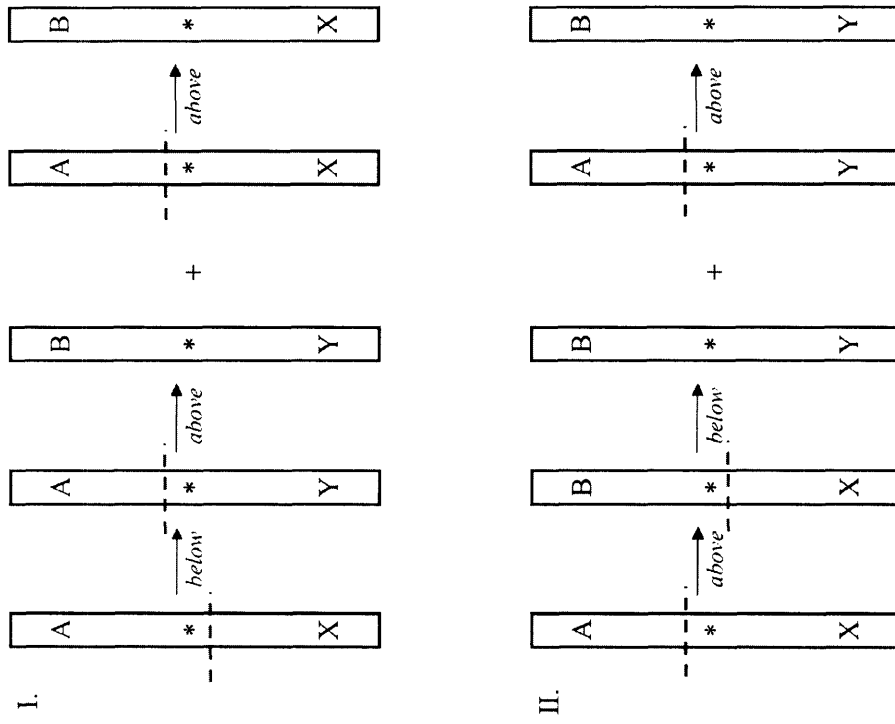


Figure 13: Schematic to illustrate how two sequential crossovers are required if crossover events alone are used to explain the four major haplotypes observed in CEPH data given a single occurrence of H63D.

In order to explain the four major haplotypes (AX, AY, BX, BY) seen in CEPH data (see Figure 12) using crossover events alone, three crossover events are required as shown to move the H63D mutation from any one haplotype to the other three haplotypes. Two examples (beginning with AX) are shown. The H63D mutation is represented by a * and each site of crossover necessary to produce the haplotype combination to the right is represented by a dotted line.

Figure 13



etc.

Chapter 3

Association analysis of the HFE locus with residual age of onset in Huntington's disease

ABSTRACT

Huntington's disease is an adult-onset neurodegenerative disease caused by an unstable, expanded CAG repeat in exon 1 of the huntingtin protein. While age of onset is directly correlated with the length of the CAG expansion, there is significant variation between individuals with a given CAG repeat which has been shown to be highly heritable (Djousse et al., 2003; Wexler et al., 2004). There is thus great interest in identifying genetic factors that are associated with this residual age of onset. We investigate several genes with proteins that have known involvement with iron homeostasis to test the hypothesis that iron regulation may influence age of onset of HD in a large cohort of kindred in Venezuela. We enriched our assay specifically to study the HFE gene on chromosome 6p21.3, where we found evidence for association with this region in a male-specific model. We identified two major candidate genes for association in this region and show that a single haplotype captures alleles that lead to later age of onset of the disease.

INTRODUCTION

Huntington's disease (HD) is a devastating adult-onset neurodegenerative disorder characterized by choreiform movements, emotional disturbances, and cognitive decline. Symptoms typically appear in mid-adulthood, with neurodegeneration that is progressive and most pronounced in the striatum and cortex. An autosomal dominant disease, HD is caused by the expansion of an unstable polyglutamine repeat in exon 1 of the huntingtin gene on chromosome 4p16.3 (HDCRG, 1993). While unaffected individuals have less than 26 repeats, incomplete penetrance results from 36-39 repeats, and penetrance of the disease results from greater than 40 repeats (Myers, 2004). HD also exhibits the phenomenon of genetic anticipation, in which a greater number of repeats results in a more severe phenotype and earlier age of onset (Bates et al., 2002).

HD has been well characterized in a large Venezuelan cohort comprised of individuals affected by HD and their unaffected family members. This well-studied cohort spans 10 generations and includes 83 kindreds, most of whom are from the Lake Maracaibo region of Zulia state (Wexler et al., 2004). These kindreds were instrumental in the isolation of the HD gene and the characterization of the causative mutation (HDCRG, 1993). Samples and extensive pedigree information have been collected alongside data from neurological, neuropsychological, and cognitive examinations performed almost yearly. Age of onset of HD is determined by motor assessment, either prospectively by examination or retrospectively by oral history or by assessment of current motor function severity.

Age of onset in HD is variable, and CAG repeat length accounts for approximately 70% of this variation (Li et al., 2003). At any given repeat length,

however, there is a considerable range of age of onset. This variability has been shown to be strongly heritable both in a large North American cohort as well as in the Venezuela cohort (Djousse et al., 2003; Wexler et al., 2004). Other genetic modifiers are thus thought to affect age of onset of this disease.

It is thus of great interest to find genetic candidates that modify age of onset. Two complementary genetic approaches are underway to address this issue. First, genome-wide linkage scans use markers across the genome to narrow down candidate regions of chromosomes without *a priori* hypotheses regarding gene product functionality. In a large North American cohort, evidence for with residual age of onset was reported to map to three regions of the genome: 4p16, 6p21-23, and 6q24-26 (Li et al., 2003). The second genetic approach is to test for association in candidate genes with potential biological relevance to age of onset. One such category of candidates associated with neurodegenerative diseases, especially those involving extrapyramidal symptoms is genes involved with iron homeostasis (Moos and Morgan, 2004; Thompson et al., 2001).

Higher concentrations of iron in the brain are associated with regions involved with motor function, and iron in these regions increases with advancing age (Zecca et al., 2004). Neurodegenerative diseases such as Parkinson's disease (PD) and Alzheimer's disease (AD) have been associated with iron misregulation and toxicity. The association of these diseases with HFE has thus been an area of increasing investigation (Dekker et al., 2003; Zecca et al., 2004). In AD, for example, some studies report increased oxidative stress and an earlier age of onset associated with HFE mutations although this

association is still being investigated (Berlin et al., 2004; Candore et al., 2003; Pulliam et al., 2003).

Additionally, disruptions in genes important for iron metabolism have been associated with motor symptoms and neuronal phenotypes. For example, a mutation in the ferritin light chain gene, which codes for a polypeptide subunit of the iron storage protein ferritin, has recently been characterized and associated with a dominant, adult-onset movement disorder known as neuroferritinopathy that affects the basal ganglia (Curtis et al., 2001). Other evidence that suggests a possible connection between iron metabolism and motor phenotypes is a knockout mouse model of the IRP-2 gene, whose protein product regulates the expression of iron homeostasis genes. In one model, the IRP-2 knockout mouse has a neurodegenerative movement disorder characterized by bradykinesia, ataxia, and tremors (LaVaute et al., 2001). This movement disorder, however, was not seen in a different IRP-2 knockout model, and therefore needs further investigation (Galy et al., 2005).

Iron homeostasis has also been linked to HD pathogenesis. *In vivo* MRI analysis of HD patients confirms postmortem studies showing increased iron deposition in HD brains. In these studies, increased ferritin iron has been shown in the basal ganglia of patients with HD and detected as early as 9 months from symptom onset (Bartzokis et al., 1999). Recently, the iron and copper chelator clioquinol, has also been shown to be effective in an *in vitro* assay as well as in a mouse model of HD (Nguyen et al., 2005). A previous study in ES cells also suggested that iron depletion can lead to huntingtin upregulation (Hilditch-Maguire et al., 2000).

In order to investigate the potential effects of iron homeostasis on age of onset in HD, we tested key variants in five genes involved in this process: transferrin, IRP-2, ferritin light polypeptide, ferritin heavy polypeptide, and HFE. We focused specifically on the HFE gene locus on chromosome 6p21.3, where a robust peak was reported in the North American genome-wide linkage scan for HD residual age of onset.

MATERIALS AND METHODS

DNA Samples. Samples were obtained from 755 individuals from Venezuela who are part of a large cohort of HD kindreds from the Lake Maracaibo area in Zulia State (Wexler et al., 2004). DNA was previously isolated by phenol chloroform extraction or with an anion exchange column (Qiagen), from lymphoblast cell lines that were originally prepared from whole blood.

Genotyping. PCR amplification of a 500-1000 bp region surrounding each polymorphism was performed in 96-well format using a thermocycler with a final reaction volume of 12-50 μ l. PCR reactions were then denatured and spotted onto Hybond N+ membranes (two identical membranes per PCR plate). Allelic discrimination was performed by using allele-specific oligo (ASO) hybridizations. Each membrane was then probed using an allele-specific oligo labeled with $\gamma^{33}\text{P}$. Membranes were hybridized for 1.5 hr-overnight, washed, and exposed to phosphor screens for subsequent visualization. Images were acquired using a Storm Phosphoimager $\text{\textcircled{R}}$ (GE) after 24 hours of exposure and analyzed visually or with ImageQuant $\text{\textcircled{R}}$ software (GE).

Determination of residual age of onset. Age of onset for each sample was determined previously (Wexler, 2004) by prospective and retrospective examination. Age of onset (referring specifically to the onset of motor symptoms), was determined using a combination of neurological tests either directly by examination or inferred from symptom severity or from patient history. Residual age of onset was also calculated previously (Wexler, 2004). Briefly, age of onset was modeled against the individual's longer repeat number using linear regression. The curvilinear relationship was fitted using the log transform age of onset. The predicted log transform age of onset was subtracted from the log transform of the observed age of onset to determine residual age of onset.

Statistical tests. Initial statistical analyses on all genotype data were performed by J. Gayan (Wellcome Trust Centre for Human Genetics, Oxford, UK) using 1290 informative subjects from the larger pedigree arranged into 45 family groups. 17 families had only a single member, while the remaining 28 had between 4-892 individuals. Association for total age of onset and residual age of onset were examined using a Total Association test, and an Orthogonal test as described in Abecassis et. al (2000a and b). Parent of origin tests were also performed. Significant results were given by a $p < 0.05$, and trends were reported for $p < 0.1$. Results from the tests above formed the basis of the analysis presented in this work. Highly concordant results were obtained using both analyses.

The statistical analyses presented in this work were performed as follows: Deviations from expected Hardy-Weinberg frequencies were calculated using the Chi-square test and a Yate's correction with one degree of freedom. In order to compare each pair of residual means, the equality of each pair of variances was assumed and the two-sample t test for independent samples with equal variances was used to compute p values (two-tailed test). Degrees of freedom for each pair of residual means was estimated by: $(s_1^2/n_1 + s_2^2/n_2) / ((s_1^2/n_1) / (n_1 - 1) + (s_2^2/n_2) / (n_2 - 1))$, where s represents the sample variances and n represents the number of individuals in each group.

RESULTS

Genotyping of major polymorphisms in transferrin and IRP2 genes

We genotyped 6 polymorphisms in the transferrin gene and 2 polymorphisms in the IRP-2 gene, all of which are known to lead to nonsynonymous amino acid changes in the resulting proteins. These polymorphisms and their chromosomal locations are presented in Table 1a. Four of these SNPs, (G142S, W37C, and T645P in transferrin; and F272L in IRP2), were not polymorphic within our population. The other four SNPs (G277S, I448V, and P589S in transferrin; and A852A in IRP2) were analyzed further for association. Allele frequencies in our population are shown in Table 1b. Only Tf4 (I448V) deviated significantly from Hardy-Weinberg equilibrium. Genotyping was performed in 755 individuals, including those with expanded CAG repeats and familial controls. Association analyses for age of onset was performed on 425 individuals from the CAG expanded group (Orthogonal and X test: Wellcome Trust, Oxford), whose

disease had become clinically significant (i.e. they had a documented age at onset). Tf5 showed a trend toward association with $p=0.05$, however, this fell above our cut-off of $p<0.05$.

Sequencing of ferritin gene loci

In order to search for candidate polymorphisms specifically relevant to our population, we sequenced regions of the ferritin light polypeptide (FTL) and ferritin heavy polypeptide 1 (FTH1) genes. We sequenced 2.2 kb of the FTL gene on chromosome 19q13 (chr19: 54160006-54161850, including exons 1, 2, 3 and partial exon 4 sequence) and 1.2 kb of the FTH1 gene on chromosome 11q12 (chr11: 61488603-61489721, including part of exon 1 and all of exons 2 and 3). For this analysis, we sequenced samples from 16 individuals, whose residual ages of onset ranged from 17.11 years earlier to 22.86 years later than expected based on CAG repeat length. These individuals are listed in Table 2. Samples were selected from different familial branches of the larger pedigree as shown. We found only a single polymorphism among these samples (rs 8108882) that does not result in an amino acid change (L55L). There was some suggestion of association in the cohort sequenced; however, further analysis in a larger cohort is needed to confirm these results.

High density SNP analysis of HFE gene and surrounding 43.6 kb

We then focused on a high density SNP analysis of the HFE locus, using 29 markers spanning 43.6 kb centered on the HFE gene. These SNPs are presented in Table 3 and shown schematically in Figure 1. Included among these SNPs are four

polymorphisms that result in nonsynonymous changes in the resulting amino acid. These SNPs are listed with the amino acid changes shown next to the SNP name. Three of these SNPs are found in HFE: 532-1(H63D), 532b-3(S65C), and 536(C282Y), while one of these SNPs is found in histone 1H1t: 565-2(Q178K). Initial statistical tests (performed by J. Gayan, Wellcome Trust, Oxford) showed evidence for association among a selection of these SNPs that was most robust in a male-specific model (Abecasis et al., 2000a; Abecasis et al., 2000b). In order to examine this more closely, the mean residual ages of onset in males was calculated for each genotype class for all 29 SNPs. Each combination of two SNP alleles gave a total of three genotype classes: one homozygote class for each SNP and a third heterozygote class. The means from these three classes were compared in pairs. Thus, for a hypothetical diallelic SNP in which alleles A=1 and G=2, for example, mean residual age of onset for the three genotype classes of 11, 22, and 12 (AA, GG, and AG respectively) were compared in the groupings: 11 vs. 22, 11 vs. 12, and 22 vs. 12. In all cases except where indicated, the SNP allele corresponding to the ancestral allele as inferred from *pan troglodyte* (chimp) sequence was designated the reference allele, 1. The remaining alternate allele was designated 2. The mean residual ages of onset as well as these comparisons are shown in Table 4. The statistical significance of the difference between each mean was determined by calculating a p value using a two-sample t test (two-tailed). A p value of < 0.05 was considered significant (highlighted in yellow in Table 4), while a p value of <0.1 was considered suggestive of association (highlighted in orange in Table 4). A total of seven SNPs showed significant (p<0.05) differences in mean age of onset between genotype classes.

The location of these SNPs in relation to the HFE locus are shown schematically in Figure 1.

Six SNP markers showing the strongest association (largest differences between residual means among genotype classes with smallest p values) are shown in Table 5. Also noted in this Table are any deviations from Hardy-Weinberg equilibrium among these SNPs in the population. To demonstrate that the association present between age of onset and these SNP markers is most robust in a male-specific manner, Table 5 also presents the differences in mean residual ages of onset by genotype class for the same SNPs in both sexes combined (males + females, top panel), as well as for females only (bottom panel) for comparison. As shown, some statistical significance was noted when males and females were combined. However, no significant differences in mean age of onset was noted for any genotype class combinations in the female group. Therefore, the associations seen in both males and females combined, therefore, appear to derive from the male-only associations, where the most robust effect is seen.

To demonstrate graphically the effect that different SNP allele combinations can have on age of onset, the mean residual ages of onset by genotype class are shown for SNPs 558, 565, and 575 in Figures 2 a,b, and c, respectively. Error bars represent standard deviations. SNP 565 demonstrated the largest difference in age of onset, with homozygotes of one allele (GG) developing onset of the disease 7.97 years later ($p=0.009$) when compared with homozygotes of the other allele (AA).

We reasoned that the SNPs showing significant association with age of onset in HD most likely had alleles that could be found in linkage disequilibrium with each other. Therefore, we wanted to determine if a “protective” haplotype could be found within the

population. We resolved haplotypes using pedigree information from 102 chromosomes within the cohort. A summary showing a representative illustration of each haplotype found in the population is shown in Figure 3. Using the seven SNPs that showed the highest association with age of onset (indicated by arrows in Figure 3), we found a single haplotype carrying “favorable” alleles at each of these SNPs (the presence of each favorable allele resulted in a later age of onset when the SNP was tested individually for association). This haplotype was found in 10 of the 204 haplotypes we examined and is shown in Figure 3. Conversely, a haplotype with “unfavorable” alleles at all seven most associated SNPs was also found in 41 of the 204 haplotypes we examined.

One of the SNPs that demonstrated significant association with age of onset with HD is SNP565-2(Q178K). The effect of this SNP on age of onset is shown graphically in Figure 4. The reference SNP allele corresponds to the first position of a glutamine codon. The alternate allele changes this glutamine (uncharged) to a lysine (basic) in the resulting histone H1t protein (Figure 5a). The reference allele sequence is conserved in chimp, mouse, rat, dog, chicken, and zebrafish, as shown in Figure 5b. In addition, chimp, mouse, and rat sequences all have a corresponding glutamine at that position.

DISCUSSION

We present here evidence for association between a 43.6 kb region on 6p21.3 and residual age of onset in Huntington’s disease in a male-specific manner. The statistical analyses presented in this report have been corroborated by further rigorous testing. First, regression analysis was used to replace the comparison of each allele class (as

shown in this report), in order to reduce the effects of multiple testing. Second, to account for the fact that many affected individuals in our cohort are related by birth, a modified TDT test (Abecasis et al., 2000b) was performed to control for possible stratification bias. These tests (performed by J. Gayan, Wellcome Trust, Oxford) placed more stringent requirements for association but nevertheless gave highly concordant results with those shown in this report, thus validating our findings of a male-specific association with HD age on onset in this region.

A previous study using a genome-wide linkage scan in a large North American HD cohort (Myers) reported a peak on 6p21-23 suggesting linkage to residual age of onset this region. We found multiple markers within 43.6 kb that showed significant association ($p < 0.05$) with residual age of onset. Seven markers with the most significant association had alleles that could be grouped together as “favorable” or “unfavorable” depending on whether age of onset occurred later or earlier than expected, respectively, according to CAG repeat length. All seven “favorable” alleles could be found together on a single haplotype in the population, as could all seven unfavorable alleles.

Here we present two candidate genes for association in this region. The first candidate gene is the HFE gene. Six of the seven most associated SNPs flank this gene, while a seventh SNP, SNP 532-2, falls within the gene between exons 2 and 3. HFE is a biological candidate for modification of HD age of onset due to its involvement with iron homeostasis. Iron increases in the brain with advancing age in humans, and it has been suggested that this may be correlated with the adult-onset nature of HD. Additionally, increased iron deposition is seen in regions of the brain in HD patients that are most

affected by the disease (Bartzokis et al., 1999). We detect an association which is most robust in males.

The most obvious candidate variants in the HFE gene are previously identified mutations known to disrupt normal function of the HFE protein in iron homeostasis. We tested three of these mutations for association: C282Y(SNP 536), H63D(SNP 532-1), and S65C(SNP 532b-3). Interestingly, the haplotype containing the “favorable” alleles of the seven associated markers in this region and thus conferring a “protective” haplotype, also contains the alternate allele for SNP 532-1, which corresponds to the H63D mutation. H63D was also tested for association directly, but only five homozygotes for the mutation were found in our cohort. This small number likely led to a lack of power to achieve statistical significance. H63D thus remains a viable candidate since the other significantly favorable alleles are in linkage disequilibrium with this marker. A case-control study specifically focused on selecting HD patients with and without this mutation will directly address this issue. The male-specific nature of our association is interesting in light of reports that males with HFE mutations present with more severe iron overload phenotypes when compared with premenopausal women (Deugnier et al., 2002). It would be of interest to investigate how systemic iron overload and brain iron overload are correlated, or possibly inversely correlated.

The second candidate gene we identified is histone H1t, found downstream from HFE in the locus we examined. Our data reveals that SNP 565-2(Q178K), has alleles which show significant association with age of onset. Specifically, we found that a single T allele, which changes the resulting amino acid from a glutamine to a lysine, results in a 2.37 year earlier age of onset ($p=0.004$) in heterozygotes with the genotype GT.

Interestingly, this nucleotide sequence is conserved in other organisms (chimpanzee, mouse, rat, dog, chicken, and zebrafish), and the glutamine at that position also appears in the chimpanzee, mouse, and rat proteins. Homozygotes of the T allele also develop an earlier age of onset, although this did not reach statistical significance in our cohort likely due to the fact that the small number (n=10) of homozygotes in our population. The alternate allele (C) at this SNP, changes an uncharged amino acid (Q) to an acidic residue (K). The significantly earlier age of onset observed in heterozygotes when compared to homozygotes of the G allele, suggests a dominant-negative mechanism resulting from this amino acid change.

We have demonstrated a male-specific association with residual age of onset in HD within a 43.6 kb region of 6p21.3 using a large cohort of HD kindreds from Venezuela. We present two candidate genes, HFE and H1t in this region, that singly or together may explain this association. Multiple markers in the region on defined haplotypes demonstrating significant association argue against a statistical artifact and instead suggest a real association that should be confirmed in other studies. Additional testing of these candidates in other populations as well as further genetic and biological tests should help to explain the physiological impact of these candidate genes on age of onset in HD.

REFERENCES

- Abecasis, G. R., Cardon, L. R., and Cookson, W. O. (2000a). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66, 279-292.
- Abecasis, G. R., Cookson, W. O., and Cardon, L. R. (2000b). Pedigree tests of transmission disequilibrium. *Eur J Hum Genet* 8, 545-551.
- Bartzokis, G., Cummings, J., Perlman, S., Hance, D. B., and Mintz, J. (1999). Increased basal ganglia iron levels in Huntington disease. *Arch Neurol* 56, 569-574.
- Bates, G., Harper, P., and Jones, L. (2002). *Huntington's Disease*, Third edn: Oxford University Press).
- Berlin, D., Chong, G., Chertkow, H., Bergman, H., Phillips, N. A., and Schipper, H. M. (2004). Evaluation of HFE (hemochromatosis) mutations as genetic modifiers in sporadic AD and MCI. *Neurobiol Aging* 25, 465-474.
- Candore, G., Licastro, F., Chiappelli, M., Franceschi, C., Lio, D., Rita Balistreri, C., Piazza, G., Colonna-Romano, G., Grimaldi, L. M., and Caruso, C. (2003). Association between the HFE mutations and unsuccessful ageing: a study in Alzheimer's disease patients from Northern Italy. *Mech Ageing Dev* 124, 525-528.
- Curtis, A. R., Fey, C., Morris, C. M., Bindoff, L. A., Ince, P. G., Chinnery, P. F., Coulthard, A., Jackson, M. J., Jackson, A. P., McHale, D. P., *et al.* (2001). Mutation in the gene encoding ferritin light polypeptide causes dominant adult-onset basal ganglia disease. *Nat Genet* 28, 350-354.
- Dekker, M. C., Giesbergen, P. C., Njajou, O. T., van Swieten, J. C., Hofman, A., Breteler, M. M., and van Duijn, C. M. (2003). Mutations in the hemochromatosis gene (HFE), Parkinson's disease and parkinsonism. *Neurosci Lett* 348, 117-119.
- Deugnier, Y., Jouanolle, A. M., Chaperon, J., Moirand, R., Pithois, C., Meyer, J. F., Pouchard, M., Lafraise, B., Brigand, A., Caserio-Schoenemann, C., *et al.* (2002). Gender-specific phenotypic expression and screening strategies in C282Y-linked haemochromatosis: a study of 9396 French people. *Br J Haematol* 118, 1170-1178.
- Djousse, L., Knowlton, B., Hayden, M., Almqvist, E. W., Brinkman, R., Ross, C., Margolis, R., Rosenblatt, A., Durr, A., Dode, C., *et al.* (2003). Interaction of normal and expanded CAG repeat sizes influences age at onset of Huntington disease. *Am J Med Genet A* 119, 279-282.
- Galy, B., Ferring, D., Minana, B., Bell, O., Janser, H. G., Muckenthaler, M., Schumann, K., and Hentze, M. W. (2005). Altered body iron distribution and microcytosis in mice deficient in iron regulatory protein 2 (IRP2). *Blood* 106, 2580-2589.

Table 1: Summary of SNPs in transferrin and IRP2 genes tested for association with residual age of onset.

a: SNPs selected for analysis.

SNPs producing an amino acid change in the resulting protein were chosen for analysis. Only four of these SNPs were polymorphic in the population we examined. The name of each SNP used in our study is shown alongside the reference SNP ID (rs#) assigned by the NCBI database.

b: Genotypes of polymorphic SNPs selected for analysis.

The number of genotypes for each allele class (11, 12, or 22) in polymorphic SNPs selected for association analysis in the transferrin and IRP2 genes are shown. Results from a Chi square test for deviations from Hardy-Weinberg equilibrium (with and without the Yates correction) is shown to the right. None of these SNPs showed significant association with age of onset in the population (not shown).

Table 1

a

SNP Name	Amino acid changes	rs	Gene	Polymorphic in Vz	Chr:position	Alleles
Tf1	G142S	rs1799830	Transferrin	No	3:134956135	A/G
Tf2	G277S	rs1799899	Transferrin	Yes	3:134958510	A/G
Tf3	W37C	rs1804498	Transferrin	No	3:134960799	A/G
Tf4	I448V	rs2692696	Transferrin	Yes	3:134967831	A/G
Tf5	P589S	rs1049296	Transferrin	Yes	3:134977052	C/T
Tf6	T645P	rs1130537	Transferrin	No	3:134978651	A/C
IRP2-2	F272L	N/A	IRP2	No	15:76551254	C/G
IRP2-5	A852A	rs13180	IRP2	Yes	15:76576543	C/T

b

SNP name	Genotypes (n)		Total n	Chi sq	Yates	
	11	12				22
Tf2	699	78	0	777	2.170	1.246
Tf4*	749	29	2	780	8.150	4.040 *
Tf5	563	190	18	771	0.172	0.084
IRP2	107	202	81	390	0.640	0.525

*deviates from HWE

Table 2: Summary of samples chosen for sequencing in ferritin genes.

Individuals were chosen for sequencing sequencing 2.2 kb on the ferritin light polypeptide (FTL) and ferritin heavy polypeptide 1 (FTH1) gene. These individuals were chosen from the larger cohort for their residual ages of onset ranging from 17.11 years earlier to 22.86 years later than expected based on CAG repeat length. Each individual is represented by a code with characteristics as shown.

Table 2

Family	CODE	Sex	Age of Onset	Allele 1	Allele 2	Residual Age of Onset
2	12	M	N/A	15	15	
2	30	F	45	43	19	5.24
2	150	F	N/A	29	18	
2	179	F	32	42	42	-9.89
1	441	F	N/A	38	15	
1	598	M	30	42	21	-11.89
2	2122	M	47	42	18	5.11
2	3967	M	N/A	41	17	
3002	5981	M	38	41	20	-6.14
100	6279	M	54	42	24	12.11
100	6280	F	59	42	17	17.11
3044	6779	M	67	41	20	22.86
3044	6817	M	54	41	22	9.86
100	8170	M	27	41	23	-17.14
3044	14788	F	63	41	21	18.86
3114	17647	F	58	43	15	18.24

Table 3: Summary of SNP markers on 6p21.3 used in analysis.

The SNP markers used in this analysis is shown. The name of each SNP used in our study is shown alongside the reference SNP ID (rs#) assigned by the NCBI database. The allele corresponding to the ancestral or chimp sequence, was designated as the reference allele (1) in all cases except where indicated (SNP 501 and SNP 512). The alleles shown correspond to the (+) strand on chromosome 6p21.3.

Table 3

SNP name	Reference (chimp) allele: "1"	Alternate allele: "2"	rs #	Chr 6 position	Dist from prev SNP (bp)	100kb Contig position
487	G	A	rs9358904	26176544	-	27180
489-1	T	C	rs9295683	26177473	929	28109
489-2	C	T	rs9295684	26177647	174	28283
493	G	C	rs9295685	26179703	2056	30339
495-2	A	G	rs6942196	26180782	1079	31418
500	G	C	rs807205	26183013	2231	33649
500-2	G	C	rs1539183	26183029	16	33665
501 [#]	C	G	rs9393684	26183509	480	34145
505	A	T	rs9358905	26185817	2308	36453
515	C	T	rs9295687	26190688	4871	41324
516 [#]	G	C	rs4529296	26191113	425	41749
517	C	A	rs9379825	26191849	736	42485
522	T	C	rs2006736	26193995	2146	44631
524	G	C	rs2794720	26195180	1185	45816
525	T	A	rs2858993	26195834	654	46470
532-1(H63D)	C	G	rs1799945	26199157	3323	49793
532b-3	A	T	rs1800730	26199163	6	49799
532-2	T	C	rs2071303	26199314	151	49950
534	C**	T	rs807208	26200125	811	50761
536(C282Y)	G	A	rs1800562	26201119	994	51755
538	G	T	rs2858996	26202004	885	52640
542	G	A	rs707889	26203909	1905	54545
550	G	A	rs1150659	26208001	4092	58637
556	G	A	rs1543680	26211155	3154	61791
557	T	A	rs198855	26211376	221	62012
558	T	C	rs198854	26212035	659	62671
565	G	A	rs198846	26215441	3406	66077
565-2(Q79K)	G	T	rs198845	26215768	327	66404
575	T	G	rs198840	26220142	4374	70778

#: in these SNPs only, the non-chimp allele is designated as reference (1)

** : chimp sequence not available at this nucleotide, reference allele chosen arbitrarily

Figure 1: Schematic representation of SNPs in a locus on 6p21.3 used in analysis and significantly associated.

29 SNP markers (see also Table 3) spanning a 43.6 kb region on 6p21.3 as shown were used in this analysis. Each SNP marker is represented by an open circle and shown from left to right. The distribution of these SNPs in relation to the genes present in this locus are shown. In addition to the HFE gene, this region also includes two histone genes: 1H4C and 1H1t, respectively, downstream from HFE as shown. SNPs showing significant association with residual age of onset in HD are labeled with a filled arrow. The SNPs used in this analysis that lead to non-synonymous amino acid changes in the corresponding protein are shown above each relevant SNP (*).

Table 4: Differences between mean residual age of onset in males by genotype class for 29 SNP markers analyzed.

For each SNP, the number (n) of individuals in each genotype class (11, 12, or 22) is shown (left panel). The corresponding mean residual age of onset and standard deviation is shown beneath each genotype class. The differences between these residual means (22-11, 11-12, 22-12), is shown in the right panel with a p value computed using a t-sample t test (two-tailed). p values significant for association ($p < 0.05$) are highlighted in yellow. p values suggestive of association ($p < 0.1$) are highlighted in orange.

Figure 1

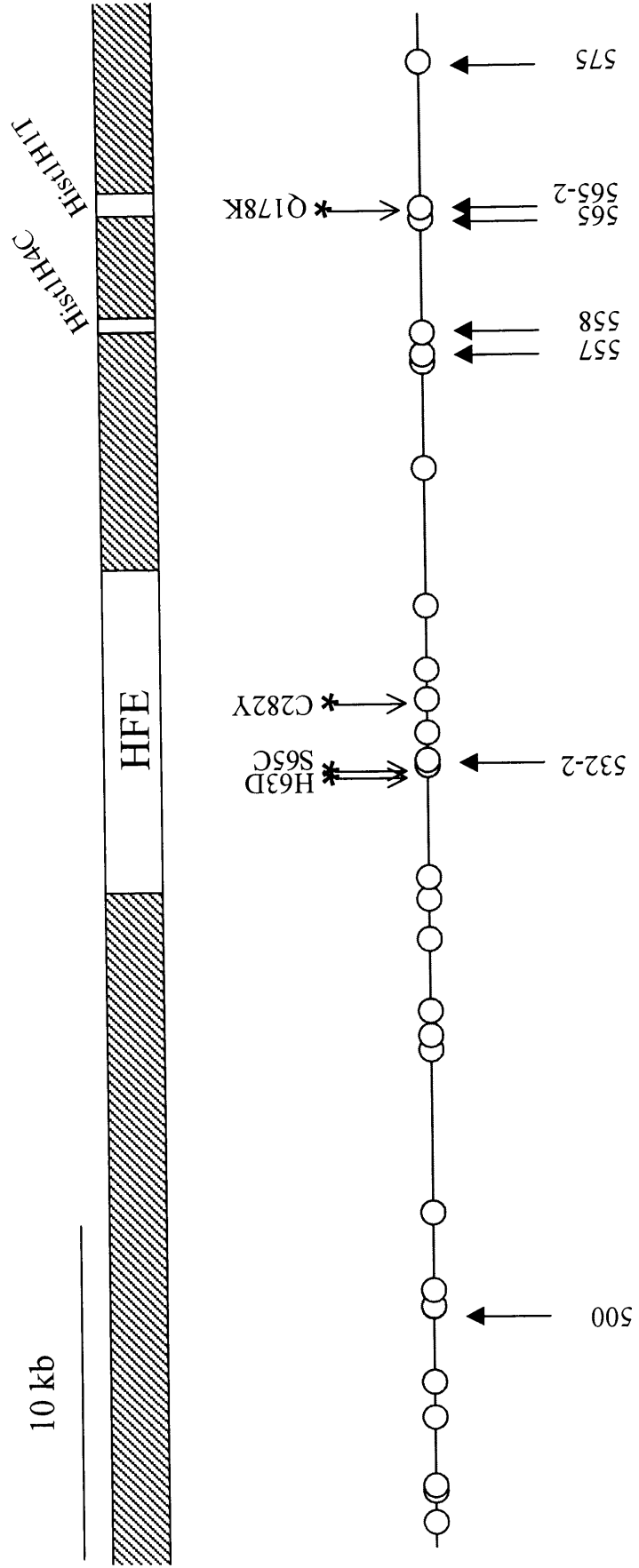


Table 4

MALES		SNP name			Genotypes			Total n	Difference between residual means by genotype					
		11	12	22	22-11		p	11-12		p	22-12		p	
487	n	86	91	19	196	0.85	0.5112	-0.23	0.7875	0.62	0.6891			
	Mean	0.05	0.28	0.90										
	STD	5.03	6.27	5.18										
489-1	n	65	58	19	142	0.12	0.9325	1.76	0.0538	1.88	0.2368			
	Mean	0.78	-0.97	0.90										
	STD	5.42	6.10	5.18										
489-2	n	21	96	88	205	-1.42	0.2590	1.02	0.4952	-0.40	0.6360			
	Mean	1.48	0.46	0.05										
	STD	5.32	6.31	5.03										
493	n	87	92	21	200	1.41	0.2676	-0.27	0.7534	1.14	0.4485			
	Mean	0.06	0.34	1.48										
	STD	5.10	6.33	5.32										
495-2	n	81	96	19	196	1.34	0.3207	-0.63	0.4712	0.71	0.6521			
	Mean	-0.19	0.44	1.15										
	STD	5.11	6.29	5.43										
500	n	77	89	36	202	-0.80	0.4687	1.82	0.0371	1.02	0.3689			
	Mean	1.41	-0.41	0.61										
	STD	5.34	5.77	5.64										
500-2	n	83	89	20	192	2.16	0.0891	-0.37	0.6685	1.79	0.2369			
	Mean	-0.06	0.31	2.10										
	STD	5.01	6.27	4.61										
501	n	84	90	19	193	0.68	0.5899	0.14	0.8688	0.82	0.5995			
	Mean	0.22	0.08	0.90										
	STD	4.83	6.29	5.18										
505	n	21	97	89	207	-1.49	0.2369	0.99	0.5060	-0.50	0.5547			
	Mean	1.48	0.49	-0.01										
	STD	5.32	6.28	5.02										
515	n	20	84	82	186	-1.50	0.2527	1.01	0.4879	-0.49	0.5674			
	Mean	1.49	0.48	-0.01										
	STD	5.46	5.88	5.09										
516	n	86	98	21	205	1.61	0.2034	-0.57	0.4996	1.04	0.4830			
	Mean	-0.14	0.44	1.48										
	STD	5.04	6.25	5.32										
517	n	85	95	22	202	1.59	0.1865	-0.70	0.4079	0.89	0.5445			
	Mean	-0.18	0.52	1.40										
	STD	4.84	6.31	5.20										
522	n	18	78	77	173	-1.52	0.2559	1.11	0.4924	-0.41	0.6599			
	Mean	1.57	0.45	0.05										
	STD	4.86	6.39	5.04										
524	n	89	97	20	206	1.34	0.2989	-0.34	0.6764	0.99	0.5073			
	Mean	0.01	0.36	1.35										
	STD	5.02	6.12	5.43										
525	n	81	86	25	192	0.24	0.8406	-0.26	0.7732	-0.02	0.9890			
	Mean	0.28	0.54	0.52										
	STD	4.87	6.43	5.77										
532-1(H63D)	n	140	62	5	207	3.91	0.1711	-1.05	0.2240	2.86	0.4034			
	Mean	-0.05	1.00	3.86										
	STD	5.11	6.60	6.44										
532b-3(S65C)	n	197	3		200	-0.33	N/A	-1.61	0.6693	-1.94	N/A			
	Mean	0.33	1.94											
	STD	5.60	2.17											
532-2	n	77	88	34	199	2.49	0.0257	-2.21	0.0091	0.28	0.8138			
	Mean	-0.98	1.23	1.51										
	STD	4.90	5.73	6.01										
534	n	191	9		200	-0.18	N/A	-1.45	0.4563	-1.63	N/A			
	Mean	0.18	1.63											
	STD	5.51	4.48											
536(C282Y)	n	179	4	1	184	10.79	N/A	-1.79	0.5703	9.00	N/A			
	Mean	0.23	2.01	11.01										
	STD	5.60	2.56	N/A										
538	n	150	42		192	-0.31	N/A	0.53	0.5860	0.22	N/A			
	Mean	0.31	-0.22											
	STD	5.68	4.99											
542	n	152	45		197	-0.44	N/A	0.73	0.4375	0.29	N/A			
	Mean	0.44	-0.29											
	STD	5.74	4.83											
550	n	155	47		202	-0.48	N/A	0.41	0.6602	-0.06	N/A			
	Mean	0.48	0.06											
	STD	5.76	5.06											
556	n	177	30		207	-0.40	N/A	0.24	0.8280	-0.16	N/A			
	Mean	0.40	0.16											
	STD	5.71	5.42											
557	n	60	118	22	200	3.36	0.0147	-0.10	0.9078	3.26	0.0250			
	Mean	-0.04	0.06	3.33										
	STD	4.72	5.85	6.35										
558	n	71	119	17	207	-3.31	0.0415	2.87	0.0007	-0.44	0.7529			
	Mean	2.28	-0.59	-1.03										
	STD	5.87	5.33	5.06										
565	n	134	67	6	207	7.97	0.0154	-0.39	0.6256	7.58	0.0335			
	Mean	0.07	0.46	8.04										
	STD	5.10	5.84	9.00										
565-2	n	104	90	10	204	-2.10	0.2578	2.37	0.0037	0.27	0.8877			
	Mean	1.58	-0.79	-0.52										
	STD	5.48	5.73	3.66										
575	n	71	120	17	208	-3.31	0.0415	2.85	0.0008	-0.46	0.7404			
	Mean	2.28	-0.56	-1.03										
	STD	5.87	5.34	5.06										

Table 5: Differences between mean residual age of onset by genotype class shown for six most significantly associated SNPs for each gender.

The six SNP markers showing the largest differences in age of onset with the smallest p values from Table 4 are shown here. The same analysis was performed in males and females combined (ALL, top panel), Males only (middle panel), and Females only (bottom panel). p values significant for association ($p < 0.05$) are highlighted in yellow. p values suggestive of association ($p < 0.1$) are highlighted in orange. No association was found among these SNPs in females only, suggesting the association seen in males and females combined derived from robust male-specific associations.

Table 5

ALL (MALES + FEMALES)

SNP name	Genotypes			Total n	Chi sq	Yates	Difference between residual means by genotype						
	11	12	22				22-11	p	11-12	p	22-12	p	
532-2	n	150	200	77	427	0.524	0.421	1.63	0.0527	-1.14	0.0827	0.49	0.5547
	Mean	-0.18	0.96	1.45									
	STD	5.88	6.21	6.13									
557*	n	146	226	54	426	5.434	5.080 *	1.43	0.1477	0.07	0.9158	1.50	0.1109
	Mean	0.54	0.47	1.97									
	STD	6.15	6.14	6.21									
558	n	163	223	53	439	3.098	2.830	-1.61	0.1202	1.64	0.0085	0.04	0.9676
	Mean	1.64	-0.01	0.03									
	STD	6.49	5.66	6.47									
565	n	279	144	16	439	0.239	0.133	4.16	0.0172	-0.37	0.5428	3.78	0.0303
	Mean	0.29	0.67	4.45									
	STD	6.01	5.92	7.37									
565-2	n	211	191	36	438	0.626	0.493	-0.95	0.4068	1.35	0.0263	0.40	0.7188
	Mean	1.27	-0.08	0.32									
	STD	6.16	5.91	6.87									
575*	n	163	228	53	444	9.729	9.411 *	-1.61	0.1202	1.66	0.0079	0.06	0.9509
	Mean	1.64	-0.03	0.03									
	STD	6.49	5.76	6.47									

*deviates from HWE

MALES ONLY

SNP name	Genotypes			Total n	Difference between residual means by genotype						
	11	12	22		22-11	p	11-12	p	22-12	p	
532-2	n	77	88	34	199	2.49	0.0257	-2.21	0.0091	0.28	0.8138
	Mean	-0.98	1.23	1.51							
	STD	4.90	5.73	6.01							
557	n	60	118	22	200	3.36	0.0147	-0.10	0.9078	3.26	0.0250
	Mean	-0.04	0.06	3.33							
	STD	4.72	5.85	6.35							
558	n	71	119	17	207	-3.31	0.0415	2.87	0.0007	-0.44	0.7529
	Mean	2.28	-0.59	-1.03							
	STD	5.87	5.33	5.06							
565	n	134	67	6	207	7.97	0.0154	-0.39	0.6256	7.58	0.0335
	Mean	0.07	0.46	8.04							
	STD	5.10	5.84	9.00							
565-2	n	104	90	10	204	-2.10	0.2578	2.37	0.0037	0.27	0.8877
	Mean	1.58	-0.79	-0.52							
	STD	5.48	5.73	3.66							
575	n	71	120	17	208	-3.31	0.0415	2.85	0.0008	-0.46	0.7404
	Mean	2.28	-0.56	-1.03							
	STD	5.87	5.34	5.06							

FEMALES ONLY

SNP name	Genotypes			Total n	Difference between residual means by genotype						
	11	12	22		22-11	p	11-12	p	22-12	p	
532-2	n	74	112	43	229	0.77	0.5417	-0.11	0.9149	0.66	0.5737
	Mean	0.64	0.75	1.41							
	STD	6.64	6.58	6.29							
557	n	86	108	32	226	0.10	0.9417	0.03	0.9761	0.13	0.9185
	Mean	0.93	0.91	1.04							
	STD	6.98	6.45	6.03							
558	n	92	105	36	233	-0.61	0.6565	0.49	0.5899	-0.12	0.9237
	Mean	1.14	0.64	0.53							
	STD	6.92	5.95	7.05							
565	n	146	77	10	233	1.80	0.4274	-0.35	0.6990	1.45	0.4861
	Mean	0.49	0.84	2.29							
	STD	6.73	6.02	5.65							
565-2	n	107	101	26	234	-0.32	0.8333	0.42	0.6380	0.10	0.9461
	Mean	0.97	0.55	0.65							
	STD	6.77	6.04	7.80							
575	n	92	108	36	236	-0.61	0.6565	0.57	0.5355	-0.04	0.9762
	Mean	1.14	0.56	0.53							
	STD	6.92	6.16	7.05							

Figure 2(a-c): Graphical representation of mean residual ages of onset in males sorted by genotype class for SNP 558 (a), SNP 565 (b), and SNP 575 (c).

The mean residual age of onset for each genotype class (11, 12, or 22) is shown by a colored circle. Error bars represent standard deviations. The number (n) of individuals in each class is shown to the right. The number of years between each mean is shown below (brackets) with p values derived from a two-sample t test (see also Table 5).

Figure 2a

SNP 558: Comparison of Mean Residual Ages of Onset in Males

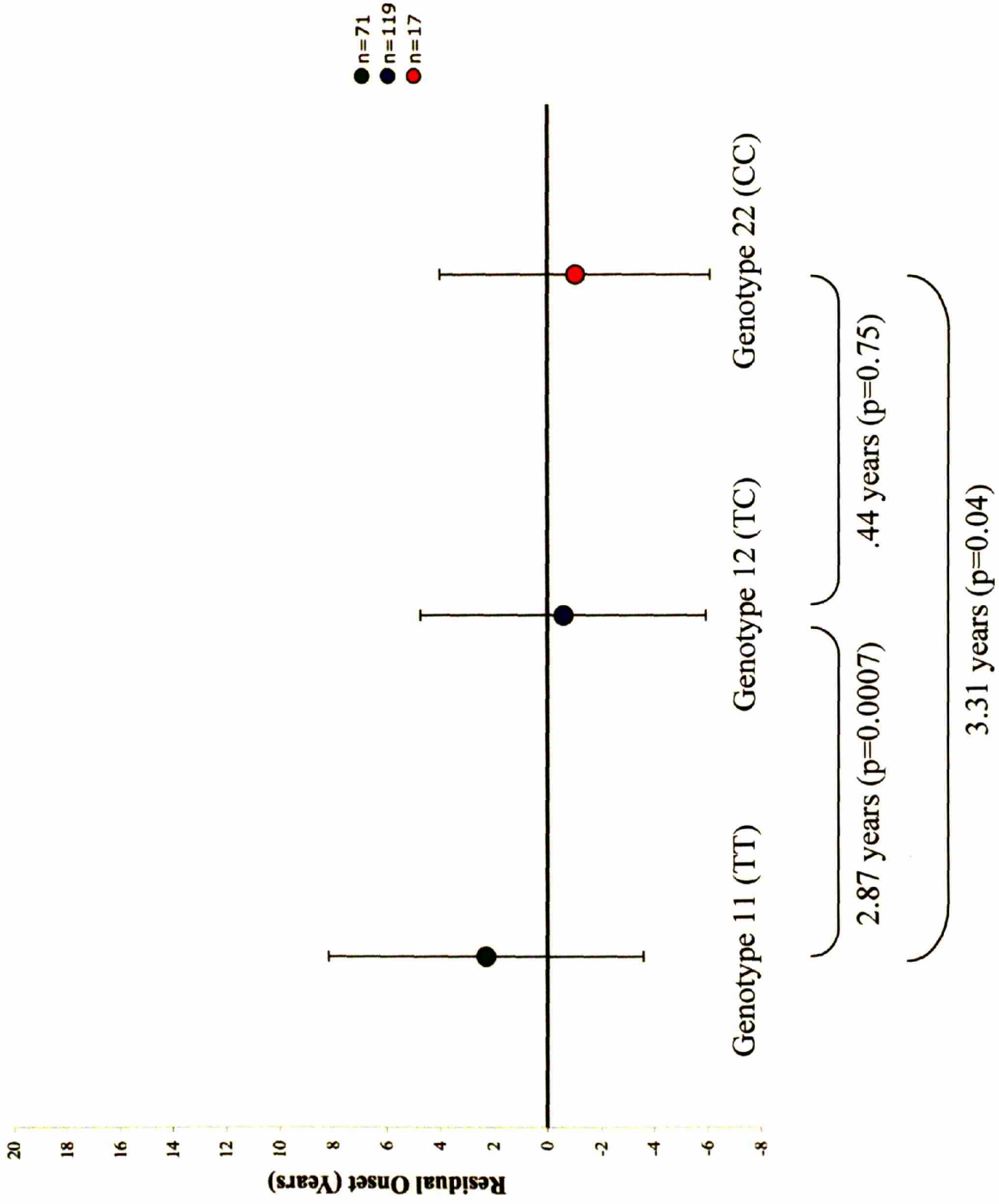


Figure 2b

SNP 565: Comparison of Mean Residual Ages of Onset in Males

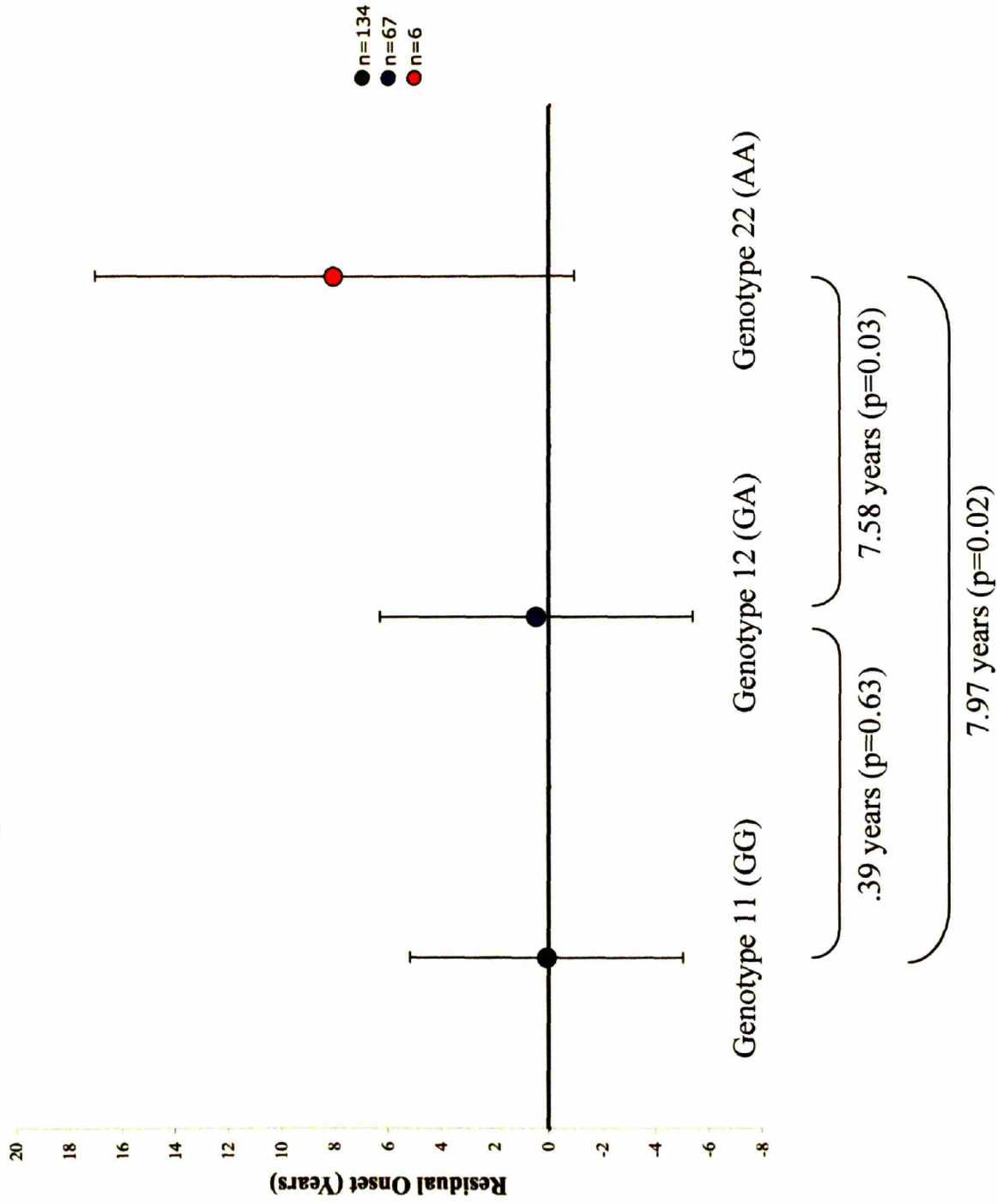


Figure 2c

SNP 575: Comparison of Mean Residual Ages of Onset in Males

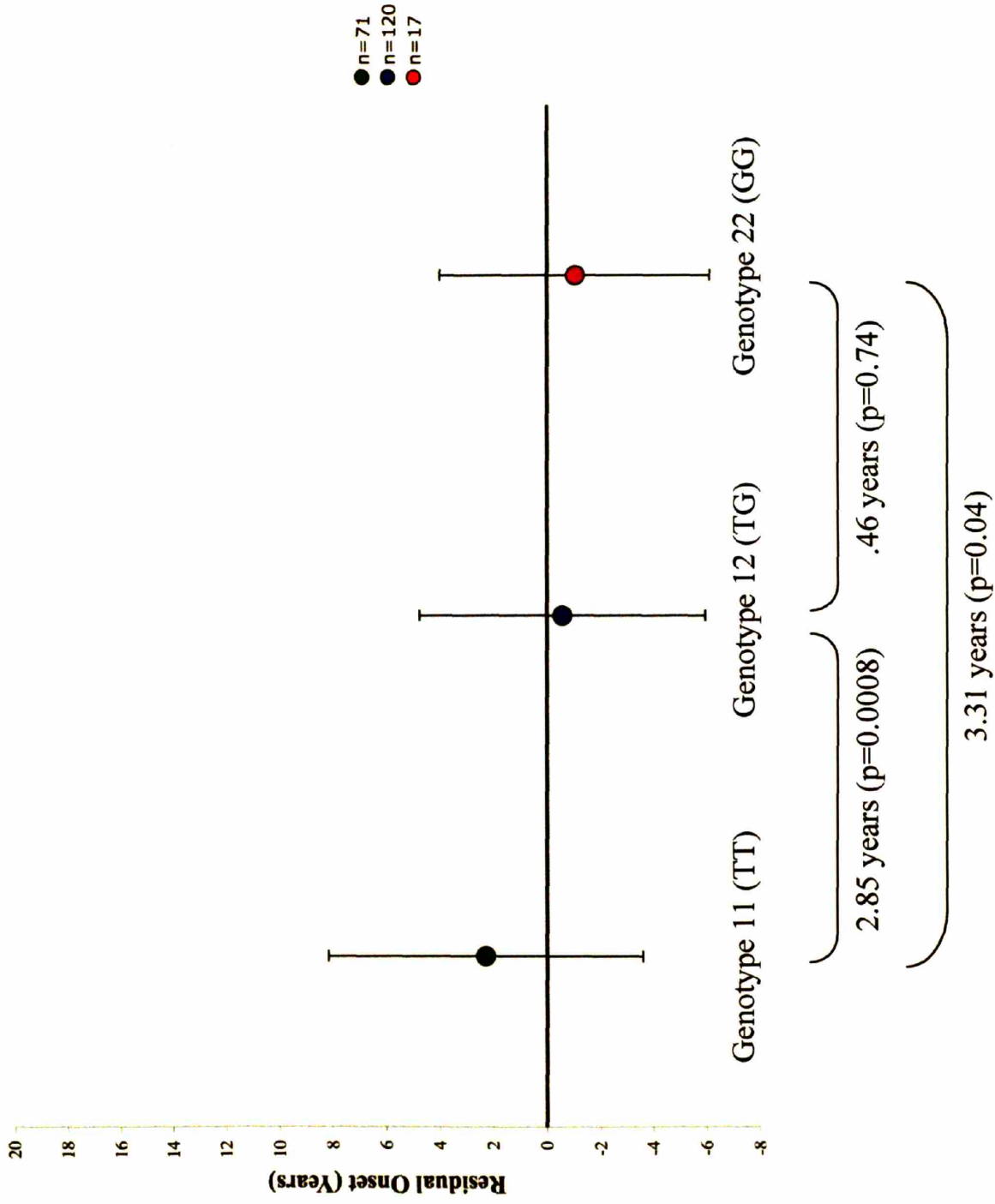


Figure 3: Representative haplotypes from the population showing haplotypes containing SNP alleles associated with a later or earlier age of onset.

204 haplotypes were resolved from a selection of 102 chromosomes in the population. A representation of each type of haplotype is shown. SNPs showing significant association with male-specific age of onset are highlighted (arrows). The “favorable” alleles for each of these SNPs producing a later age of onset in males can be found together on a single haplotype as shown found in 10 of the 204 haplotypes studied. The “unfavorable” alleles leading to an earlier age of onset in males is also found together on a single haplotype as shown and was observed in 41 of the 204 haplotypes examined.

Figure 3

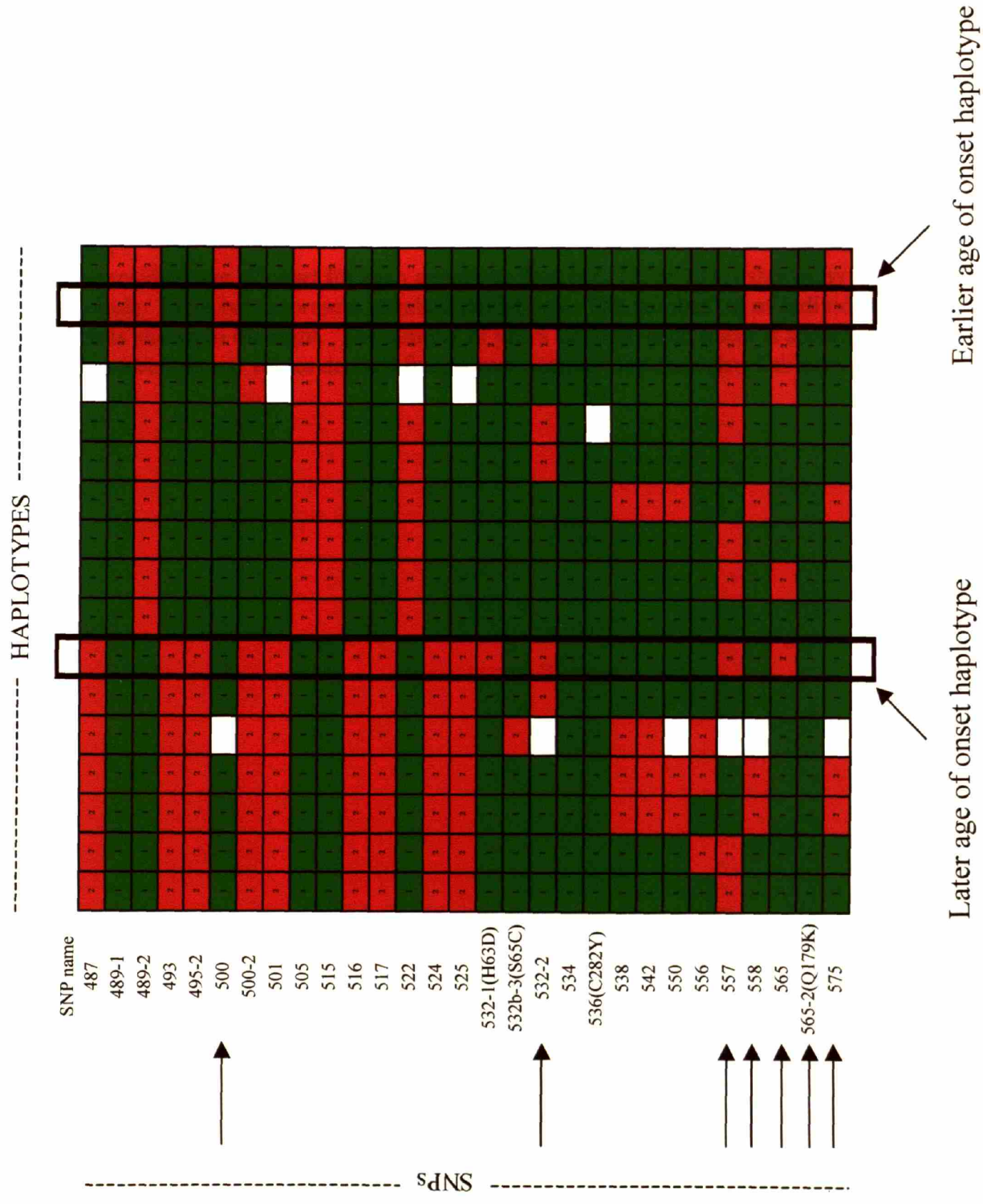


Figure 4: Graphical representation of mean residual ages of onset for SNP 565-2(Q178K).

The mean residual age of onset for each genotype class (11, 12, or 22) is shown by a colored circle. Error bars represent standard deviations. The number (n) of individuals in each class is shown to the right. The number of years between each mean is shown below (brackets) with p values derived from a two-sample t test (see also Table 5).

Figure 4

SNP 565-2(Q178K) : Comparison of Mean Residual Ages of Onset in Males

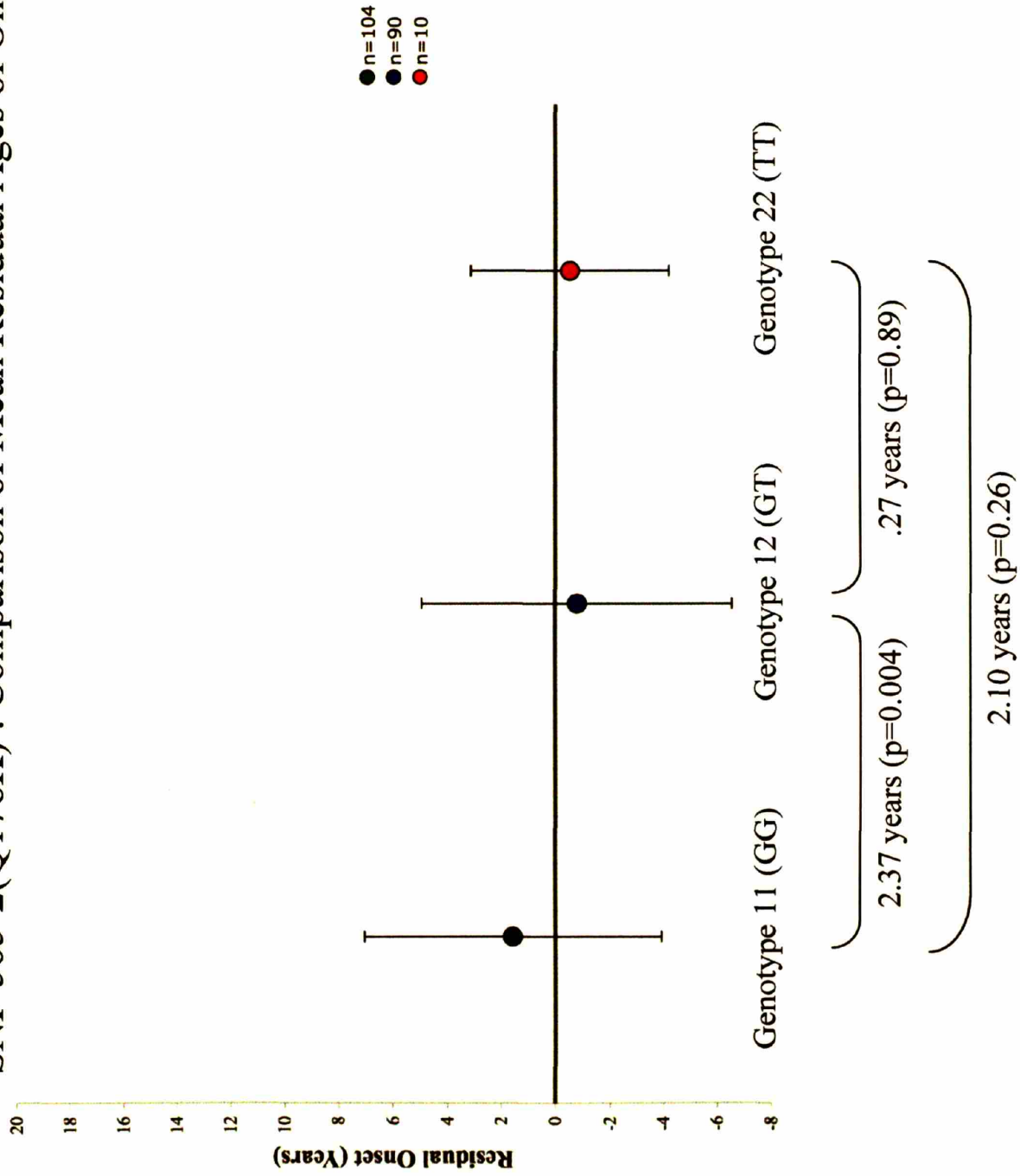


Figure 5: A single nucleotide change at SNP 565-2 affects an amino acid which is conserved in other species.

a: The SNP alleles at SNP 565-2 result in a non-synonymous amino acid change.

A single nucleotide change at SNP 565-2 to the alternate allele (A) changes a glutamine to a lysine in the resulting H1t protein. Codon 179 is shown in green, with the position of SNP 565-2 highlighted in red. Note that the nucleotides shown correspond to the (-) strand of chromosome 6p21.3

b: The reference allele of SNP 565-2 is conserved in other organisms.

The reference allele of SNP 565-2 (C) is shown conserved in chimp, mouse, rat, dog, chicken, and zebrafish sequence. Note that the sequence in chimp, mouse, and rat all result in a conserved glutamine codon at that site.

Figure 5

SNP 565-2

a

AAG GGT AAG CAA CAG CAG AAG AGC CCA GTG
 K G K Q Q K S P V

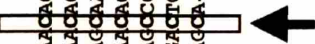
→

AAG GGT AAG CAA **AAG** CAG AAG AGC CCA GTG
 K G K Q **K** Q K S P V

b

- strand

HUMAN AGAGCGACAACTCCTAATAAATGTTAGGAGCGGGAGAAAGGCTAAAGGAGCCCAAGGGTAAGCAACAGCAGAAAGAGAGCCAGTGAAGGGCTTCGAAAGTCAAAATTTGACCCCAACATCATGAAGT
 CHIMP AGAGCGACAGCTCCTAATAAATGTTAGGAGCGGGAGAAAGGCTAAAGGAGCCCAAGGGTAAGCAACAGCAGAAAGAGAGCCAGTGAAGGGCTTCGAAAGTCAAAATTTGACCCCAACATCATGAAGT
 MOUSE AAGGCTACGCCCCACAAAAGCTTCTGGGAGCGGAAAGGAGACCCAAAGGGCCCAAGGGCTGCACACACGTAAAGAGCCCGCCAAAGCCAGGGCTTCGAAGGCCAAAATGGTCATGCA----GAAGC
 RAT AAAGCTACGCCCCACAAAAGGTTCTGGGAGCGAAGAGAACCCAAAGGGCCCAAGGGCTTGCACACAGCGCAAAAGCCCGCCAAAGCCAGGGCTTCGAAAGTCAAAATGGTCATGCA----GAAGC
 DOG GGGGGGCAGCGCACAGACTGCTTGTAGTGGCAGGAAGGCCAAAGGGCCCAAGAGCCCAAGACACAGCCCGGAAAGAGCCCC----AGGCAAGGCCCCGAAAACCCCAAGGGCCCGCCAGCA----GAAAGT
 CHICKEN GCGGCTGCCGCCACCACAAGAGCGCCCAAGAGCCCAAGAGGCTGCCAAGCCCAAGGGCTGTCG-CAAAAAGCCCGGGCCAAAGGCAAGGCGCCCAAGGGCCCGCCAAAACCCCAAGCCCAAGGCAGC
 ZEBRAFISH GCAGCTG-----CCAAGACGACCAAGA-AGGCAAGAAACCAGAGCTGCTAAGAA--AGCAGCA-----AAGAGCCCAAGAGAGGTTGA-----GAAGCCCAAAACGGCCCAACCTAAGCGCGG



Adapted from UCSC genome browser comparative genomics: <http://genome.ucsc.edu>

Chapter 4

Conclusions and Prospects for future work

Haplotype analysis and recombination events in the HFE locus

Local sites of crossover recombination in HFE

We provide evidence for a local recombination hotspot in the HFE gene, occurring in a cluster with the most frequent crossover activity evident between exons 1 and 2 in the HFE gene. While not an active site for crossing over relative to a genome-wide comparison of hotspots, we demonstrate that historical crossovers at these sites can explain the blocks of haplotype observed in this region. As expected, this site for crossing over forms the boundaries of haplotype blocks.

It is interesting to note that two of the mutations that can lead to hemochromatosis, H63D and S65C (represented by SNP 532-1(H63D) and SNP 532b-3(S65C) in our study), are located in the immediate vicinity of the hotspot we identified. Another non-CpG mutation, Q127H that has been reported in South African patients with iron overload (de Villiers et al., 1999), is also located within 1 kb of this region. At least one report has suggested increased nucleotide diversity at RHS (Jeffreys et al., 2000). Double strand breaks initiate recombination at a hotspot, so it is interesting to consider the possibility that these breaks and subsequent repair can promote a greater number of mutations in that region.

Gene conversion events in the HFE locus

We provide evidence for gene conversion events that are not limited to recombination hotspots using haplotype analysis in the human HFE locus. To understand the significance of this mechanism in influencing haplotype structure, we measured the

frequency with which these events occurred in a single generation using progeny of mouse backcrosses derived from genetically divergent strains. Our results provide a direct demonstration of gene conversion events resulting from mammalian female meioses. We detected two confirmed conversion events in 23,753 genotypes which gives a frequency even higher than our estimated gene conversion frequency of approximately 1 in 10^6 per site per generation derived from human haplotype analysis. These events were not limited to a hotspot for crossing over.

Our observation that gene conversion events are not limited to recombination hotspots has important consequences for our general understanding of the mechanism of gene conversion. Crossovers and gene conversions have long been thought to result from the alternate resolution of a single late intermediate (Szostak et al., 1983). Such a model predicts that these events occur at shared hotspots. In yeast, however, gene conversions can arise with crossover events, or can arise independent of crossovers (at sites of “noncrossovers”). A recent model proposes that these two pathways diverge shortly after a common initiating double strand break (Borner et al., 2004). By extrapolation, this supports the possibility that gene conversion events are not limited to locations where crossovers take place, i.e. RHS. If this model is applicable to mammalian meiotic events, then we could imagine punctate areas of the genome vulnerable to double strand breaks, some of which could resolve either with gene conversion or crossover, while others would consistently resolve with gene conversions only. This not only supports our findings of gene conversions not limited to hotspots but is also consistent with studies measuring the relative frequencies with which gene conversions occur in relation to crossovers. One could imagine a more transient strand invasion that might be required

for a gene conversion (modeled by several models in yeast including synthesis dependent strand annealing) (Bishop and Zickler, 2004) rather than a more committed mechanism required for crossover. This model would make gene conversion events more frequent than crossover events. This is consistent with evidence that gene conversion: crossover frequencies measured in mouse and human sperm at RHS are reported at ratios of 4:1-15:1 (Jeffreys and May, 2004). Our data also supports a high frequency of gene conversion events.

In addition to contributing to our overall understanding of gene conversion events in mammals, our data also carries important implications for the haplotype-based approach to finding genes responsible for contributing to complex traits. Using haplotype data in human chromosomes, we show that gene conversion events cause short-range sequence changes that can occur within a haplotype block (i.e. not limited to recombination hotspots at the borders of these haplotype blocks). These conversion events cause punctate changes that will lead to the eventual establishment of new haplotypes as these chromosomal segments are propagated in a population.

The frequency with which gene conversion events occur leads us to suggest that gene conversion events can be an extremely useful tool as markers of specific groups of chromosomes in the course of disease association studies. A popular approach suggested by the HapMap project is to define a set of “tag SNPs” that when used in association studies will capture most of the power of a complete set of common SNPs to identify the sites of DNA sequence variation for genes contributing to human complex traits. “Tag SNPs” are a selection of SNPs on each block that captures the information about common SNPs that identify an entire block. The idea is to reduce the complexity of haplotype

blocks and decrease the genotyping of redundant SNPs in the same block. These tag SNPs are useful for identifying a haplotype block on which a particular mutation may lie. However, this approach is effective only when the variant site leading to phenotypic effect is similar in frequency to the haplotype tagging SNPs. If, on the other hand, as some have speculated, the causal allele is less common than these tag SNPs (the causal allele may have arisen in a more recent event, for example), then they could be missed by such an approach. We propose that haplotypes created by gene conversion events will occur with a frequency and distribution ideally suited for tracking sites of DNA variation responsible for phenotypic effects in complex trait association studies in cases when these alleles are less common in the population.

The occurrence of gene conversion events would lead to the subdivision of haplotype blocks which we suggest can be used to subdivide haplotype classes allowing significantly greater power to reveal association with a disease-causing allele than can be obtained with a single set of haplotype-tagging SNPs.

Figure 1 illustrates this strategy schematically. A single haplotype block (shown in red) is present on an ancestral chromosome (topmost). Descendants of this ancestral chromosome will all carry the same haplotype block (red). On one of these, a mutation (indicated by a star) will occur that contributes to a complex trait (disease). Descendants receiving this disease-causing mutation will also receive the corresponding red haplotype on which the mutation arose. A gene conversion (represented by a *) is also shown that happens close in time to the disease-causing mutation that will be inherited by red haplotypes also carrying the mutation. Tag SNPs would be represented by any marker on the haplotype identifying it as “red”. While the disease-causing allele is found only on

red chromosomes, not all red chromosomes carry the disease allele. By only using tag SNPs that are common to the red or “framework” haplotype, therefore, we are diluting our sample and reducing our power to detect the disease-causing mutation. However, sampling these tag SNPs along with the gene conversion directly, would allow us to select only the subset of the red chromosomes that carry the disease-causing mutation as well.

How could one most efficiently identify such haplotypes created by gene conversion? In the context of a genome wide association study, the most likely way to identify such haplotypes would be to include multiple sites within a haplotype block, with varying allele frequencies, rather than relying upon a small but redundant set of haplotype-tagging SNPs. In addition, candidate regions can be identified by functional considerations or by initial genome-wide linkage or association studies. Haplotypes that arise as a consequence of gene conversions that are associated with a disease-causing variant sequence can emerge from the sequencing of these candidate regions in affected individuals with disease phenotypes. As technology moves towards the relatively lower cost sequencing of whole genomes, the potential for directly identifying high power haplotypes created by gene conversion for association studies should be considered an integral part of the analytical approach to the identification of disease causing genetic variants.

Association study: HD residual age of onset and the HFE region

We demonstrated evidence for association with residual age of onset in HD with a 43.6 kb region of 6p21.3 and have identified two major candidate genes in this region. Seven SNPs that we tested show significant association that is most robust in a male-specific context. Biological experiments will best address the physiological relevance of the two genes in relation to HD pathogenesis.

The first candidate gene we identified is the HFE gene, whose protein product is involved with iron homeostasis. A series of seven most significantly associated SNPs found in the locus we tested flanked the HFE gene. In addition, the “favorable” alleles of these seven SNPs could be found on a single haplotype in the population we examined. This haplotype with alleles associated with a later age of onset of HD also contained the H63D mutation of HFE. While the H63D mutation itself (tested as SNP 532-1) did not show statistical significance with age of onset, only five homozygotes with this mutation in the population decreased our power to detect an association considerably. The high prevalence of H63D throughout the world has led to much speculation regarding a potential protective role of this mutation. It is intriguing to consider that this mutation may affect iron handling in the aging brain.

To begin to address the physiological issue of iron homeostasis and HD pathogenesis, we are conducting a series of experiments using the R6/2 mouse model of HD. We have nearly completed an investigation into the clinical usefulness of an iron/copper chelator, clioquinol, a lipophilic molecule able to cross the blood brain barrier and investigated for its clinical usefulness in other neurodegenerative diseases

including Alzheimer's disease (in which a Phase II human clinical trial is currently in progress) (Cherny et al., 2001; Finefrock et al., 2003) and Parkinson's disease (in which clioquinol has been shown to mitigate the motor defects in an MPTP-induced Parkinson's mouse model) (Kaur et al., 2003). Thus far, we have not seen significant effects on age of onset in these animals (assessed by motor phenotypes), although a current report does suggest that this drug may ameliorate R6/2 motor phenotypes (Nguyen et al., 2005). Measurements of CAG repeat length in our animals should help to address the disparity of these results. We are also approaching this question by investigating the effects of crossing the R6/2 model to transgenic animals with alterations in genes involved with iron homeostasis to assess whether these disruptions can affect age of onset of the HD phenotype.

The second candidate gene we identified is the histone H1t gene, a testis-specific gene reported to be expressed during spermatogenesis (Drabent et al., 1991). A single variant in this gene, Q178K, corresponding to SNP 565-2, leads to an earlier age of onset in individuals carrying this mutation. Based on the fact that the largest effect is seen in heterozygotes carrying this mutation (rather than a gradient in which heterozygotes show an intermediate phenotype), we suggest that the Q178K mutation leads to a dominant negative effect on phenotype. To address how this mutation could lead to earlier age of onset we considered several possibilities. The first possibility is that the Q178K mutation directly effects expansion of the CAG repeat in spermatocytes. This would corroborate our findings of a male-specific association as well as reports of increased CAG expansion occurring through paternal transmission seen both in human patients (Bates et al., 2002), and in HD mouse models (Mangiarini et al., 1997). Our tests for association, however,

looked specifically at male patients rather than their progeny. If this were the reason for the association we detected, then it might be due to the fact that children of male parents carrying the mutation might also be expected to inherit the mutation. A more direct test to see if this mutation leads to greater CAG expansion in the next generation would be to compare the progeny of males with the Q178K mutation with the progeny of males without the mutation. We would expect, in this case, to see a greater mean CAG repeat length in children of males carrying the mutation. We performed this statistical test on our cohort, and did not observe a significant difference in mean CAG length of children of males carrying the Q178K mutation. Our sample size was considerably reduced, however, and likely too small to detect an effect with adequate statistical significance.

Other reports, however, such as studies in one HD mouse model suggest that sex-dependent factors in the embryo itself may also result in more CAG expansion in male mice vs. female mice (Kovtun et al., 2000). This would be consistent with our male-specific association, although challenging to explain if the mutation is only expressed in the testis during spermatogenesis. In addition, general transcriptional misregulation has often been associated with HD pathogenesis and for this reason histone deacetylase (HDAC) inhibitors have been the focus of several preclinical investigations in HD (Hockly et al., 2003). Again, for a mutation in histone H1t to lead to general transcriptional misregulation, it will be necessary to assess whether its expression is in fact limited to the testis or not. To address this issue, we are examining expression of H1t specifically focusing on the brain by Northern analysis. While previous reports suggest expression largely restricted to the testis, microarray analyses suggest the potential for expression in other tissues (<http://symatlas.gnf.org/SymAtlas/>, Probe set id:

207982_at). Brain expression would offer the possibility that the histone gene in some way influences CAG expansion in males in the somatic cells which are most affected by HD (albeit by a mechanism still poorly understood), leading to earlier age of onset as seen with the histone H1t variant Q178K. HD pathogenesis is still an area under intense investigation. Understanding the nature of the impact of HFE and/or histone H1t on age of onset will shed light on disease progression and provide new options for therapeutic interventions to delay age of onset of this devastating disease.

REFERENCES

- Bates, G., Harper, P., and Jones, L. (2002). *Huntington's disease*, Third edn (New York: Oxford University Press).
- Bishop, D. K., and Zickler, D. (2004). Early decision; meiotic crossover interference prior to stable strand exchange and synapsis. *Cell* *117*, 9-15.
- Borner, G. V., Kleckner, N., and Hunter, N. (2004). Crossover/noncrossover differentiation, synaptonemal complex formation, and regulatory surveillance at the leptotene/zygotene transition of meiosis. *Cell* *117*, 29-45.
- Cherny, R. A., Atwood, C. S., Xilinas, M. E., Gray, D. N., Jones, W. D., McLean, C. A., Barnham, K. J., Volitakis, I., Fraser, F. W., Kim, Y., *et al.* (2001). Treatment with a copper-zinc chelator markedly and rapidly inhibits beta-amyloid accumulation in Alzheimer's disease transgenic mice. *Neuron* *30*, 665-676.
- de Villiers, J. N., Hillermann, R., Loubser, L., and Kotze, M. J. (1999). Spectrum of mutations in the HFE gene implicated in haemochromatosis and porphyria. *Hum Mol Genet* *8*, 1517-1522.
- Drabent, B., Kardalidou, E., and Doenecke, D. (1991). Structure and expression of the human gene encoding testicular H1 histone (H1t). *Gene* *103*, 263-268.
- Finefrock, A. E., Bush, A. I., and Doraiswamy, P. M. (2003). Current status of metals as therapeutic targets in Alzheimer's disease. *J Am Geriatr Soc* *51*, 1143-1148.
- Hockly, E., Richon, V. M., Woodman, B., Smith, D. L., Zhou, X., Rosa, E., Sathasivam, K., Ghazi-Noori, S., Mahal, A., Lowden, P. A., *et al.* (2003). Suberoylanilide hydroxamic acid, a histone deacetylase inhibitor, ameliorates motor deficits in a mouse model of Huntington's disease. *Proc Natl Acad Sci U S A* *100*, 2041-2046.
- Jeffreys, A. J., and May, C. A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet* *36*, 151-156.
- Jeffreys, A. J., Ritchie, A., and Neumann, R. (2000). High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum Mol Genet* *9*, 725-733.
- Kaur, D., Yantiri, F., Rajagopalan, S., Kumar, J., Mo, J. Q., Boonplueang, R., Viswanath, V., Jacobs, R., Yang, L., Beal, M. F., *et al.* (2003). Genetic or pharmacological iron chelation prevents MPTP-induced neurotoxicity in vivo: a novel therapy for Parkinson's disease. *Neuron* *37*, 899-909.
- Kovtun, I. V., Therneau, T. M., and McMurray, C. T. (2000). Gender of the embryo contributes to CAG instability in transgenic mice containing a Huntington's disease gene. *Hum Mol Genet* *9*, 2767-2775.

Figure 1: Schematic to illustrate a strategy for subdividing haplotypes using gene conversion events.

A haplotype block (shown in red) that lies on an ancestral chromosome is illustrated. If a mutation (star) occurs on a descendent of this chromosome on the red haplotype, all chromosomes with the disease allele will also have a portion of the red haplotype. Sampling a population with tag SNPs alone will identify all “red” haplotypes, only some of which contain the disease-causing allele. Gene conversion events that happen close in time to the mutation (*) could be useful tools for the subdivision of haplotypes. Sampling for the red haplotype and the gene conversion will enrich a sample for mutation-containing haplotypes.

Figure 1

