Patterns of genetic recombination and variation in the human genome

Inaugural – Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch- Naturwissenschaftlichen Fakultät

der Universität zu Köln

Vorgelegt von

Ekaterina Shabanova

aus Astrachan, Russland

Köln 2009

Berichterstatter:	Prof. Dr. Thomas Wiehe
	Prof. Dr. Jonathan Howard

Tag der muendlichen Pruefung: 20 October 2009

Abstract

Genetic recombination plays an important role in shaping genome variation. It enhances haplotype diversity, helps to maintain genome integrity and ensures the proper segregation of the chromosomes. While participating in DNA rearrangement, recombination is not an “independent player”. It is tightly connected and influenced by other genomic features, such as nucleotide diversity. A correlation between nucleotide diversity and recombination rate was observed in the human genome, as well as in the genomes of other organisms (Arabidopsis, Drosophila, etc). The traditional view that selection has contributed to shape this pattern was questioned by the view that it may solely be due to a mutagenic effect. Extensive analysis of the data from dbSNP and broad scale recombination maps revealed a high degree of uncertainty in the inferred correlation coefficients. One goal of this study was to re-assess the magnitude and disentangle the possible reasons for this observation. The results show that there is no evidence for a presence of strong correlation between nucleotide diversity and recombination rate. In fact, the observed effect can be due to insufficient data or poor data quality in earlier studies. Analysis of the more recent data shows that the correlation between diversity and recombination may be due to sequence composition, such as sequence composition. While looking at the fine scale it became clear that recombination hotspots are very ephemeral structures, which do not have influence on long-term molecular evolution. Only detailed experimental studies can reveal, whether the apparent correlation between diversity and recombination rate has a causal connection.

In order to obtain high resolution recombination rates, a single sperm typing approach was applied to 2.5 Mb sequence consisting of four Encode regions on human chromosome 11. The regions were selected not on the basis of classical linkage studies, but in unbiased fashion. The outcome revealed a cross over rate of 0.12 cM/Mb, which was much lower compared to the expected 1.875 cM/Mb according to high resolution recombination maps. We confirm an increased gene conversion rate compared to cross over. Out of 10 recombinants, 7 had conversion haplotypes. We assume that such a low cross over rate can be due to the individual or sex-specific variation. Our data did not support genetic characteristics known to be associated with gene conversion or recombination in general. For instance, identified conversion tracts did not have a high GC content. No association between converted or cross over regions and a 13-mer degenerate motif (predicted to be associated with recombination hotspots) was observed. The observed CpG fraction was so low compared to the expected, that the possible explanation could be a extensive methylation of the converted regions. Finally, we hypothesize, that there is a difference in regulation mechanism as well as in frequency of double-strand breaks resolved as conversion events in short and long tract conversions. The frequency of the breaks can also be a reason to differentiate hotspots-associated conversions and the ones occurring in coding regions.

Kurzzusammenfassung

Genetische Rekombination spielt eine tragende Rolle in der Gestaltung der Genom-Variation. Sie fördert die Haplotyp-Diversität, unterstützt die Aufrechterhaltung der Genom-Integrität und stellt eine korrekte Segregation der Chromosomen sicher. Trotz ihrer tragenden Rolle in der Reorganisation von DNA-Sequenzen, ist Rekombination keine unabhängige evolutionäre Kraft. Sie steht im Wechselspiel mit anderen genomischen Merkmalen, und wird zum Beispiel von der Nukleotid-Diversität beeinflusst. Die Korrelation zwischen Nukleotid-Diversität und Rekombinationsrate wurde in den letzten Jahren im menschlichen Genom, aber auch in den Genomen anderer Spezies, wie zum Beispiel Arabidopsis oder Drosophila, beobachtet. Die traditionelle Sichtweise, dass natürliche Selektion diese Korrelation mitbeinflusst hat, wird durch die Ansicht, dass dies alleinig auf den „mutagenischen Effekt“ zurückzuführen sei, in Frage gestellt. Eine umfangreiche Analyse von dbSNP-Daten und grob-skalierte Rekombinationskarten offenbarten eine grosse Unsicherheit in der Richtigkeit der gewonnenen Korrelationskoeffizienten. Ein Ziel dieser Arbeit war es, die Stärke der Korrelation neu abzuschätzen und mögliche Gründe für diese Beobachtung aufzudecken. Die Resultate der Studie konnten keine starke Korrelation nachweisen. Der beobachtete Effekt kann durch unzureichende Daten oder die niedrige Datenqualität in früheren Studien erklärt werden. Die Analyse neuerer Datensätze zeigt, dass die Korrelation zwischen Diversität und Rekombination eventuell auf die DNA-Sequenz-Zusammensetzung zurückzuführen ist. In Fein-Skala Untersuchungen zeigte sich außerdem, dass Rekombinationshotspots sehr kurzlebige Erscheinungen sind, und deshalb keinen Einfluss auf die langfristige molekulare Evolution haben. Nur durch detaillierte experimentelle Studien kann diese Korrelationsfrage besser ergründet werden.

Um Rekombinationsraten in einer hohen Auflösung zu bestimmen wurde der Spermien-Typisierungs-Ansatz auf Sequenzen einer Gesamtlänge von 2.5Mb bestehend aus vier Encode-Regionen des menschlichen Chromosoms 11 angewendet. Die Regionen wurden zufällig ausgewählt ohne auf klassische Linkagestudien zurückzugreifen. Als Ergebnis zeigt sich eine Crossing-Over-Rate von 0.12 cM/Mb, die im Vergleich zu den erwarteten 1.875 cM/Mb sehr niedrig ist. Es konnte eine im Vergleich zur Cross-Over Rate erhöhte Genkonversionsrate bestätigt werden. Von zehn Rekombinanten weisen sieben Konversionshaplotypen auf. Die im Ganzen niedrige Crossing-Over-Rate ist möglicherweise auf individuelle oder sex-spezifische

Variation zurückzuführen. Unsere Daten konnten genetische Merkmale, die im Allgemeinen mit Genkonversion oder Rekombination assoziiert werden, nicht bestätigen; zum Beispiel zeigten die identifizierten Genkonversions-Regionen keinen hohen GC-Anteil auf. Zudem konnte keine Assoziation zwischen Genkonversions- beziehungsweise Rekombinations-Regionen mit einem degenerierten 13-mer-Motiv, dass für Rekombinationshotspots vorhergesagt wurde, bestätigt werden. Der beobachtete CpG-Anteil war im Vergleich zur Erwartung so gering, dass die Möglichkeit einer beträchtlichen Methylation der konvertierten Regionen in Frage kommt. Zum Abschluss stellen wir die Hypothese auf, dass es einen Unterschied in den Regulationsmechanismen und in der Häufigkeit der Doppelstrangbrüche gibt, der sich als Konversions-Ereigniss in kurzen und langen Konversionen widerspiegelt. Die Frequenz der Doppelstrangbrüche kann eine Möglichkeit darstellen, hotspot-assoziierte Konversionen von jenen in kodierenden Regionen zu unterscheiden.

Table of Contents

Abstract………………………………………………………………………………………….…1

Kurzzusammenfassung……………………………………………………………………….3

Table of Contents…………………………………………………………………………..…5

1. Introduction…………………………………………………………………………………8

1.1 Processes influencing variation of the human genome………………………………..8

1.2 Genetic recombination, forms and definition………………………………………….....9

1.3 Cross over………………………………………………………………………….........9

1.4 Importance of meiotic recombination………………………………………………...10

1.5 Mechanism of the homologous recombination…………………………….……10

1.6 Recombination rate variation and hotspots………………………………...............13

1.7 Hotspot detection…………………………………………………………………..13

1.7.1 Linkage maps based on pedigree data……………………………………...14

1.7.2 Linkage Disequilibrium in recombination mapping……………………......14

1.7.3 Sperm typing analyses……………………………………………………...15

1.8 Hotspots evolution……………………………………………………………….....15

1.9 Meiotic gene conversion…………………………………………………….....…16

1.10 Association of recombination with other genomic features……………….…..16

1.11 Technological limitations and future perspectives of the recombination analysis…………………………………………………………………………….…….18

1.2 The aims of the study……………………………………………………………...18

Table of contents

1.3 Experimental regions of interest……………………………………………………….……..19

2. Materials and Methods…………………………………………………………………20

2.1 Material (sperm collection)………………………………………………………..20

2.2 Experimental methods……………………………………………………….….20

2.3 Theoretical methods…………………………………………………………….…30

3. Results…………………………………………………………………………….…...31

3.1 Recombination and variability in public data……..………………………………..31

3.2 Specificity of sperm sorting on FACSVantage SE Cell Sorter………………………38

3.3 Estimation of the cross over rate……………………………………………………..40

3.4 Annotation of the ENCODE regions………………………………………….…...42

3.4.1 Chromosome 11 at a glance…………………………………….….…..42

3.4.2 ENm009……………………………………………………………....45

3.4.3 ENr332………………………………………………..…………47

3.4.4 ENm003……………………………………………………………….....48

3.4.5 ENr312……………………………………………………………..…51

3.5 Haplotype screen…………………………………………………………….….52

3.5 .1 “Reference” haplotypes……………………………………………….…......54

3.5.2 Recombinant haplotypes………………………………………………........54

3.5.3 Cross over and gene conversion events within and near ENm 009………...55

3.5.4 Conversion events in ENmr332…………………………………………….…60

3.5.5 Conversion tracts in ENm003………………………………………………...65

3.5.6 Cross over ENr312……………………………………………………………67

4. Discussion…………………………………………………………………………..…70

4.1 Analysis of correlation between SNP density and recombination rate in the human genome……………………………………………………………………………………...70

4.2 Analysis of experimental cross over data…………………………………………71

4.3 Analysis of gene conversion events………………………………………………73

Summary of the results…………………………………………………………77

Implication of the study……………………………………………………….....78

Outlook………………………………………………………………………..…78

Reference List…………………………………………………………………………..…79

Appendix……………………………………………………………………………………86

Acknowledgements…………………………………………………………………………96

1. Introduction

1.1 Processes influencing variation of the human genome

Genetic variability, in a way, measures how much the trait or the genotype will tend to vary in a population (Burt, 2000). The presence of variability in a population is an important factor in a number of fundamental biological processes, such as shaping biodiversity, evolutionary development of a population and susceptibility to diseases (Wills and Christopher, 1980). The role of variability for biodiversity can be seen in successful adaptation to environmental conditions and, thus, contributes to success in avoiding extinction. Variability also plays a crucial role in evolution: it stimulates individual responses to environmental stresses and can lead to survival of the fittest variants because of natural selection. Genetic variability has become one of the factors contributing to an increased interest in “personalized medicine” since it underlies differential susceptibility to diseases and sensitivity to toxins or drugs (Burt, 2000).

There are a few known sources of genetic variability in a population. One of them is genetic recombination. Recombination is also variable in frequency and location and thus, it can be selected for to increase fitness, because more recombination leads to more variability; this increased genetic variability makes it easier for a particular population to handle changes (Burt, 2000). Genetic mutations are the main source of genetic variability within a population. Mutations can have positive, negative or neutral effects on fitness (Wills and Christopher, 1980). Natural selection can propagate mutation-induced variability through the population if the mutation is beneficial or its effect will be hidden if the mutation is deleterious. The excess of deleterious mutations is typically observed in smaller populations with reduced variability levels (Wills and Christopher, 1980).

Other sources contributing directly or indirectly to genetic variability include immigration, emigration and translocation. When an individual moves in or out of a population, genetic variability in the next generation will increase if it reproduces (Linhart et al. 2003). Polyploidy in sexual organisms provides yet another source; it allows more recombination during meiosis and, therefore, leads to more genetic variability in the offspring.

In asexual organisms with limited sources of variability, diffused centromeres can become one, since they allow the chromatids to split apart in variety of ways and assists chromosome fragmentation and polyploidy to create more variability (Linhart et al. 2003).

In this study, specific attention will be given to genetic recombination as a source of genetic variability.

1.2 Genetic recombination, forms and definition

Genetic recombination is the transmission-genetic process by which chromosomes are broken and then rejoined to form a new genetic combination, different from the original (Brooker, 1999). A major type of genetic recombination is homologous recombination, which is an essential feature of all sexual organisms. It occurs between DNA segments homologous to each other. This process enhances genetic diversity, helps to maintain genome integrity (when genome is capable of multiplication, variation and heredity) and ensures the proper segregation of chromosomes. Another type is site-specific recombination, where non-homologous segments are recombined at the specific sites. This type of recombination happens within genes that encode antibody polypeptides and also occurs when certain viruses integrate their genomes into host cell DNA. The third type of recombination is known as transposition. Small segments of DNA called transposons can move themselves to multiple locations within the host’s chromosomal DNA (Brooker, 1999).

It should be noted that a number of other genetic exchanges do not fall into any of the above classes, and are named illegitimate recombination or recombination without homology (Brooker, 1999).

1.3 Cross over

Chromosomal crossover (or crossing over) is the exchange of genetic material between homologous chromosomes during prophase I of meiosis, in a process called synapsis. Synapsis begins before the synaptonemal complex develops, and is not completed until near the end of prophase I. Crossover usually occurs when matching regions on matching chromosomes break and then reconnect to the other chromosome. The result of this process is an exchange of genes, called genetic recombination. Chromosomal crossovers also occur in asexual organisms and in somatic cells, since they are important in some forms of DNA repair (Li and Heyer, 2008).

1.4 Importance of meiotic recombination

In most eukaryotes, a cell carries two copies of each gene, each referred to as an allele. Each parent passes on one allele to each offspring. An individual gamete inherits a complete haploid complement of alleles on chromosomes that are independently selected from each pair of chromatids lined up on the metaphase plate. Without recombination, all alleles for those genes linked together on the same chromosome would be inherited together. Meiotic recombination allows a more independent selection between the two alleles that occupy the positions of single genes, as recombination shuffles the allele content between homologous chromosomes (Brooker, 1999).

Meiotic recombination is one of the fundamental biological mechanisms leading to exchange of genetic material between homologous chromosomes. During meiosis, this process is also associated with a few important cellular functions such as the formation of the synaptonemal complex and proper chromosome segregation (Padhukasahasram et al. 2006). The role of meiotic recombination in evolutionary biology is to produce novel allelic combinations to increase genetic diversity within a population.

To understand the variation of the recombination rate is one of the goals in association mapping and evolutionary inference studies in particular, and in molecular biology in general. (Padhukasahasram et al. 2006).

1.5 Mechanism of the homologous recombination

The most famous central model of the molecular steps occurring during homologous recombination was proposed by R. Holliday in 1964. It is based on studies of *Ustilago maydis* (a simple eukaryote) and explains both cross over and gene conversion events (Brooker, 1999).

At the beginning of the process (Fig.1) two homologous chromosomes are aligned with each other. During the first step a break occurs at the identical sites in one strand of both parental chromosomes. During the second step the strands invade the opposite helices and base pair with the complementary strands. In the third step this event is followed by the covalent linkage of

DNA to create a *Holliday junction* (a structure, which incorporates four strands of DNA, two of each homologous chromosomes).

As shown in step six (Fig.1), the cross in the Holliday junction can migrate in a lateral direction. As it does so, a DNA strand in one helix is being swapped for the DNA strand in the other helix. This process is called branch migration, because the branch connecting the two double helices migrates laterally. Since the DNA sequences in the homologous chromosomes are similar but not identical, the swapping of the DNA strands during branch migration may produce regions in the double stranded DNA called heteroduplexes. A heteroduplex is a DNA double helix that contains mismatches. Their occurrence is explained by the fact that the DNA strands come from homologous chromosomes and their sequence is not perfectly complementary.

The next step (step 7 in Fig.1) is isomerization, during which the Holliday structure may make an 180° turn. The two structures existing in this step are structural isomers of each other (chemically identical except for the location of certain segments). The final steps of the recombination processes are called resolution, since they involve the breakage and rejoining of two DNA strands to create separate chromosomes. In other words, the entangled DNA strands become resolved into two separate structures. Resolution can occur in two ways (step 8, Fig.1). The first way is without isomerization (8b, Fig.1). The breakage could occur at the same strands that were broken in the first step. In this case at the resolution, the strands are rejoined to produce a non-recombinant pair of chromosomes. The second way is with isomerization, and then the resolution phase can involve breakage of the two DNA strands that were not broken before in the first step. In this case, the rejoining of the corresponding strands produces two recombinant chromosomes (step 8a, Fig.1). The original Holliday model was based on the results of observed cross over events in fungi where the products of meiosis are contained in the same single ascus (the sexual spore-bearing cell produced in ascomycete fungi). Nevertheless, molecular research in many other organisms has supported the central tenets of the Holliday model (Brooker, 1999).

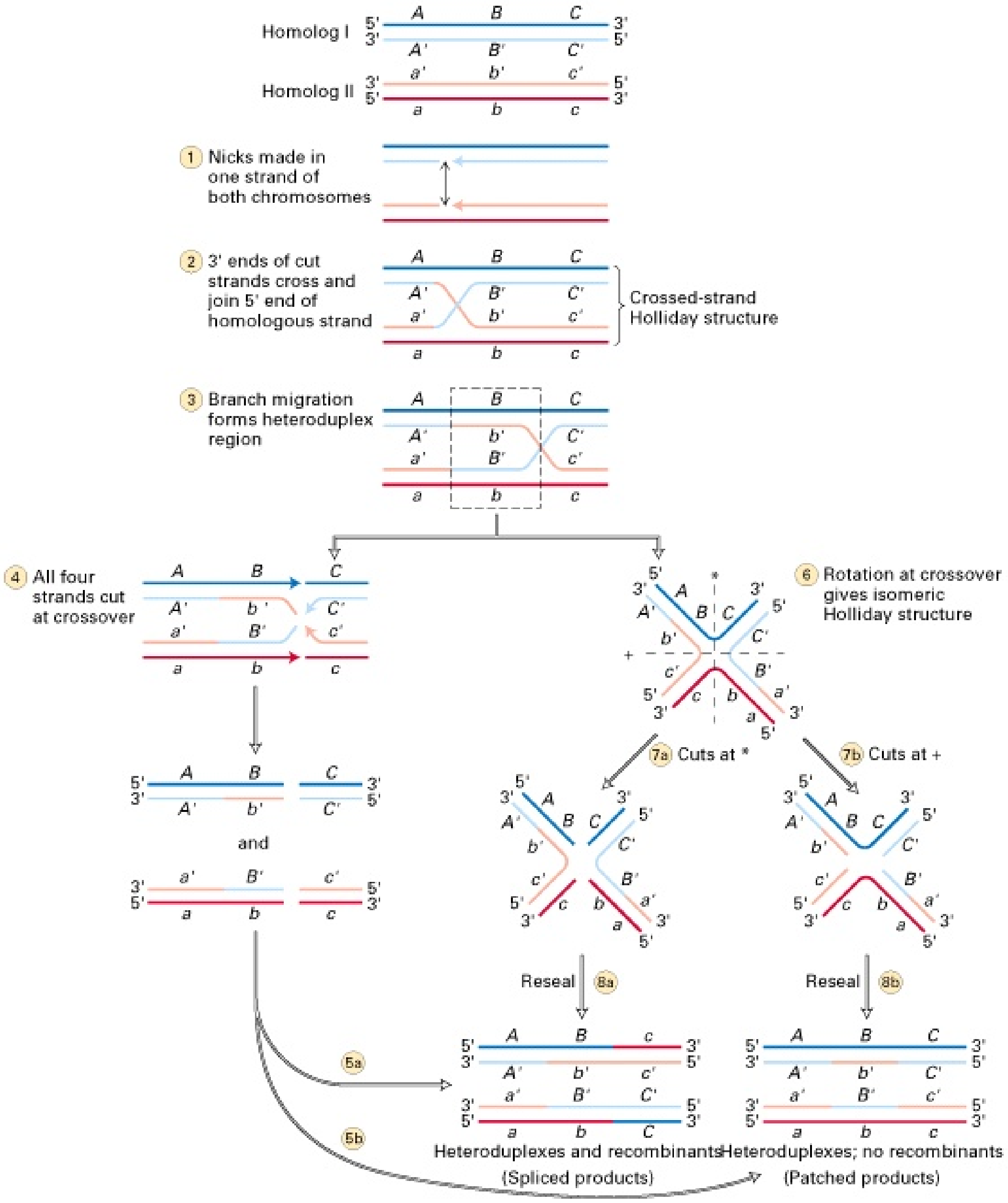

Homolog I
Homolog II
① Nicks made in one strand of both chromosomes
② 3' ends of cut strands cross and join 5' end of homologous strand
Crossed-strand Holliday structure
③ Branch migration forms heteroduplex region
④ All four strands cut at crossover
and
⑥ Rotation at crossover gives isomeric Holliday structure
7a Cuts at *
7b Cuts at +
Reseal 8a
Reseal 8b
5a
5b
Heteroduplexes and recombinants (Spliced products)
Heteroduplexes; no recombinants (Patched products)

Fig.1. **The Holliday model for the homologous recombination** from *Brooker RJ(1999);* 8a - the resolution of the Holliday complex into cross over product; 8b – the resolution of the Holliday complex into a non-recombinant heteroduplex.

1.6 Recombination rate variation and hotspots

The rate of meiotic recombination (i.e., crossing over) in humans varies on different physical scales (Lynn et al. 2004). While the broad-scale rates are relatively similar across populations and possibly across species (Serre et al. 2005, Ptak et al. 2005), on the fine scale, the situation is different: most cross over events tend to be concentrated into 1-2 kb regions named “recombination hotspots”. In addition to this, “cold spots” or regions of low recombination rate have also been noted (Petes 2001). Crossover initiation sites can be distributed randomly throughout the genome, but the rate of initiation at particular sites can be significantly higher at some genomic regions compared to others (Arnheim et al. 2003).

There is much evidence that shows variation of hotspots within a population through time (Jeffreys et al. 2005), as well as a lack of hotspot conservation between human and chimpanzees (Wall et al. 2005, Ptak et al. 2005). There are a few reasons why it is important to understand the mechanistic underpinning of hotspots (Petes 2001). One such reason is to understand the recombination initiating mechanism, which requires knowledge of factors regulating both hotspots and cold spots. Other reasons include: the number of recombination events per chromosome is relevant to the accuracy of chromosome segregation; distribution of exchanges influences the probability of assembling new configurations of physically linked genes during evolution and, finally, understanding hotspots and cold spots will assist in understanding other DNA-related processes affected by chromosome context, e.g., transcription and replication (Petes 2001).

1.7 Hotspot detection

There are a few approaches developed to infer recombination rates and map genomic hotspots. Two approaches have gained particular popularity: estimates can be obtained either indirectly through examining patterns of marker association (linkage disequilibrium), established through population dynamic processes, or directly through labor-intensive screening of millions of sperm for recombinant DNA molecules within relatively short genomic intervals (Webb et al. 2008).

1.7.1 Linkage maps based on pedigree data

Constructing a linkage map is one of the common approaches to obtain recombination rates. The cross over events are inferred from the transmission patterns of polymorphic markers in a large pedigree or in extensive crosses (Coop and Przeworski, 2007). The rates of genetic exchange between markers are converted into a linkage map by use of a function that takes into account a model of cross over interference. To obtain the cross over rates, it is necessary to compare the genetic map with the physical one. This approach is widely used to characterize variation among individuals, but, since it relies on estimates of recombination from transmitted chromosomes, this method misses half of the cross over events that occur in meiosis or those any gametes that are selected against (Coop and Przeworski, 2007).

1.7.2 Linkage Disequilibrium in recombination mapping

This approache utilizes a population genetic model to relate the patterns found in a current polymorphism sample to the historical rate of recombination (McVean et al. 2004). The estimates of historic rates encompass thousands of meioses of the sample's history, so the rate is estimated as an average of male and female rates over time (Coop 2005).

Recent studies identified considerable variation of linkage disequilibrium across the human genome (Gabriel et al. 2002; Phillips et al. 2003, Arnheim et al. 2003), which reflects variation in the underlying recombination rate (Reich et al. 2002; Wang et al. 2002a and b; Innan et al. 2003; Phillips et al. 2003). There is a certain correlation between uneven recombination rate distribution and existence of LD blocks in certain genomic areas, although not all LD blocks can be explained by recombination variation (Arnheim et al. 2003). Recombination maps available by 2002-2003 were very inaccurate for intervals smaller than 1 cM (Kong et al. 2002; Weber 2002). This inaccuracy complicated revealing to what extent LD variation is determined by variation in recombination. Recently, the International HapMap project has mapped the LD landscape on a kilobase level fo the entire genome. The data allowed detection of the global recombination landscape at high resolution by using coalescent analyses and observed haplotypes were explained through *in silico* reconstruction with variable recombination rates (Webb et al. 2008). These studies revealed approximately 33 thousand putative cross over hotspots, their distribution and population-specific activity, as well as provided data on motifs possibly associated with them (Altshuler et.al., 2005).

1.7.3 Sperm typing analyses

A sperm-typing approach allows direct observation of cross over events in male individuals and produces a uniquely detailed picture of the current recombination rate at a finer-scale (Coop 2005). Sperm typing enabled for the first time to measure human recombination fractions in a single individual (male), because the large number of sperm available from a single donor permits a high level of accuracy to be achieved (Arnheim et al. 2003). In contrast to LD hotspot mapping, few human recombination hotspots have been directly characterized in sperm and the obtained data does not completely explain whether linkage disequilibrium data predicts genuine hotspots accurately, or correctly determines their historical activity (Webb et al. 2008). Due to Webb *et al* data (2008) up to date sperm typing covered approximately 0.6 Mb of the human genome. One major goal was the identification of hotspots with special focus on the MHC region, where approximately seven hotspots were characterized (Jeffreys et al. 2000, Jeffreys and Neumann 2002, Kauppi et al. 2005). In contrast, the entire chromosome 1 contains only eight hotspots (Jeffreys et al. 2000, Jeffreys et al. 2005). These studies revealed additional phenomena such as variation in hotspot activity between different male individuals, as well as polymorphic sites driving activity of the hotspot (Neumann, Jeffreys 2006). In most cases, hotspots obtained by sperm typing and linkage disequilibrium hotspots are in good concordance (Webb et al. 2008).

1.8 Hotspots evolution

The availability of high-resolution approaches mentioned above allowed the characterization of hotspots as ephemeral structures that are quick to evolve within a single population and even between individuals of the same population. There is experimental evidence for differences in recombination activity in the *CAP10* genomic region between African ancestral and European and Asian populations (Clark et al. 2007). Sperm typing showed different intensities in the MSTM 1a hotspot in 26 male individuals (Neumann and Jeffreys 2006). The difference between female and male rates should also be noted (Coop and Przeworski, 2007). One of the possible explanations for a hotspot's quick development could be the presence of a certain motif, polymorphic site or other "control element" (modifier) that can be associated with the hotspot. The loss of such an element (e.g., during DNA repair) can potentially result in further decrease of recombination activity within the region.

1.9 Meiotic gene conversion

Homologous gene conversion is considered as a poorly characterized form of recombination (Gay et al. 2007), a non-reciprocal process acting on short lengths of DNA, where genetic material from one parental chromosome is incorporated into an alternate chromosome during meiotic exchange (Szostak et al. 1983). Cross over events are believed to include gene conversion tracts as well, which can not be detected by current population-based methods. Therefore, initially the term gene conversion defined events not accompanied by cross over (Gay et al. 2007). Gene conversion in the human genome is believed to be up to 15 times as frequent as cross overs (Jeffreys and May 2004). These events are difficult to track because of the short length of DNA transferred – with an average size of approximately 350 bp – which represents a difficulty in marker selection. When the close association between cross overs and gene conversion events was revealed, some of the characterized hotspots with higher polymorphism levels were investigated for conversion activity (Jeffreys and May 2004). The results revealed the more active the hotpot was the more conversion events were observed.

1.10 Association of recombination with other genomic features

The extent to which adaptive evolution shaped the recent evolutionary history in humans is much debated (Spencer et al. 2006). However, adaptive evolution is expected to leave its footprint in the patterns of genetic variation. In regions of high recombination, the footprint is expected to be smaller, since recombination moves a beneficial mutation onto different genetic backgrounds allowing linked diversity to occur. Therefore, the observed positive correlation between nucleotide diversity and recombination rate suggests that many loci were or still are the targets of adaptive evolution (Spencer et al. 2006).

In 1992, Aquadro and Begun explained the presence of a correlation between nucleotide diversity and recombination rate in Drosophila by positive selection (hitchhiking) acting on certain regions in the genome. In 1998, M. Nachman observed a correlation between diversity and recombination at certain coding loci in humans. Selection (background or hitchhiking) was proposed as the possible explanation for this phenomenon. In 2002, Lercher and Hurst noticed a correlation at the genome-wide scale. They also observed a correlation between human-mouse divergence and recombination rate. They revealed that the correlation holds not only for coding regions or control elements, but across the entire genome. Thus, they assumed that increased diversity in the regions of higher recombination can be due to higher mutation rates. Based on these data the neutral hypothesis was proposed stating that recombination can be mutagenic.

In 2003, the new human, as well as the chimpanzee data supported the neutral explanation (Hellmann et al, 2003). The correlation between nucleotide diversity and recombination rate was observed within the human population, as well as the correlation between human-chimpanzee (human-macaque) divergence and recombination rate. They showed that regions with less recombination have reduced divergence to chimpanzee and baboon; diversity levels within these regions are also lower. This observation suggested an association between recombination and mutation and supported the neutral explanation for the presence of the positive correlation.

In 2005, with the availability of diversity and recombination data from different human deep-sequencing projects (HapMap and Encode), the presence of the correlation between nucleotide diversity and recombination rates was confirmed on the broad (megabase) scale, but since recombination rates were known to change rapidly on the finer sclale, they became better predictors of diversity than divergence (Hellmann et al. 2005). Recombination, diversity and divergence were found to be correlated with such genomic features as GC and CpG content, as well as with gene expression and gene density. As a result of that study, 12 parameters were investigated and confirmed to be correlated with both diversity (divergence) and recombination rates. Association with these features appeared to be a better explanation, rather than mutagenicity of recombination.

In 2006, the association between genomic features was confirmed (Spencer et al. 2006). Recombination was shown to influence genetic diversity at the hotspot level, since both diversity and recombination were positively correlated with sequence composition. This fact was used to explain the broad-scale association as well. However, the hotspots were confirmed to have no influence on the substitution rate, since they are ephemeral on an evolutionary time scale and have little influence on broader scale patterns of base composition and long-term molecular evolution (Spencer et al. 2006).

1.11 Technological limitations and future perspectives of the recombination analysis

One of the most important and obvious limitations of sperm typing is sex specificity: sperm typing measures recombination only in males, whereas haplotypes in human populations are the products of recombination events happening in both sexes (Arnheim et al, 2003). Although making conclusions about haplotype formation based only on male autosomes is still accepted, it is necessary to keep in mind that recombination rate in females is, on average, higher (Arnheim et al, 2003; Ptak et al, 2005). Moreover, there is substantial sex-dependent regional genomic variation. Thus, in the mouse, one of the MHC hotspots was identified only in female individuals (Isobe et al, 2002). These types of data are not available in humans, because to date, it is not possible to isolate female gametes in sufficient amounts to conduct a high-resolution study at the same level as in sperm cells (Arnheim et al., 2003).

Another possible limitation is the formation of a new haplotype by gene conversion that can be cross over-associated or not (Arnheim et al, 2003). Technical limitations of the sperm-typing methods should be taken into account as well. The most important requirement is the presence of heterozygous polymorphic markers (at least two flanking the region of interest). As a condition for future recombination analysis, deeper sequencing and mapping of polymorphic markers to infer population and individual specific Linkage Disequilibrium patterns and, thus, recombination rates, is needed.

Also, most of the hotspots characterized experimentally are located within MHC, PAR-1, beta-globin and other highly recombining areas. All together, they belong to approximately 0.6 % of the genome. How representative these regions are remains unknown (Arnheim et al. 2003).

Clearly, high-resolution recombination analysis should be performed in an unbiased fashion on different chromosomal segments. Choosing regions based on classical linkage studies can be misleading since most of predicted hotspots may not be found due to individual specific variation (Arnheim et al. 2003).

1.12 The aims of the study

The general aim of this study is to estimate on both fine and broad scale to what extent recombination influences genetic variation patterns. The broad scale analysis consisted of investigating the presence or absence of a positive correlation between nucleotide diversity and recombination rate, based on public whole genome SNP data and available recombination maps.

The primary goal of the fine scale analysis is to experimentally obtain recombination rates for certain genomic regions not so well known for their high recombination rates, compare them with population broad-scale rates from public databases (DeCode, Marshfield, Genethon), also compare with the finer scale rates, especially at the locations of potential “super” hotspots (like MHC, PAR-1 or beta-globin) and see if these estimates, obtained in an unbiased fashion, can correlate with diversity and other features of genome composition.

Theoretical and experimental approaches were expected to provide better insights into the field of recombination and evolution of hotspots and gene conversion.

1.13 Experimental regions of interest

Encode regions on chromosome 11 were selected as experimental targets. Encode is an abbreviation for **Enc**yclopedia **O**f **D**NA **E**lements, another human genome consortium, started in September 2003 to carry out a project to identify all functional elements in the human genome sequence. The pilot phase tested and compared existing methods to rigorously analyze a defined portion of the human genome sequence (1%). For this purpose, certain regions in the human genome were picked randomly or manually. As our regions of interest, we chose four Encode regions on chromosome 11. The primary reason was the variation in fine scale and broad scale recombination rates among those regions. Also, the variation of the base composition, gene density, and location of the regions on the chromosome was taken into account.

2. Materials and Methods

2.1 Material:

The sperm samples for trials and experiments were provided by a single male donor of European origin.

Sperm collection

Semen sample was collected into 14 ml BD falcon and divided into several aliquots: for genomic DNA preparation and for fluorescent activated cytometry sorting (FACS) procedure.

2.2 Experimental methods

Salt extraction of Genomic DNA

Genomic DNA was extracted from bulk sperm cells using standard salt extraction method (Aljanabi and Martinez, 1997).

500 µl of homogenizing (HOM) buffer (80 mM EDTA, 100 mM Tris and 0.5% SDS) were added to 50 µl of the sperm sample and incubated at 55º C for at least 3 hours.

Then the sample was mixed gently with 500 µl of sodium chloride solution (4.5 Molar) and 300 µl Chloroform and mixed gently for 15 min; spun for 10 minutes at 1000 rpm.

After centrifugation the upper phase (850 µl) was transferred into a new tube and mixed carefully with 595 µl of pure Isopropanol (0.7 volume); immediately spun for 10 minutes at 13000 rpm. After that the supernatant was removed. The sediment was diluted in 0.5 ml 70% Ethanol, incubated for 5 minutes and spun for 10 minutes at 13000 rpm. After the removal of supernatant the previous step was repeated. After another removal of supernatant the pellet was dried and dissolved in 10 µl of TE buffer.

DNA concentration was determined using NanoDrop® ND-1000 spectrophotometer.

The sample was diluted and used for PCR to identify heterozygous markers

Sperm sorting and lysis

Prior sorting semen aliquots were diluted with PBS buffer (1:1) stained with Hoechst fluorochrome (5 µl of dye per 1 ml of the sample) and incubated for 1 hour at 35 ºC (Shapiro, 2003).

Single human sperm were sorted by flow cytometry using FACSVantage SE Cell Sorter (sperm sorting profiles are described in chapter 3.1) into 96-well plates containing 2.5 µl of freshly-prepared lysis solution (described in Jiang et al. 2005).

Lysis solution

For 1 ml total volume: 100 µl DTT (1M stock), 400 µl KOH (1M stock), 20-10 µl EDTA (stock either 0.5 or 1 M) and 480-490 µl H_2O (depending on EDTA concentration: 0.5 or 1 M).

After sorting cells were incubated for 10 min at 65 ºC and then the lyses was stopped by adding 3.5 µl of neutralization buffer (Jiang et al. 2005).

Neutralization buffer

Stop solution from REPLI-g mini-kit for whole genome amplification (Qiagen Inc.). The recipe is not available.

After neutralization, the cells were frozen at -20 ºC overnight or immediately continued with whole genome amplification.

Whole genome amplification (WGA)

WGA was accomplished with REPLI-g screening kit according to the manufacture’s manual (Qiagen Inc.). DNA was amplified using Multiple Displacement Amplification. Blank lysis (water sample) was used as a negative control.

The sample was lysed and the DNA was denatured by incubating in the Buffer SB1 at 65 ºC. After the denaturating was stopped by cooling the solution down to the room temperature, Buffer SB2 and DNA polymerase Phi29 were added to the reaction. The isothermal amplification reaction of total 40 µl for 16 hours at 30 ºC and then terminated for 10 min at 65 ºC. Amplified

DNA products were then stored at -20 ºC. 2-fold dilutions were used for allele screening (SNPstream) or allele amplification (confirmation of the recombinant haplotypes).

Selection of the markers

The major prerequisite for marker selection was heterozygosity for the particular donor. Initial preference of microsatellites was abandoned after inability to observe both single alleles in equal quantities in a set of 30 sperms. SNPs were selected as markers also due to a well-known fact that majority of them are biallelic (Petkovski et al. 2005).

Prior to marker selection the four Encode regions were divided into 5 or 9 (depending on the size of the region) 100 kb intervals. The preference for the length of the interval was based on assumption that the regions have intermediate recombination rate of 10^{-8} recombination events per bp and the following calculation:

For 1000 sperm cells: $10^{-8} * 1000 = 10^{-5}$ events per bp. In order to recover at least 1 recombination event between adjacent SNPs, markers should be roughly located 10^{5} bp or 100 kb away from each other.

As soon as the intervals were confirmed, all SNPs residing in intermediate zones of 5-7 kb between the intervals were downloaded from HapMap database. The preference was given to the SNPs with high allele frequency (close to 0.5:0.5 ratio) in the American Utah population of European origin (CEU) to increase the possibility to identify heterozygous markers.

Development of marker loci

The primers for 500 bp loci with allele of interest in the middle were developed using Primer3 program, ordered from Sigma-Aldrich, the loci were amplified by polymerase chain reaction.

Amplification of the marker loci

PCR was run using Qiagen Multiplex PCR Kit and 10 ng of high quality bulk sperm DNA. The total reaction of 10 µl containing 5 µl of Qiagen Multiplex PCR Master Mix with HotStarTaq DNA Polymerase and a unique PCR buffer containing the novel synthetic Factor MP (which stabilizes specifically bound primers and enables efficient extension of all primers in the reaction without optimization Ref: Qiagen Inc), 0.5 µl of forward and reverse primers, 1µl of bulk sperm

DNA and 3 µl of ddH_2O. The reaction was run on Peltier Thermal Cycler (DNA Engine Tetrad2, BRB-Langertechnik GmbH) under the following cycling conditions:

Initial denaturation – 5 min at 96 ºC

30 cycles:

Denaturation: 30 sec at 95 ºC

Annealing: 45 sec at 50-70 ºC according to the T_m of the primer

Elongation: 1 min at 72 ºC

Final elongation: 5 min at 72 ºC

The PCR products were immediately run on 1% agarose gels containing EtBr and visually inspected for products under the UV light.

Purification of the PCR product

Before sequencing the PCR products were purified by Exonuclease I and Shrimp Alkaline Phosphatase. The clean up was conducted using the kit ExoSAP-IT (Affymetrix / USB). ExoSAP-IT was added directly to the PCR product and incubated at 37 °C for 15 min. After PCR treatment, ExoSAP-IT was inactivated simply by heating to 80°C for 15 min.

DNA sequencing

The DNA products were sequenced directly, using the same primers as for PCR. Samples for both marker and haplotype confirmation were sequenced at the sequencing service facility of Cologne Center for Genomics at the Institute for Genetics, Cologne in an Applied Biosystems 3730 capillary DNA analyzer.

Sequencing reaction was performed using Big Dye Terminator Sequencing Kit version 3.1 (ABI, Applied Biosystems, Foster City USA) according to the manufacturer protocol in a total volume of 10 µL :

10 μM sequencing primer – 0.5 μl

10-30 ng purified PCR product

Sequencing Buffer - 2 μl

Big Dye Terminator – 2 μl

ddH_2O to 10 μl final volume

The reactions were run in a GeneAmp PCR System 9700 (ABI, Applied Biosystems) thermocycler under the following conditions:

Initial denaturation: 1 min at 96 ºC

40 cycles:

Denaturation: 10 sec. at 96 ºC

Annealing: 15 sec at 50 ºC

Elongation: 4 min at 60 ºC

Final elongation: 5 min at 72 ºC

Reactions were stored at 12 ºC. Prior to capillary analysis 10 μl of ddH_2O was added to every sample (total 20 μl volume). The output trace files were analyzed with FinchTV software.

Haplotype screen

The initial haplotype screen over 2.5 Mb total of 4 Encode regions flanked by 48 SNP markers (2 SNPs per locus, total 25 loci of interest) was performed by using 48 GenomeLab SNPstream (Beckman Coulter, California, USA) at Cologne Center for Genomics. All materials (SNP ware reagent kits, mastermixes, etc) and services (primer development, software) were provided by Beckman Coulter, Inc.

After confirmation of the heterozygous markers, primers were developed and ordered.

After whole genome amplification cells were diluted 1:8 and 1µl of the dilution (20-100 ng of the amplificon) were used for the SNPstream screen. The samples were placed in 384-well plates. Total of 1536 single cells were analyzed.

Principle of SNPstream

Using the SNPstream system (Beckman Coulter), multiplex genotyping can be performed in a 384-well microtiter plate format (Syvänen, 2005). SNPs with the same nucleotide variation are combined into multiplex reactions with 12 - 48 SNPs per reaction. Thus, up to 48 SNPs can be genotyped in 384 samples to generate a maximum of 18,400 genotypes on a single plate. Following multiplex PCR, the remaining primers and dNTPs are inactivated enzymatically by Exonuclease I and shrimp alkaline phosphatase treatment. Cyclic minisequencing reactions are performed in solution using primers with 5'-Tag sequences and fluorescent ddNTPs, labelled with Tamra for one allele and Fluorescein for the other (Bell et al. 2002). The minisequencing reaction products are captured by hybridisation to complementary Tag-sequence immobilised on thebottom of 384-well glass plates.

The fluorescent nucleotides incorporated into the primers are detected using the CCD camera of the SNPstream system (Bell et al. 2002). The fluorescence signals from each sample are extracted from the images and displayed in a scatter plot where three clusters define a homozygous or a heterozygous genotype for each sample (Fig. 37, Appendix).

SNPstream procedure

The first stage of SNPstream analysis consisted of 48-plex PCR reaction for every sample provided with 96 unique primers per every reaction. PCR conditions:

Initial denaturation: 1 min at 94 ºC

39 cycles:

Denaturation: 30 sec. at 94 ºC

Annealing: 30 sec at 55 ºC

Elongation: 1 min at 72 ºC

4 ºC final hold temperature

Following PCR each sample was treated with 3 µl of SBE clean up reagent (USB, Inc), consisting of Endonuclease 1 and Shrimp Alkaline Phosphatase mixture. Then the 384-well plates were sealed with Microseal A film and incubated on PTC-225 Tetrad at 37 ºC for 30 min. A final incubation at 100 ºC ensured complete inactivation of the enzymes.

The second stage consisted of single base primer extension. The primers were developed and ordered together with multiplex PCR primers. Special extension mastermix was provided by Beckman Coulter. 7 µl of the mastermix were added to every sample. Then the plates were sealed and thermal cycled according to the following protocol:

Initial denaturation: 3 min at 96 ºC

45 cycles:

Denaturation: 20 sec. at 94 ºC

Annealing: 11 sec at 40 ºC

4 ºC final hold temperature

After the extension step, 48-plex SNPware plates were washed with Wash Buffer 1 (diluted in 1:20 in ddH$_2$O).

The third stage was hybridization. 10 µl of hybridization solution (SNPware hybridization, Beckman Coulter) were added to each well. The SNPware plates were sealed and incubated at 42 ºC for two hours at close to 100% humidity. During this time extension products were hybridized to specific microarray spots on the bottom of the plate with specific tags. After the incubation, the plates were washed with Wash buffer 2 and dried by inverting in a Beckman Coulter TJ-25 centrofuge and spinning at 1000 x g for 3 minutes. The SNPware plates were then imaged on the SNPstream Imager for duration of 7 minutes.

The forth stage was to capture data from SNPware plate by the Run Manager software module and transfer as raw images to the image application (Syvänen 2005). The image software module analyzed the images from the each well and determined the positions of the microarray spots to assess individual spot quality and extract their average intensity values. These data were uploaded to the SNPstream database and then combined with the sample set up data (Syvänen 2005). The GetGenos/QC review converted the numbers and sample position data into scatter plots, which were then automatically separated into statistical genotype clusters (Fig. 37, Appendix). Genotype clusters passing user-definable quality parameters were accepted and the genotypes were exported into a result file generated by the Report module.

Allele confirmation

Samples showing the recombinant haplotype were assessed and additional markers were selected to confirm them. The main criterion for these markers was heterozygosity condition and location as close to the initial marker as possible.

Primers for initial and nested PCR were developed with Primer3 software and ordered from Sigma-Aldrich. The loci containing allele of interest in the middle were 500 bp long for first PCR and 250 for the nested.

The products of whole genome amplification were diluted 1:2 and 1 µl (5-10 ng) were used for amplification of the marker loci.

PCR was run using Qiagen Multiplex PCR Kit and 10 ng of high quality bulk sperm DNA. The total reaction of 10 µl containing 5 µl of Qiagen Multiplex PCR Master Mix with HotStarTaq DNA Polymerase and a unique PCR buffer containing the novel synthetic Factor MP (which stabilizes specifically bound primers and enables efficient extension of all primers in the reaction without optimization Ref: Qiagen Inc), 0.5 µl of forward and reverse primers (10 µM – 0.5 ml), 1 µl of single sperm DNA and 3 µl of ddH_2O. The reaction was run on Peltier Thermal Cycler (DNA Engine Tetrad2, BRB-Langertechnik GmbH) under the following cycling conditions:

Initial denaturation – 5 min at 96 ºC

30 cycles:

Denaturation: 30 sec at 95 ºC

Annealing: 45 sec at 50-70 ºC according to the T_m of the primer

Elongation: 1 min at 72 ºC

Final elongation: 5 min at 72 ºC

The PCR products were immediately run on the 1% agarose gels containing EtBr and visually inspected for products under the UV light.

In case of product absence on the gel, a nested PCR was performed using ordered nested primers. The total reaction consisted of 10 µl containing 5 µl of Qiagen Multiplex PCR Master Mix, forward and reverse primers (10 µM – 0.5 mL), 1 µl of single sperm DNA and 3 µl of ddH_2O. The reaction was run on Peltier Thermal Cycler (DNA Engine Tetrad2, BRB-Langertechnik GmbH) under the following cycling conditions:

Initial denaturation – 5 min at 96 ºC

30 cycles:

Denaturation: 30 sec at 95 ºC

Annealing: 45 sec at 50-70 ºC according to the T_m of the primer

Elongation: 1 min at 72 ºC

Final elongation: 5 min at 72 ºC

The PCR products were immediately run on 1% agarose gels containing EtBr and visually inspected for products under the UV light.

Purification of the PCR products and DNA sequencing were conducted the same way as described on page 23.

Buffers

HOM (homogenizing buffer)	160 mM Sucrose
	80 mM EDTA (ph 8.0)
	100 mM Tris (ph 8.0)
	0.5 % SDS
	0.10 mg/ml Proteinase K
10x PBS (phosphate buffered saline)	1.3 M NaCl
	70 mM Na_2HPO_4
	30 mM NaH_2PO_4
	pH 7.2
TAE (tris/acentic acid/EDTA) buffer	40 mM Tris, ph 8.5
	2 mM EDTA
	0.114 % glacial acetic acid
TE (tris/ EDTA)	10 mM Tris – HCl
	2 mM EDTA

2.2. Theoretical methods

Correlation between nucleotide diversity and recombination rate

9.4 Million human SNPs were extracted from SNPdb (for the years 1998, 1999, 2000, 2001, 2002, 2003, 2004) at NCBI and analyzed according to their entry date, validation status, chromosomal location and compared their genomic positions with available recombination maps (Dib et al., 1996; Yu et al., 2001; and Kong et al., 2002). Recombination rates are given for bins of 1MB. Therefore, this bin size was used to determine SNP density by accumulating the number of SNPs within each bin. Heterozygosity was estimated from SNP density and average sample size within each bin (Watterson's estimate of *theta,* Waterson, 1975):

$$\hat{\theta}_W = S / \sum_{i=1}^{n-1} \frac{1}{i}$$

Where **S** is a number of the segregating sites and **n** is a sample size

Displaying obtained data

Conversion and cross over data, as well as data on 13-mer degenerate motif and core motif (Myers et al, 2008) were displayed as custom annotation tracks in the UCSC genome browser (https://cgwb.nci.nih.gov/goldenPath/help/customTrack.html). All files were organized in “BED” format and displayed either as BED lines (conversion tracts and motif) or as BED wiggle (fine scale recombination rates from International HapMap Consortium, 2003 and 2005).

SNP density (for four Encode regions) was expressed as an average number of SNPs in a one or fifty kb sequence.

The CG content percentage of the conversion tracts was calculated as:

[(G+C)/(A+T+G+C)] ×100

Expected and observed CpG fraction was calculated for conversion tracts as well.

3. Results

3. 1 Recombination and variability in public data

To investigate the presence of a correlation between variation data and recombination, human SNPs (bulk genome-wide data, year 1999-2004, roughly 9.4 million) were downloaded from dbSNP and analyzed according to entry date, validation status and chromosomal location. They were also compared to genomic positions with available recombination maps from Genethon (Dib et al. 1996), deCode (Kong et al. 2002) and Marshfield Center (Yu et al. 2001).

First, SNP density for the two sets of SNPs was plotted against recombination rate (deCode, sex-average, 2002). The first set contained all available SNPs from SNPdb, the second contained the set of SNPs with a defined validation status (Fig.6).

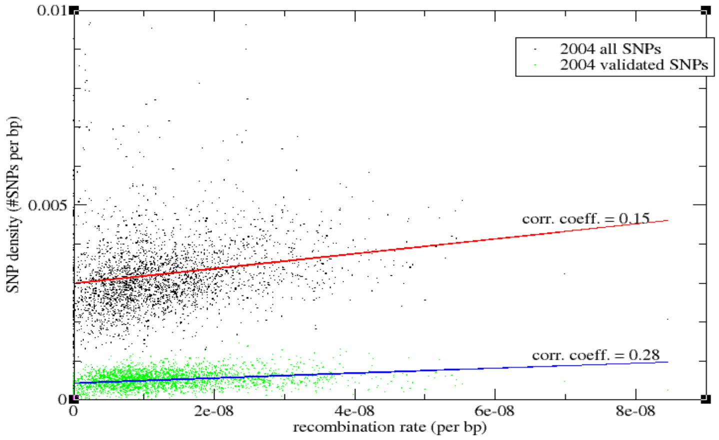


Fig.6. **Scatterplot of recombination rate** (deCode) **and SNP density** based on data from SNPdb, build 124 (2004). Correlation coefficient for all SNPs is r = 0.15 and, for validated SNPs, r = 0.28. The red and the blue lines indicate the regression lines for the two data sets.

In the case of whole genome bulk data, the correlation coefficient between SNP density and recombination rate was r = 0.15. When only validated SNPs (those, for which the validation status in the database was higher or equal to 2; validated SNPs have annotated experimental evidence, which confirms that the particular SNP is not a sequencing error) are considered, the observed correlation became stronger (r = 0.28). However, the picture changed when a different measure of variability was compared to the recombination map. Heterozygosity (SNP density corrected by the sample size) plotted against the recombination map revealed either weaker correlation (r = 0.14) compared to SNP density in the case of validated SNPs, or even negative correlation when bulk data (all SNPs with available sample size) were considered. Thus, these data did not support the initial expectation that positive correlation would be present independent of the quality and quantity of the SNP sets. The possible explanation for this observation is higher sample size of SNPs in the higher recombining regions of functional interest (e.g., major histocompatibility complex – MHC). In such regions of increased diversity and recombination rate, correlation coefficient (based on data up to 2004) was strongly positive, whereas the strength of correlation on the genome-wide scale was weakly positive or negative due to incomplete sampling in non-functional genomic regions.

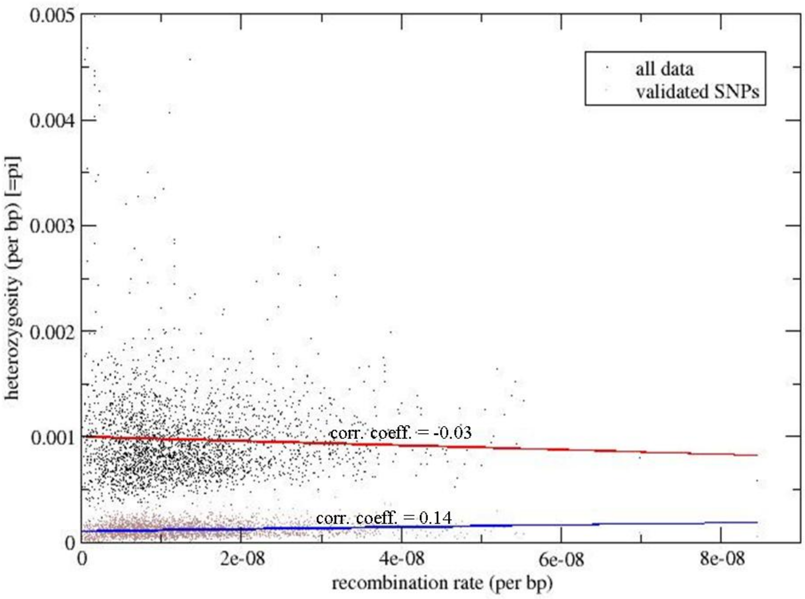


Fig. 7. **Scatterplot of the recombination rate** (deCode) **and heterozygosity** based on data from SNPdb, build 124 (2004). Correlation coefficient for all SNPs is r = -0.03 and for validated SNPs r = 0.14. The red and the blue lines indicate the regression lines for the two data sets.

The strength of correlation varied on both broad- and fine-scale levels. The correlation coefficient varied greatly when SNP density, as well as recombination rate was plotted for each chromosome separately (Fig.8).

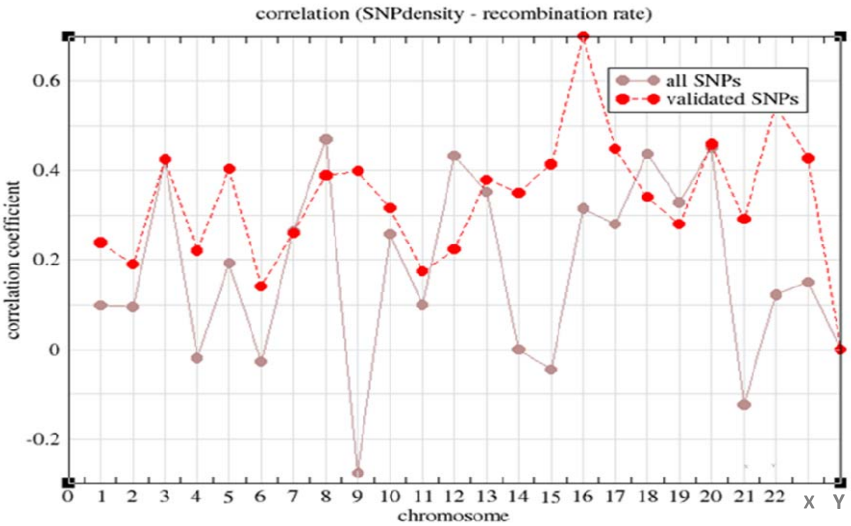


Fig.8. **Correlation coefficient between SNP density and recombination rate for each chromosome (**SNPdb, build 124 and deCode recombination map).

Again, there was no consistency between the bulk SNPs coefficient and the coefficient calculated for validated SNPs. Coefficients were in agreement only for chromosomes 3 and Y. Chromosome 9 shows a good example of sampling bias due to the discrepancy of correlation coefficients for two different data sets (Fig.8). Correlation is strongly positive for validated SNPs and negative for all available SNPs (SNPdb, build 124) on the chromosome.

The magnitude of correlation between SNP density and recombination rate changed across years as well (Fig. 9). SNP data from 1998 and 1999 were scarce and did not uniformly cover the entire genome. Starting from 2000, the accumulation of SNPs was accompanied by a decrease in the correlation coefficient, which could be due to the availability of higher resolution recombination maps, while resolution of SNPs lagged behind. With additional sequencing data from 2003 and 2004, the correlation coefficient increased. Another reason for the coefficient decline could be the change in sample size (number of chromosome investigated per SNP). From 1998 up to 2002 the sample size per SNP has increased, but the correlation coefficient seemed to decline.

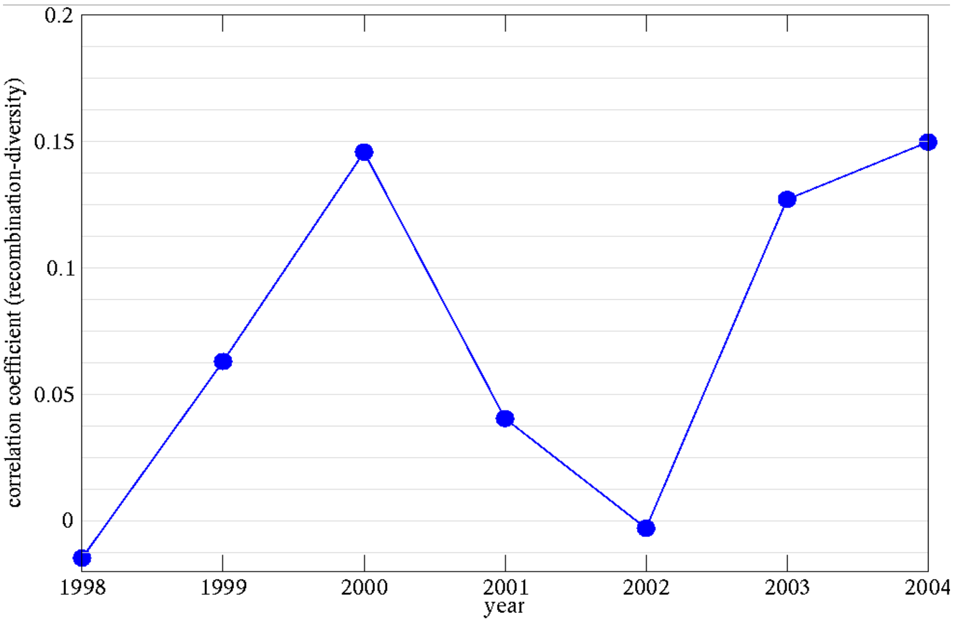


Fig.9. **Correlation Coefficient between nucleotide density and recombination rate across seven years,** all SNPs from SNPdb, build 124 (2004).

The strength of correlation between SNP density and recombination rate also depends on the recombination map. Fig.10 shows how the magnitude of correlation varied with the different recombination data while SNP density remained the same. Up to 2004 both precision and resolution of these maps were of low quality (Genethon, DeCode and Marshfield genetic maps were estimated from pedigree studies with resolution at the CentiMorgan scale from 1 to 2 MB, although recombination rates are known to vary at the kilobase scale – Spencer et al., 2006) and, thus, their reliability is questionable.

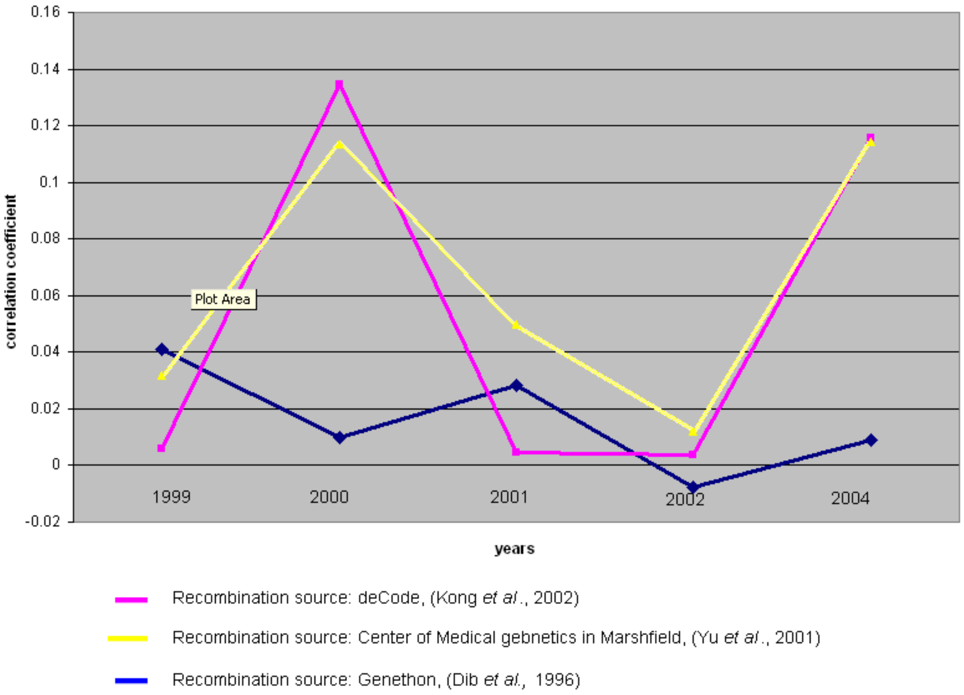


Fig.10. **Correlation coefficient plotted for different recombination maps across the years**

Data available from 2005 and later, after the HapMap project had started, provided a better resolution and a larger sample size and was assumed to give better insights into the observed correlations. The plot of SNP density and fine scale recombination rates from Phase 1 of the HapMap project revealed a negative correlation between the SNP density and recombination map (McVean et al 2004; Fig.11).

Generation of additional HapMap data sets (Phase 1 and 2) starting from 2006 provided new insights into the correlation between SNP density and recombination rate. Fig.11 shows the correlation between SNP density and recombination rate on chromosome 11; the extracted data is from 2008. The correlation coefficient is markedly weak (r =0.029) and when compared to the plot with correlation magnitude for chromosome 11 across the years, the decrease in correlation strength is obvious. Similar observations for 1998 – 2002 in correlation magnitude leads to the hypothesis that generation of new data are accompanied with decrease in coefficient strength.

These observations of correlation strength discrepancy in 1998-2002 compared to the period of 2002-2004 can raise a question of what is a “true signal” and whether strong positive correlation was indeed observed or was this observation due to incomplete data?

According to the higher resolution data, such a correlation decrease with arrival of a new dataset can again indicate that incomplete sampling and prevalence of SNPs from functional highly recombining regions was the possible reason to observe strong positive correlation on the genome-wide scale.

Recombination data used in this analysis were obtained mainly from broad-scale maps (Dib et al. 1996, Yu et al. 2001, Kong et al. 2002) based on population genetic data, which were scarce at the beginning of the century. Although the current data set from HapMap is of much better resolution, the highest possible resolution can only be achieved by direct experiments. The experimental approach of single-sperm typing was modified according to available facilities and conducted to study current individual (not population-wide) cross over rates. Sperm typing was performed for four Encode regions on chromosome 11. The following chapter demonstrates why chromosome 11 and four Encode regions were selected to investigate cross over and gene conversion rates.

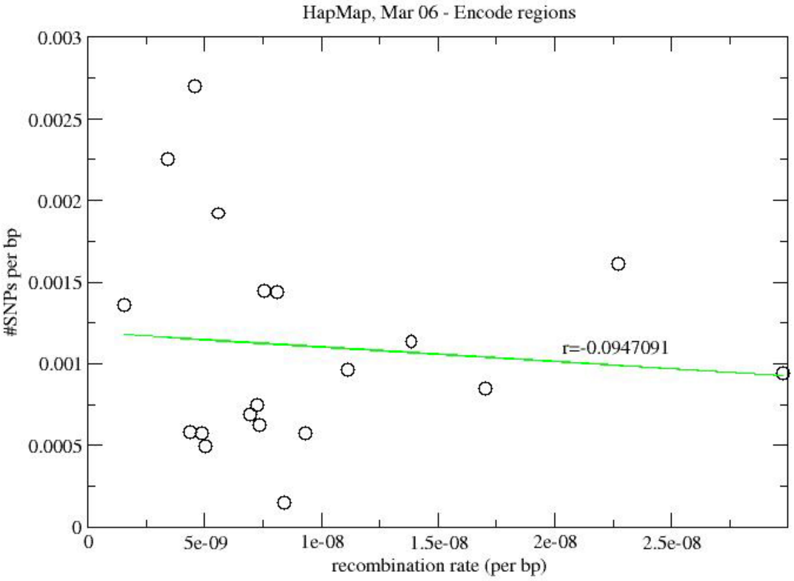


Fig.11. **Correlation between SNP density and recombination rate.** Both SNPs and recombination rates were obtained from HapMap (A haplotype map of human genome, 2005). The correlation coefficient r = -0.094. Recombination rates come from finer scale data (per base pair)

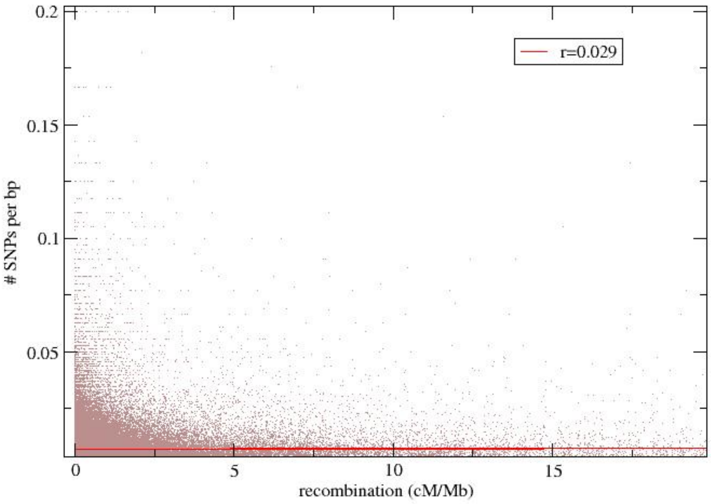


Fig.12. **Correlation between SNP density and recombination rate for chromosome 11** (Conrad et al., 2006). The correlation coefficient r = 0.029. Recombination rates come from broad scale data (centiMorgan per Megabase).

3.2 Specificity of sperm sorting on FACSVantage SE Cell Sorter

In sorting experiments, purity, yield, sorting speed, and count accuracy must be balanced depending on the desired outcome. The critical point for sperm sorting on a typical FACSVantage SE system is to setup an appropriate drop-delay.

The drop-delay is the time it takes for a cell to travel between the laser intersection point down to the end of the liquid jet before the jet breaks off in droplets (breakoff distance, Fig.13). At the laser intersection point, the optical cell parameters are measured. If the cell fulfills the sorting criteria, a voltage pulse is generated after the drop-delay and applied to the liquid jet during the separation of the droplet from the liquid jet. Thus, the droplet containing the cell is charged and ready for deflection in a static electric field generated by two deflection plates. If the drop delay setting is incorrect, droplets are charged at the wrong time and the sorter deflects empty droplets or droplets with erroneous cells at great precision. To set up the desired drop-delay, a sample with fluorescent beads is run through the sorter. This configuration is called “AccuDrop” and it is indispensable for single cell sorting. A sorting window is set in such a way that all beads are included and sorted to the deflected droplet stream. A fluorescence emission filter is moved in front of the video camera to block the laser light (the camera is attached under the deflection plates to illuminate the deflected and un-deflected droplet stream). On the monitor one or two fluorescence spots can be visualized from the fluorescent beads inside the droplets. With the right drop-delay setting no fluorescence should be detected in the un-deflected droplet stream. By watching the monitor and adjusting the drop-delay setting up or down, an optimal value for the drop-delay can be found. The fluorescence spot from the deflected droplet stream should be as bright as possible and other spots as dark as possible. As an example, the AccuDrop method is sufficient for sorting of embryonic cells (Shapiro, 2003).

Sperm cells require a better optimized drop-delay setting due to certain morphological peculiarities. The tail position (straight or curled) and smooth shape of the cell change the fixed speed necessary to cover the break off distance, even with the AccuDrop method. Also, they tend to form clumps and often arrive too close together to be individually sorted. In this case, droplets containing particles are aborted and replaced by empty droplets in the plate well (Shapiro, 2003). The optimal drop-delay option can be determined by performing a Drop Delay Profile (Table 1). For this procedure, a specified number of droplets (for instance, 20) are sorted onto a microscope slide at different drop-delay settings. This procedure requires the use of a fluorescent microscope and the pre-staining of sperm with Hoechst flourochrome. It begins with the lowest drop-delay from the bead sorting and this value is increased at each iteration by as low as 0.1 drops (Table 1). The setting that provides the highest sperm count is the optimal drop-delay setting.

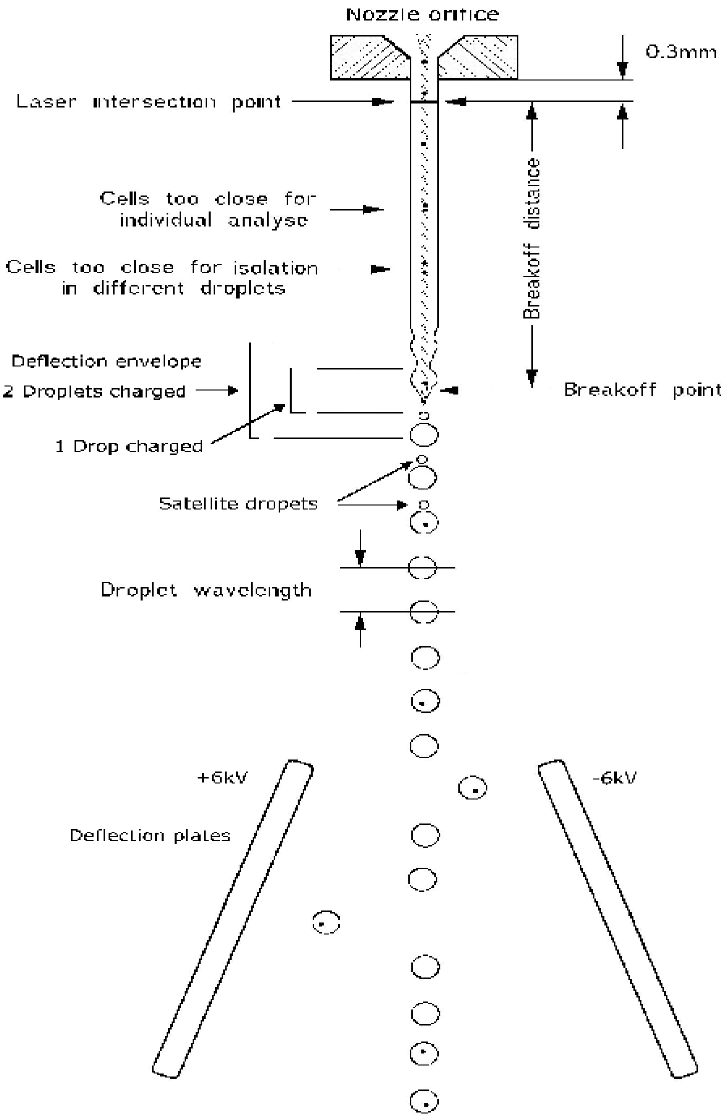


Fig 13. **The sketch of a sorting nozzle** (Shapiro 2003).

Table 1. **Example of a drop delay profile in a sperm sorting experiment using FACSVantage SE sorting system**.

drop-delay	-0.3	-0.2	-0.1	AccuDrop measurement[*] 19.1	+0.1	+0.2	+0.3	+0.4	+0.5
Sperm cells count	-	1	5	10	11	19	19	20	20

[*] - optimal setting determined by AccuDrop method for the particular experiment prior profile start.

3.3 Estimation of the cross over rate:

According to the hotspot detection strategy by single-sperm analysis (Arnheim et al., 2003): if a pair of markers flanking the interval of interest is 1 Mb in size, then 1% of sperm are expected to have a recombinant haplotype for these two markers, assuming that recombination rate is average.

Expected cross over rate in all four regions. Considering the length of every interval flanked by two markers in this study (all Encode region were divided into intervals of 100kb) and the sample size of sperm (1000 single cells), 0.1% of sperm are expected to have recombinant haplotypes within every interval. In all four Encode regions (2.5 Mb long), 25 recombinant haplotypes are expected in total if 1000 sperm cells are typed.

Expected cross over rate per region. According to theoretical expectation based on the assumption that the average rate is 1 cM/Mb (Arnheim et al., 2003), 0.1% of cells are expected to have recombinant haplotypes per every 100 kb interval. Three out of four Encode regions (ENr332, ENm003, ENr312) 0.5 Mb long will be expected to have 0.5% of recombinant haplotypes. In the sample of 1000 sperm, each of these three regions is expected to have five recombinants. ENm009, which is 1 Mb long, is expected to have 10 recombinants. According to the population data expectation (LD-based, deCode map, UCSC browser), the average recombination rates per region are 1.95 (ENm009), 0.9 (ENr332), 1 (ENm003) and 2.5 (ENr312) cM/Mb. Fig. 14 shows the comparison of the expected (LD-based, broad scale) with the observed (experimental) cross over events.

Observed cross over rate in four regions. In this study, only three cross over events were identified experimentally, which makes the cross over rate (considering the total sample of 1000 cells) 0.3% per 2.5 Mb. For every 1 Mb of these 2.5 Mb, the rate is 0.12 %. Thus, the final estimate is 0.12 cM/Mb.

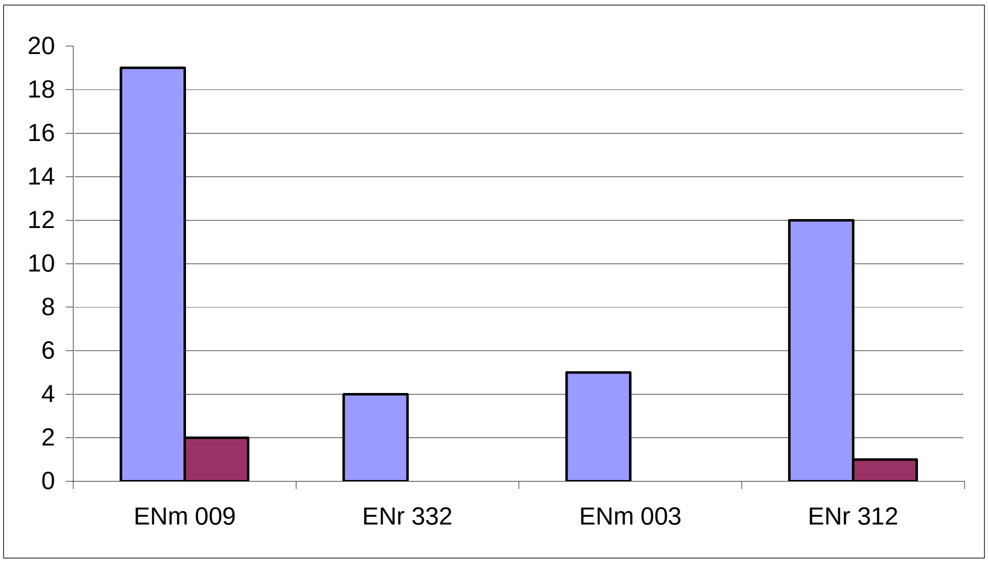


Fig. 14. **Expected (from population data from DeCode map) versus observed (experimental) number of cross over events per Encode region.** Knowing the average recombination rates per Encode region (deCode map, UCSC browser) 1.95, 0.9, 1 and 2.5 cM/Mb, the numbers of expected cross over events per region were calculated as 19, 4, 5 and 12, respectively. Observed number of cross over events were 2 in ENm009 and 1 in ENr312. The expected amount of cross overs is shown in blue and observed amount is in red.

3.4 Annotation of the ENCODE regions

3.4.1 Chromosome 11 at a glance

Human chromosome 11 (HSA11) represents approximately 4.4% of the human genome (Taylor et al, 2006). Being average in size, it is one of the most gene-rich chromosomes: there are approximately 11.6 genes per megabase. This ranks the HSA11 as the fourth highest in gene number among the human chromosomes, after 1, 2 and 19 (Taylor et al. 2006). The recent high-resolution sequencing data from HSA11 (Taylor et al. 2006) is in concordance with existing physical and genetic maps (Genethon, deCODE and Marshfield). Available recombination data reflected the linear relationship between recombination rate and physical distance (Taylor et al. 2006). Analysis of the broad-scale (Fig.15, deCODE, sex-averaged) recombination data revealed variation from 0 to 4.7 cM/Mb, with the maximum concentrated at 11q, the most telomeric region. Overall, in HSA11 there are only ten 1 Mb long regions (deCode sex-average: 130 Mb of the chromosome split into 1 Mb regions, for which recombination data are available) showing a recombination rate exceeding 2.5 cM/Mb. These regions either have an 11q telomeric location or are distributed along 11p (Fig.15B)

The fine-scale rate (A haplotype map of human genome, 2005) varies from 0 to 89.5 cM/Mb. The hotspot map is in general concordance with the broad-scale data: uniform distribution on 11p and telomeric location on 11q, with the “highest peaks” on 11p (Fig.15A).

Genes, apart from recombination, are not uniformly distributed. The majority of coding genes, including the olfactory cluster, are concentrated on the 11p telomeric or 11q centromeric parts of the chromosome. The gene catalogue consists of 1,524 potentially active protein-coding genes and 765 pseudogenes (Taylor et al. 2006). The gene coverage of the chromosome is approximately 47%, whereas exon coverage is only 2.7%. More than 50% of the genes (805) have two or more splice variants. Also, most of these genes contain CpG islands (Taylor et al. 2006). The study of correlation between the presence of a CpG island and the number of variant transcripts for a gene revealed a significant correlation: 70% of genes with CpG islands have two or more variants (Taylor et al. 2006).

HAS 11 has the largest olfactory receptor gene clusters (approximately 166 genes). All human chromosomes except for 20 and Y contain olfactory receptor genes, but HSA11 is by far the richest in these clusters (Taylor et al. 2006). Most of them are concentrated in 11p near the telomere and 11q near the centromere, where a concentration of coding sequence is observed.

Diversity rates also vary along the chromosome. SNP density of 4 SNPs per 1 kb and larger is found on the 11q centromere and the 11p telomere with the highest rates within these blocks. Diversity distribution is in positive correlation with the recombination rate distribution along 11p on the broad scale (Taylor et al. 2006), except for two "hot" regions (in terms of recombination and diversity): On 11q recombination rates were higher near the telomere, SNP density was higher near the 11p telomere and 11q centromere, with the highest gene density in the same regions (Fig. 15C). SNP density, presented by dbSNP, reflects the SNP density of HapMap SNPs with European origin.

Tajima's D estimate distinguishes between DNA evolving randomly ("neutrally" if $D = 0$) versus one evolving under a non-random process, e.g., balancing selection or population decline if $D > 0$, and either positive or purifying selection, or population expansion if $D < 0$. Over all HSA 11 Tajima's D tended to remain positive with a maximum value of 3 and a minimum of -2.5 (Fig. 15E). There are ten regions (larger than 500 kb), where Tajima's D had a negative value; six of them lie on 11q, and four others are on 11p. The question of significance of Tajima's D, obtained from UCSC genome browser without coalescent simulations, remains open. In UCSC Tajima's D estimates were calculated based on Assembly 2004 of human genome, whereas SNP density and recombination data came from Assembly 2006. In addition Tajima's D values do not reveal any particular force shaping variation patterns in certain genomic regions. Both positive and negative values mean that either a certain type of selection is acting on the genome or population structure forms the pattern.

Chromosome 11 contains five Encode regions: two of which are on 11p, closer to the telomere and three are on 11q. In this study, four of these Encode regions are considered in detail.

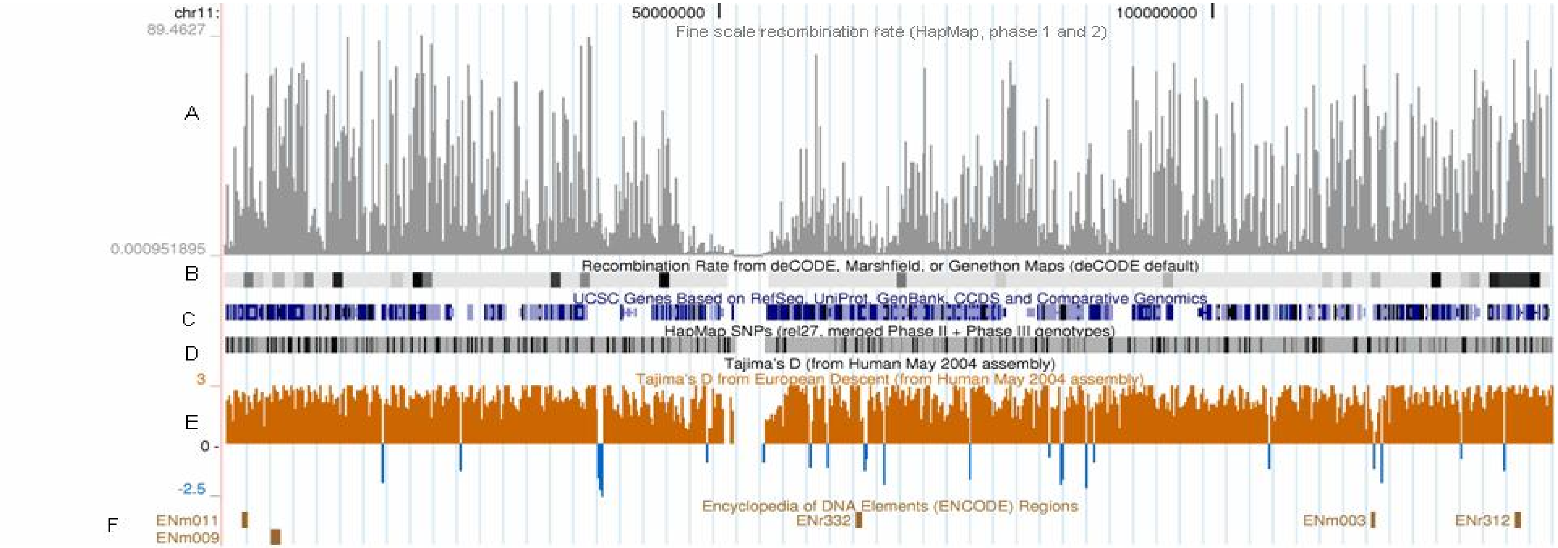


Fig.15. **Overview of the Chromosome 11 from UCSC browser. A**. International HapMap Consortium fine scale recombination rates, (Altshuler et al., 2005); **B**. deCode recombination rates, sex-average; **C**. The genes from UCSC; **D**. SNP density for the American Utah population of the European origin, **E**. Tajima's D estimate for the European population (2004 Assembly); **F**. location of the Encode regions along the chromosome are indicated.

3.4.2 ENm009

ENm009, the largest of all Encode regions, is 1 Mb long (Fig. 16F). This region contains hemoglobin Beta gene, which encodes one of the best studied proteins in the human genome (Taylor et al. 2006). There were several reasons to include this region into the sperm typing screen. Besides telomeric location, this region displays variation in the recombination rate (A haplotype map of human genome, 2005). It has a higher SNP density than the other chosen regions. Furthermore, the gene density is higher compared to other Encode regions. ENm009 is located on the above mentioned 11p with high gene and pseudogene density compared to the other parts of the chromosome. Fifty percent of genes in ENm009 belong to olfactory receptor gene clusters and are short in length (one exon). The rest consists of hemoglobin family members, tripartite motif (TRIM) family members and others.

The broad-scale recombination rates are 1.8 and 2.1 cM/Mb in its 5' and 3' parts (Fig.15B). The high-resolution recombination rates (A haplotype map of human genome, 2005) range from almost 0 to 46.8 cM/Mb or, on a base pair scale of resolution, even up to 75.5 cM/Mb. The hotspot map is quite dense, presenting increased amounts of hotspots interrupted by relatively short coldspot “valleys”.

Another genomic feature - the SNP density – varies greatly in the region. The mean value of the SNP density was calculated per 50 kb and per 1 kb with variance for all regions. For the ENm009, the average SNP density along the entire region was 2.59 SNPs per 1 kb and 129.625 per 50kb with the variance 2707.23. The variance of SNP density within this region is the highest compared to the other three regions. One of the reasons could be the high density of pseudogenes in the this region. Particularly, 24 genes from olfactory clusters are pseudogenes (Zheng and Gerstain, 2006) The SNP density per 50 kb varied from 49 up to 199 in ENm009.

HapMap polymorphism data confirmed dbSNP data in terms of variation distribution: it was not even, but formed clusters or polymorphic “blocks”. However, according to the HapMap data, higher rates are at the ends of the region rather than in the middle.

With respect to other nucleotide properties, the estimated GC content, with 38.2 %, content is low compared to all other regions.

Tajima's D estimate is mainly positive, except at the end of the region in the centromeric direction. In some parts of the ENm009 the estimate was higher than 2, indicating the possibility of balancing selection or population decline.

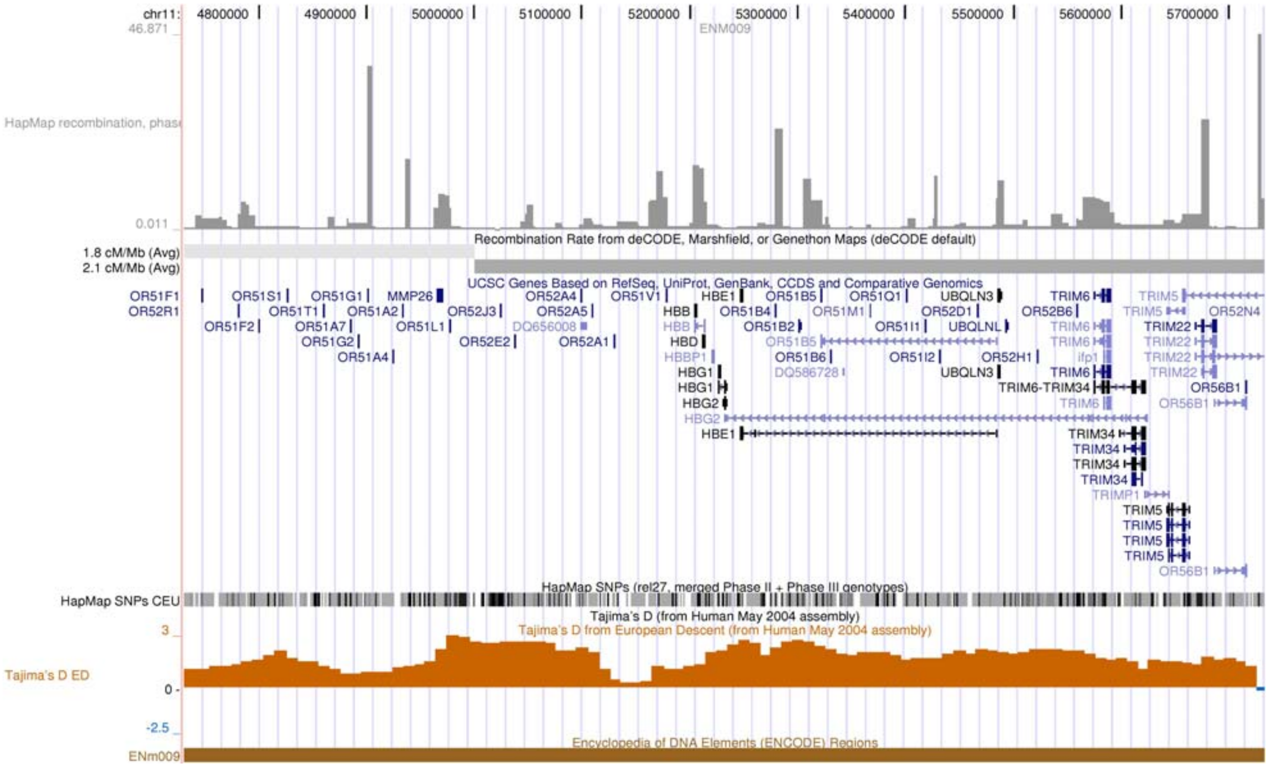


Fig.16. **Overview of the ENm009 region from UCSC browser**. International HapMap Consortium fine-scale rates (Altshuler et al 2005) are in grey; deCode recombination rates, sex-average; the genes from UCSC; SNP density for the American Utah population of the European origin (CEU); Tajima's D estimate for the European population (2004 Assembly) and the current Encode region are indicated.

3.4.3 ENr332

ENr332 is an Encode region of standard size (0.5 Mb), chosen by the Encode project due to its non-exonic conservation level (15.9%) and gene density (2.9%) (fig. 16). We have selected this region because of its relative low recombination rate (compared to ENm009). The region is remarkable for its chromosomal position as well: it is the only Encode region on HSA11 located near the centromere.

Among the genes located in this region, there are members of the integral membrane protein family, members of the Neurexin family, brain-enriched nucleotide exchange factor, members of the serine/threonine protein kinase family, a gene encoding Menin, a putative tumor suppressor among others. Gene density in ENr332 is highly variable.

Moreover, the recombination rate is also highly variable. The broad-scale rate is 0.9 cM/Mb, which is below average. The fine scale rates varied from 0 to 23.2 cM/Mb. The hotspot map consisted of two “superhot” (above 10 cM/Mb) and approximately four hotspots, all of them interrupted by vast coldspot “valleys”.

The average SNP density was calculated per 50 kb and per 1 kb. The mean value for the 50 kb intervals is 51.8 SNPs with a variance of 189.16. The SNP density per 1 kb is 1.715 with a variance of 0.9785. The SNP number in 50 kb intervals varies from 33 up to 69 SNPs, which is considerably less compared to ENm009 maximum (199).

There are 12 genes in this region per region, but interestingly, each of them has at least three splice variants.

The CG content estimated for the region is the highest among all four regions in this study: 53.2%. The prediction is that there would be a higher conversion rate in this region.

Furthermore, the estimate of Tajima’s D was variable in this region as well. Half of the region had positive values up to 3, indicating either balancing selection or population decline. The other half had low positive values (1 – 1.5) interrupted by two “valleys” with negative values of 20 kb each in coding sequence, indicating possibility of positive or purifying selection acting on the loci or population expansion.

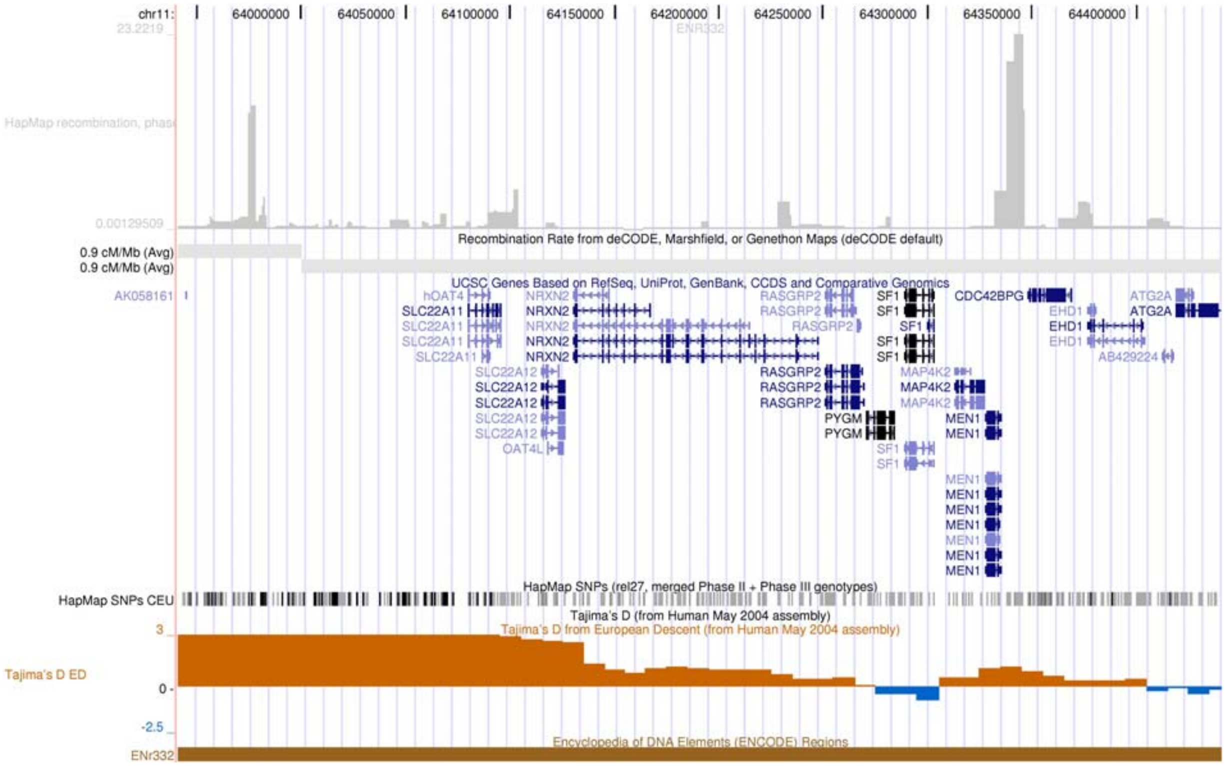


Fig.16. **Overview of the ENr332 region from UCSC browser**. International HapMap Consortium fine-scale rates (Altshuler et al., 2005) are in grey, deCode recombination rates, sex-average, the genes from UCSC, SNP density for the American Utah population of the European origin (CEU), Tajima's D estimate for the European population (2004 Assembly) and the current Encode region are indicated.

3.4.4 ENm003

ENm003 is anothe Encode region on chromosome 11 (Fig.17) and it was picked manually because of the apolipoprotein cluster family. The region is of interest for this study due to the lower recombination rates (below 1.5 cM/Mb) and its telomeric location.

Average in size (0.5 Mb), it has the physical position on 11q23.3 near the telomere. Compared to other Encode regions, this region has only nine genes. Among them, there is the BUD 13 human homolog, the apolipoprotein gene family and KIAA0999, which is Serine/threonine-protein kinase gene expressed in human fetal brain. Most of these genes contain more than one transcript variant.

Roughly 200 kb at the proximal end of ENm003 contain coding sequence.

Results

The recombination rate is low and not uniformly distributed. The broad-scale rate ranged from 0.9 to 2.1 cM/Mb (deCode sex-average). The fine-scale rate varied from 0.02 up to 47.7 cM/Mb (HapMap fine-scale rates, phase 1 and 2, 2006). The hotspot map is represented by three immense hotspots interrupted by vast coldspots “valleys”, most of which are localized at the proximal end of ENm003.

As for the SNP density, the mean value was 75.7 SNPs per 50 kb with a variance of 491.61. At the fine-scale, the average density was 2.28 SNPs per 1 kb with a variance of 2.12. The number of SNPs per 50 kb intervals varied from 27 up to 190, which is the reason for the high variance value.

Tajima’s D had a low value within this region compared to other regions. Half of the region had a low positive value (proximal direction), whereas the other half had a slightly negative value, indicating the possibility of positive or purifying selection or the population expansion. It is one of the few relatively large (230 kb) regions that showed negative D values. The first 20 kb of this negative “valley” was also characterized by a very low SNP density – another indication of selection (hitchhiking can reduce the variability in the region).

The GC content for the region is 43.8%, which is lower than the expected value (above 50%) for conversion-biased activity, but which is close to human average of 41 %.

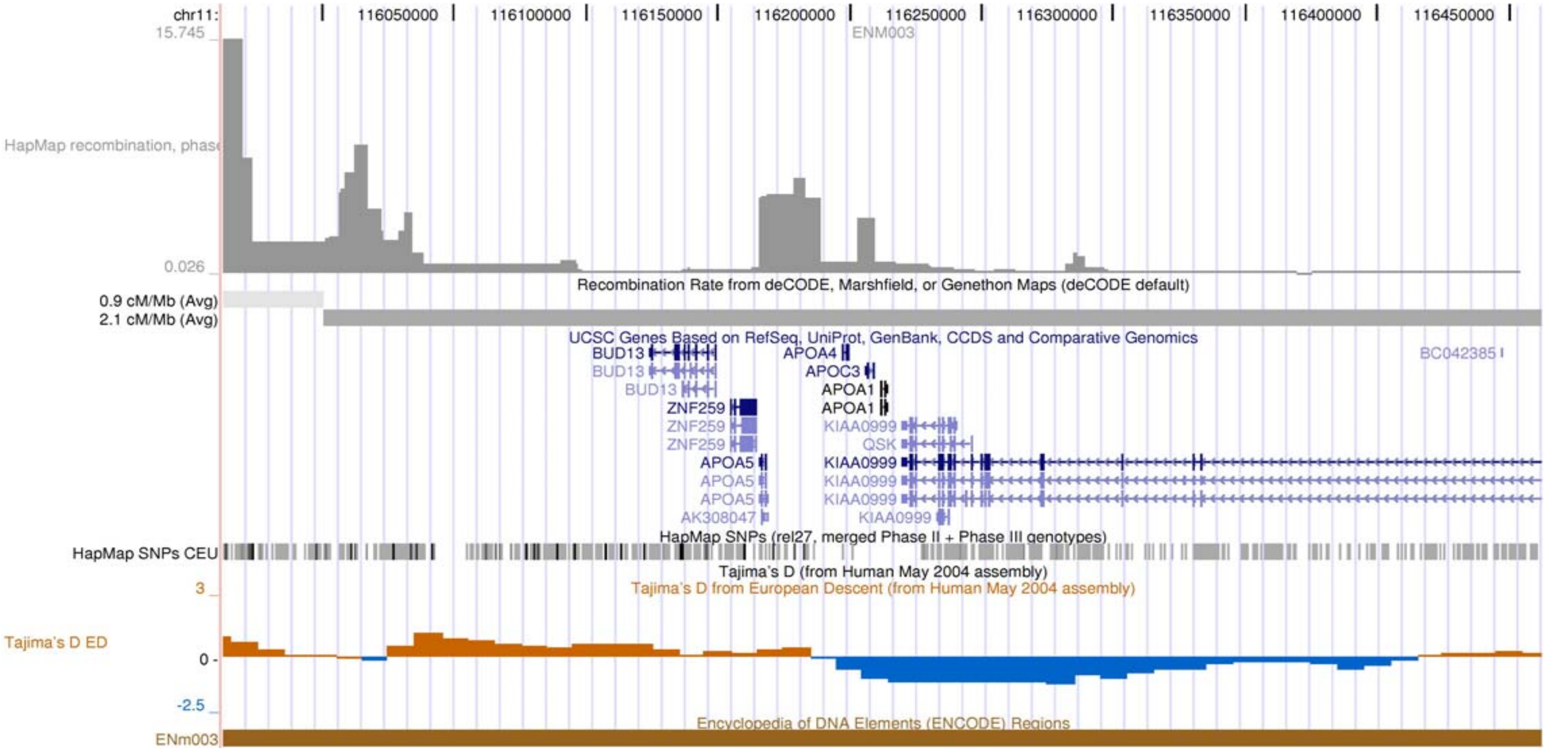

chr11:
116050000
116100000
116150000
116200000
116250000
116300000
116350000
116400000
116450000
ENM003
15.745
HapMap recombination, phase
0.026
Recombination Rate from deCODE, Marshfield, or Genethon Maps (deCODE default)
0.9 cM/Mb (Avg)
2.1 cM/Mb (Avg)
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
BUD13
BUD13
BUD13
APOA4
APOC3
APOA1
APOA1
ZNF259
ZNF259
ZNF259
APOA5
APOA5
APOA5
AK308047
KIAA0999
QSK
KIAA0999
KIAA0999
KIAA0999
KIAA0999
BC042385
HapMap SNPs (rel27, merged Phase II + Phase III genotypes)
HapMap SNPs CEU
Tajima's D (from Human May 2004 assembly)
Tajima's D from European Descent (from Human May 2004 assembly)
3
Tajima's D ED
0
-2.5
Encyclopedia of DNA Elements (ENCODE) Regions
ENm003

Fig.17. **Overview of the ENm003 region from UCSC browser**. International HapMap Consortium fine-scale rates (Altshuler et al., 2005), deCode recombination rates, sex-average, the genes from UCSC , SNP density for the American Utah population of the European origin (CEU), Tajima's D estimate for the European population (May 2004 Assembly) and the current Encode regions are indicated.

3.4.5 ENr312

ENr312 is the most distal of the four Encode regions on chromosome 11, located on 11q25 close to the telomere (Fig. 18). It was picked by Encyclopedia of DNA Elements due to its low non-exonic conservation (13.5%) and gene density of 0.3%. For this study, the region was chosen due its telomeric location and increased recombination rates. The size of the region is 0.5 Mb.

The lowest observed gene density of the region – only 2 genes are there – can be explained by the heterochromatic position of the region.

The recombination rates were variable. The broad-scale rates were above average: 2.7 cM/Mb.

The fine-scale rates (HapMap fine-scale rates, phase 1 and 2, 2006) varied greatly from 0.01 to 44.3 cM/Mb. The hotspot map was very intense, and the coldspot “valleys” were not so frequent. Approximately five “superhotspots” lay within the region. Some of these hotspots were longer than 4 kb.

The SNP density mean value was 103 SNPs per 50 kb with a considerable variance of 839.4. The SNP number per 50 kb varied from 63 to 161 SNPs, which is higher than in ENr332 and ENm003 but still lower than in ENm009. The fine-scale data (HapMap fine scale rates, phase 1 and 2, 2006) had an average SNP density of 2.614 per 1 kb and a variance of 3.0745.

The CG content for the region did not exceed the 50% level (43.9%).

Tajima’s D was estimated and had considerably positive values (from 1.5 to 3), indicating balancing selection or population decline.

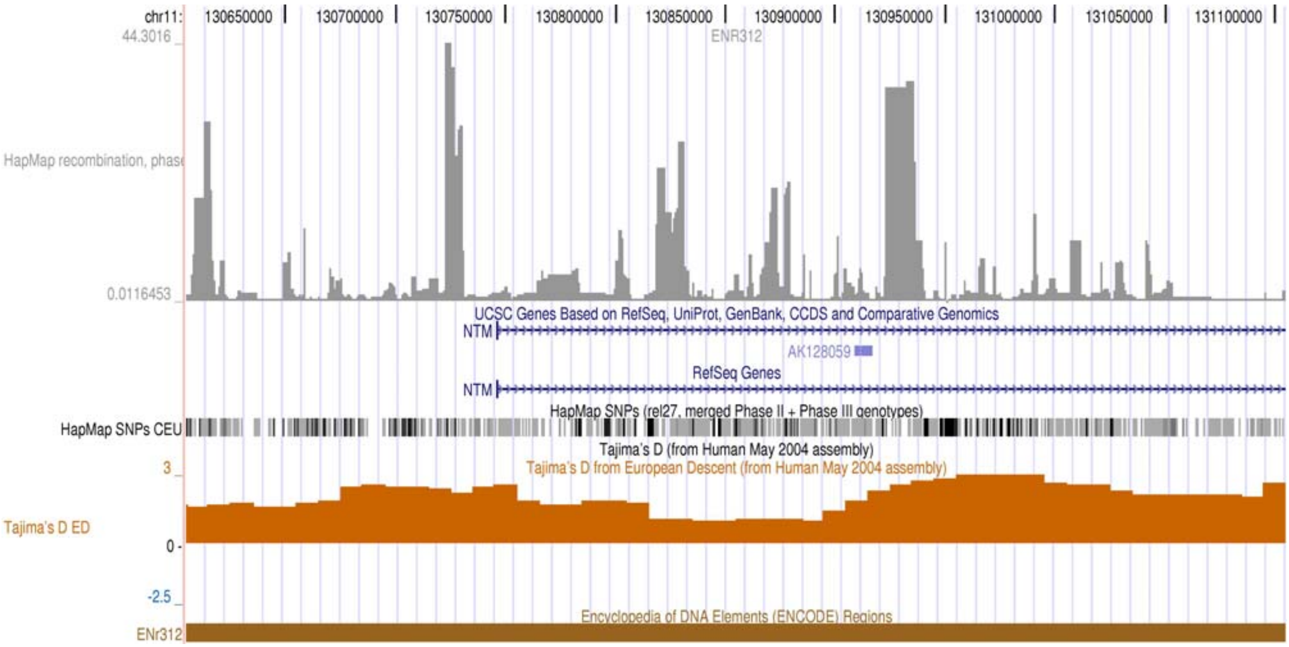

chr11:
130650000
130700000
130750000
130800000
130850000
130900000
130950000
131000000
131050000
131100000
44.3016
ENR312
HapMap recombination, phas
0.0116453
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
NTM
AK128059
RefSeq Genes
NTM
HapMap SNPs (rel27, merged Phase II + Phase III genotypes)
HapMap SNPs CEU
Tajima's D (from Human May 2004 assembly)
3
Tajima's D from European Descent (from Human May 2004 assembly)
Tajima's D ED
0
-2.5
Encyclopedia of DNA Elements (ENCODE) Regions
ENr312

Fig.19. **Overview of the ENr312 region from UCSC browser**. International HapMap Consortium fine scale rates (Altshuler et al., 2005) are in grey, deCode recombination rates, sex-average, the genes from UCSC), the genes from UCSC, SNP density for the American Utah population of the European origin (CEU), Tajima's D estimate for the European population (2004 Assembly) and the current Encode regions are indicated.

3.5 Haplotype screen

Approximately 1000 single genomes were screened with markers flanking 100 kb intervals within four Encode regions on chromosome 11 covering in total a non-consecutive 2.5 Mb (Fig.20). Analysis of theses genomes revealed two types of "reference" haplotypes (further referred to as "black" and "white") present in most of the cells (except for 10 recombinants) based on allele associations. These haplotypes were constructed, first, for each Encode region and then for all regions per every genome. Figure 21 shows allele affiliation with a particular haplotype for ENm009 along with distribution of major/minor alleles between haplotypes. Figure 20 presents a general picture of the "reference" haplotypes for all four Encode regions compared with recombinant haplotypes (which deviated from the "reference" showing the exchange of one or more alleles between the haplotypes).

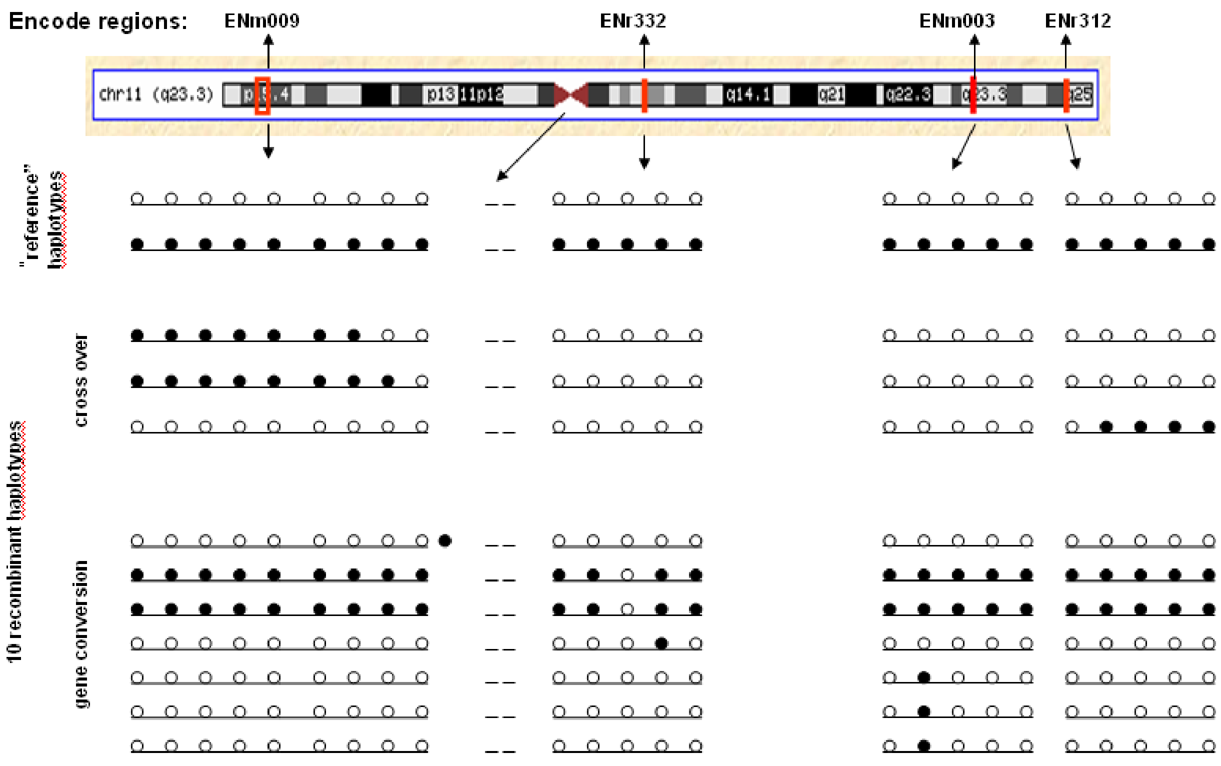


Fig 20. **Overview of the haplotypes**. The circles represent markers flanking 100kb intervals for the genotyping screen. Two reference haplotypes are shown in black and white. Recombinant haplotypes are combinations of black and white either as a “one-site exchange” (conversions) or haplotype “switch” (cross over).

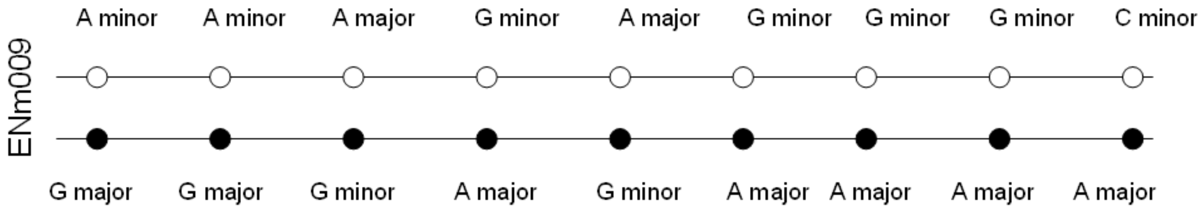


Fig. 21. **Allele affiliation with two haplotypes for ENm009** with distribution of major/minor alleles between these haplotypes.

3.5 .1 “Reference” haplotypes

The markers selected for the genotype screen for which the donor was heterozygous were HapMap annotated SNPs with calculated allele frequencies close to 0.5:0.5 ratios of distribution within the population. However, based on genotyped individuals, each allele was assigned as major or minor for a certain population. In this study, since the sperm donor was of European origin, only CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) allele frequencies of markers were taken into consideration.

There is no strict definition that one of the reference haplotypes was “common” or consists only of major alleles and the other was “rare” (has only minor alleles). Allele affiliation showed that both “reference” haplotypes were a mixture of both allele types with some tendency (bias) (Fig. 21., Fig. 34, 35, 36 Appendix). Analysis of allele combinations present in both haplotypes showed prevalence of major alleles in the “black” haplotype. This fact is especially clear in ENm009 (Fig. 21): seven out of nine. In other regions, on average three out of five alleles were major ones in the “black” haplotypes.

3.5.2 Recombinant haplotypes

Based on the comparison with the two prevalent “reference” haplotypes, recombinants were obtained. The criteria to treat the haplotype as a “recombinant” were either one site exchange within one of the “reference” haplotypes (e.g., when a certain marker in a single genome showed the presence of “white” allele in the “black” haplotype or “black” allele in the “white” haplotype), or switch from one of the “reference” haplotypes to the other within the same single genome (e.g., in one sperm having the “black “ haplotype from the beginning until a certain marker, then starting from the next position, only alleles of “white” haplotype are present) . The former case was considered as a gene conversion event, the latter as a cross over.

Using a “single-sperm recombination detection strategy”, ten recombination events in total were obtained and later confirmed by additional markers. The confirmation consisted of finding one or more markers heterozygous for the male donor of the project. These markers had to confirm the “switch” (in case of cross over) or “one-site exchange” (gene conversion) of the haplotype at the particular position in the recombinant candidate.

The confirmation markers were chosen such that they were no more than 2-3 kb (or closer – depending on the presence of heterozygous SNP) away from the initial screening marker.

For gene conversion, we also estimated the tract length. The region upstream and downstream from the screening marker was investigated for heterozygous SNPs in the recombinant candidate genome. All heterozygotes found were analyzed according to allele affiliation and compared to "reference" haplotypes in other sperm cells. The distance between two or more heterozygotes showing allele affiliation different from the expected "reference" haplotype was defined as a minimum conversion tract. The distance between the markers showing "recombinant" allele affiliation and markers showing "reference" allele affiliation was not considered, since the exact position of conversion start/end is impossible to identify by genotyping.

Out of ten recombinant candidates, three were identified as cross over events (Fig.20) and six as gene conversion within the four Encode regions. The seventh conversion event was discovered during the confirmation process for one of the cross over candidates in the ENm009 region. The position of the conversion tract was 66 kb distal to ENm009.

All recombinant haplotypes found had only one event per single genome. However, each of the mentioned Encode regions had at least one cross over or gene conversion event among the thousand typed sperm cells.

3.5.3 Cross over and gene conversion events within and near ENm 009

According to the haplotype map, ENm009 is the only region with two cross over events and one identified gene conversion outside of the region. Cross over events were detected in between markers seven and eight in one single sperm and between markers eight and nine for another sperm. In total, these are two 100kb intervals most distal within ENm009. To confirm the recombinant haplotype shown by marker nine, another pair of informative markers (SNPs) were chosen 50 kb distal to this Encode region. These markers have confirmed the recombinant haplotype within ENm009 and helped to identify a gene conversion event in a third sperm cell.

The two cross over events were located in the region of increased recombination rate at both broad- and fine-scale (Fig. 22). The broad-scale rate for these two 100 kb intervals with cross over is 2.1 cM/Mb and can be classified as an above average rate. On the fine-scale there were five “super” hotspots within these 200kb. The upper bound was 46.871 cM/Mb for cross over site two and about 11 cM/Mb for cross over site one. The observed cross overs might have occurred within these hotspots.

At the same time, this Encode region is densely packed with genes; the majority of them belongs to the olfactory receptor family. Within the area where the cross overs were identified in the first sperm cell, there were three olfactory receptor members and two members of the tripartite motif (TRIM) cluster. The TRIM family members have a few splice variants each and they cluster within 200kb at the end of the ENm009 covering partially or completely the recombination hotspot provided by the fine-scale map. Both events identified in the two cells could have occurred within this area.

It has been already described that SNP density varies significantly along the region: from 49 up to 199 SNPs per 50 kb with the variance 2707.23.

Since the two observed cross over events occurred within the last two 100kb intervals of the region (Fig.22, purple part of the “haplotype” line for the first and second sperm cells), the mean and the variance were calculated for these intervals as well. Within the first cross over interval, the average SNP density is 3.456 SNPs per 1 kb with a variance of 5.32. The second cross over interval had the lower average density of 2.349 SNPs per 1kb and a variance of 2.349. Both these intervals did not show any significant increase in SNP density either in general or where (according to the Hapmap, phase 1 and 2, 2006) the hotspots were located. Therefore, the increase in SNP density was not observed concomitantly with the increase HapMap recombination rate.

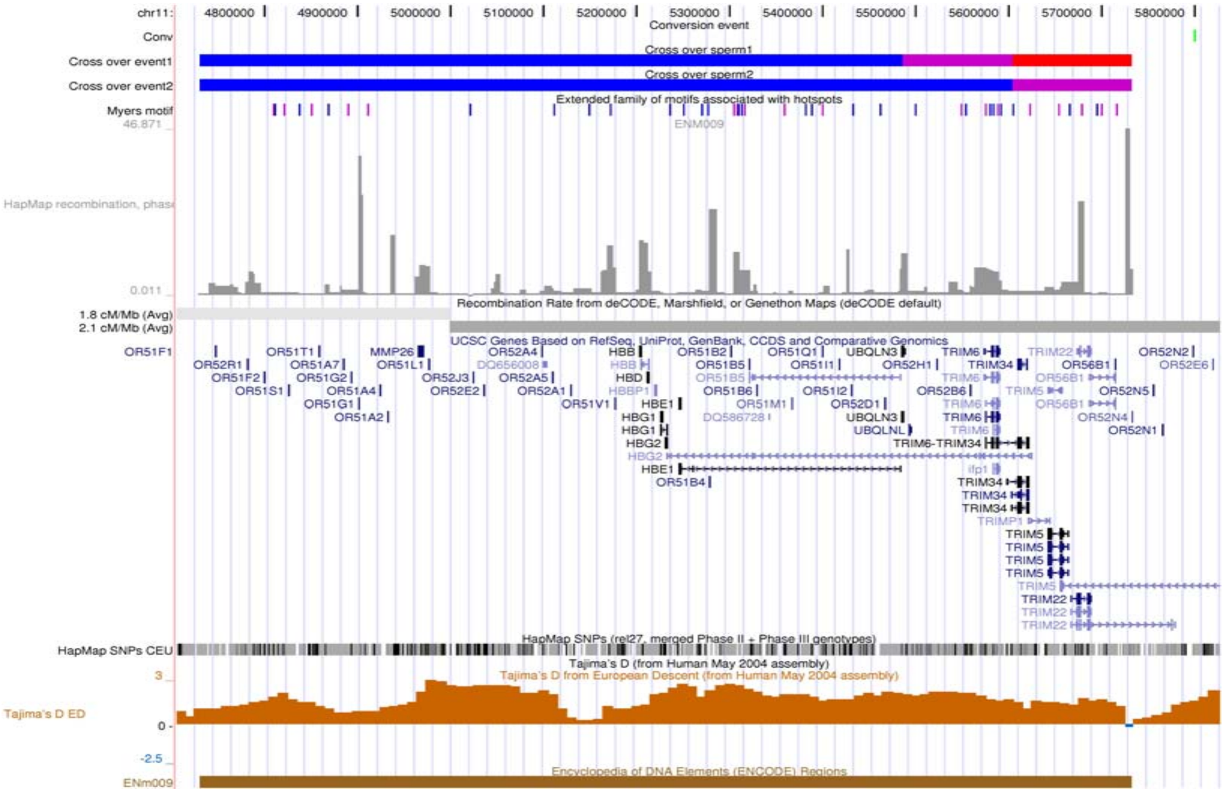


Fig.22. **Cross over and gene conversion within ENm009**. Conversion event (green) 67 kb outside of ENm009); two cross over haplotypes in the last two 100 kb intervals: the putative cross over are painted in purple, initial two non-recombinant haplotypes are in red and blue; distribution of degenerate motif associated with recombination hotspots, International HapMap Consortium fine scale rates (Altshuler et al., 2005) are in grey, deCode recombination rates, sex-average (dense mode), the genes from UCSC (pack mode), the genes from UCSC, SNP density for the American Utah population of the European origin (CEU), Tajima's D estimate for the European population (2004 Assembly) and the current Encode region are indicated.

Tajima's D values within these regions were mostly positive (above 1.5), with a negative non-significant value (-0.2) along 1.5 kb just at the distal end of the region within the hotspot of intensity 46.871 cM/Mb, indicating either a balancing selection or population decline.

ENm009 was screened for the presence of the extended family of motifs based around the degenerate 13-mer CCNCCNTNNCCNC, which was shown to recruit cross over events to at least 40% of all human hotspots (Myers et al. 2008). The expected association between the motif accumulation and hotpots location (HapMap fine-scale rates, phase 1 and 2, 2006) was not observed. Out of 43 total motif hits for this region only 16 (37%) appeared within last two 100 kb intervals where the experimental cross over events were identified. Considering that these two intervals are only 20% of entire ENm009, the amount of hits is quite high. There is a cluster of these motifs on the border between the first and the second intervals near the recombination hotspot of 7.343 cM/Mb (HapMap fine-scale rates, phase 1 and 2, 2006).

ENm009 was also screened for the presence of the core motif, CCTCCCTNNCCAC (Myers et al. 2008). Only one hit of this motif was identified at the beginning of the region, 700kb away from the first cross over interval (Fig 23).

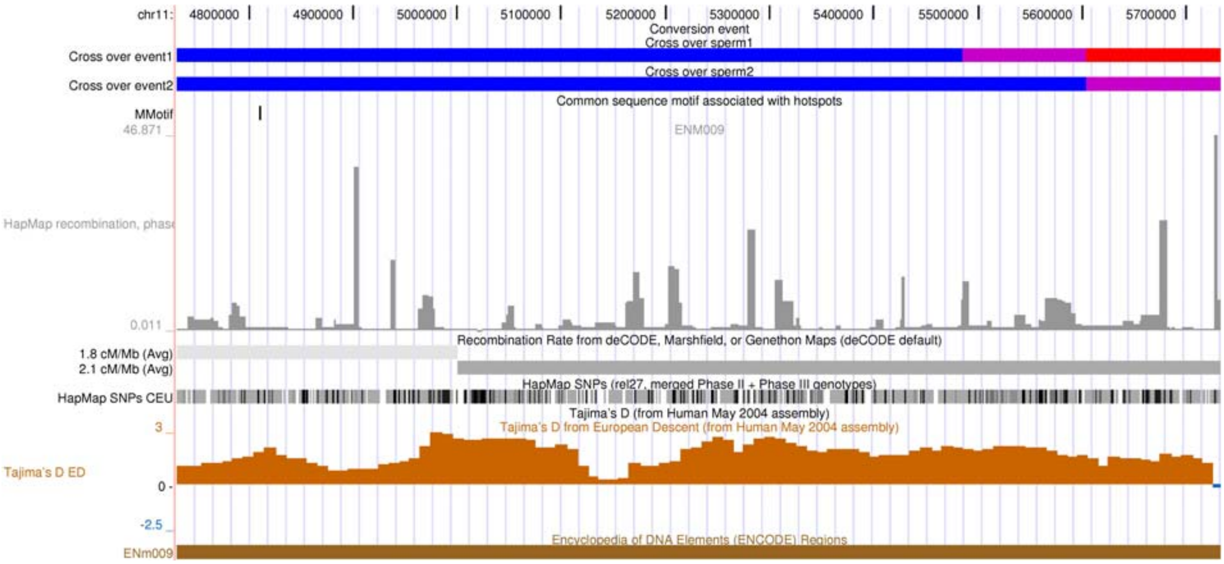

chr11:
4800000
4900000
5000000
5100000
5200000
5300000
5400000
5500000
5600000
5700000
Conversion event
Cross over sperm1
Cross over event1
Cross over sperm2
Cross over event2
Common sequence motif associated with hotspots
MMotif
46.871
ENM009
HapMap recombination, phase
0.011
Recombination Rate from deCODE, Marshfield, or Genethon Maps (deCODE default)
1.8 cM/Mb (Avg)
2.1 cM/Mb (Avg)
HapMap SNPs (rel27, merged Phase II + Phase III genotypes)
HapMap SNPs CEU
Tajima's D (from Human May 2004 assembly)
Tajima's D from European Descent (from Human May 2004 assembly)
3
Tajima's D ED
0
-2.5
Encyclopedia of DNA Elements (ENCODE) Regions
ENm009

Fig.23. **Common sequence motif in ENm009**. Two cross over haplotypes in the last two 100 kb intervals in blue and red, the putative cross over are painted in purple, initial two non-recombinant haplotypes are in red and blue; one hit of common sequence motif associated with recombination hotspots, International HapMap Consortium fine scale rates (Altshuler 2005) are in grey, deCode recombination rates, sex-average, the genes from UCSC, the genes from UCSC, SNP density for the American Utah population of the European origin (CEU), Tajima's D estimate for the European population (2004 Assembly) and the current Encode region are indicated.

The conversion event was identified 66.89 kb distal to the end of ENm009. The estimated minimum tract length between two adjacent SNPs showing a converted haplotype is 2095 bp (table 1), which is by far, larger than the estimated average conversion length of 55 to 300 bp (Jeffreys and May, 2004).

It was located within the longest of the five splice variants of the TRIM5 gene: tripartite motif protein TRIM5 isoform, which localizes to cytoplasmic bodies and whose function has not been identified. The conversion was positioned within the second intron of the transcript (fig. 24). At the same time, the conversion tract is 378 bp downstream from OR52N2, a very short gene of less than 1 kb with a single coding exon. The gene is a member of the olfactory family coding for one of the olfactory receptors. The SNP density within the tract was 4.5 SNPs per 1 kb. The average decode recombination estimate was 2.1 cM/Mb. Tajima's D value was positive, below 1.5. The conversion tract was investigated for CG content and expected/observed fraction of CpG dinucletides. The GC content for the tract was 37.3%, which is quite low especially when considering that conversion tracts are hypothesized to be CG biased.

The CpG observed fraction was 0.0057 compared to expected 0.035.

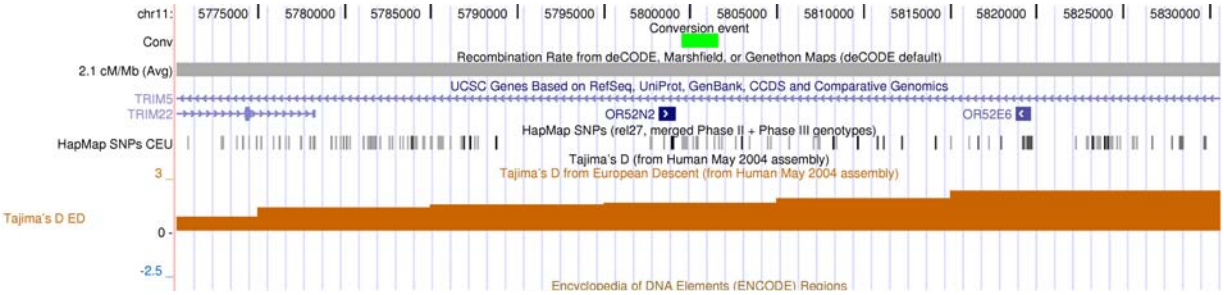

chr11:
5775000
5780000
5785000
5790000
5795000
5800000
5805000
5810000
5815000
5820000
5825000
5830000
Conversion event
Conv
Recombination Rate from deCODE, Marshfield, or Genethon Maps (deCODE default)
2.1 cM/Mb (Avg)
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
TRIM5
TRIM22
OR52N2
OR52E6
HapMap SNPs (rel27, merged Phase II + Phase III genotypes)
HapMap SNPs CEU
Tajima's D (from Human May 2004 assembly)
3
Tajima's D from European Descent (from Human May 2004 assembly)
Tajima's D ED
0
-2.5
Encyclopedia of DNA Elements (ENCODE) Regions

Fig.24. **Conversion event outside of ENm009**. The identified conversion tract (in green); deCode recombination rates, sex-average, the genes from UCSC, the genes from UCSC, SNP density for the American Utah population of the European origin (CEU), Tajima's D estimate for the European population (2004 Assembly) are indicated.

3.5.4 Conversion events in ENr332

In the centromeric ENr332 region, three gene conversion events were identified. Two of them were 8223 bp away from each other and a third one was more than 100kb away from the first two. All of the events resided in the regions of reduced recombination rate. The first two were within 0.9 cM/Mb on the broad-scale (deCode map) and from 0.2 up to 0.4 cM/Mb on the fine-scale (HapMap fine scale rates, phase 1 and 2, 2006); the third was within 0.9cM/Mb and 0.314 cM/Mb, respectively (Fig. 25).

The first gene conversion event had a length between two identified markers of 3104 bp and was located in the second intron of NRXN2, neurexin 2 isoform alpha-1 precursor gene, which belongs to a family of proteins that function in the vertebrate nervous system as cell adhesion molecules and receptors. NRXN2 has two splice variants (Fig.25).

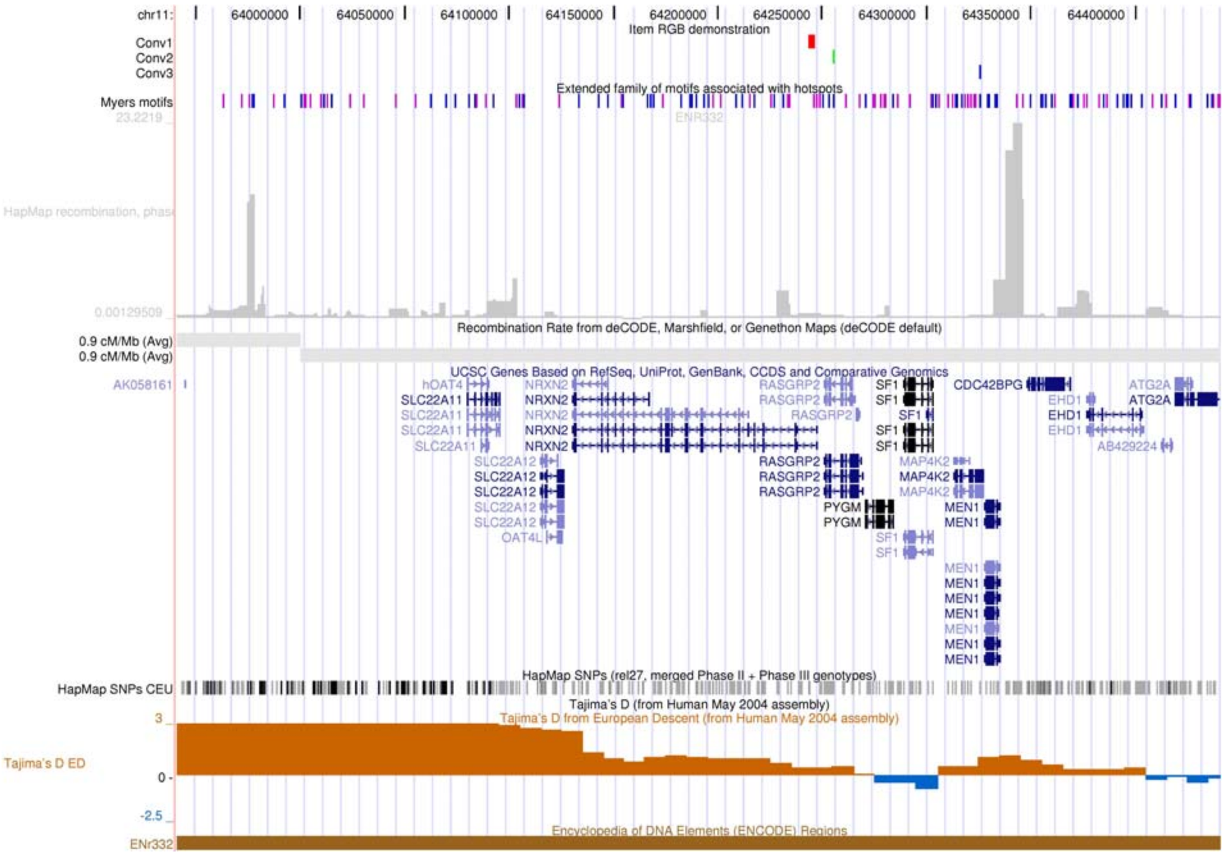

chr11:
64000000
64050000
64100000
64150000
64200000
64250000
64300000
64350000
64400000
Item RGB demonstration
Conv1
Conv2
Conv3
Extended family of motifs associated with hotspots
Myers motifs
23.2219
ENR332
HapMap recombination, phase
0.00129509
Recombination Rate from deCODE, Marshfield, or Genethon Maps (deCODE default)
0.9 cM/Mb (Avg)
0.9 cM/Mb (Avg)
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
AK058161
hOAT4
SLC22A11
NRXN2
RASGRP2
SF1
CDC42BPG
EHD1
ATG2A
AB429224
SLC22A12
OAT4L
MAP4K2
PYGM
MEN1
HapMap SNPs (rel27, merged Phase II + Phase III genotypes)
HapMap SNPs CEU
Tajima's D (from Human May 2004 assembly)
Tajima's D from European Descent (from Human May 2004 assembly)
3
Tajima's D ED
0
-2.5
Encyclopedia of DNA Elements (ENCODE) Regions
ENr332

Fig.25. **Gene conversion within ENr332**. Conversion events are in red, green and blue; distribution of degenerate motif associated with recombination hotspots, International HapMap Consortium fine scale rates (Altshuler et al., 2005) are in grey, deCode recombination rates, sex-average, the genes from UCSC, the genes from UCSC, SNP density for the American Utah population of the European origin (CEU), Tajima's D estimate for the European population (2004 Assembly) and the current Encode region are indicated.

The second conversion event was 1136 bp long (distance between two adjacent markers) and was located in the seventh intron of RASGRP2, a RAS guanyl releasing protein 2, a brain-enriched nucleotide exchanged factor that contains an N-terminal GEF domain, 2 tandem repeats of EF-hand calcium-binding motifs, and a C-terminal diacylglycerol/phorbol ester-binding domain. This protein can activate small GTPases, including RAS and RAP1/RAS3 (Fig.26)

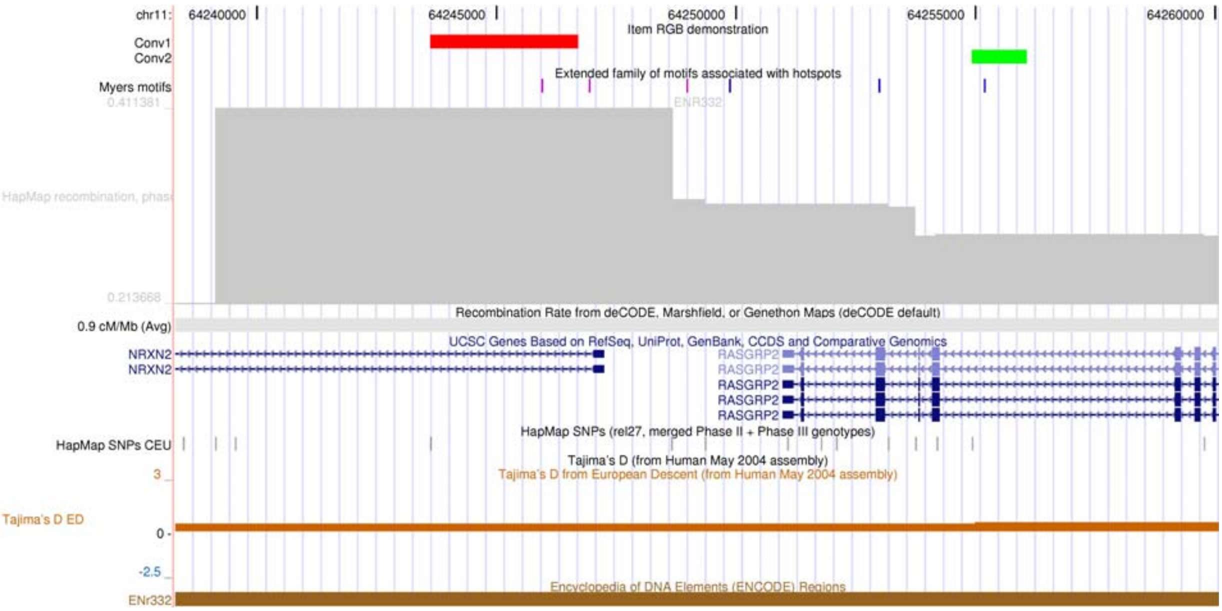

chr11:
64240000
64245000
64250000
64255000
64260000
Item RGB demonstration
Conv1
Conv2
Extended family of motifs associated with hotspots
Myers motifs
0.411381
ENR332
HapMap recombination, phas
0.213668
Recombination Rate from deCODE, Marshfield, or Genethon Maps (deCODE default)
0.9 cM/Mb (Avg)
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
NRXN2
NRXN2
RASGRP2
RASGRP2
RASGRP2
RASGRP2
RASGRP2
HapMap SNPs (rel27, merged Phase II + Phase III genotypes)
HapMap SNPs CEU
Tajima's D (from Human May 2004 assembly)
Tajima's D from European Descent (from Human May 2004 assembly)
3
Tajima's D ED
0
-2.5
Encyclopedia of DNA Elements (ENCODE) Regions
ENr332

Fig.26. **Two gene conversion events within ENr332**. Conversion tracts are in red, green; distribution of the degenerate motif associated with recombination hotspots, International HapMap Consortium fine scale rates (Altshuler et al., 2005) are in grey, deCode recombination rates, sex-average, the genes from UCSC, the genes from UCSC, SNP density for the American Utah population of the European origin (CEU), Tajima's D estimate for the European population (2004 Assembly) and the current Encode region are indicated.

The third conversion event was located within the MAP4K2 gene, mitogen-activated protein kinase, covering the seventh intron and eighth exon. The protein encoded by this gene is a member of the serine/threonine protein kinase family. This kinase can be activated by TNF-alpha, and has been shown to specifically activate MAP kinases. It was the only conversion covering a coding sequence. It is potentially the shortest when compared to the other conversions: 470 bp between the two adjacent markers (Fig.27).

The average SNP density was low compared to the other regions as well: 1.715 SNPs per 1 kb with a variance of 0.9785 for the entire region. All three conversion tracts were located within regions having on average 1 SNP per 1 kb (A haplotype map of human genome, 2005). To determine the minimum length of the third conversion tract, the tract was sequenced to find heterozygous markers.

The Tajima's D value within conversion tracts was positive, but not significant (below 1). Interestingly, the first two and the third conversions are separated by a 100 kb interval, which contains a 30 kb region of negative Tajima’s D values (above - 1.5). Thus far, there has been no indication that this region is a candidate for a selective sweep.

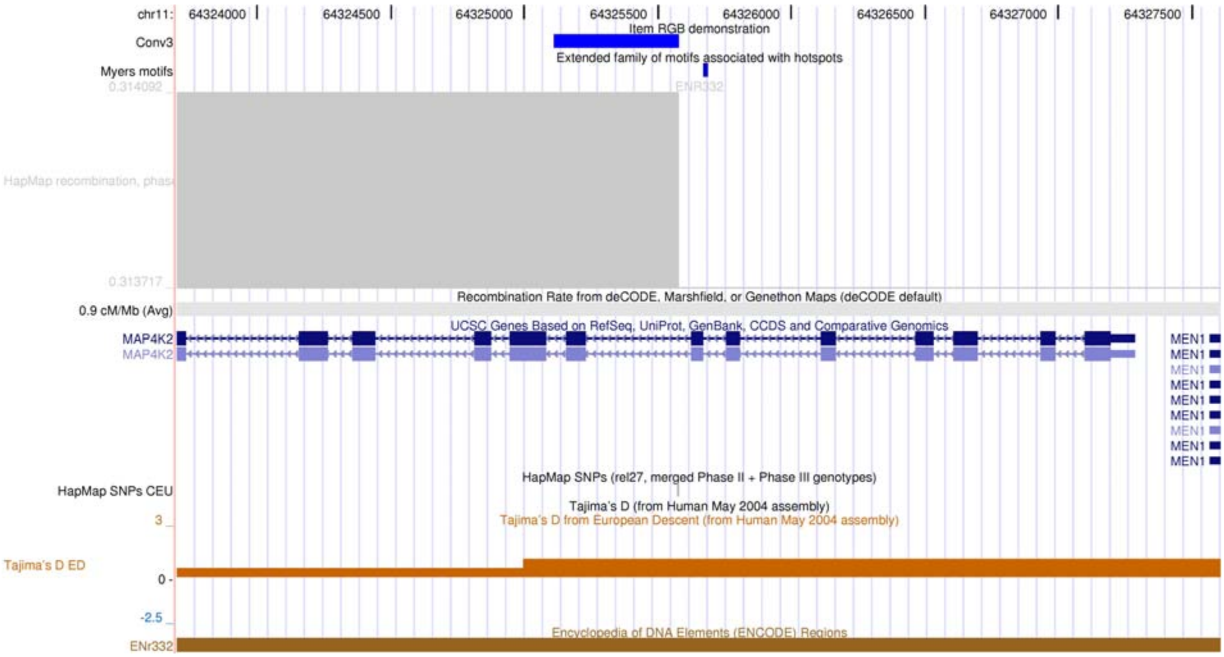

chr11:
64324000
64324500
64325000
64325500
64326000
64326500
64327000
64327500
Item RGB demonstration
Conv3
Extended family of motifs associated with hotspots
Myers motifs
0.314092
ENR332
HapMap recombination, phas
0.313717
Recombination Rate from deCODE, Marshfield, or Genethon Maps (deCODE default)
0.9 cM/Mb (Avg)
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
MAP4K2
MAP4K2
MEN1
MEN1
MEN1
MEN1
MEN1
MEN1
MEN1
MEN1
MEN1
HapMap SNPs (rel27, merged Phase II + Phase III genotypes)
HapMap SNPs CEU
Tajima's D (from Human May 2004 assembly)
3
Tajima's D from European Descent (from Human May 2004 assembly)
Tajima's D ED
0
-2.5
Encyclopedia of DNA Elements (ENCODE) Regions
ENr332

Fig.27. **Third gene conversion tract within ENr332**. Conversion tract is in blue; distribution of degenerate motif associated with recombination hotspots, International HapMap Consortium fine scale rates (Altshuler et al., 2005) are in grey; deCode recombination rates, sex-average; the genes from UCSC, the genes from UCSC, SNP density for the American Utah population of the European origin (CEU), Tajima’s D estimate for the European population (2004 Assembly) and the current Encode region are indicated.

The ENr332 region was also screened for the presence of a common sequence degenerate motif and the core motif associated with the hotspots. Again, the distribution of the degenerate motif hits was uniform, unassociated with the few hotspots present in this region, according to the fine-scale map (A haplotype map of human genome, 2005). Although the ENr332 average recombination rate was low compared to the other three Encode regions, 122 degenerate motif hits were detected – the highest number of the motifs identified among all four Encode regions.

The first two conversion tracts had at least one motif hit within the tracts and the third one had a motif hit 50 bp outside of the tract.

Moreover, the ENr332 region had the highest number of the core motif hits among other the regions (four hits). All of them were outside of conversion tracts and one was located within a 100kb interval separating the first two and the third (Fig. 28).

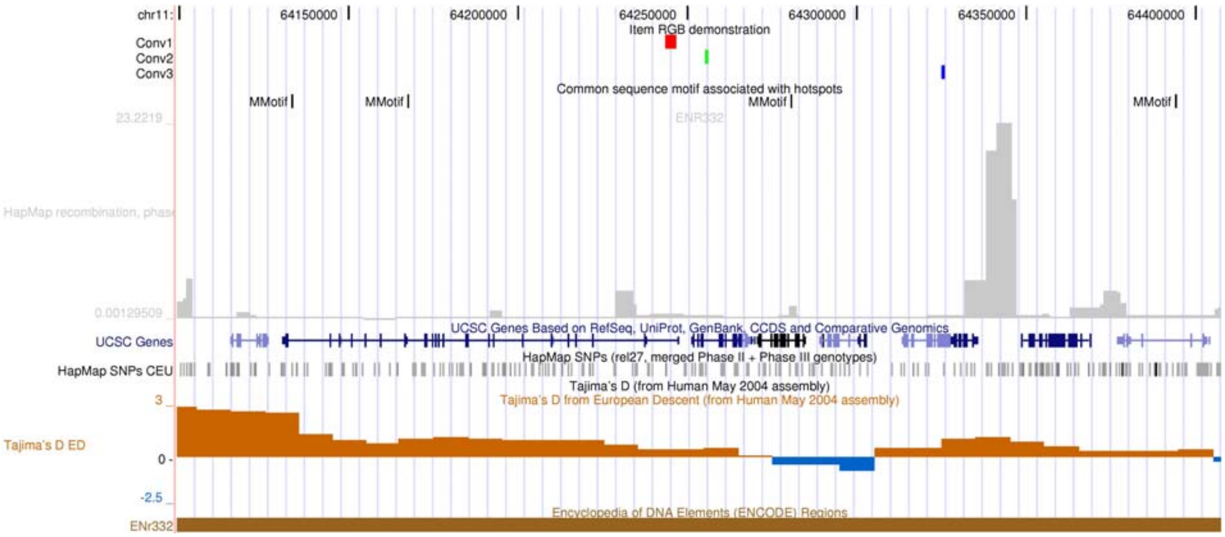

chr11:
64150000
64200000
64250000
64300000
64350000
64400000
Item RGB demonstration
Conv1
Conv2
Conv3
Common sequence motif associated with hotspots
MMotif
MMotif
MMotif
MMotif
23.2219
ENr332
HapMap recombination, phase
0.00129509
UCSC Genes
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
HapMap SNPs CEU
HapMap SNPs (rel27, merged Phase II + Phase III genotypes)
Tajima's D (from Human May 2004 assembly)
Tajima's D from European Descent (from Human May 2004 assembly)
3
Tajima's D ED
0
-2.5
Encyclopedia of DNA Elements (ENCODE) Regions
ENr332

Fig.28. **Common sequence motif associated with hotspots in ENr332**. Conversion events are in red, green and blue; distribution of the common motif associated with recombination hotspots, International HapMap Consortium fine scale rates (Altshuler et al., 2005) are in grey, deCode recombination rates, sex-average, the genes from UCSC, the genes from UCSC, SNP density for the American Utah population of the European origin (CEU), Tajima's D estimate for the European population (2004 Assembly) and the current Encode region are indicated.

Among all conversion events, the three identified within this region had the highest GC content. Within the first conversion tract, GC content was 56%, within the second 42%, and within the third 62%, which is significantly high. One of the possible explanations for the highest CG rate in the third conversion is that the first two are located in introns, whereas the third is in the intron and the exon.

The CpG observed fraction for the tracts was on average from three to eight fold lower than the expected (Table 2).

3.5.5 Conversion tracts in ENm003

The three conversion tracts found within this region represent the most intriguing finding among all the others. In three sperm cells, these events were presumably initiated or terminated within the same site, which was 1.5 kb away from the BUD13 Homo sapiens homolog transcription termination site. For the shortest tract, two markers identified a "converted" pattern relative to the other haplotypes. For the middle tract, three markers identified such a pattern and five markers in the long tract. Starting from the long tract, the sizes of the tracts were 9599 bp, 3580 bp and 1996 bp, respectively (Fig 29).

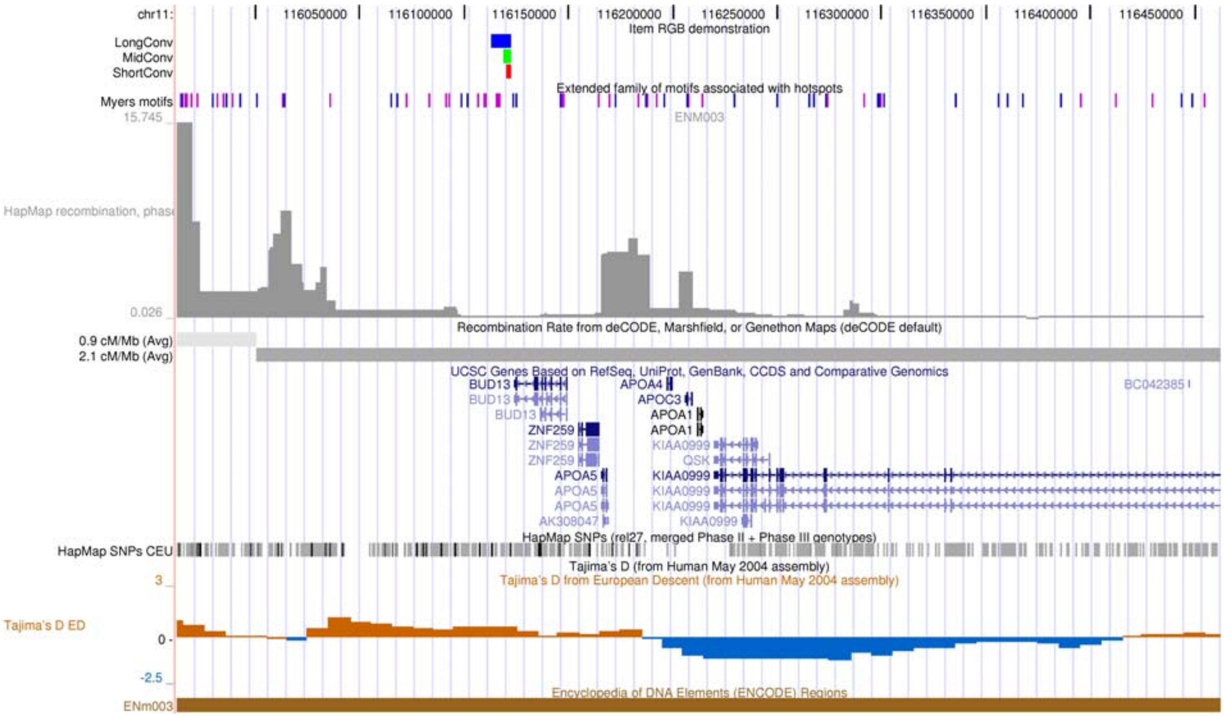

chr11:
116050000
116100000
116150000
116200000
116250000
116300000
116350000
116400000
116450000
Item RGB demonstration
LongConv
MidConv
ShortConv
Extended family of motifs associated with hotspots
Myers motifs
15.745
ENM003
HapMap recombination, phase
0.026
Recombination Rate from deCODE, Marshfield, or Genethon Maps (deCODE default)
0.9 cM/Mb (Avg)
2.1 cM/Mb (Avg)
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
BUD13
APOA4
BC042385
APOC3
APOA1
ZNF259
KIAA0999
QSK
APOA5
AK308047
HapMap SNPs (rel27, merged Phase II + Phase III genotypes)
HapMap SNPs CEU
Tajima's D (from Human May 2004 assembly)
Tajima's D from European Descent (from Human May 2004 assembly)
3
Tajima's D ED
0
-2.5
Encyclopedia of DNA Elements (ENCODE) Regions
ENm003

Fig.29.**Gene conversion events in ENm003**. Conversion events are in red, green and blue (full mode); distribution of degenerate motif associated with recombination hotspots (dense mode), International HapMap Consortium fine scale rates (Altshuler et al., 2005) are in grey (full mode), deCode recombination rates, sex-average (dense mode), the genes from UCSC (pack mode), the genes from UCSC (pack mode), SNP density for the American Utah population of the European origin (CEU), Tajima's D estimate for the European population (2004 Assembly) and the current Encode region are indicated.

The recombination rates provided by maps of different resolution yielded different estimates for the region where the conversions were: 2.1 cM/Mb according to the broad-scale (deCode map), which is an average rate, and 0.115 cM/Mb according to the fine-scale map (HapMap fine scale rates, phase 1 and 2, 2006), which identifies cold spot “valley” within this area.

The SNP density was 2.28 SNP per 1 kb with a variance of 2.12, which is also in line with the low recombination rate observation. The number of SNPs per 50 kb within this region was 87. Tajima’s D was positive (below 1) in part of the region and negative (down to -2) in the other part, which can indicate positive selection or population expansion.

The region was also screened for the presence of degenerate and core motifs associated with the hotspots. Sixty-five total hits were identified for the degenerate motif. Clustering was not observed as in previous cases. Although the conversions were detected in the area with very low recombination rate (A haplotype map of human genome, 2005), there were four motif hits within the long tract. For the core motif there was only one hit per ENm003 and it was outside of the conversion tracts (Fig.30).

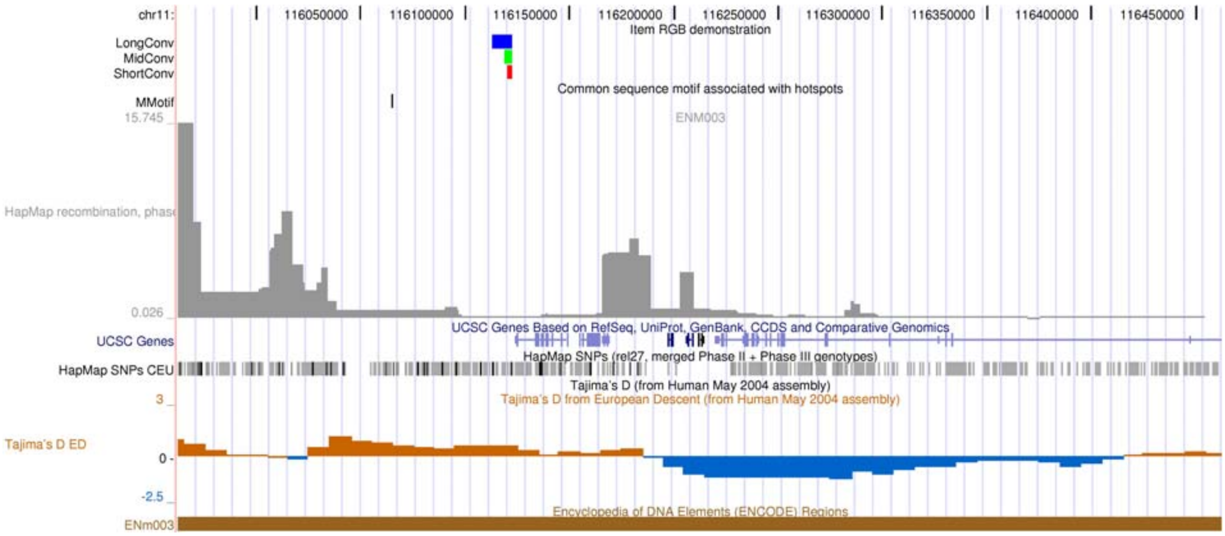


Fig.30. **Common sequence motif in ENr332**. Conversion events are in red, green and blue; distribution of common motif associated with recombination hotspots, International HapMap Consortium fine scale rates (Altshuler et al., 2005) are in grey, deCode recombination rates, sex-average, the genes from UCSC, the genes from UCSC, SNP density for the American Utah population of the European origin (CEU), Tajima’s D estimate for the European population (2004 Assembly) and the current Encode region are indicated.

The CG content within these conversions was 44%, 40% and 38% for the long, middle-size and short, respectively. There was no evidence that these tracts are CG biased.

The CpG observed fraction in all the tracts was on average three to five fold lower compared to the observed (table 2).

The increased conversion rate within this site can indicate the presence of individual conversion hotspot. The distribution of haplotypes (Fig. 35, appendix) shows presence of major allele in the the haplotype with minor allele prevalence.

3.5.6 Cross over ENr312

The third cross over haplotype was identified in ENr312. The event occurred within the first 100 kb interval (Fig. 31).

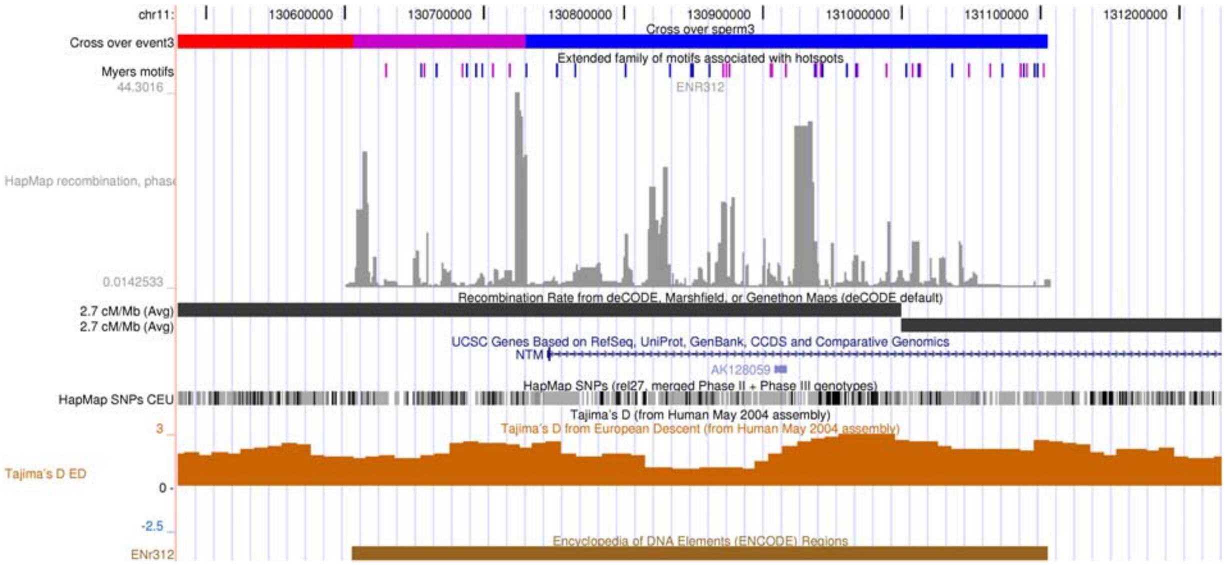


Fig.31. **Cross over event in ENr312**. Cross over haplotype in the first 100 kb interval, the putative cross over are painted in purple, initial two non-recombinant haplotypes are in red and blue; distribution of degenerate motif associated with recombination hotspots, International HapMap Consortium fine scale rates (Altshuler et al., 2005) are in grey, deCode recombination rates, sex-average, the genes from UCSC, the genes from UCSC, SNP density for the American Utah population of the European origin (CEU), Tajima's D estimate for the European population (2004 Assembly) and the current Encode region are indicated.

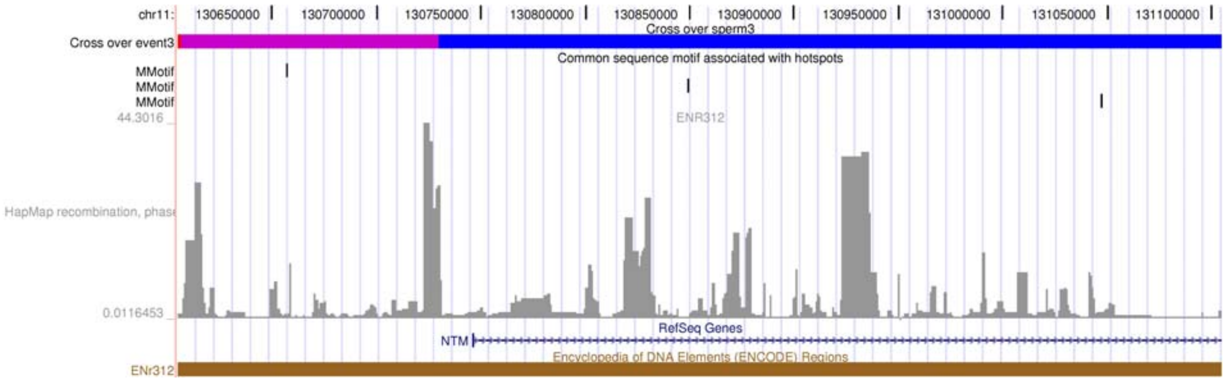
The recombination rates from available maps are quite high within this region: 2.7cM/Mb broad-scale (deCode map) and from 0.0116 up to 44.301 cM/Mb on the fine-scale (HapMap fine scale rates, phase 1 and 2, 2006). The interval had no coding sequence, but a HNT gene starts right outside the boundary of this interval.

The SNP density in this first 100kb interval was 2.62 SNPs per 1 kb with a variance of 2.462. The average number of SNPs per 50 kb was 94.

Tajima's D value was considerably positive (above 1.5), indicating balancing selection or population decline.

Although the region presented high variation in the recombination rate, only 47 hits for the degenerate motif associated with the hotspots were detected. Twenty-five percent (12) of these motifs were identified within the first 100kb interval showing a cross over pattern. The distribution was uniform without association to any particular hotspot.

Three core motif hits were identified for this region. One of them was located in the first cross over interval (Fig.32).



.

Fig.32.**Common sequence motif in ENr312**. Cross over haplotype in the first 100 kb interval, distribution of the common sequence motif associated with recombination hotspots, International HapMap Consortium fine scale rates (Altshuler et al., 2005) are in grey, deCode recombination rates, sex-average, the genes from UCSC, the genes from UCSC, SNP density for the American Utah population of the European origin (CEU), Tajima's D estimate for the European population (2004 Assembly) and the current Encode region are indicated.

Table 2. Summary of identified conversion tracts and genomic features

Encode region	Minimum length (between 2 adjacent markers), bp	Beginning position	End position	Locus specific location	GC content (within each tract) %	CpG fraction (within each tract)		Average SNP density (per kb)		#SNP per 50 kb
						observed	expected	mean	variance	
Near ENm009	2095	5799485	5801580	Second intron Trim5	37.3	0.0056	0.0334	2.47	2.349	91
ENr332	3104	64243595	64246699	second intron of NRXN2	57.9	0.016	0.0802	1.715	0.9785	41
ENr332	1136	64254922	64256058	7th intron of RASGRP2	43.4	0.0121	0.045	1.715	0.9785	35
ENr332	470	64325109	64325579	7th intron and 8th exon MAP4K2	63.3	0.029	0.0942	1.715	0.9785	33
ENm003	9599	116112851	116122450	1.5 kb away from BUD13	44.9	0.0097	0.0483	2.28	2.12	87
ENm003	3580	116118870	116122450	1.5 kb away from BUD13	41.3	0.0107	0.0409	2.28	2.12	87
ENm003	1996	116120454	116122450	1.5 kb away from BUD13	39	0.0079	0.0366	2.28	2.12	87

4. Discussion

4.1 Analysis of correlation between SNP density and recombination rate in the human genome

Our analysis of SNP data sets available from 1998 and different broad-scale recombination maps revealed the presence of weak positive correlation between SNP density and recombination rate and negative correlation between heterozygosity and recombination rate, which can be explained by limited amount of validated SNPs with defined sample size used to calculate the correlation coefficient. The cause of the correlation is highly debated and uncertain. Yet another uncertainty is the magnitude of the correlation, due to the high variation that is observed. The correlation depends largely upon the amount and quality of SNP data, as well as the quality of the recombination maps. The analyzed data can be divided into two groups.

First is the data from early years (1998-2004). This period is characterized by a sparse amount of publicly available SNPs in 1998-2001, intense accumulation of SNPs in databases (e.g., dbSNP) in 2002-2004 and low resolution recombination maps (genetic maps from pedigree studies). Second is the data from later years (2004-2007), which has better annotation of polymorphic alleles and a higher resolution of recombination.

In the early years, despite the quality of SNP data, there is a strong positive correlation between SNP density and recombination rate, especially using the data from 2000-2001. The magnitude of correlation decreases with the arrival of new data sets with larger sample sizes. One possible explanation is that SNPs with rare alleles are abundant in the genome, but can be detected only in large samples. Another reason that could influence the fluctuation of the correlation is the quality of the recombination maps. Although SNP data varied greatly from year to year, the recombination sources remained the same: broad scale and low resolution. For the SNP data from 2006 and 2008, a better resolution quality map was used (HapMap), which showed slightly negative or slightly positive (in the case of chromosome 11) correlation. In other words, high resolution and fine scale data demonstrate no correlation. This observation supports the hypothesis from Spencer et al. 2006 showing that recombination influences genetic diversity at the hotspot level, since both diversity and recombination were positively correlated with sequence composition.. Positive correlation can be observed with the broad scale data, but is very weak or absent at a finer scale, as well as in cases when correlation was calculated for the separate chromosomes. The broad scale association between diversity and recombination can be explained by the fact that both genomic features (diversity and recombination) are associated with base composition (such as GC content) and, thus, correlate with each other through base composition biases. Fine scale recombination events are known to cluster into narrow hotspots. They evolve fast and according to the hotspot paradox, “die” fast on the evolutionary scale (Spencer et al. 2006). Thus, they cannot influence the substitution rate and do not change long-term molecular evolution.

4.2 Analysis of experimental cross over data

The experimental aim of this study was to estimate the cross over rate in four Encode regions on human chromosome 11 by high resolution sperm typing and compare the results with those from LD-based maps. The regions were selected by their chromosomal positions (telomeric, centromeric) and predicted recombination rate (high, low). All Encode regions were divided into 100 kb intervals with informative SNP markers flanking each interval. 24 SNP markers were genotyped at each locus, determining the phase of each of the two markers and categorizing it as having a recombinant or non-recombinant haplotype.

One of the intriguing findings regarding our experimental observations is the lower cross over rate identified in all four Encode regions (0.12 cM/Mb), in contrast to the expected 1.875 cM/Mb (deCode sex-averaged).

A few studies (Cullen et al. 2002, Clark et al. 2007) reported decreased estimates by an order of magnitude for cross overs detected by sperm typing, in contrast to rates from patterns of linkage disequilibrium (LD).

One explanation of this phenomenon could be the significant individual variation in the recombination rates (Coop and Przeworski, 2007). Some sperm typing reports (Jeffreys and Neumann 2002, Carrington and Cullen, 2004) have demonstrated variation in fine scale rates, as well as in the intensities of individual hotspots. According to this assumption, our observation could be the individual genomic characteristic of a particular donor. The size of the entire screened area (2.5 Mb) and location of separate Encode regions are ambiguous factors for this hypothesis. The entire area is large enough to show both differences and similarities in experimental rates compared with LD data.

Another explanation could be sex-specific differences in recombination rates. Males have been confirmed to have lower cross over rates compared to females (Coop and Przeworski, 2007). To test this hypothesis, we obtained the average male-specific rate from the deCode recombination map for the four Encode regions (1.9 cM/Mb), which appeared to be even higher compared to the sex-average (1.875 cM/Mb) for the same Encode regions. However, the distribution of the male cross over rates is uneven, from 0 cM/Mb in the centromeric regions to 3.0 and 4.7 cM/Mb in the telomeric regions, when looking across chromosome 11. Such a distribution supports the hypothesis that females have higher cross over rates in centromeric regions and lower rates in telomeric regions, whereas the converse is true for males. Our data is in line with this hypothesis, since three experimentally obtained cross over events were detected in telomeric regions on both chromosome arms. In the centromeric ENr332 and ENm003, only gene conversions were identified.

The lower recombination rate 0.12 cM/Mb compared to the expected theoretical rate 2.5 cM/Mb (based on the assumption that at least 1 % of recombinants are expected in 1 Mb, see chapter 3.3) can partially be explained by the relatively small sample size; 1000 cells did not produce a resolution high enough to detect fine-scale rates, although the markers were spaced to identify at least 1% of recombinant haplotypes per 1 Mb. In case both conversion and cross over events are considered the rate is 0.4 cM/Mb.

What, then, is considered as a “recombination rate”? Is it the initiation of the double strand break (DSB), or the resolution of the Holliday structure and number of cross over events that actually form? The current hypothesis (Kauppi, Jeffreys and Keeney, 2004) is that meiotic cells make a large excess of DSBs relative to the number of cross overs that actually form. Since every DSB must be repaired, this means all excess breaks must be processed and be resolved in a non-cross over configuration (such as gene conversion). Thus, local variation in crossover frequency could reflect not only variation in DSB frequency, but also variation in how probable it is that a DSB will give rise to a cross over instead of non-cross over outcome (Kauppi, Jeffreys and Keeney, 2004).

4.3 Analysis of gene conversion events

Since the initial aim of this study was to obtain experimental cross over rates and compare them with diversity levels, the identification an excess of gene conversion events in recombinant haplotypes was quite surprising. Nevertheless, all conversion events were confirmed with at least two heterozygous markers. The conversion minimum tract length was estimated as the distance between two adjacent SNPs which have the converted haplotype.

Another finding is that conversion events outnumber cross overs by two-fold. This is especially interesting when considering conversion events that were identified around markers flanking 100kb intervals. For our purposes, we considered only those conversions for which at least two informative markers were identified to confirm the event as a conversion. According to our observations, heterozygous alleles are more rare compared to non-heterozygous SNPs; as such, only confirmed conversions with an average length larger than 300 bp were considered in this study. The mean length of conversion events was calculated before as 55-290 bp (Jeffreys and May, 2004). It is likely that in our study most of the events are missed due to the unavailability of markers to confirm them. In addition, the data of all conversion events within a 100kb interval are missing as well; therefore, conversion rates can be only calculated for the small fraction (25-30 kb per region), where markers were placed.

The regulation of gene conversion length is another controversial aspect. It remains unknown whether short–tract gene conversion and long-tract gene conversion arise by distinct mechanisms in mammalian cells. Only in one of the studies did the Rad51 paralog Rad51C show itself to be involved in the control of homologous recombination (Innen et. al, 2006). In the absence of Rad51C in hamster cells, gene conversion seemed to be biased for long tracts. The authors proposed a hypothesis that long conversion tracts (at least in the cells they studied) are the products of a distinct mechanism of a repair synthesis (such as recombination dependent replication in *E. coli*), whereas the short tracts were the outcome of conventional gene conversion.

Nevertheless, the difference between cross over and conversion frequency is in agreement with current observations within the recombination hotspots. Sperm typing of “super” hotspots revealed increased conversion rates when compared to cross over intensity at an average ratio of 10:1 (Jeffreys and May, 2004). A plausible explanation for the data could be that conversion and cross over are initiated at the same sites. From this perspective, it is possible to consider gene conversion as a way of Holliday structure resolution at the sites where the exchange of alleles is not favorable, but still some information transfer to increase allele variability is required. These sites primarily include coding sequences, which have to maintain peptide conservation, and where linkage distortion could lead to undesirable consequences. All of the conversion events identified in this study are either intergenic (as in case of conversions in ENm003) or located in the introns of genes.

The contribution of gene conversion to formation of local haplotype patterns is considered to be more significant than the reciprocal recombination contribution (Jeffreys and May, 2004).

Yet another reason to favor gene conversion over crossover could be a shorter conversion tract transmitted from one DNA molecule to another. However, our results do not agree with this, since most of identified tracts longer than 1 kb. One reason for this disagreement could be our consideration of only a limited number of conversions that were confirmed by heterozygous markers. Thus, the data are more in line with our observations. At present, there have been only a limited number of gene conversion long tracts reported (Losekoot et al, 1997), and they come from rare germline inversions or double recombination events. In light of this, the conversion tract of almost 10kb long, which was identified in the ENm003 region, is a candidate for a double recombination event, although no direct evidence was found to confirm the double recombination between flanking markers specific for this tract.

The average conversion tract length identified was 1.5 kb (with the upper bound of almost 10 kb and the lower below 500 bp), which means that our data exceeds by at least a 1 kb the current experimental observations of other sperm typing investigators. The explanation for this phenomenon could lie in the following: other investigators attempted to look at conversion activity inside hotpots and, thus, concentrated specifically on observations within 1-2 kb intervals. Our analysis covers 2.5 Mb of chromosome 11 in total and we did not target specific hotspots, but one of the goals of this study was to track the conversion length identified only at the sites of first set of markers spaced 100 kb from each other. One of the reasons why converted tracts are shorter in the hotspots is because of high recombination initiation frequency within “super” hotspots, converted regions become smaller as a hotspot evolves. Consequently, nearly conserved coding sequences where non-frequent resolution of the recombination initiation will favor information transmission from one allele to the other, rather than the breakage of linkage association and rejoining between molecules -- the conversion tract could be large enough to contribute sufficiently to some degree of allele diversity.

At a broad scale, recombination rates are known to increase with gene density due to the proposed association between recombination and transcription (Petes, 2001). On a finer scale, LD studies have shown that hotspots and coding sequences do not prefer to share the same location sites (Coop and Przeworski, 2007): the rates decrease near the genes. Besides the fact that population data can be influenced by selection and, thus, are ambiguous to interpret, another explanation could be that broad scale rates count only recombination initiation sites without resolution of Holliday structures. Therefore, the increased conversion rate accompanying cross over events can significantly increase the rate near coding sequences and influence the broad scale estimates.

Another reason for increased gene conversion rates compared to cross over rates could be the distorted transmission ratio (Webb et. al., 2008) of haplotypes. The over transmission of haplotypes is not surprising: it was previously shown in a few human and mouse hotspots (Jeffreys et al., 2005). Our results could be considered as support for this theory in the sense that we revealed the prevalence of one haplotype over the other. First, analysis of the two reference haplotypes revealed that one of them (“black”, Fig. 21) consisted primarily of major alleles of heterozygous markers used for initial screening to detect cross over and conversion events. Thus, the other haplotype consisted primarily of minor alleles. Interestingly, five out of seven sperm cells with indentified conversion events had a “white” haplotype with a “black” allele transmitted as a ”one-site exchange”, whereas only two had the “black” haplotype with a transmitted “white” allele. Also, in all three sperm cells with cross over events, the prevailing haplotype was the “white” one.

Such a distortion could lend credence to the existing variation in cross over events between males. Such prevalence of one allele over the other could result from transmission of information from one allele to the other more frequently compared to cross over. In the context of our data, this means that the “black” haplotype could be suppressed, with systematic over-transmission of markers into cross over progeny (Webb et. al., 2008).

Discussion

One reason for the segregating distortion observed is the subtle disparity between haplotypes in the recombination initiation rates (Webb et. al., 2008). Alternatively, subtle mismatch repair biases could lead to biased gene conversion in cross overs. The alternative explanation could reflect why so many "black"/major alleles are transmitted onto "white" haplotypes in the identified conversion tracts.

Another aspect is whether gene conversion is GC biased or not. According to Galtier (2003), gene conversion is GC biased not only in humans, but also in amphibians and birds. This is connected to fact that GC→AT mutations are more frequent than AT→GC mutations. We investigated GC content within all identified conversion tracts. Except for the shortest tract in ENr332, there was no significant increase in the GC content. The longer conversion apparently had lower GC content, whereas in some shorter tracts, GC content was up to 53%. In the shortest tract, GC content had the highest rate at 62%. Since it is located in the MAP4K2, which covers both an intron and an exon, and as Galtier reported, he used histone genes for GC content comparisons, the fact that this conversion tract covers the coding sequence in contrast to other tracts could have contributed to the increase in CG content. Half of the identified tract was found in the exon.

CpG dinucleotide expected and observed fractions were also investigated for all conversion tracts. The observed fraction was always (table1) lower when compared to the expected. Since CpG islands are known to be concentrated in the upstream regions of genes and to be anti-correlated with DNA methylation, we hypothesize that a decrease in CpG dinucleotides could be associated with extensive methylation of the DNA region where the conversion tracts are. However, two studies contradict our hypothesis. Hoegstrand and Boehme (1999a) detected gene conversion in the major histocompatibility complex genes, which were associated with CpG-rich regions. Matyasek et al. (2003) found increased gene conversion in under-methylated DNA regions in *Nicotiana rustica.*

Although we observed very few cross over events, they were identified in the regions with a density of more than four SNPs per 1 kb (ENm009, ENr312). In contrast, conversion events were found in the regions with lower SNP density (2-3 SNPs per kb). Most of the converted tracts were intergenic or in introns – close to coding sequence with expected decrease of SNP density. Based on the scarse amount of data, there is little can be said about correlation between SNP density and recombination rates.

We were also unable to find any association between conversions or crossovers and a common sequence associated motif; neither a common motif family, nor a specific core motif, which is very rare in our considered Encode regions (maximum four).

Summary of the results

We demonstrated that the positive correlation between SNP density and recombination rate observed with broad scale data is the result of an association of both features with base composition, in contrast to the observation that with respect to fine scale data, this correlation is close to zero.

The low experimental cross over rate is either due to sex specific variation or due to individual variation in a particular donor.

The cross over rate can reflect the variation of the cross over initiation that can be later resolved into reciprocal and non-reciprocal outcomes. Based on this assumption, our rate estimate could very well exceed 0.12 cM/Mb if conversion events are also taken into account.

We confirm the higher number of gene conversion events compared to cross overs (seven versus three). These conversion events were identified at the sites of marker location. The 100kb intervals are not taken into consideration as well as those conversion events which were impossible to confirm with other heterozygous markers. Thus, the ratio between conversion and cross over should be much higher for conversions.

The segregation distortion could be the reason for the higher transmission of major alleles from one of the haplotypes (“black” Fig. 20) to be transmitted to the “white” haplotype, with greater prevalence for minor alleles.

The identified conversion tracts are not CG biased, except a single tract with the size below 500bp. Underrepresentation of CpG islands within the conversion tracts could be suggestive of extensive DNA methylation within these regions, but as of yet, this is mostly conjecture as at this point, no other studies have either confirmed or refuted such a hypothesis.

Lastly, we observed neither a correlation between the SNP density, nor an association with common sequence motifs defining hotspots.

Implication of the study

The theoretical and experimental results in the context of current knowledge regarding the correlation between nucleotide diversity and recombination rates, and in the context of recombination and gene conversion activity provides better insights into the inherent variation of the genomic landscape.

Outlook

1. Theoretical approach: to study the association of diversity and recombination with base composition.
2. Experimental approach: to avoid individual bias in cross over and gene conversion; another donor of the same origin (European) must be genotyped to confirm the obtained results. For differences in cross over and gene conversion at a population level, donors from different populations (non-european) must be genotyped.

Reference List

(2003). The International HapMap Consortium. The International HapMap Project Nature *426*, 789-796.

Aljanabi,S.M. and Martinez,I. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. Nucleic Acids Res. *25*, 4692-4693.

Altshuler,D., Brooks,L.D., Chakravarti,A., Collins,F.S., Daly,M.J., Donnelly, P. (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

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.

Auton,A. and McVean,G. (2007). Recombination rate estimation in the presence of hotspots. Genome Res. *17*, 1219-1227.

Begun,D.J. and Aquadro,C.F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature *356*, 519-520.

Bell PA, Chaturvedi S, Gelfand CA, Huang CY, Kochersperger M, Kopla R, Modica F, Pohl M, Varde S, Zhao R, Zhao X, Boyce-Jacino MT (2002) SNPstream UHT: ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery. Biotechniques *30*: S70-770

Brooker,R.J. (1999). Genetics. Principles and Analyses. Addison Wesely Longman Inc..

Burt,A. (2000). Perspective: sex, recombination, and the efficacy of selection--was Weismann right? Evolution *54*, 337-351.

Carrington,M. and Cullen,M. (2004). Justified chauvinism: advances in defining meiotic recombination through sperm typing. Trends Genet. *20*, 196-205.

Clark,V.J., Ptak,S.E., Tiemann,I., Qian,Y., Coop,G., Stone,A.C., Przeworski,M., Arnheim,N., and Di,R.A. (2007). Combining sperm typing and linkage disequilibrium analyses reveals differences in selective pressures or recombination rates across human populations. Genetics *175*, 795-804.

Conrad,D.F., Jakobsson,M., Coop,G., Wen,X., Wall,J.D., Rosenberg,N.A., and Pritchard,J.K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet. *38*, 1251-1260.

Coop,G. (2005). Can a genome change its (hot)spots? Trends Ecol. Evol. *20*, 643-645.

Reference list

Coop,G. and Przeworski,M. (2007). An evolutionary view of human recombination. Nat. Rev. Genet. *8*, 23-34.

Cullen,M., Perfetto,S.P., Klitz,W., Nelson,G., and Carrington,M. (2002). High-resolution patterns of meiotic recombination across the human major histocompatibility complex. Am. J. Hum. Genet. *71*, 759-776.

Dib,C., Faure,S., Fizames,C., Samson,D., Drouot,N., Vignal,A., Millasseau,P., Marc,S., Hazan,J., Seboun,E., Lathrop,M., Gyapay,G., Morissette,J., and Weissenbach,J. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature *380*, 152-154.

Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P., Leal,S.M., Pasternak,S., Wheeler,D.A., Willis,T.D., Yu,F., Yang,H., Zeng,C., Gao,Y., Hu,H., Hu,W., Li,C., Lin,W., Liu,S., Pan,H., Tang,X., Wang,J., Wang,W., Yu,J., Zhang,B., Zhang,Q., Zhao,H., Zhao,H., Zhou,J., Gabriel,S.B., Barry,R., Blumenstiel,B., Camargo,A., DeFelice,M., Faggart,M., Goyette,M., Gupta,S., Moore,J., Nguyen,H., Onofrio,R.C., Parkin,M., Roy,J., Stahl,E., Winchester,E., Ziaugra,L., Altshuler,D., Shen,Y., Yao,Z., Huang,W., Chu,X., He,Y., Jin,L., Liu,Y., Shen,Y., Sun,W., Wang,H., Wang,Y., Wang,Y., Xiong,X., Xu,L., Waye,M.M., Tsui,S.K., Xue,H., Wong,J.T., Galver,L.M., Fan,J.B., Gunderson,K., Murray,S.S., Oliphant,A.R., Chee,M.S., Montpetit,A., Chagnon,F., Ferretti,V., Leboeuf,M., Olivier,J.F., Phillips,M.S., Roumy,S., Sallee,C., Verner,A., Hudson,T.J., Kwok,P.Y., Cai,D., Koboldt,D.C., Miller,R.D., Pawlikowska,L., Taillon-Miller,P., Xiao,M., Tsui,L.C., Mak,W., Song,Y.Q., Tam,P.K., Nakamura,Y., Kawaguchi,T., Kitamoto,T., Morizono,T., Nagashima,A., Ohnishi,Y., Sekine,A., Tanaka,T., Tsunoda,T., Deloukas,P., Bird,C.P., Delgado,M., Dermitzakis,E.T., Gwilliam,R., Hunt,S., Morrison,J., Powell,D., Stranger,B.E., Whittaker,P., Bentley,D.R., Daly,M.J., de Bakker,P.I., Barrett,J., Chretien,Y.R., Maller,J., McCarroll,S., Patterson,N., Pe'er,I., Price,A., Purcell,S., Richter,D.J., Sabeti,P., Saxena,R., Schaffner,S.F., Sham,P.C., Varilly,P., Altshuler,D., Stein,L.D., Krishnan,L., Smith,A.V., Tello-Ruiz,M.K., Thorisson,G.A., Chakravarti,A., Chen,P.E., Cutler,D.J., Kashuk,C.S., Lin,S., Abecasis,G.R., Guan,W., Li,Y., Munro,H.M., Qin,Z.S., Thomas,D.J., McVean,G., Auton,A., Bottolo,L., Cardin,N., Eyheramendy,S., Freeman,C., Marchini,J., Myers,S., Spencer,C., Stephens,M., Donnelly,P., Cardon,L.R., Clarke,G., Evans,D.M., Morris,A.P., Weir,B.S., Tsunoda,T., Mullikin,J.C., Sherry,S.T., Feolo,M., Skol,A., Zhang,H., Zeng,C., Zhao,H., Matsuda,I., Fukushima,Y., Macer,D.R., Suda,E., Rotimi,C.N., Adebamowo,C.A., Ajayi,I., Aniagwu,T., Marshall,P.A., Nkwodimmah,C., Royal,C.D., Leppert,M.F., Dixon,M., Peiffer,A., Qiu,R., Kent,A., Kato,K., Niikawa,N., Adewole,I.F., Knoppers,B.M., Foster,M.W., Clayton,E.W., Watkin,J., Gibbs,R.A., Belmont,J.W., Muzny,D., Nazareth,L., Sodergren,E., Weinstock,G.M., Wheeler,D.A., Yakub,I., Gabriel,S.B., Onofrio,R.C., Richter,D.J., Ziaugra,L., Birren,B.W., Daly,M.J., Altshuler,D., Wilson,R.K., Fulton,L.L., Rogers,J., Burton,J., Carter,N.P., Clee,C.M., Griffiths,M., Jones,M.C., McLay,K., Plumb,R.W., Ross,M.T., Sims,S.K., Willey,D.L., Chen,Z., Han,H., Kang,L., Godbout,M., Wallenburg,J.C., L'Archeveque,P., Bellemare,G., Saeki,K., Wang,H., An,D., Fu,H., Li,Q., Wang,Z., Wang,R., Holden,A.L.,

Reference list

Brooks,L.D., McEwen,J.E., Guyer,M.S., Wang,V.O., Peterson,J.L., Shi,M., Spiegel,J., Sung,L.M., Zacharia,L.F., Collins,F.S., Kennedy,K., Jamieson,R., and Stewart,J. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

Gabriel,S.B., Schaffner,S.F., Nguyen,H., Moore,J.M., Roy,J., Blumenstiel,B., Higgins,J., DeFelice,M., Lochner,A., Faggart,M., Liu-Cordero,S.N., Rotimi,C., Adeyemo,A., Cooper,R., Ward,R., Lander,E.S., Daly,M.J., and Altshuler,D. (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225-2229.

Gay,J., Myers,S., and McVean,G. (2007). Estimating meiotic gene conversion rates from population genetic data. Genetics *177*, 881-894.

Hellmann,I., Ebersberger,I., Ptak,S.E., Paabo,S., and Przeworski,M. (2003). A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. *72*, 1527-1535.

Hogstrand,K. and Bohme,J. (1999). Gene conversion of major histocompatibility complex genes is associated with CpG-rich regions. Immunogenetics *49*, 446-455.

Hogstrand,K. and Bohme,J. (1999). Gene conversion can create new MHC alleles. Immunol. Rev. *167*, 305-317.

Innan,H. and Nordborg,M. (2003). The extent of linkage disequilibrium and haplotype sharing around a polymorphic site. Genetics *165*, 437-444.

Innan,H. (2003). The coalescent and infinite-site model of a small multigene family. Genetics *163*, 803-810.

Innan,H., Padhukasahasram,B., and Nordborg,M. (2003). The pattern of polymorphism on human chromosome 21. Genome Res. *13*, 1158-1168.

Innan,H. (2003). A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. Proc. Natl. Acad. Sci. USA *100*, 8793-8798.

Innan,H. and Stephan,W. (2003). Distinguishing the hitchhiking and background selection models. Genetics *165*, 2307-2312.

Ireland,J., Carlton,V.E., Falkowski,M., Moorhead,M., Tran,K., Useche,F., Hardenbol,P., Erbilgin,A., Fitzgerald,R., Willis,T.D., and Faham,M. (2006). Large-scale characterization of public database SNPs causing non-synonymous changes in three ethnic groups. Hum. Genet. *119*, 75-83.

Reference list

Isobe,T., Yoshino,M., Mizuno,K., Lindahl,K.F., Koide,T., Gaudieri,S., Gojobori,T., and Shiroishi,T. (2002). Molecular characterization of the Pb recombination hotspot in the mouse major histocompatibility complex class II region. Genomics *80*, 229-235.

Jeffreys,A.J., Craig,I.W., and Francke,U. (1979). Localisation of the G gamma-, A gamma-, delta- and beta-globin genes on the short arm of human chromosome 11. Nature *281*, 606-608.

Jeffreys,A.J. and Barrie,P.A. (1981). Sequence variation and evolution of nuclear DNA in man and the primates. Philos. Trans. R. Soc. Lond B Biol. Sci. *292*, 133-142.

Jeffreys,A.J., Tamaki,K., MacLeod,A., Monckton,D.G., Neil,D.L., and Armour,J.A. (1994). Complex gene conversion events in germline mutation at human minisatellites. Nat. Genet. *6*, 136-145.

Jeffreys,A.J., Murray,J., and Neumann,R. (1998). High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. Mol. Cell *2*, 267-273.

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.

Jeffreys,A.J. and Neumann,R. (2002). Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nat. Genet. *31*, 267-271.

Jeffreys,A.J., Holloway,J.K., Kauppi,L., May,C.A., Neumann,R., Slingsby,M.T., and Webb,A.J. (2004). Meiotic recombination hot spots and human DNA diversity. Philos. Trans. R. Soc. Lond B Biol. Sci. *359*, 141-152.

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., Neumann,R., Panayi,M., Myers,S., and Donnelly,P. (2005). Human recombination hot spots hidden in regions of strong marker association. Nat. Genet. *37*, 601-606.

Jeffreys,A.J. and Neumann,R. (2005). Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. Hum. Mol. Genet. *14*, 2277-2287.

Jeffreys,A.J. and Neumann,R. (2009). The rise and fall of a human recombination hot spot. Nat. Genet. *41*, 625-629.

Jiang,Z., Zhang,X., Deka,R., and Jin,L. (2005). Genome amplification of single sperm using multiple displacement amplification. Nucleic Acids Res. *33*, e91.

Kauppi,L., Sajantila,A., and Jeffreys,A.J. (2003). Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. Hum. Mol. Genet. *12*, 33-40.

Reference list

Kauppi,L., Jeffreys,A.J., and Keeney,S. (2004). Where the crossovers are: recombination distributions in mammals. Nat. Rev. Genet. *5*, 413-424.

Kauppi,L., Stumpf,M.P., and Jeffreys,A.J. (2005). Localized breakdown in linkage disequilibrium does not always predict sperm crossover hot spots in the human MHC class II region. Genomics *86*, 13-24.

Kong,A., Gudbjartsson,D.F., Sainz,J., Jonsdottir,G.M., Gudjonsson,S.A., Richardsson,B., Sigurdardottir,S., Barnard,J., Hallbeck,B., Masson,G., Shlien,A., Palsson,S.T., Frigge,M.L., Thorgeirsson,T.E., Gulcher,J.R., and Stefansson,K. (2002). A high-resolution recombination map of the human genome. Nat. Genet. *31*, 241-247.

Lercher,M.J., Williams,E.J., and Hurst,L.D. (2001). Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. Mol. Biol. Evol. *18*, 2032-2039.

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

Li,X. and Heyer,W.D. (2008). Homologous recombination in DNA repair and DNA damage tolerance. Cell Res. *18*, 99-113.

Linhart,W., Briem,D., Schmitz,N.D., Priemel,M., Lehmann,W., and Rueger,J.M. (2003). Treatment of metaphyseal bone defects after fractures of the distal radius. Medium-term results using a calcium-phosphate cement (BIOBON). Unfallchirurg *106*, 618-624.

Losekoot,M., Hoogendoorn,E., Olmer,R., Jansen,C.C., Oosterwijk,J.C., van den Ouweland,A.M., Halley,D.J., Warren,S.T., Willemsen,R., Oostra,B.A., and Bakker,E. (1997). Prenatal diagnosis of the fragile X syndrome: loss of mutation owing to a double recombinant or gene conversion event at the FMR1 locus. J. Med. Genet. *34*, 924-926.

Lynn,A., Ashley,T., and Hassold,T. (2004). Variation in human meiotic recombination. Annu. Rev. Genomics Hum. Genet. *5*, 317-349.

Matyasek,R., Lim,K.Y., Kovarik,A., and Leitch,A.R. (2003). Ribosomal DNA evolution and gene conversion in Nicotiana rustica. Heredity *91*, 268-275.

McVean,G. (2007). The structure of linkage disequilibrium around a selective sweep. Genetics *175*, 1395-1406.

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.

Reference list

Nachman,M.W., Bauer,V.L., Crowell,S.L., and Aquadro,C.F. (1998). DNA variability and recombination rates at X-linked loci in humans. Genetics *150*, 1133-1141.

Neumann,R. and Jeffreys,A.J. (2006). Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation. Hum. Mol. Genet. *15*, 1401-1411.

Padhukasahasram,B., Wall,J.D., Marjoram,P., and Nordborg,M. (2006). Estimating recombination rates from single-nucleotide polymorphisms using summary statistics. Genetics *174*, 1517-1528.

Petes,T.D. (2001). Meiotic recombination hot spots and cold spots. Nat. Rev. Genet. *2*, 360-369.

Petkovski,E., Keyser-Tracqui,C., Hienne,R., and Ludes,B. (2005). SNPs and MALDI-TOF MS: tools for DNA typing in forensic paternity testing and anthropology. J. Forensic Sci. *50*, 535-541.

Phillips,M.S., Lawrence,R., Sachidanandam,R., Morris,A.P., Balding,D.J., Donaldson,M.A., Studebaker,J.F., Ankener,W.M., Alfisi,S.V., Kuo,F.S., Camisa,A.L., Pazorov,V., Scott,K.E., Carey,B.J., Faith,J., Katari,G., Bhatti,H.A., Cyr,J.M., Derohannessian,V., Elosua,C., Forman,A.M., Grecco,N.M., Hock,C.R., Kuebler,J.M., Lathrop,J.A., Mockler,M.A., Nachtman,E.P., Restine,S.L., Varde,S.A., Hozza,M.J., Gelfand,C.A., Broxholme,J., Abecasis,G.R., Boyce-Jacino,M.T., and Cardon,L.R. (2003). Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat. Genet. *33*, 382-387.

Przeworski,M., Coop,G., and Wall,J.D. (2005). The signature of positive selection on standing genetic variation. Evolution *59*, 2312-2323.

Ptak,S.E., Hinds,D.A., Koehler,K., Nickel,B., Patil,N., Ballinger,D.G., Przeworski,M., Frazer,K.A., and Paabo,S. (2005). Fine-scale recombination patterns differ between chimpanzees and humans. Nat. Genet. *37*, 429-434.

Reich,D.E., Schaffner,S.F., Daly,M.J., McVean,G., Mullikin,J.C., Higgins,J.M., Richter,D.J., Lander,E.S., and Altshuler,D. (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. Nat. Genet. *32*, 135-142.

Shapiro H.M., (2003). Practical flow cytometry. Fourth edition. Wiley Liss Edition.

Serre,D., Nadon,R., and Hudson,T.J. (2005). Large-scale recombination rate patterns are conserved among human populations. Genome Res. *15*, 1547-1552.

Spencer,C.C., Deloukas,P., Hunt,S., Mullikin,J., Myers,S., Silverman,B., Donnelly,P., Bentley,D., and McVean,G. (2006). The influence of recombination on human genetic diversity. PLoS. Genet. *2*, e148.

Syvänen,A.C. (2005). Towards genome-wide SNP genotyping. Nat. Genet. *37* Suppl: S5-10.

Reference list

Szostak,J.W. (1983). Replication and resolution of telomeres in yeast. Cold Spring Harb. Symp. Quant. Biol. *47 Pt 2*, 1187-1194.

Taylor,T.D., Noguchi,H., Totoki,Y., Toyoda,A., Kuroki,Y., Dewar,K., Lloyd,C., Itoh,T., Takeda,T., Kim,D.W., She,X., Barlow,K.F., Bloom,T., Bruford,E., Chang,J.L., Cuomo,C.A., Eichler,E., FitzGerald,M.G., Jaffe,D.B., LaButti,K., Nicol,R., Park,H.S., Seaman,C., Sougnez,C., Yang,X., Zimmer,A.R., Zody,M.C., Birren,B.W., Nusbaum,C., Fujiyama,A., Hattori,M., Rogers,J., Lander,E.S., and Sakaki,Y. (2006). Human chromosome 11 DNA sequence and analysis including novel gene identification. Nature *440*, 497-500.

Wall,D.P., Hirsh,A.E., Fraser,H.B., Kumm,J., Giaever,G., Eisen,M.B., and Feldman,M.W. (2005). Functional genomic analysis of the rates of protein evolution. Proc. Natl. Acad. Sci. USA *102*, 5483-5488.

Wang,P., Lee,J.W., Yu,Y., Turner,K., Zou,Y., Jackson-Cook,C.K., and Povirk,L.F. (2002). Gene rearrangements induced by the DNA double-strand cleaving agent neocarzinostatin: conservative non-homologous reciprocal exchanges in an otherwise stable genome. Nucleic Acids Res. *30*, 2639-2646.

Wang,W., Thornton,K., Berry,A., and Long,M. (2002). Nucleotide variation along the Drosophila melanogaster fourth chromosome. Science *295*, 134-137.

Webb,A.J., Berg,I.L., and Jeffreys,A. (2008). Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. Proc. Natl. Acad. Sci. USA *105*, 10471-10476.

Weber,J.L. (2002). The Iceland map. Nat. Genet. *31*, 225-226.

Willer,C.J., Scott,L.J., Bonnycastle,L.L., Jackson,A.U., Chines,P., Pruim,R., Bark,C.W., Tsai,Y.Y., Pugh,E.W., Doheny,K.F., Kinnunen,L., Mohlke,K.L., Valle,T.T., Bergman,R.N., Tuomilehto,J., Collins,F.S., and Boehnke,M. (2006). Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. Genet. Epidemiol. *30*, 180-190.

Wills,A., Christopher,J.D (1980). Genetic Variability. New York: Oxford University Press.

Yu,A., Zhao,C., Fan,Y., Jang,W., Mungall,A.J., Deloukas,P., Olsen,A., Doggett,N.A., Ghebranious,N., Broman,K.W., and Weber,J.L. (2001). Comparison of human genetic and sequence-based physical maps. Nature *409*, 951-953.

Zheng,D. and Gerstein,M.B. (2006). A computational approach for identifying pseudogenes in the ENCODE regions. Genome Biol. *7 Suppl 1*, S13-10.

Appendix

Table 1. Haplotype screening loci for SNPstream analysis

Region	Marker ID	Major Allele	Minor Allele	Forward primer	Reverse primer	Single base extension primer
ENm009	Rs 10836571	G	A	cccttgaggacttcgagaga	cacatgaagacttccctcct	gcgcccttcttcgagaga
ENm009	Rs12273277	G	A	gttccattacctcgcgggcg	ttacgaggacatcgactga	aaattttggagttcgcgcg
ENm009	Rs 7933046	A	G	aactcggagtaccccgagc	cggtcaagcactccgaaa	gacttgatcccccaagag
ENm009	Rs 7120612	A	G	tttgaaagcactcccatatac	caaggaggacttcgcgcgc	agcctgagtactacgactg
ENm009	Rs 2855039	G	A	cggattgattacggcgagat	agctagaagacatcgcccga	cgcggatgacttctcagta
ENm009	Rs 4462380	A	G	aaaccatgaggattggcctg	tgcaagaggacctgttgactg	aaggttacgacaacgttg
ENm009	Rs 435810	A	G	gtccacgtcattcgtcagtca	tcattgcatacttcgcctgagg	ctccgattacggactccag
ENm009	Rs2047459	A	G	acctgttgactaccgtattcg	aaatggcctactctggactca	atgcatgagacatggacga
ENm009	Rs 12787013	A	G	cttgactacttcgagagagca	actgacatgaacgatcagtac	cgtacagtacagtccagta
ENm009	Rs 2133266	A	C	ggatagcgtactatcgagaga	agtcagtatagcagacacagt	ttgcactgcattcccagtgc
ENm009	Rs 1083658	T	A	gctgacgtgagcccagtgcc	caagtgccgatgatccgatga	tgacctatcgagagatgca
ENm009	Rs 1433917	T	C	gctccagaatactaccgtaac	aaacttcaagtcgggaattcag	cgatcatgactgcagtcgt
ENm009	Rs 11035066	C	T	cagtgacgtcttgcaccagt	aagctcctcgattgacgatgca	gcagtcattatcagtctcg
ENm009	Rs 2499948	C	T	gccagtaccatacttgcaac	cgtagcagtcgatgcagtccc	cgtgcacttcgggaactt
ENm009	Rs 2011051	C	A	tctacgatgcgggtactccga	ccagacgtgaactacccgtg	cagtagccaaactatcac
ENm009	Rs 2723375	A	G	aacgtgacggtgccagtcgt	tgacgtacgctacgtatcagcc	cacacatccgtagcagttt
ENm009	Rs 11821912	A	T	gacgtagcgttcccgtcacag	cagtgaccgtagcgatgacag	tcacgtgtttcagtgatcg
ENm009	Rs 11038414	A	T	acatccgccgacttgacccga	tgtggatatccagtccgtgga	aacgtgacagtgccgtga
ENm003	Rs 10790152	A	G	agagatgattggccgttgcag	cgtgccacagtgcgtgcgtgt	cgtgcagtaaaccgtgac
ENm003	Rs 17518841	G	A	gcctagcgtcgatgcccgtg	ggtgactccgtgcagtgcgtg	tatatgtgacgtgacttgc
ENm003	Rs 675	T	A	ccagtgacgccagtgaacgt	cacagtgccgtgtgtacacgtg	ttgatgcgactgatgccgt
ENm003	Rs 10502221	C	T	cacgtgacgtgtgcacccgga	cgcctgagtcacactggtggc	accgtgatccgatgatga
ENm003	Rs 11216267	C	T	gatgcagatcccgtgacacag	ggagtccgtcgcttgctggcc	cgtaggtgtccgttggag
ENm003	Rs 7116825	A	C	ggcagtcgcacgtgcgccccg	cacgtgacgtgttgcagtgcg	tatataacttgcgtgacgt

ENm003	Rs 1145196	A	G	atgacgtgacccgtgaccgtg	gtgacgctgattcgtgaccgat	cgatcagttccgtagcca
ENm003	Rs 1263055	T	C	gaccacgtactccagtgtatga	cgatgcacagtgccgatgccc	ttgtgacttgcagtgccac
ENm003	Rs 6589582	C	A	ccgcgctgacctgaggacgttt	tctcacttgacccgtgcgtgttg	atgaccgttcttcagtggc
ENm003	Rs 5678626	A	G	agagttgtccacttccgatgaa	cgtagcaaatggtagacgttaa	cgatgagacgtgaaagg
ENr332	Rs 11601686	C	T	gtgcacgtgaaccgtggcgtg	tatgctcgtgacgtggtcaaaa	cgtgaacgacttccgttac
ENr332	Rs 11231816	G	A	caacacactgcacaggtgccc	cgacgtgaccacgttgacata	cgtgactgaccgtacgtg
ENr332	Rs 500531	T	G	cgtgactagcccacagagatga	gatcactgaacgtcccgtgaac	ttcaacacagtgagagcct
ENr332	Rs 10897526	T	C	gccactgactcacgtgtgcttac	agtgcccgttaactataccacgt	ccgtgactgtgtttccacga
ENr332	Rs 618006	T	C	tagtgcacgagacacttggtgca	ttgtgcacccggtgcactgaccc	cccaccactctgtgactga
ENr332	Rs 1601686	C	T	gcttaacgtgaaccacttgatag	tgtcatcatatcaggtgcccctg	acacgtgcattgcatagttc
ENr332	Rs 4930423	T	C	gcacttccacggtgtatagcggc	cactgtgcatagctccgtgcac	cggggtgactagtctgcaa
ENr332	Rs 7124676	A	G	cttgtgacatgactcccggtaag	tgtgcaactagcgcgctgactta	gtgcaacgtcccgtgaatg
ENr332	Rs 2666559	A	C	gcattgtgccaacccgtgggct	aactggataccgtgtacgtactt	acgtaacgtagcgtacgta
ENr332	Rs 660543	A	G	cgtgactagtcgccgtgacgtg	cacccaaactggtgcccgtgtc	accgtgaacccgtgatagc
ENr312	Rs 1628402	G	A	ccagtgaccaccagtgaaaatg	tgatgcaatgccgttatagcact	ccgtgaatgcatgcatgctg
ENr312	Rs 11222584	G	C	cactggcgtgaccagttgtggt	cggtgaccgtgacgtgacgtgt	atgccgtggaactgacctg
ENr312	Rs 10894402	G	T	gtgtgtgacccccacaccata	cactggcctgaacgcgtagcc	ttattatgagtcgcgcggg
ENr312	Rs 10894420	G	A	cgtgttgttgaaaacgtgttgtg	cgtgagttgtgaacgtggcctgt	gtggtgaccgtgagcctat
ENr312	Rs 704629	A	G	gactgattgcgcgtagccagtt	ccatgagcccgtaaagtgctgtc	tagccccgtcctaggccca
ENr312	Rs 1113928	G	A	gcgtgatcgaacagtgaattag	ttatagaagtgccacatgatgaa	atatgcagatagcaccaca
ENr312	Rs 1241896	C	T	atctggatgcgctggcatgaaa	tagcctgagagccagtagagag	ccacgtgaaaccgtgaga
ENr312	Rs 10791148	A	G	gactttggtgtccgtgctgattg	tagatgatgatagtcccgtgatga	ccgtgatcccgtgatggct
ENr312	Rs 7121221	G	A	cagtggtgaccgcagtcgtgc	ctgtcacgtgaacaccaggtgc	tgatccagttgcgtgaccgt
ENr312	Rs 2658873	G	A	atgaccgtgacccagtgacca	ggtttaccgtgaccggtgatccg	ccgtgttaccgtgagtggtt

Table 2. Confirmation SNPs for cross over and gene conversion events (with recombinant haplotype)

Region	Event	Marker position	Major Allele	Minor Allele	Forward primer	Reverse primer	Nested forward primer
ENm009	Cross over	5611969	A	G	cgtttgagggtttcgatgccc	accgtgatagcgatgatgtcgt	caccgtgatgcgtgagtcg
ENm009	Cross over	5604295	A	G	cagtgatgctaccgtttgacca	agtgctgtcaccccgtggtacc	cccgtgagctatgtgatccg
ENm009	Cross over	5728880	T	C	gcatatgacatgcgtgacgtcg	acccataagtgatatattacg	cgtactgccagcgttgccga
ENm009	Cross over	5733786	A	G	ggttaccagtagttgaccagtt	tgaccagtggcccacatgaa	agagcctaaccatgtgaaat
ENm009	Conversion	5799485	G	A	atgccgatgaccagtgaccgt	aaccagtgaaagaacttgga	ccagtggcaaacctgtggg
ENm009	Conversion	5801580	G	A	tttttattgagcctgccaaaatt	accatgatcccgtgattcacg	acacgtccgtgatggaata
ENr332	Conversion	64243595	T	G	tacgtgtctactgtgaagaaa	tacactgagctaacctgagtt	ccaactggactgacttataa
ENr332	Conversion	64246699	G	A	ttcagtcgtagtcgatgatacc	acctcgtgagcggatgatag	gaattgcgatccctagccgt
ENr332	Conversion	64254922	A	G	tccgatcagtgagagcgtgt	cccatgatgcggcacaccagt	cgtagccgtgatcggatgcg
ENr332	Conversion	64256058	T	C	ccagtgccagtgcccgatgc	cagtgaattgcggtaggattgt	cggtaccgtgagcgtaggc
ENr332	Conversion	64325109	A	G	cagtgggaccgcgctaggtc	acgtggcgcgtgcggtcgggt	ggtgccacgtgcacagtgg
ENr332	Conversion	64325579	G	A	cggtgaccgcgtgaccgctg	cggtgaccagtcgtaggcgt	cgtagccagtgaaaccacctg
ENm003	Conversion	116122450	A	C	gtagccagtaggtgaccgtga	ctgagcccgtgaggcttgaa	ccgtgctgctaaaaaaccgtg
ENm003	Conversion	116120454	G	A	cgcgtaggcgtccggtaggc	acacgtagacgcgtggccc	cacacgatgaccccggctga
ENm003	Conversion	116118870	A	G	ttcgcgccgtgagaaatgcgc	cgcatcgtgataacacgacca	attgcgtcgaccgtaggcccc
ENm003	Conversion	116115657	G	A	acacgtcggtacgatgcgtgg	acacacccgtttgatgcccc	ctcgagcgtagaatatgcgttt
ENm003	Conversion	116113369	C	T	agagatgcgtccgtgagccc	acacagtagacgccgtagcg	accagtgagaccgtgagctcg
ENm003	Conversion	116112851	A	G	attacgtagaggtgccgtgag	ggctgacacaggcgtagcagt	ccgtagagtgagagtggttgaa
ENr312	Cross over	130605008	G	A	gagcgtgccgctggcgccc	tatatgagcatattccagtgaa	acacgtgacctccgtgagaaa
ENr312	Cross over	130606132	A	T	cggatgagcaacgatgtcgt	acacgtcgtgaggctttgag	ccgtaggagacgctgaggatgt

Table 3. Confiramtion SNPs (with non-recombinant haplotype)

Region	Event	Marker position	Major Allele	Minor Allele	Forward primer	Reverse primer	Nested forward primer
ENm009	Cross over	5475554	A	G	cagtgttcacgtgaccagtga	gacagtgctgaccgtgatgc	ttgatccgtagcaccagtga
ENm009	Cross over	5485494	G	A	ggcagtttgacacacactgg	tatgcgtgatccgatgttgagc	accagtgatccgtgagggc
ENm009	Cross over	5799485	A	G	gatgcggtgcggatggcgc	taccagtagacgatggcccc	accagtaggcgtgcgtagg
ENm009	Conversion	5798932	C	T	acgatgagcgtagccgtagcg	gagaggttgagcgtgagcc	acacgatagatcgctgagca
ENr332	Conversion	64242485	G	A	tgacacacggcaacgctgagc	aactcgatgctccgtagcttc	accatgacgtcgctgacccc
ENr332	Conversion	64247135	A	G	ccatgcgtcgatgccccagatg	tccgtcgacgctaggaccgt	cagctgcgctacgatagcag
ENr332	Conversion	64254003	T	C	cgctaagtggccatatatgcgtc	ccatgtagcgccgtaccgtg	aacgtggacaggttaccacga
ENr332	Conversion	64256178	A	G	acactaagaacagtgcgaaa	acacgatagagcgctagcgt	aacacgatgcaaacacagtga
ENr332	Conversion	64325036	G	A	cgtgcactagagctcgctgag	ggcagttaccatgaatactg	cagtagcccttgacgttgacc
ENr332	Conversion	64325690	A	G	cgtgaccgtacacacactgg	acacggaccaaattggatg	aacctgatgcgctgacgttgaa
ENm003	Conversion	116120364	A	G	gcgtactgacagtcgtcgctagc	ctgacgtactagcgattttac	actgacgtgacaatcgctagcg
ENm003	Conversion	116119354	G	A	atgaggcgatcgatgatcggatg	cagttcggatcgatggcgt	ggtggccgatgaaccacattg
ENm003	Conversion	116117359	A	G	cgtagcggtcgctagagcgtttga	tatgcgatgctagcgatttga	acctgagcgattcgtgacgtga
ENm003	Conversion	116112803	G	C	agctgacgcgtgacgtcgaa	tatgcgtgatcgagcggaaa	aacgaacaccaaacttggccac
ENr312	Cross over	130727143	G	C	atcgagatgcacgatagacaga	cgatcggatcgctaggccaa	taggcgtagccgatcgaccttt
ENr312	Cross over	130729798	A	G	ggtcgctgaccaggctggatta	atatgatgcgtgagacgatg	agcgtgagaagcgcgcgaaaa

Fig. 34. **Allele affiliation with two haplotypes for ENr332** with distribution of major/minor alleles between these haplotypes

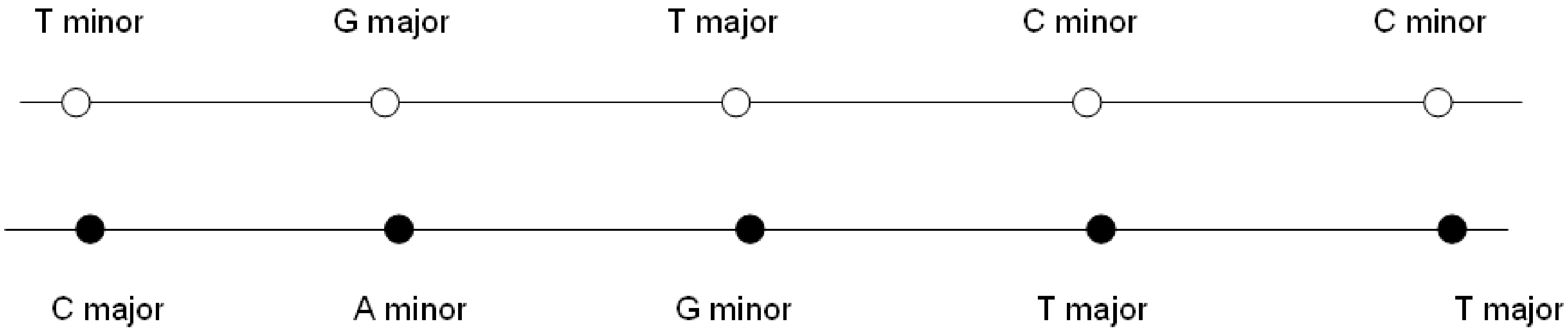


Fig. 35. **Allele affiliation with two haplotypes for ENm003** with distribution of major/minor alleles between these haplotypes

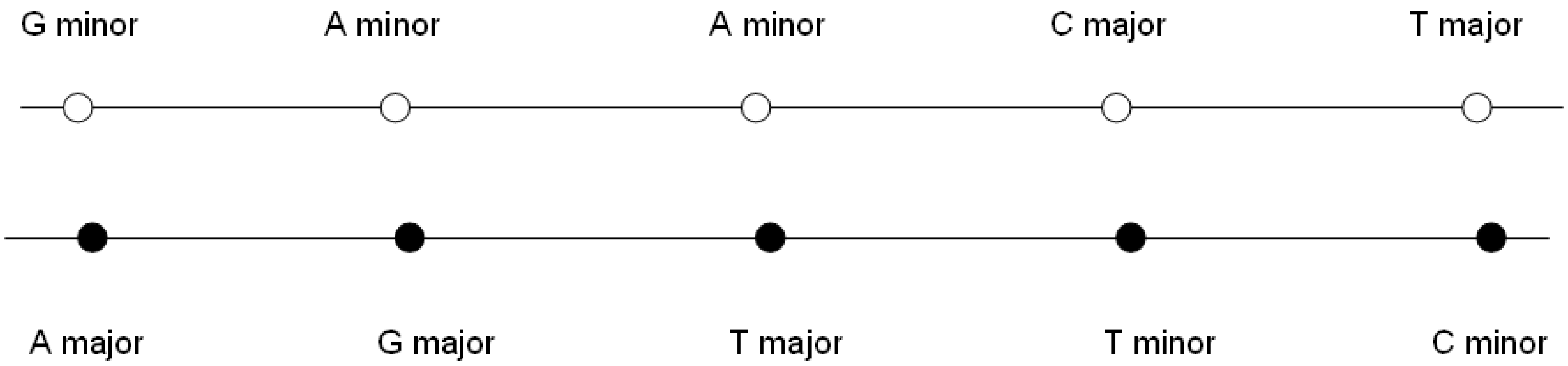


Fig. 36. **Allele affiliation with two haplotypes for ENr312** with distribution of major/minor alleles between these haplotypes (confirmation alleles come on top in the loci where recombinant haplotype was detected).

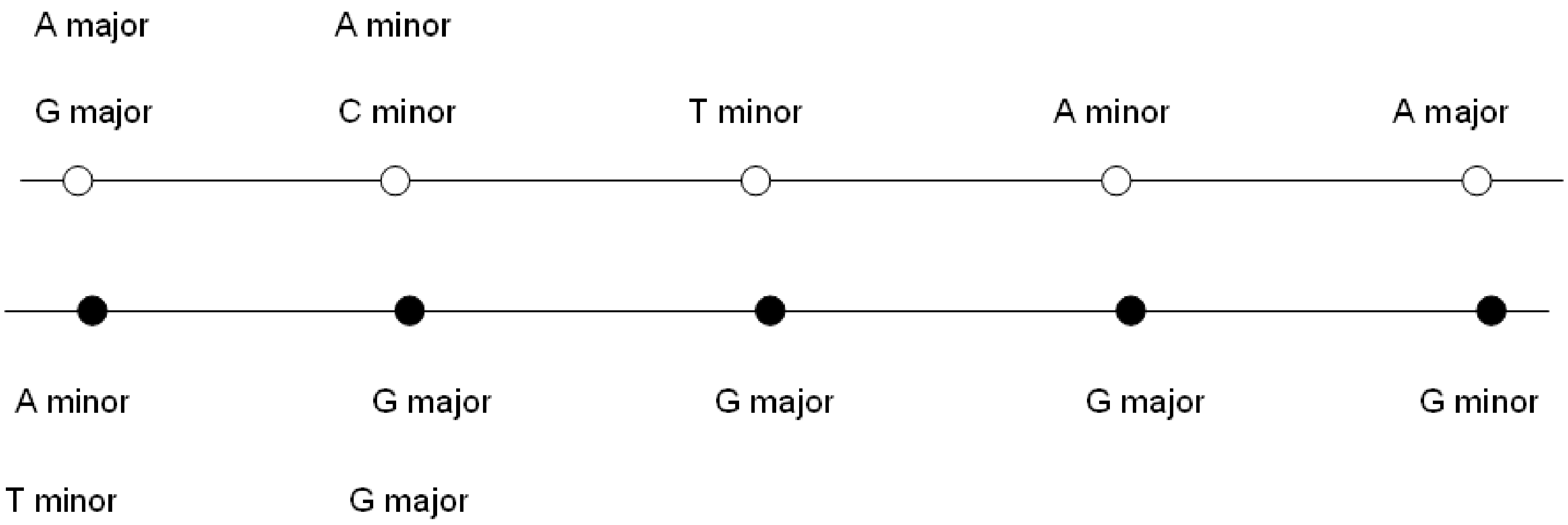
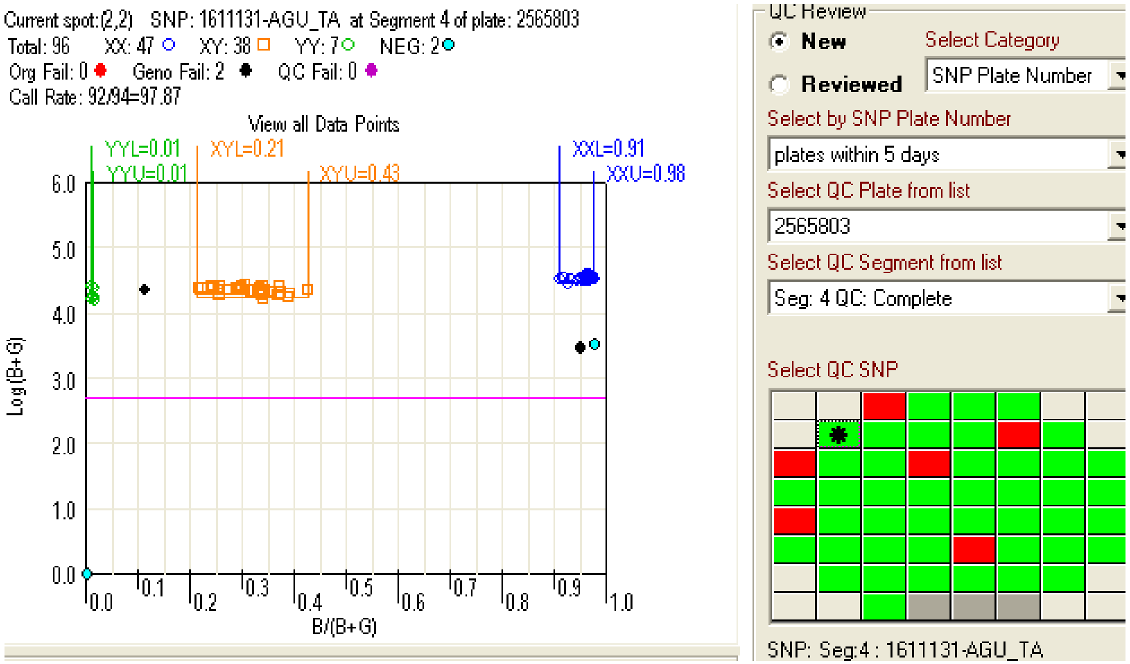


Fig. 37. An example of the SNPstream scatterplot in GenGenos/QC review format with fluorescent signals on the vertical axis and fluorescent signal ratios on the horizontal axis. Three clusters represent three genotypes of the sample (Syvänen 2005).



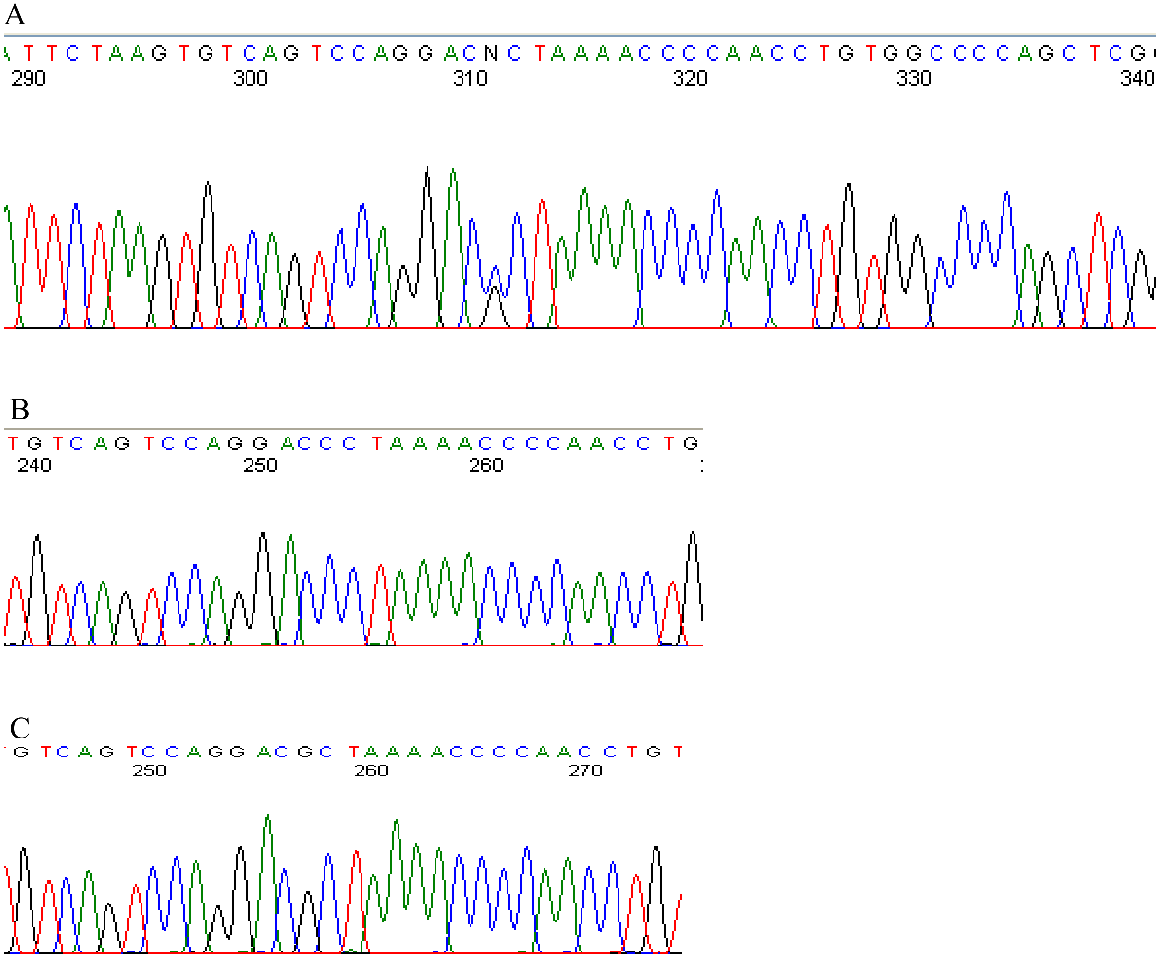
Current spot:(2,2) SNP: 1611131-AGU_TA at Segment 4 of plate: 2565803
Total: 96 XX: 47 ○ XY: 38 □ YY: 7 ○ NEG: 2 ●
Org Fail: 0 ● Geno Fail: 2 ● QC Fail: 0 ●
Call Rate: 92/94=97.87

QC Review
● New
○ Reviewed
Select Category: SNP Plate Number
Select by SNP Plate Number: plates within 5 days
Select QC Plate from list: 2565803
Select QC Segment from list: Seg: 4 QC: Complete
Select QC SNP

		🟥	🟩	🟩	🟩		
	✱	🟩	🟩	🟩	🟥	🟩	
🟥	🟩	🟩	🟥	🟩	🟩	🟩	🟩
🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩
🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩
	🟩	🟩	🟩	🟩	🟩	🟩	
		🟩	▨	▨	▨		

SNP: Seg:4 : 1611131-AGU_TA

Fig.38. FinchTV representation of alleles in bulk sperm (A) and single sperm samples (B, C)



Appendix

List of abbreviations

bp – base pair

cM/Mb – centiMorgan per Megabase

CpG – phosphodiester bond between the cytosine and guanine

ddH$_2$O – biodistilled water

ddNTP – dideoxynucleotide

dNTP - deoxyribonucleotide

DNA - Deoxyribonucleic acid

DSB – double strand break

DTT - Dithiothreitol

EDTA – ethylenediamine – tetraacetic acid disodium salt

EtBr – ethidium bromide

h - hours

kb – kilo base

KOH - potassium hydroxide

LD – linkage disequllibrium

M - molar

Mb – megabase

mM – milli Molar

min – minutes

ng - nanogram

Appendix

PBS - phosphate buffered saline

PCR – polymerase chain reaction

rpm – rotation per minute

SDS – sodium dodecyl sulfate

sec – seconds

TE – tris EDTA buffer

Tm – melting temperature

μl – micro liter

% - percentage

Acknowledgements

During the last four years of my work on this project, it didn't go entirely smooth. There were "ups" and "downs", disappointment and excitement; and whatever period it was, there were lessons to learn. So at the end I would like to thank people who helped, supported and encouraged me.

First of all, I would like to thank Dr. Bettina Harr for the project idea, project design and experimental guidance.

Then, I would like to express the gratitude to my supervisor, Prof. Dr. Thomas Wiehe for theoretical guidance, support and understanding and enormous patience (I really was not a good student all the time).

Also, I would like to thank Prof. Dr. Jonathan Howard and Diethard Tautz for support and invaluable advice in the difficult times of this project.

The next ones on my list are Christoph Goettlinger and Mohammad Reza Toliat. I would like to thank them for helping me with two methods, which, I believe, contributed a lot to single sperm typing technology – sperm sorting with FACs and SNPstream.

Huge gratitude goes to International Graduate School in Genetics and Functional Genomics for funding, support, advice and help (especially, Isabell Witt and Brigitte von Wilken-Bergmann).

I addition, I would like to mention my colleagues (former and current), who are all interesting personalities and created special atmosphere on he 4th floor in the lab of "Bioinformatics and population genetics": Nora Pierstorff, Haipeng Li, Ivana Vukusic, Daniel Zivkovic, Andreas Wollstein, Robert Fuerst, Alex Klassman, Anton Malina, Frank Lill and others (whom I might have forgotten to write about, but will always keep in my memories). Special thanks for Dale Richardson for correcting this thesis and Sabari Thirupathy for valuable advice and assistance.

Acknowledgements

Finally, I would like to thank the most important people in my life – my family. First of all, my son, Kirill (who changed from baby into a smart boy during the thesis time), for being patient and loving, although mummy was so often away; second, my husband – for understanding, patience and support in realizing professional ambitions; and the last but not the least (actually foremost), my parents, whose help with Kirill allowed me to work during PhD time.

Thanks to all of you once again!

Katya Shabanova

24.08.2009

Erklärung

Ich versichere, daß ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; daß diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; daß sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, daß ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Thomas Wiehe betreut worden.

Koeln, August 25 2009 Ekaterina Shabanova

Curriculum vitae

Persönliche Angaben

Name	Ekaterina Shabanova
Adresse	Beethovenstr., 28 69214 Eppelheim
Tel	+49-17621312156
Geburtsdatum	20.2.1981
Geburtsort	Astrachan Russland
Staatsbürgerschaft	russisch

Schulausbildung

09/1988 – 06/1998	Staatschule Nr. 36 in Astrachan
06/1998	Abitur

Hochschulausbildung

09/1998 – 07/2003	Studium der Biologie am Institut für Biologie und Umweltkunde der Technischen Staats-Universität von Astrachan
09/2000 – 05/2001	Studienaufenthalt an der St. Lawrence Universität in New York; Schwerpunkte: Genetik und Molekularbiologie
07/2003	Diplom
04/2005 – 10/2009	Promotion am Institut für Genetik, Arbeitsgruppe für Populationsgenetik und Bioinformatik der Universität zu Köln
10/2009	Voraussichtlicher Abschluß der Promotion

Koeln, August 2009 Ekaterina Shabanova