Linkage Disequilibrium and Transmission Distortion Affecting Human Chromosome 6p

Dissertation

zur Erlangung des Grades des

Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht am Fachbereich Biologie, Chemie, Pharmazie

der Freien Universität Berlin

vorgelegt von

Pablo Sandro Carvalho Santos

aus Brasília

Februar 2010

Diese Dissertation entspricht der zwischen Februar 2006 und Februar 2010 im Institut für Immungenetik, Charité – Universitätsmedizin Berlin, unter Betreuung von Prof. Dr. rer. nat. Andreas Ziegler sowie von Dr. rer. nat. Barbara Uchanska-Ziegler durchgeführten Forschungsarbeit.

This thesis is based on research conducted between February 2006 and February 2010 at the Institut für Immungenetik, Charité – Universitätsmedizin Berlin, in Berlin, Germany, under supervision of Prof. Dr. rer. nat. Andreas Ziegler and Dr. rer. nat. Barbara Uchanska-Ziegler.

Gutachter

Erster Gutachter:	Prof. Dr. Andreas Ziegler Institut für Immungenetik Charité-Universitätsmedizin Berlin Freie Universität Berlin
Zweiter Gutachter:	Prof. Dr. Heribert Hofer Institut für Zoo- und Wildtierforschung (IZW) im Forschungsverbund Berlin e.V. Fachbereich Veterinärmedizin Freie Universität Berlin
Tag der Disputation:	9.9.2010

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig und ohne unzulässige Hilfe oder Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die vorliegende Arbeit wurde weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde zum Zweck einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt. Alles aus anderen Quellen und von anderen Personen übernommene Material, das in der Arbeit verwendet wurde oder auf das direkt Bezug genommen wird, wurde als solches gekennzeichnet.

Berlin, den 9. Februar 2010 Pablo Sandro Carvalho Santos

Danksagung, Acknowledgements, Agradecimentos

An erster Stelle möchte ich mich bei meinem Betreuer Prof. Andreas Ziegler bedanken. Ich bedanke mich bei Ihnen, dass Sie mir in der Nacht des 14. August 2005 auf meine allererste Email, die ich aus Curitiba mit vielen Rechtschreibfehlern verfasst habe, antworteten, und damit den Anfang dieser Doktorarbeit bereiteten. In den vier Jahren hätte ich mir keine bessere Betreuung wünschen können. Durch Sie habe ich mich zum Beispiel für die t-Haplotypen begeistern lassen. Es war jedoch viel mehr als nur eine wissenschaftliche Betreuung und hat mir überdies auch auch viel Freude bereitet. Mittlerweile kann ich mich in der deutschen Sprache besser ausdrücken, und auch dafür muss ich mich bei Ihnen bedanken.

Ich möchte mich auch bei Frau Dr. Uchanska-Ziegler bedanken. Die Zeit mit Ihnen wird mir in guter Erinnerung bleiben. Ihre Kreativität hat mir so manches Mal neue Wege eröffnet, und wird mich auch in meiner professionellen und persönlichen Zukunft weiterhin begleiten. Ich danke Ihnen für die vielen guten Ideen, die Liebenswürdigkeit und Ihre Betreuung. Die Freiheit, die ich im Institut für Immungenetik hatte, war für diese Arbeit ausschlaggebend.

Herzlich bedanke ich mich auch bei Euch, den Mitarbeiterinnen und Mitarbeitern des Instituts für Immungenetik, Angelika Zank (Du warst ein Engel für mich), Christina Schnick, Hans Huser, Rolf Misselwitz, Caroline Backhaus und Alexander Ziegler, für Eure Freundschaft und Kompetenz.

Dem Berliner Senat bin ich für die finanzielle Unterstützung durch das NaFöG Stipendium zu Dank verpflichtet. In besonderer Weise, möchte ich hiermit die ewige Dankbarkeit, die ich der Berliner Krebsgesellschaft e.V. schuldig bin, ausdrücken. Für das keineswegs selbstverständliche Vertrauen und die finanzielle Unterstützung, möchte ich mich herzlich bedanken, insbesondere bei Frau Dr. Barbara Fey.

I am very thankful to the LIGH staff in Curitiba, Brazil, especially Rafael Vargas, Fernanda Ribas, Fabiana Poerner, Prof. Juarez Gabardo and my dear Prof. Maria da Graça Bicalho.

My dear Chee Seng Hee, Pravin Kumar and Thomas Kellermann, you have often been a light for me in these gray winter days of Berlin. I thank you for teaching me so much, and for your friendship. I feel lucky and grateful to have shared this PhD time with the three of you.

I am thankful to my co-authors George Füst, Zoltán Prohászka, Roger Horton, Chack-Yung Yu, Stephan Beck, Peter Schlattmann, Inke König, Andreas Ziegler (in Lübeck), Carolina Sens-Abuázar, Maria Luiza Petzl-Erler, Valéria Sperandio-Roxo, Fabiana Poerner and Clineu Julien Seki Uehara, with special mention to Armin Volz, Johannes Höhne, and Marcos Miretti, for your cooperation and kindness.

Danke an meine kleine Familie Katrin Krschak, Tales Alessandro, Annalisa Scozzari, Paola Gisela, Maria Alzira, Érgio Messias: ihr habt mir alles gegeben. Vocês me deram e são tudo o que eu tenho. Milan, você é o meu pequeno **sol**, um sol que completa hoje 9 anos! Worte sind zu wenig, auch die 32.660 dieser Arbeit.

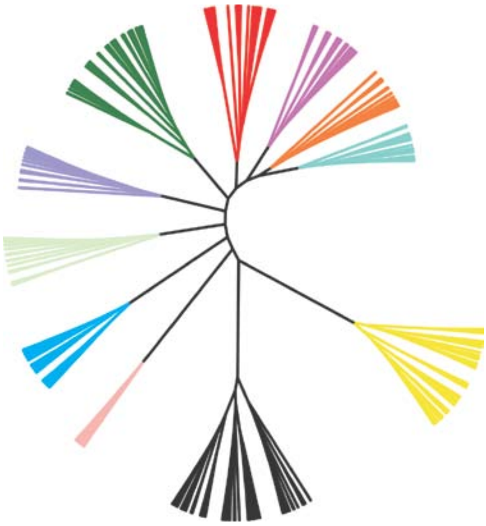
I am eternally thankful to my dearest friends spread through this and other planets, for your love: Renata Oliveira de Souza, Rebeca und Bruna Pupatto Ruano, Jana Kruspe, Ulrike Vetter, Nora Rother, Nastasia Hase, Gabriel Nogueira, Maíra Tito, und die kleinen Levin, Kolja und Joschka aus der Lehrter Kommune. Sandra Kulik, for being a special chapter in my life, and for having taught me so many important things beyond genetics. Esta tese não teria sido escrita sem o amor de vocês – this thesis would not have been written without your support and love.

I also wish to thank all the colleagues at Metrinomics GmbH, und alle im Platzhirsch AG, for being so kind to me, and for being there.

Finally, from far away, you have accompanied me through these four years as well, giving me inspiration to look at the beauty of this world: Ms. Chan Marshall, Ms. Jenny Lou Ziegel, and Mr. Leonard Cohen.

Even if this is allegedly the most widely read part of a PhD thesis, I do encourage you all to dedicate some time to the next chapters. It took me four years to write it, and it is an interesting work, in my opinion.

Thank you all.



Zusammenfassung

Sowohl die jüngste Entwicklung neuer Technologien für die Hochdurchsatzsequenzierung und -genotypisierung von Organismen als auch die Erzeugung neuer Werkzeuge für großflächige genomische Analysen, die in letzter Zeit von verschiedenen Forschungsinstituten weltweit zur Verfügung gestellt wurden, haben die Art der Fragestellungen in den verschiedenen Arbeitsfeldern der Genetik stark beeinflußt. Die vorliegende Arbeit ist ein Beispiel für die Nutzung dieser Ressourcen in Verbindung mit traditionellen Labormethoden, um sich zwei Themenfeldern in den Bereichen Immun- und Populationsgenetik zu nähern: einerseits den mit dem Haupthistokompatibilitätskomplex gekoppelten Geruchsrezeptorgenen (aus dem Englischen, *major histocompatibility complex-linked olfactory receptor genes*, MHC-linked OR), und andererseits dem Phänomen der Allelweitergabedistorsion (*transmission distortion*, TD). Das Konzept des Kopplungsungleichgewichts (*linkage disequilibrium*, LD) ist für die Diskussion beider Themenfelder grundlegend.

Mit Bezug auf die MHC-OR Gene präsentiert diese Arbeit eine detaillierte Beschreibung des LD Profils innerhalb der OR Gene sowie zwischen ihnen und dem MHC. Durch Analyse der OR Polymorphismen und konservierter MHC-Haplotypen sowie der Genotypisierung einer Kohorte von Raucherinnen und Nichtraucherinnen konnten wir den Zusammenhang zwischen einem OR Polymorphismus und dem Rauchverhalten europäischstämmiger Frauen beschreiben. Wir schilderten genotypische und phänotypische Variationen der MHC-OR Gene bei achtzehn menschlichen Zelllinien, indem wir neue Polymorphismen und OR Allele bestimmten, und damit die Grundlage für die Einschätzung der funktionalen Partizipation von MHC-linked OR bei der Partnerwahl schufen. Darüber hinaus führten wir einen chromosomalen und phylogenetischen Vergleich von MHC-OR Genen bei vierzehn Wirbeltierspezien durch und konnten Schlussfolgerungen über die phylogenetische Geschichte der Genfamilien und Tierarten ziehen.

Bezüglich TD konnten wir dieses Phänomen durch eine reine *in silico*-analyse in einer Region des menschlichen Chromosoms 6p ermitteln, welche die Transkriptionsfaktorgene *SUPT3H* (suppressor of Ty 3 homolog), *RUNX2* (runt-related transcription factor 2) sowie mikro RNA *MIR586* beherbergt. In einer anschließenden Untersuchung, in der wir eine große unabhängige südamerikanische Kohorte genotypisierten, konnten wir die Anwesenheit von TD in diesem chromosomalen Segment, zumindest bei europäischstämmigen Populationen, bestätigen. Aufgrund der hohen Bedeutung von *RUNX2* (und den entsprechenden

Proteinisoformen) haben unsere Erkenntnisse Auswirkungen auf die Interpretation einer Reihe von früheren Studien, die dieses Gen mit verschiedenen Phänotypen wie Krebs, Knochendichte und Entwicklungsstörungen in Verbindung brachten. Diese Ergebnisse wurden in elektronischen Dateien für den Internet-basierenden Genome Browser des HapMap Projekts zusammengestellt, um sie im genomischen Kontext sichtbar werden zu lassen und folglich einen ersten Schritt in Richtung einer genomweiten TD Ressource zu unternehmen. Zusätzlich fanden wir heraus, dass TD mit LD korrespondiert und stellten die Hypothese auf, TD könnte die LD „Landschaft“ des Genoms aktiv gestalten. Schließlich weisen unsere Ergebnisse auf ein potenzielles technisches Problem mit der Datenbank des HapMap Projektes hin, das wir vorschlugen bei zukünftigen Erweiterungen der Datenbank zu berücksichtigen.

Summary

The recent development of new technologies for high throughput sequencing and genotyping of organisms and individuals, as well as the generation of new tools for large-scale genomic analysis, all available through different research institutions world-wide, have changed the way questions from the different fields of genetics are addressed. This thesis is an example of the use of such resources in combination with traditional laboratory methods, in order to approach two themes belonging to immunogenetics and population genetics: on the one hand, the olfactory receptor genes linked to the major histocompatibility complex (MHC-linked OR) and, on the other hand, the allele transmission distortion (TD) phenomenon. The concept of linkage disequilibrium (LD) is regarded as a connecting subject, being fundamental for the discussion of both themes.

This thesis presents a detailed description of LD patterns both within MHC-linked OR genes, and between OR loci and the MHC. Through the analysis of OR polymorphisms and conserved MHC haplotypes and the genotyping of a known cohort of smokers and non smokers, we were able to describe the association of one OR polymorphism with smoking habits in Caucasian women. Furthermore, we analyzed genotypic and phenotypic variation of MHC-linked OR genes for 18 human cell lines based on the description of new polymorphisms and OR alleles, thereby forming a basis for the functional assessment of the participation of MHC-linked OR loci in mate choice. Moreover, we developed the most comprehensive comparison performed to date – both chromosomal and phylogenetic – of MHC-linked OR genes in fourteen vertebrate species, enabling us to derive conclusions about the phylogenetic history of gene families and species.

Considering TD, we were successful in detecting this phenomenon, through a purely *in silico* analysis of healthy family trios, in a region of human chromosome 6p that harbours the transcription factor encoding loci *SUPT3H* (suppressor of Ty 3 homolog), *RUNX2* (runt-related transcription factor 2), and the microRNA *MIR586*. In a follow-up investigation in which we genotyped a large, independent South American cohort, we were able to confirm the presence of TD in that chromosomal segment, at least for populations of Caucasian ancestry. Given the high medial relevance of *RUNX2* and the corresponding encoded protein isoforms, our findings have considerable implications for the interpretation of the many studies that found this locus to be associated with different phenotypes such as cancer, bone density and developmental disorders. These results were compiled into track files for uploading into internet-based genome browsers. These can thus be visualized in the genomic context, and provide a first step towards a genome-wide TD resource. We additionally found TD to be intimately associated to LD in the loci assessed, leading us to hypothesise that TD might be actively shaping the LD landscape of genomes. Finally, our results indicate one potential technical problem with the database of the International HapMap Project, which we suggest should be addressed in future updates of this database.

Resumo

O desenvolvimento de novas ferramentas para análises genômicas em larga escala, assim como a disponibilidade de uma grande quantidade de dados provenientes do sequenciamento e da genotipagem de vários organismos, produzidos e ofertados por diferentes instituições do mundo, tornaram possível uma nova maneira de abordar e de responder questões de diferents áreas da genética. Este trabalho é um exemplo do uso integrado de técnicas laboratoriais tradicionais com novos recursos eletrônicos disponíveis à comunidade científica, para abordar dois temas principais que se encontram entre a imunogenética e a genética de populações: de um lado, os genes de receptores olfatórios ligados ao complexo principal de histocompatibilidade (do inglês, *major histocompatibility complex-linked olfactory receptor genes*, MHC-linked OR), e de outro, o fenômeno da distorção de transmissão alélica

(*transmission distortion*, TD). O conceito de desequilíbrio de ligação (DL) fundamenta a discussão dos resultados e age assim como tema transversal.

Quanto aos genes MHC-linked OR, obtivemos evidência da associação de um haplótipo com o hábito de fumar em mulheres com ancestralidade européia. Geramos um panorama detalhado do desequilíbrio de ligação e variação genotípica/fenotípica entre haplótipos humanos, e desenvolvemos uma abrangente análise comparativa destes *loci* – tanto estrutural quanto filogenética – entre quatorze espécies de vertebrados.

Quanto a TD, tivemos sucesso em detectar evidência da presença deste fenômeno em uma região do braço curto do cromossomo 6 humano, correspondendo aos *loci SUPT3H*, *RUNX2* e *MIR586*, sendo os dois primeiros, fatores de transcrição, e o último, um microRNA. Estes resultados foram primeiramente observados em uma população já genotipada e com genótipos disponíveis através do banco de dados do projeto *HapMap*, e posteriormente confirmados em uma população independente. Os resultados mencionados têm implicações tanto para a gênese e manutenção do desequilíbrio de ligação em populações humanas, como também para a interpretação dos vários estudos que ligaram o *locus RUNX2* a diversos fenótipos, ignorando o fato de que esta região gênica está sob TD em populações de ancestralidade européia.

Esta dissertação está organizada de forma cumulativa, apresentada por meio de cinco artigos submetidos a periódicos científicos, e de um capítulo com resultados ainda não submetidos para publicação.

List of Publications that are Part of this Cumulative Thesis

1. **Santos PSC**, Füst G, Prohászka Z, Volz A, Horton R, Miretti M, Yu CY, Beck S, Uchanska-Ziegler B, and Ziegler A (2008). Association of smoking behavior with an odorant receptor allele telomeric to the human major histocompatibility complex. *Genetic Testing*, 12: 481-486.
2. **Santos PSC**, Höhne J, Schlattmann P, König IR, Ziegler A, Uchanska-Ziegler B, Ziegler A (2009). Assessment of transmission distortion on chromosome 6p in healthy individuals using tagSNPs. *European Journal of Human Genetics*, 17:1182-1189.
3. Sens-Abuázar C, **Santos PSC**, Bicalho MG, Petzl-Erler ML, Sperandio-Roxo V (2009). MHC microsatellites in a Southern Brazilian population. *International Journal of Immunogenetics*, 36:269-274.
4. **Santos PSC**, Höhne J, Schlattmann P, Poerner F, Bicalho MG, Ziegler A, and Uchanska-Ziegler B (2010). Presence of Transmission Distortion on Human Chromosome 6p Revealed by SNP Genotyping of Southern Brazilian Families. Submitted.
5. **Santos PSC**, Uehara CJS, Ziegler A, Uchanska-Ziegler B and Bicalho MG (2010). Variation and linkage disequilibrium within olfactory receptor gene clusters linked to the human major histocompatibility complex. Submitted.

List of Abbreviations

ASW	African ancestry in Southwest USA (HapMap population)
Bta	*Bos taurus* (cow)
BLAST	basic local alignment search tool
bp	base pair
cen	centromere
CEU / CEPH	Utah residents with ancestry from Northern and Western Europe (HapMap population), from the *Centre d'Etude du Polymorphisme Humain (CEPH)* panel.
Cfa	*Canis familiaris* (dog)
CHB	Han Chinese in Beijing, China (HapMap population)
Chr	chromosome
Chr6p / Hsa6p	short arm of human chromosome 6
CNV	copy number variation
CP	cytoplasmic domain
Dre	*Danio rerio* (zebra fish)
EC	extracellular domain
Eca	*Equus caballus* (horse)
F	frequency
Fca	*Felis catus* (cat)
Fig.	figure
GPCR	G-protein-coupled receptor
HapMap	international HapMap project
HLA	human leukocyte antigen
Hsa	*Homo sapiens* (human)
JPT	Japanese in Tokyo, Japan (HapMap population)
Kb	Kilobase (one thousand base pairs)
LD	linkage disequilibrium
LIGH	Laboratório de Immunogética e Histocompatibilidade (Immunogenetics and Histocompatibility Laboratory in the Federal University of Paraná)
MAF	minor allele frequency
Mamu	*Macaca mulatta* (rhesus macaque)
Mb	Megabase (one million base pairs)
Mdo	*Monodelphis domestica* (opossum)
MEX	Mexican ancestry in Los Angeles, California (HapMap population)

MHC	major histocompatibility complex
MHCHP	major histocompatibility complex haplotype project
MKK	Maasai in Kinyawa, Kenya (HapMap population)
µl	microliter (10^{-6} L)
Mumu	*Mus musculus* (mouse)
nsyn	non synonymous (nucleotide substitution)
OR	olfactory receptor / odorant receptor
PCR	polymerase chain reaction
Ppy	*Pongo pygmaeus* (orangutan)
Ptr	*Pan troglodytes* (chimpanzee)
QC	quality control
Rno	*Rattus norvegicus* (rat)
SE	shared epitope
S-M-R	genomic region including the loci *SUPT3H*, *MIR586* and *RUNX2*
SNP	single nucleotide polymorphism
Ssc	*Sus scrofa* (pig)
STR	short tandem repeat
tel	telomere
TD	transmission distortion
TDT	transmission/disequilibrium test
TM	transmembrane domain
trans	transcript specific (single nucleotide polymorphism)
Xtr	*Xenopus tropicalis* (frog)
YRI	Yoruba in Ibadan, Nigeria (HapMap population)

List of Online Resources

BioMart Project	http://www.biomart.org/
BLAST Alignment Tool	www.ncbi.nlm.nih.gov/BLAST/
Database of Drosophila Genes & Genomes	http://flybase.org/
dbSNP SNP Database	http://www.ncbi.nlm.nih.gov/snp
EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl/
ENSEMBL Genome Browser	http://www.ensembl.org/
ePCR	http://www.ncbi.nlm.nih.gov/sutils/e-pcr
European Bioinformatics Institute	http://www.ebi.ac.uk/
GARField Cat Genome Browser	http://lgd.abcc.ncifcrf.gov/cgi-bin/gbrowse/cat/
GenBank Sequence Database	http://www.ncbi.nlm.nih.gov/Genbank
GeneCards Human Gene Database	http://www.genecards.org/
HUGO Gene Nomenclature Committee	http://www.genenames.org/
Human Olfactory Data Explorer (HORDE)	http://genome.weizmann.ac.il/horde/
IMGT/HLA Database	http://www.ebi.ac.uk/imgt/hla/
International HapMap Project	http://hapmap.ncbi.nlm.nih.gov/
InterPro	http://www.ebi.ac.uk/interpro/
JGI Frog Genome Browser	http://genome.jgi-psf.org/cgi-bin/browserLoad/?db=Xentr4
MUSCLE Sequence Comparison Tool	http://www.ebi.ac.uk/Tools/muscle/
NCBI’s Genome Project Resources	http://www.ncbi.nlm. nih.gov/ genomeprj
Online Mendelian Inheritance in Man (OMIM)	http://www.ncbi.nlm.nih.gov/omim/
PubMed Life Science Database	http://www.ncbi.nlm.nih.gov/sites/entrez
Reference Sequence (RefSeq) Database	http://www.ncbi.nlm.nih.gov/RefSeq/
STRING Database for Protein-Protein Interactions	http://string.embl.de/
The MHC Haplotype Project	https://www.sanger.ac.uk/HGP/Chr6/MHC/
UCSC Genome Browser	http://genome.ucsc.edu/
VEGA Genome Browser	http://vega.sanger.ac.uk/
VISTA Genome Browser	http://pipeline.lbl.gov/

Table of Contents

1. Introduction

This doctoral thesis is organized as a cumulative work presented in the form of five different manuscripts and one additional chapter with results not yet submitted for publication, having the subjects “linkage disequilibrium” and “transmission distortion” as a common focus. The five manuscripts, each of which are preceded by a short introductory summary, were written in the context of this doctoral work and submitted for publication to peer-reviewed journals within the last three years. While three of these articles have already been published and will therefore be shown here with the respective journal’s layout, two further manuscripts are currently under review by the corresponding editorial boards.

The aim of this chapter is to present general concepts that will be referred to throughout this thesis.

1.1. Linkage Disequilibrium in the Human Genome

The term “linkage disequilibrium” (LD) is used in population genetics to refer to the non-random pattern of association between alleles at different loci. LD was first described fifty years ago [Lewontin & Kojima, 1960], and the term derives from the fact that when present, LD will prevent the combination of alleles from two or more neighbouring loci on a single chromosome, also termed haplotype, to reach that expected on the basis of the frequency of each individual allele [Slatkin, 2008]. In other words, LD can be described as the difference between the observed frequency of a haplotype and the frequency it would be expected to have, based on the individual allele frequencies [Nordborg and Tavaré, 2002; Slatkin, 2008].

LD is generally denoted by the letter D:

$$D_{AB} = F_{AB} - F_A F_B \qquad (1)$$

where A and B are alleles from two different loci, and F stands for frequency.

However, because the numeric value of D is strongly dependent on the individual allele frequencies, it is often not the best way to express LD when one is interested in comparing levels of LD between different regions or different pairs of alleles. In order to make this kind of comparison possible, the concept of D′ was proposed four years after the first LD description, taking into account the highest possible value that D can reach, given the frequencies of the alleles being considered [Lewontin, 1964]:

$$D' = \frac{D}{D_{max}} \tag{2}$$

D′ is therefore a normalization of D, so that it takes values between 0 and 1 (or 0 and –1) without regard to the allele frequencies. D′ has the property that when $\mathrm{D}' = 1$, linkage of at least one of the four possible allele combinations will be absolute, and at least one of the four possible haplotypes will be absent. This way to quantify LD is often used in assessments of polyallelic loci, although many others have been proposed for these situations [Ardlie et al., 2002; Zaykin et al., 2008]. In the case of the assessment of single nucleotide polymorphisms (SNP), which are generally biallelic, an alternative way to measure LD is more commonly used, the correlation coefficient r^2 [Hill and Robertson, 1968], which is given by:

$$r^2 = \frac{D^2}{F_A(1 - F_A)F_B(1 - F_B)} \tag{3}$$

Similarly to D′, r^2 is based on D, and $r^2 = 1$ when alleles A and B are in complete LD with each other, but, differently from D’, r^2 additionally requires A and B to have similar allele frequencies in order to equal 1 [VanLiere & Rosenberg, 2008]. In summary, r^2 is a generally more stringent LD measure than D′, and it is the current standard measure of LD within genome-wide and other large scale LD assessments [Ardlie et al., 2002; Zaykin et al., 2008]. An example of an LD plot is given in Fig. 1.1.

LD is understood as a product of many interacting evolutionary forces such as natural selection, genetic drift, population bottlenecks, mixing of subpopulations, inbreeding, genomic inversions and gene conversion [Slatkin, 2008]. LD in humans in often understood from a genealogical point of view, according to which LD is the result of “remainders” of ancestral haplotypes, and tends to disappear as a function of time and recombination events. In this context, LD is traditionally believed to be stronger within population isolates (in which genetic drift and inbreeding are pronounced), but one recent study reveals that this might be a mistaken assumption [Bosch et al., 2009].

Based on LD, the human genome has been described as an assembly of more or less discrete blocks of high LD named haplotype blocks [Daly et al., 2001; Johnson et al., 2001; Patil et al., 2001; Gabriel et al., 2002]. The low LD segments separating such blocks are predicted to correspond to recombination hotspots [Reich et al., 2001, Daly et al., 2001, Patil et al., 2001; Gabriel et al., 2002]. However, it is essential to bear in mind that haplotype blocks are rather a theoretical structure than a biological phenomenon [Blomhoff et al., 2006], since the borders of a haplotype block will diverge considerably, depending on the density of assessed markers [Ke et al., 2004] and on the algorithm used to define a haplotype block [Schulze et al., 2004].

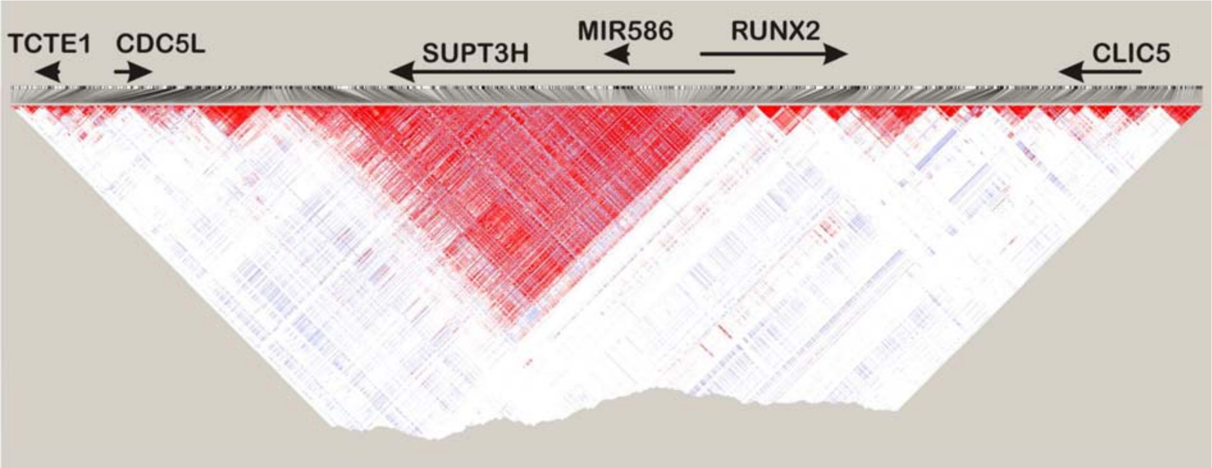


Figure 1.1: LD plot of 1170 SNPs for a region of human chromosome 6p. The positions and lengths of the arrows correspond to those of the six loci (designations and transcriptional orientation are indicated) harboured in the region. Within the plot, each diagonal represents one single SNP, and each point in the plot (the intersection of two diagonals) represents LD between two loci. Red spots indicate strong LD (statistically significant $D' = 1$ and $r^2 > 0.8$), white indicates absence of statistically significant LD, while intermediate values (high D′ but lower r^2) are indicated by the interspersed bluish spots. According to this graph, the whole reading frames of the loci *SUPT3H* and *MIR586*, as well as about one third of *RUNX2* are within one LD block. This plot was generated with genotyping data from the CEU population. (Modified from Santos et al., Eur J Hum Genet 2009).

Since LD provokes the association of alleles from different loci, it represents an obstacle for phenotype association studies, as it confounds the interpretation of results linking a region with high LD to a given phenotype. Depending on the extension of the LD block, a locus found to be linked to a disease, for instance, can be several megabases (Mb) distant from the locus responsible for the primary association [Zhang et al., 2004]. A classical example highlighting this problem is the case of the gene causing the disease haemochromatosis type 1 disease (hereditary iron overload). This medical condition was long believed to be associated with a gene-dense region on the short arm of human chromosome 6 (Chr6p) known as the human leukocyte antigen (HLA complex), characterized by the high number of genes involved in immune responses and by extreme levels of LD. The real causative locus for haemochromatosis was later identified, and located four Mb telomeric to the HLA complex [Feder et al., 1996].

Despite this "drawback", the intelligent use of LD together with the development of new computational and statistical methods as well as the genotyping and the sequencing of the genomes of a large number of individuals over the last few years has opened new possibilities for association studies and population genetics based on LD. For example, LD can be used for selection of so-called tagSNPs. These are SNPs that are informative for a whole genomic region exhibiting high LD, in which many other SNPs with similar allele frequencies could be present [Johnson et al., 2001; Miretti et al., 2005]. With this approach, one is able to assess a given genomic segment through genotyping, using only a small subset of the many polymorphisms described for that region.

1.2. The Major Histocompatibility Complex (MHC)

The major histocompatibility complex (MHC) was first described over fifty years ago [Snell, 1968; Dausset, 1981]. It is a gene dense-region present in the genomes of all vertebrates studied so far, harbouring genes that make it the most important region with regard to normal immune responses, but also for autoimmune and infectious diseases [Ryder et al., 1981; Lie et al., 2005; Trowsdale, 2005; Fernando et al., 2008; Vandiedonck and Knight, 2009]. Extreme degrees of genetic diversity and LD are also features of the MHC [Horton et al., 2004; Lie et al., 2005; Trowsdale, 2005; Vandiedonck and Knight, 2009]. The number of scientific articles published on a subject is often used as a measure of the general interest end effort dedicated to it. A current literature search on the life science online database PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez) yielded 10.252 articles published with the expression "Major Histocompatibility Complex" on their titles or abstracts within the last ten years, revealing that the MHC is one of the most intensively studied areas in biomedical research today.

The classical MHC loci encode molecules that present auto- and alloantigens to T cells, building thus the immunological basis of self/nonself recognition. The human MHC is termed HLA complex. It is a ~ 3.5 Mb long region harboured on chromosome 6p (Fig. 1.2), associated with far more diseases than any other region of comparable size in the genome [Horton et al., 2004; Trowsdale, 2005; Fernando et al., 2008].

The medical relevance of the HLA is additionally enhanced due to the role played by this gene system in allogenic tissue transplantations. Especially in the cases of kidney and stem cell transplantation, the selection of a donor with HLA alleles matching those of the recipient plays a determinant role in graft survival and transplant success. However, the exuberant polymorphism of this genomic segment has the consequence that the probability of two unrelated individuals to carry the same class I and class II alleles is extremely low, and the task of finding a compatible unrelated donor is a very hard one. For the three most polymorphic loci, 965 HLA-A, 1.543 HLA-B and 762 HLA-DRB1 alleles have been described, as of January 2010 [IMGT/HLA Database, 2010].

These alleles do not appear, however, in all theoretically possible combinations. Due to the strong level of LD that characterizes the MHC, some allele combinations are more frequent than expected under free recombination. In fact, some MHC haplotypes are so remarkably conserved that LD ranges beyond the borders of the MHC [Alper et al., 1989; Yunis et al., 2003]. Based on LD, a wider area including the MHC has been designated extended MHC (xMHC), including the "core" MHC, a large telomeric neighbouring segment (extended class I), and a shorter centromeric region (extended class II) [Alper et al., 1989; 1992; Ziegler, 1997; Horton et al., 2004]. This nomenclature is based on the traditional subdivision of the core MHC into the tree subregions, class I, class III and class II (Fig. 1.2). One example of a conserved haplotype is the case of the combination HLA-A1, HLA-B8, HLA-DR3, for which LD has been reported to be extreme over the entire length of the xMHC [Alper et al., 2006], and is therefore considered one of the ancient haplotypes of the human MHC.

Another key feature of the MHC is the role played by this gene complex for reproductive patterns such as pre- and post-copulatory mate choice. The first report associating the MHC with mating patterns was observed over thirty years ago in mice [Yamazaki et al., 1976]. According to this and other related studies, females are able to distinguish, most probably through olfactory cues, males that are MHC-similar to themselves from those which are MHC-dissimilar. Mate choice may thus be driven towards maximization or an optimization of the MHC heterozygosity of the offspring [Sommer, 2005; Ziegler et al., 2005; Milinsky, 2006; Eizaguirre et al., 2009; Woelfing et al., 2009]. Similar observations have been made since the initial observation by Yamazaki and co-workers [1976] for fish [Reusch et al., 2001], lizards [Olsson et al., 2003], rats [Singh et al., 1987; Brown et al., 1989], lemurs [Schwensow et al., 2008] and humans [Wedekind et al., 1995, Ober et al., 1997; Jacob et al., 2002, Santos et al., 2005]. In this context, it is plausible that the olfactory receptor genes present within the class I region of the xMHC play an important role in odour- and MHC-dependent patterns of mate choice [Ziegler, 1997; 2000a; 2000b; 2002; Eklund et al., 2000; Thompson et al., 2010].

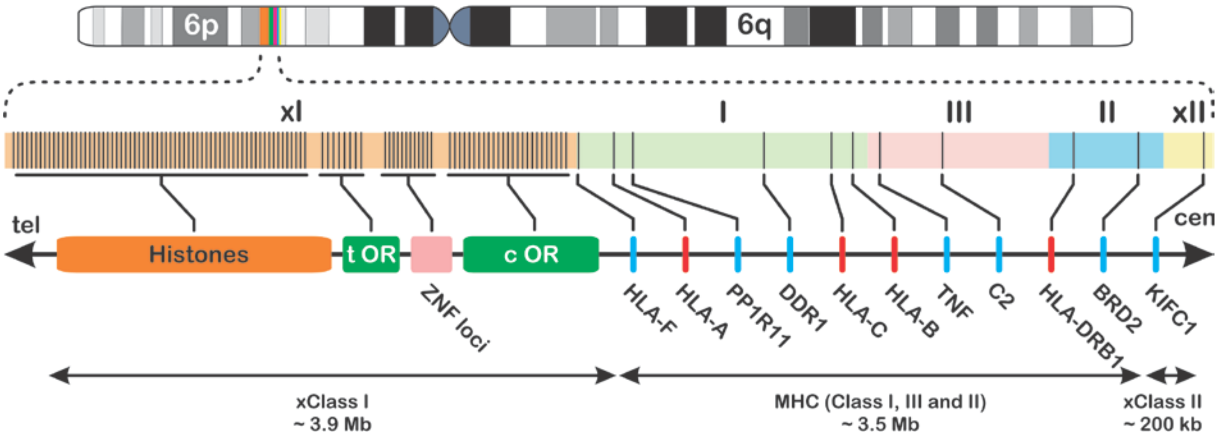


Figure 1.2: Map of the xMHC on human chromosome 6p. The localization of the region within chromosome 6 is given in the upper panel, while a representation of the five subregions: extended class I (xI, orange), class I (I, green), class III (III, pink), class II (II, blue), and extended class II (xII, yellow) is depicted in the middle panel. Selected loci or gene clusters are displayed in the lowest panel: a large histone cluster (orange box), two olfactory receptor clusters (green boxes), a zinc finger gene cluster (pink box), classical MHC genes (red ticks), and selected other selected MHC genes (blue ticks). The approximate length of each subregion is also indicated. (Generated according to genomic coordinates given by the ENSEMBL genome database).

1.3. MHC-linked Olfactory Receptor Genes

Olfactory receptors (OR) are G-protein-coupled membrane receptors that are responsible for the molecular basis of the ability to recognize odours. When expressed on the membrane of olfactory neurons present within the olfactory epithelium, these receptors are able to interact with odorant molecules dispersed in the air and provoke neuron firing and signal transduction into the olfactory bulb. OR consist of seven transmembrane domains, interspersed by four extracellular and four cytoplasmic regions (Fig. 1.3) [Lancet and Pace, 1987; Buck and Axel, 1991; Mombaerts, 1999].

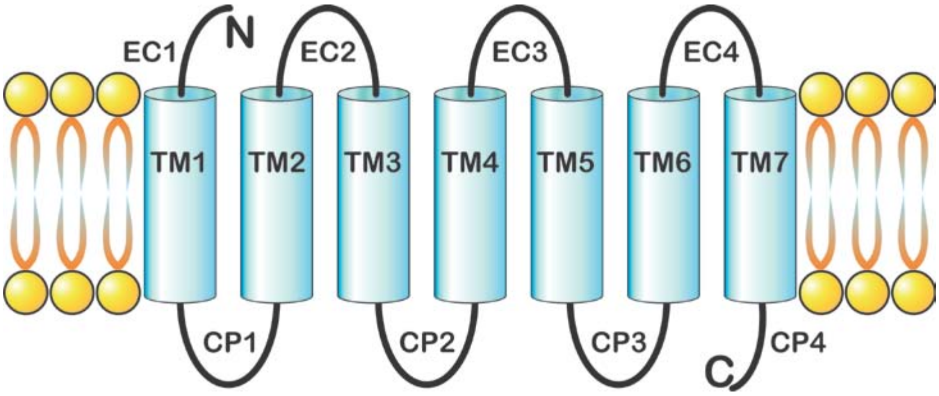


Figure 1.3: Schematic representation of an OR within the cytoplasmic membrane, with the N terminus within the extra cellular space and the C terminus within the cytoplasm. The seven transmembrane domains (TM1 to TM7) are shown as blue cylinders. The four extracellular (EC1 to EC4) and the four cytoplasmic (CP1 to CP4) loops are also indicated. (Adapted from Buck and Axel, 1991).

Olfactory receptor genes belong to the largest gene families in vertebrate genomes, with around 100 genes in fish [Ngai et al., 1993], over one thousand in mice and rats [Zhang and Firestein, 2001; Young et al., 2002], and ~ 800 in humans [Glusman et al., 2001; Zozulya et al., 2001; Malnic et al., 2004; Nei et al., 2008]. These loci are normally organized in gene clusters which, in humans, are distributed throughout the genome, over almost all chromosomes [Glusman et al., 2001; Zozulya et al., 2001; Nei et al., 2008]. Two human OR gene clusters with 34 loci are located in the extended MHC class I region, between one large histone cluster and the telomeric border of the HLA class I region (Fig. 1.4) [Younger et al., 2001].

Following the completion of the human genome project, the sequencing of genomes of other vertebrates revealed not only strong homology between members of mammalian OR gene families in different species [Lane et al., 2001; Aloni et al., 2006], but also that the presence of OR clusters linked with the MHC is remarkably conserved, as in the case of human and mouse [Amadou et al., 2003]. As previously mentioned, the biological relevance of MHC-linked OR genes has increasingly been the focus of research due to its possible role for MHC-dependent mate choice [Ziegler, 1997; 2000a; 2000b; 2002; Eklund et al., 2000; Thompson et al., 2010]. Because of the close proximity and high LD with the MHC, it is plausible that MHC variations, which have been found to correlate with odour-driven mate choice patterns or odour preferences [Yamazaki et al., 1976], correspond to variations within the linked OR clusters. Although a recent report [Thompson et al., 2010] indicates that MHC-linked OR genes do not directly drive odour preferences in humans, the suggestion that MHC-based cryptic or post-copulatory mate choice might be under the influence of these OR genes seems plausible. Several studies have shown that OR and other GPCR genes (including MHC-linked OR genes) are expressed in testis [Parmentier et al., 1992; Vanderhaeghen et al., 1997a; 1997b; Walensky et al., 1998; Tatsura et al., 2001; Volz et al., 2003; Fukuda et al., 2004], and sperm cells [Vanderhaeghen et al., 1993, Spehr et al., 2003; Fukuda et al., 2004].

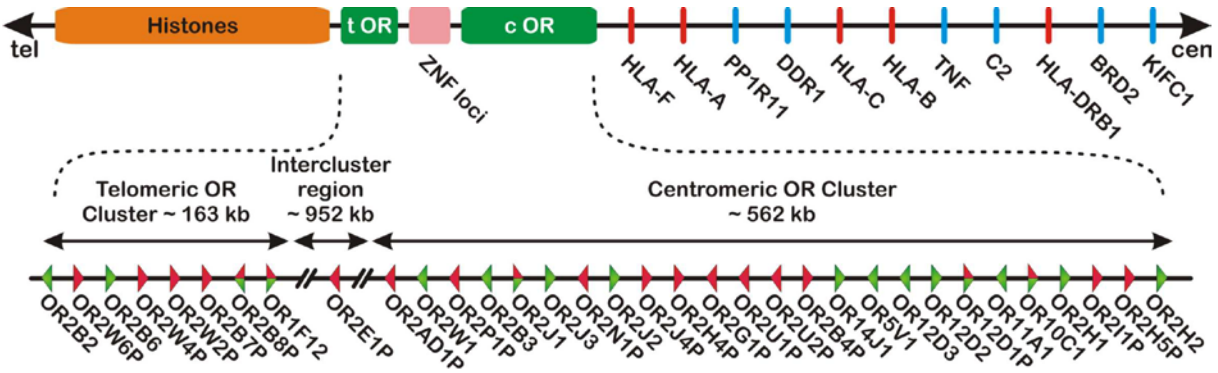


Figure 1.4: Map (not to scale) of the region encompassing the two human MHC-linked OR gene clusters on chromosome 6p. In the upper panel, the relative position of the two OR gene clusters within the xMHC is given (green boxes), while the lower part of the figure shows the schematic positions of OR genes and pseudogenes (suppressing most of the intercluster region). Triangles indicate transcriptional orientation, and are filled in green (genes), red (pseudogenes) or in both colours (segregating pseudogenes). tel: direction to the telomere; cen: direction to the centromere; tOR: telomeric OR gene cluster; cOR: centromeric OR gene cluster. (Generated according to genomic coordinates given by the ENSEMBL genome database).

1.4. Online Resources for Genomic Variation Analyses

The genomic data available through the World Wide Web developed in the last few years to an indispensable resource for genetics research. Since the sequencing of the human genome [International Human Genome Sequencing Consortium, 2001], ongoing efforts of several institutions worldwide are currently directed at the generation of sequence data from many different organisms. Once the genome assembly of one species is complete, additional sequence and genotyping data are commonly generated, in order to provide more data on the variation among individuals of that species. This huge, and continuously growing amount of data is organized in different ways through databases, genome browsers, and online resources that combine both. In this context, the ability to extract specific data out of the immense datasets available plays a critical role, and can be aided by so-called “data mining” tools.

1.4.1. Genome Browsers

Genome browsers are online resources through which the user is able to interact with databases in a visual way, being able to observe a genomic region in its genomic context. Different browsers focus on different features of genomes, while some are specialized for different species. The genome browsers used in this work were ENSEMBL (comprehensive database gathering automatic annotated genomic sequences of over 40 vertebrate species, www.ensembl.org/), VEGA (“Vertebrate Genome Annotation”, a resource based on manual annotation of genes from the genomes of human, mouse and other species, http://vega.sanger.ac.uk/), the Genome Bioinformatics Site of the University of California in Santa Cruz UCSC (http://genome.ucsc.edu/), the VISTA genome browser for comparative analyses (http://pipeline.lbl.gov/), the Genome Annotation Resource Field for *Felis catus* GARField (http://lgd.abcc.ncifcrf.gov/cgi-bin/gbrowse/cat/), the genome project resource of the National Center for Biotechnology Information (NCBI), which gathers information of over fifty vertebrate genome projects from different institutions (http://www.ncbi.nlm.nih.gov/ genomeprj), the Frog Genome Browser of the Energy Joint Genome Institute (http://genome.jgi-psf.org/cgi-bin/browserLoad/?db=Xentr4), as well as the genome browser of the international HapMap Project (http://hapmap.ncbi.nlm.nih.gov/).

1.4.2. The International HapMap Project

The International HapMap Project (HapMap) is the result of an international venture aiming to generate a comprehensive online resource of haplotype variation in different human populations [International HapMap Consortium, 2003]. It is thus intended to promote an efficient approach for the discovery of markers associated with diseases through the use of a population-specific mapping of LD blocks and tagSNPs [International HapMap Consortium, 2003; 2005; 2007].

Until the final release of the last HapMap version (HapMap2), the project was focusing on individuals from the four following human populations: CEU (Utah residents with ancestry from Northern and Western Europe, 90 individuals from 30 family trios), YRI (Yoruba in Ibadan, Nigeria, 90 individuals from 30 family trios), CHB (Han Chinese in Beijing, 45 unrelated individuals) and JPT (Japanese in Tokyo, 45 unrelated individuals). All individuals were assessed with a high throughput SNP genotyping platform, for more than 3.1 million SNPs genome-wide [International HapMap Consortium, 2007]. This data has served as a reference for several population studies [Skelding et al., 2007; Mägi et al., 2007; Manolio et al., 2008; Gu et al., 2008], including some addressing the HLA complex, concerning its LD profile, haplotype variation, as well as genotyping efficiency [Miretti et al., 2005; de Bakker et al., 2006].

By the end of 2008, the HapMap entered a new phase (termed HapMap3). This new HapMap release, which is currently not completely established, involved an increase in the number of individuals of the "old" populations, as well as the inclusion of over one thousand new individuals from seven new populations: ASW (African ancestry in Southwest USA), CHD (Chinese in Metropolitan Denver, CO, USA), GIH (Gujarati Indians in Houston, TX, USA), LWK (Luhya in Webuye, Kenya), MEX (Mexican ancestry in Los Angeles, CA, USA), MKK (Maasai in Kinyawa, Kenya), and TSI (Tuscans in Italy). However, the panel of markers genotyped in HapMap 3 is somewhat reduced (around 50%), as compared to that of HapMap2.

1.5. Transmission Distortion

Transmission Distortion (TD) describes the situation in which the "normal" random segregation of alleles from parents to their offspring according to the Mendelian law of the independent assortment of alleles is violated [Lyttle, 1993; Pardo-Manuel de Villena and Sapienza, 2001]. The term is widely used as a synonym of "transmission ratio distortion" and "segregation distortion", although there is some dispute regarding which expression should be used in a given case of TD [Lyon, 2003].

There are several processes that can potentially distort segregation rates of alleles, LD blocks or whole chromosomes and thereby provoke TD. These include gamete selection (generally referring to the preferential fertilization of eggs depending on alleles or haplotypes carried by the sperm cells [Lyon, 2003]), meiotic drive (the asymmetric segregation of alleles during gametogenesis [Lyttle, 1993; Axelsson et al., 2010]), and embryonic selection (the genotype-dependent survival of embryos or foetuses [Murphy et al., 2008]).

The most notable case of TD has been described for the mouse t haplotypes, already over 70 years ago [Chesley and Dunn, 1936]. These are variants of the proximal one third of mouse chromosome 17, consisting generally of four non-overlapping inversions which appear with a frequency of around 15% in wild mice populations and which suppress recombination between t and wild-type haplotypes [Bennett, 1975; Willison and Lyon, 2000; Lyon, 2003]. The most important feature of t haplotypes is, however, their ability to violate Mendelian transmission in their favour. Although females segregate t- and wild-type haplotypes according to Mendelian expectations, more than 50% (in fact up to 99%) of the offspring from heterozygous t/+ males inherit the t-haplotype [Willison and Lyon, 2000; Lyon, 2003]. The balance equilibrates, keeping the locus in heterozygous state for basically two reasons: t/t homozygosity is either lethal or causes sterility [Lyon, 2003] and, as has only recently been demonstrated, t/+ heterozygous males face a strong reproductive disadvantage in comparison to wild-type homozygous males regarding their territorial behaviour [Carroll et al., 2004]. The genomic region “affected” by t haplotypes comprises up to 40 Mb of mouse chromosome 17 [Lyon et al., 1988]. This segment is largely syntenic to the short arm of human chromosome 6 and generally includes both the mouse MHC and the MHC-linked OR genes.

Other examples of TD have since been described for plants [Fishman et al., 2008], Drosophila [Ganetzky, 1977], birds [Aparicio et al., 2010; Axelsson et al., 2010], mice [Wu et al., 2005; Haston et al., 2007; Kriz et al., 2007; Purushothaman et al., 2008; Girirajan et al., 2009], cattle [Murphy et al., 2008] and also humans [Evans et al., 1994; Chakraborty et al., 1996; Naumova et al., 1998; 2001; Eaves et al., 1999; Lemire et al., 2004; Zöllner et al., 2004; Dean et al., 2006], including genes harboured within the HLA complex [Hanchard et al., 2006]. The statistical problem of multiple testing is, however, a constant obstacle preventing low levels of TD in large-scale investigations to be detected [Zöllner et al., 2004; International HapMap Consortium, 2005].

While the molecular mechanisms leading to TD in humans remain unknown, mouse t haplotypes are much better understood. In this case, the transcription of a t haplotype-associated variant of the sperm motility kinase-1 gene (*Smok1*) is enhanced specifically in t haplotype-carrying sperm cells, rescuing them from a “poisonous” effect that other t haplotype-associated genes (called t complex distorters) have caused in parallel, affecting the development of all germ cells [Herrmann et al., 1999]. An explanation of how the t complex-related *Smok1* gene exerts its specificity selectively for t-bearing sperm has only recently been obtained [Véron et al., 2009]. The consequence of this selective activity of *Smok1* is that t sperm and wild type sperm differ in flagellar motility, with an advantage for t sperm regarding egg cell fertilization [Herrmann et al., 1999; Véron et al., 2009].

In humans, however, TD is not only of interest as a biological phenomenon, but is also of critical importance for numerous disease association studies that are currently performed. As discussed elsewhere [Greenwood et al., 2000; Paterson et al., 2003; 2009], TD can, if present in a given genomic region in the general population, introduce significant bias into association studies regarding that region when TD is disregarded. According to this notion, family-based association studies that assess only the affected offspring are prone to falsely conclude that a locus is linked with the disease under investigation, while it might actually be under TD in the general population [Greenwood et al., 2000; Paterson et al., 2003; 2009].

1.6. Aims and Scope of this Work.

As shown in this chapter, the major histocompatibility complex, linkage disequilibrium, MHC-linked olfactory receptor genes and allelic transmission distortion are subjects that are intimately related.

The specific objectives of the studies assessing OR genes were the following:

1. Investigate the genetic structure and phylogenetic relationships of MHC-linked OR genes among all animals for which enough sequence data has been produced and made available through public databases.
2. Investigate, through direct DNA sequencing and the use of publicly available sequence data, the LD structure and the presence of unknown polymorphisms within human MHC-linked OR genes, in order to generate an overview of human genotypic and phenotypic variation of this genomic region. Additionally, assess the possibility that new cases of loci can be found in which intact genes and pseudogenes segregate in different haplotypes.
3. Investigate the degree of association between polymorphisms within human MHC-linked OR genes with smoking habits, focusing on HLA haplotypes that were previously described to be associated with diseases that are strongly influenced by smoking.
4. Generate a panel of LD blocks and tagSNPs for the human MHC-linked OR genes, using population data available through public databases.

The specific objective of the study assessing the MHC was the following:

1. Investigate the possibility of increasing genotyping efficiency of HLA complex alleles through the description of high levels of TD between these alleles and microsatellites.

The specific objectives of the studies assessing TD were the following:

1. Investigate the presence of TD within the extended MHC and surrounding genomic regions, using population data available through public databases.
2. Seek for confirmation of TD, trough direct genotyping of an independent population cohort.
3. Investigate possible relationships between LD and TD.

2. MHC-linked Olfactory Receptor Genes and Smoking

2.1. Summary

The work published by Füst and co-workers [Füst et al., 2004] described a correlation between a specific haplotype of the HLA complex (A1-B8-DR3) with smoking habits in Caucasian women. As the genes assessed lacked any clear link to smoking, as they were neither involved in nicotine metabolism or predisposition to addiction, etc, the primary association responsible for the findings could in fact be due to genes that are part of this haplotype, but located in the telomeric vicinity of the HLA, namely on one or both the olfactory receptor gene clusters. The work described in the following article aimed to test this hypothesis, as well as the possible association of smoking with two other haplotypes involved in smoking-related diseases that are known to be influenced by the HLA complex. The methodology integrated publicly available SNP genotyping information from 180 Caucasian haplotypes (CEU population) from the international HapMap project, as well as traditional "in house" genotyping methods. It could be determined, *in silico*, which polymorphism, among the hundreds available, should be assessed. A sample from the Füst study [Füst et al., 2004] was then genotyped, and the results revealed that one single non-synonymous SNP within the gene *OR12D3* was stronger correlated with smoking habits than the originally found association between smoking and the A1-B8-DR3 haplotype. For the first time, a human behaviour was found to be correlated to an olfactory receptor polymorphism.

2.2. Publication

Santos PS, Füst G, Prohászka Z, Volz A, Horton R, Miretti M, Yu CY, Beck S, Uchanska-Ziegler B, Ziegler A (2008): Association of smoking behavior with an odorant receptor allele telomeric to the human major histocompatibility complex. Genetic Testing 12: 481-486.

Pages 29-42

GENETIC TESTING
Volume 12, Number 4, 2008
© Mary Ann Liebert, Inc.
Pp. 481–486
DOI: 10.1089/gte.2008.0029

Association of Smoking Behavior with an Odorant Receptor Allele Telomeric to the Human Major Histocompatibility Complex

Pablo Sandro Carvalho Santos,[1] George Füst,[2] Zoltán Prohászka,[2,3] Armin Volz,[1] Roger Horton,[4] Marcos Miretti,[4] Chack-Yung Yu,[5] Stephan Beck,[6] Barbara Uchanska-Ziegler,[1] and Andreas Ziegler[1]

Smoking behavior has been associated in two independent European cohorts with the most common Caucasian human leukocyte antigen (HLA) haplotype (A1-B8-DR3). We aimed to test whether polymorphic members of the two odorant receptor (OR) clusters within the extended HLA complex might be responsible for the observed association, by genotyping a cohort of Hungarian women in which the mentioned association had been found. One hundred and eighty HLA haplotypes from Centre d'Etude du Polymorphisme Humain families were analyzed *in silico* to identify single-nucleotide polymorphisms (SNPs) within OR genes that are in linkage disequilibrium with the A1-B8-DR3 haplotype, as well as with two other haplotypes indirectly linked to smoking behavior. A nonsynonymous SNP within the *OR12D3* gene (rs3749971^{T}) was found to be linked to the A1-B8-DR3 haplotype. This polymorphism leads to a ^{97}Thr → Ile exchange that affects a putative ligand binding region of the OR12D3 protein. Smoking was found to be associated in the Hungarian cohort with the rs3749971^{T} allele ($p = 1.05 \times 10^{-2}$), with higher significance than with A1-B8-DR3 ($p = 2.38 \times 10^{-2}$). Our results link smoking to a distinct OR allele, and demonstrate that the rs3749971^{T} polymorphism is associated with the HLA haplotype–dependent differential recognition of cigarette smoke components, at least among Caucasian women.

Introduction

There were nearly 1.3 billion smokers worldwide in the year 2003, and this number is expected to rise to 1.7 billion (~1.2 billion males and 500 million females) by 2025, with the number of female smokers contributing most to the increase (American Cancer Society, 2003). In nearly all investigated regions of the world, the ratio of female to male smokers among young people was found to be higher than the ratio among adults, suggesting a global trend for an increase in smoking habits among female adolescents and young women (Global Youth Tobacco Survey Collaborating Group, 2003). Smoking is associated with many serious health problems, including cancer of various organs, coronary artery disease, as well as several autoimmune disorders (Hegediüs *et al.*, 2004; Klareskog *et al.*, 2006; Warren *et al.*, 2006; American Cancer Society, 2007; Koch *et al.*, 2007; Hawkes, 2007), and it is thus considered a leading cause of death and disability worldwide. Although there is general agreement that nicotine is the core addictive component of cigarette smoke (Jarvis, 2004), there are hundreds of further substances that may influence the initiation and continuation of tobacco abuse (Baker *et al.*, 2004), independent of nicotine (Franklin *et al.*, 2007). Smoking is also modulated by genetic factors, as demonstrated by epidemiological and twin studies (Sullivan and Kendler, 1999; Li *et al.*, 2003). In support, a haplotype of the major histocompatibility (human leukocyte antigen, HLA) complex was found to be associated with smoking behavior, stronger in women (odds ratio: 13.6) than in men (odds ratio: 2.79) (Füst *et al.*, 2004). This haplotype, –HLA-A1-B8-DR3–, the most common among Caucasians (Alper *et al.*, 2006), is also associated with autoimmune disorders, of which some are clearly connected with tobacco abuse, such as Graves' ophthalmopathy (Weetman, 2000; Hunt *et al.*, 2001; Hegediüs *et al.*, 2004; Holm *et al.*, 2005).

At least two further HLA haplotypes have also been linked to autoimmune diseases that are triggered or heavily influenced by tobacco smoking. The HLA-A3-B7-DR15 haplotype

[1] Institut für Immungenetik, Charité-Universitätsmedizin Berlin, Berlin, Germany.
[2] Third Department of Internal Medicine and Szentagothai János Knowledge Center, Semmelweis University, Budapest, Hungary.
[3] Research Group of Inflammation Biology and Immunogenomics, National Academy of Sciences, Budapest, Hungary.
[4] Genome Campus, Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom.
[5] Center for Molecular and Human Genetics, Columbus Children's Research Institute and College of Medicine and Public Health, The Ohio State University, Columbus, Ohio.
[6] UCL Cancer Institute, University College London, London, United Kingdom.

is overrepresented in individuals with multiple sclerosis (Dyment *et al.*, 2004; Herrera *et al.*, 2006), while HLA-DR "Shared Epitope" (SE) haplotypes are overrepresented in individuals who smoke, possess antibodies against citrullinated proteins, and suffer from rheumatoid arthritis (Klareskog *et al.*, 2006; Linn-Rasker *et al.*, 2006). SE haplotypes are characterized by the presence of the HLA-DRB1*01, -DRB1*04, or -DRB1*10 alleles.

The exceptionally strong linkage disequilibrium (LD) that is typical for certain haplotypes of the xHLA encompasses a chromosomal segment with a length of about 7 Mb between the gene *HFE* and loci within the HLA class II region (Horton *et al.*, 2004). In case of the A1-B8-DR3 haplotype, LD is extreme over its entire length (Alper *et al.*, 2006), suggesting that an allele of any genes within the region of LD could in principle predispose to smoking. Therefore, polymorphic members of the two odorant receptor (OR) gene clusters within the telomeric xHLA (Ehlers *et al.*, 2000; Younger *et al.*, 2001; Ziegler and Uchańska-Ziegler, 2006) must be considered as plausible candidate genes.

The Centre d'Etude du Polymorphisme Humain (CEPH) panel of families (Miretti *et al.*, 2005) offers unique opportunities for genetic association studies comprising the HLA region. The samples analyzed here comprise 180 founder chromosome 6 that have already been analyzed with regard to their single-nucleotide polymorphism (SNP) alleles and tagSNPs (representative SNPs for a genomic region exhibiting high LD) (Miretti *et al.*, 2005). Recently, the analysis has been extended to both OR gene clusters at the telomeric section of the xHLA, permitting the correlation with alleles within the HLA class I, II, and III regions (de Bakker *et al.*, 2006).

The present study was conducted in two steps: (i) *in silico*, we aimed to identify SNP alleles within the two HLA-linked OR gene clusters that are characteristic for the HLA haplotypes mentioned before; and (ii) *in vitro* we wanted to test whether their possible association with tobacco abuse is stronger or weaker than the one observed with A1-B8-DR3, by genotyping a subsample of the cohort in which an association between smoking and loci at the HLA class III region had been found before.

Materials and Methods

In silico *tagSNP selection and assessment of HLA haplotype–dependent OR SNP alleles*

All *in silico* analyses were based on xHLA high-density SNP genotyping data from 180 founder chromosome 6 from the CEPH collection (Miretti *et al.*, 2005). Details on marker selection, CEPH subjects, and on their genotypings are given elsewhere (Miretti *et al.*, 2005; de Bakker *et al.*, 2006). A total of 1170 SNPs spanning the region between *OR2B2* and *MOG* (Fig. 1) were initially considered and analyzed with regard to their allelic diversity (supplemental Fig. S1, available online at www.liebertpub.com). A list of all 1170 SNPs with their genomic coordinates is given in the supplemental Table S1 (available online at www.liebertpub.com). From this set of markers, we chose 110 tagSNPs capturing the haplotypic information from the whole genomic region. SNP tagging was performed using a pairwise tagging algorithm (de Bakker *et al.*, 2006), as implemented in HAPLOVIEW v. 4 (Barrett *et al.*, 2005), and considering a maximum intermarker distance of 650 kb and an LD coefficient (r^2) threshold of ≥0.8. Loci with a minor allele frequency of less than 1%, those not conforming to Hardy–Weinberg equilibrium, or not reported by the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP) were excluded.

We tested the 110 tagSNPs characterizing the *OR2B2-MOG* segment for association with three groups of haplotypes (A1-B8-DR3, A3-B7-DR15, and SE), the aim being to determine which tagSNPs were characteristic for each of the three haplotype groups.

Genotyping of SNP rs3749971

The genotyping of the SNP *rs3749971* was performed by real-time PCR in a sample of 32 Hungarian female Caucasians (average age 46.75 years, ranging from 24 to 76 years), which

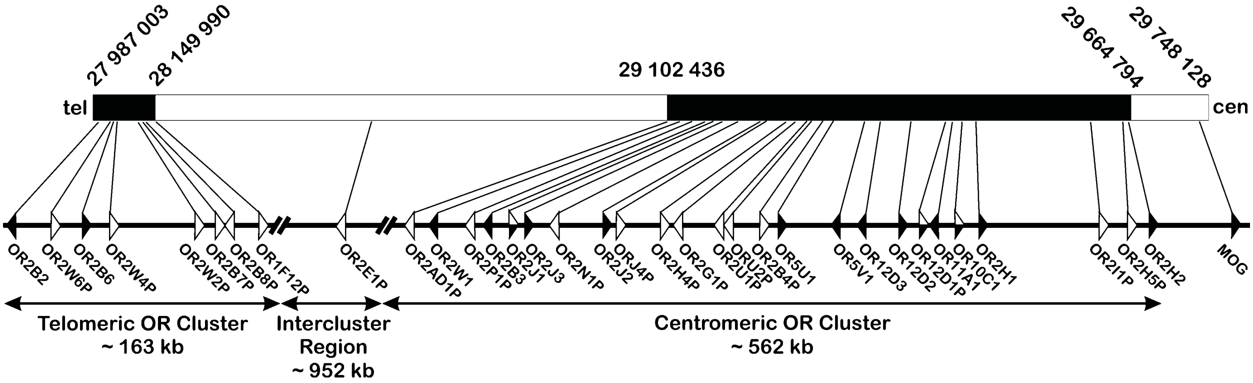


FIG. 1. Map (not to scale) of the region encompassing the two OR clusters on chromosome 6p. Above the figure, the NCBI 36 coordinates (bp) of the chromosomal segments analyzed here are depicted. The upper plot is filled in black for OR clusters and in white for regions outside the OR clusters, while the lower part of the figure shows the OR genes/pseudogenes at their relative chromosomal locations (suppressing most of the intercluster region). Triangles indicate transcriptional orientation, and are filled in black (genes), white (pseudogenes), or in both colors (haplotype-dependent genes/pseudogenes). tel: direction to the telomere; cen: direction to the centromere.

was a subsample of a cohort in which a correlation had previously been found between smoking and loci of the HLA class III region (Füst *et al.*, 2004). Based on LD, women who carried the C4A*Q0 (mono-S) genotype as well as the AGER-429C, HSPA1B-1267G, and TNF-308A alleles were considered carriers of the A1-B8-DR3 haplotype (Füst *et al.*, 2004). Other genotypings of these subjects, details on registration of smoking habits, preparation of genomic DNA, as well as informed consent from the cohort have been provided before (Füst *et al.*, 2004).

The PCR reactions were performed using a Stratagene realtime PCR instrument (Stratagene, Amsterdam, NL) with the following cycle conditions: 1 cycle of 94°C for 1 min and 80 cycles of 94°C for 20 s, 62°C for 20 s, 72°C for 30 s, and 78°C for 10 s. Fluorescence was measured during the 62°C (annealing) step. Twenty-microliter PCR reactions contained 100 ng of human genomic DNA; 1.5 pmol C- or T-specific reverse primers (details below); 3 pmol forward primer (rs3749971-For: AGCGAAGAGGATTGCAGATGGC); 2 μL Genetherm polymerase buffer; 0.7 mM each of dATP, dCTP, dGTP, and dTTP; 2 mM MgCl; and 2 U of GenTherm™ Taq Polymerase. Allele-specific primers were designed as molecular beacons (Jordens *et al.*, 2000) (rs3749971-C: Fam-atacagc CTATATCTTTTCTAGGCTGTA$\boldsymbol{T}_{\mathrm{BHQ}}$CA**C** and rs3749971-T: Hex-atacagcCTATATCTTTTCTAGGCTGTA$\boldsymbol{T}_{\mathrm{BHQ}}$CA**T**), labeled either with FAM (6-carboxyfluorescein) or Hex (4,7,2′,4′,5′,7′-hexachloro-6-carboxyfluorescein), and quenched with Black Hole Quencher (BHQ™) attached to a ***T*** natively present within the primer binding site. Seven bases were added to the 5′ end (displayed in lower-case letters) to allow the formation of a 29-bp "hairpin" with the 10-bp complementary region to ensure fluorescence quenching of the unused primers. To assess the reproducibility of this genotyping approach, control DNA of seven individuals (including three CEPH samples) was typed 15 times independently, with 100% reproducibility.

Statistical analyses

Nucleotide diversity (π) was calculated as described by Nei (1987), using the software DNAsp (Rozas and Rozas, 1999). The two-sided Fisher's exact test was employed for all association analyses, with a 1% level of significance. The Bonferroni correction for multiple comparisons was applied for *p*-values of the *in silico* analyses only.

Results

In silico *analyses*

The degree of nucleotide diversity as assessed by 1170 SNPs (elicited in the supplemental Table S1) was found to be highest between the genes *OR12D3* and *OR10C1* (supplemental Fig. S1), including the most polymorphic loci within this cluster (Ehlers *et al.*, 2000). The SNP diversity outside of the OR clusters did not differ substantially from that within the clusters. SNP densities within the telomeric (0.388/kb) and the centromeric OR clusters (0.877/kb) were relatively high when compared with other genomic regions (Zhao *et al.*, 2003).

Within the panel of 180 xHLA haplotypes, 11 A1-B8-DR3, 5 A3-B7-DR15, and 59 HLA-DR SE haplotypes were found. No significant association between any of the tagSNPs and the 59 HLA-DR SE haplotypes was observed, as shown in Figure 2. A similar result was obtained with regard to the A3-B7-DR15 haplotypes, although this could be due to the low number of A3-B7-DR15 haplotypes (five) in the CEPH panel. In contrast, a significant correlation was found for 12 tagSNPs (representing 81 tagged SNPs displayed in supplemental Table S2, available online at www.liebertpub.com) when the 11 A1-B8-DR3 haplotypes were compared with the 169 non–A1-B8-DR3 haplotypes for association with the 110 tagSNPs (Fig. 2), and after setting a Bonferroni cutoff for multiple comparisons. Only one of the 81 captured SNPs is a coding, nonsynonymous SNP (*rs3749971*, tagged by tagSNP #51), within the gene *OR12D3*. In A1-B8-DR3 haplotypes, the respective allele (rs3749971$^{\mathrm{T}}$) is responsible for a Thr→Ile exchange at amino acid position 97 within the OR12D3 protein. The remaining 80 SNPs lead either to synonymous exchanges or are located in intergenic or in intronic regions. Because none of these are, to our knowledge, directly involved in any, so far known, biological process, they were not considered further.

Since the A1-B8-DR3 haplotype was reported in another independent European cohorts to be overrepresented in

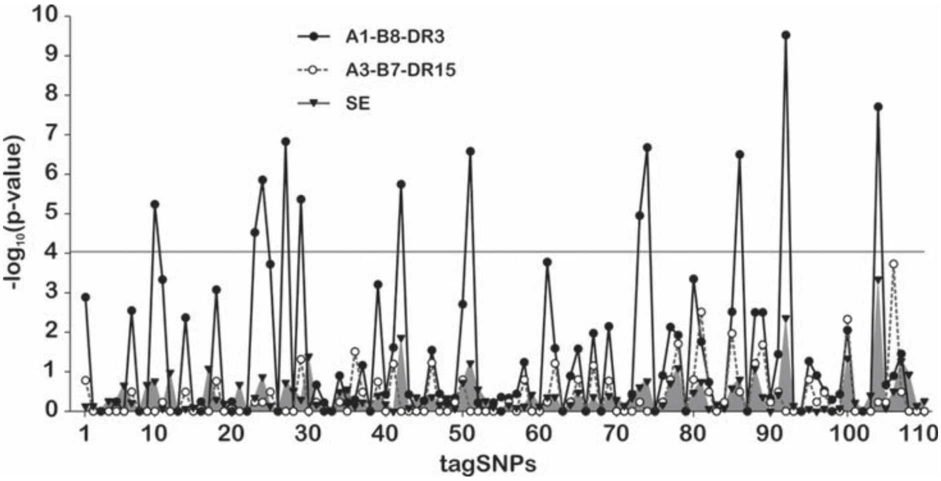


FIG. 2. Allelic association of 110 tagSNPs from the region comprising both HLA-linked OR clusters with A1-B8-DR3, A3-B7-DR15, and SE haplotypes. The marker *rs3749971* is represented by tagSNP #51. The gray horizontal line marks the Bonferroni threshold for 110 tests with a significance level of 1%.

smokers (Icelandic sample, Füst *et al.*, 2004), and is also associated with various autoimmune diseases, of which some correlate with this behavioral trait, as in Graves' disease (Hegediüs *et al.*, 2004; Holm *et al.*, 2005), this data demonstrate that these associations must also extend to the A1-B8-DR3–associated rs3749971^{T} allele.

Cohort genotyping

In order to validate this finding, comparing it to the previously found association, we genotyped *rs3749971* in 32 female individuals from the same Hungarian cohort (Füst *et al.*, 2004). These individuals differ in their smoking behavior and do not suffer from any autoimmune disease (Füst *et al.*, 2004). Here we found that the correlation between smoking and rs3749971^{T} was slightly stronger ($p = 1.05\times10^{-2}$, Fig. 3b) than the correlation between this trait and A1-B8-DR3 ($p = 2.38\times10^{-2}$, Fig. 3c). Seven and 25 subjects were rs3749971$^{C/T}$ and rs3749971$^{C/C}$ carriers, respectively. A strong association ($p = 5.87\times10^{-4}$) between the rs3749971^{T} allele and the A1-B8-DR3 haplotype was observed, as five out of seven of the rs3749971^{T} carriers but only 1/25 of the noncarriers were found to be A1-B8-DR3 positive (Fig. 3d). The group of rs3749971$^{C/C}$ carriers includes both smokers and nonsmokers, as shown in Figure 3a and b.

Discussion

HLA-A1–positive individuals, in contrast to those with HLA-A2 or HLA-A3, have been reported to exhibit a preference for the odor of bergamot (Milinski and Wedekind, 2001). The existence of a relationship between rs3749971^{T} (or OR12D3^{97Ile}) and this predilection for a perfume ingredient is supported by the strong LD between HLA-A1 and rs3749971^{T} that is described here. The fact that many tobacco brands are scented with bergamot oil components (Baker *et al.*, 2004) provides a plausible explanation for the correlation of rs3749971^{T} with smoking, and the identification of OR12D3^{97Ile} ligands will facilitate the design of volatiles that might be used as antidotes in individuals with a predisposition to tobacco abuse. The exchange of the hydrophilic amino acid threonine by isoleucine, a residue with a hydrophobic side chain, is expected to alter the physicochemical properties of the OR12D3 protein. However, as no X-ray crystallographic studies of ORs have been reported to date, the likely location of ^{97}Ile within the OR12D3 protein can only be inferred from models (Man *et al.*, 2004; Katada *et al.*, 2005; Abaffy *et al.*, 2007; Schmiedeberg *et al.*, 2007) that take the structure of rhodopsin (Palczewski *et al.*, 2000) into account. Such models locate residue 97 at the end of the first extracellular loop or at the beginning of the third transmembrane domain, close to

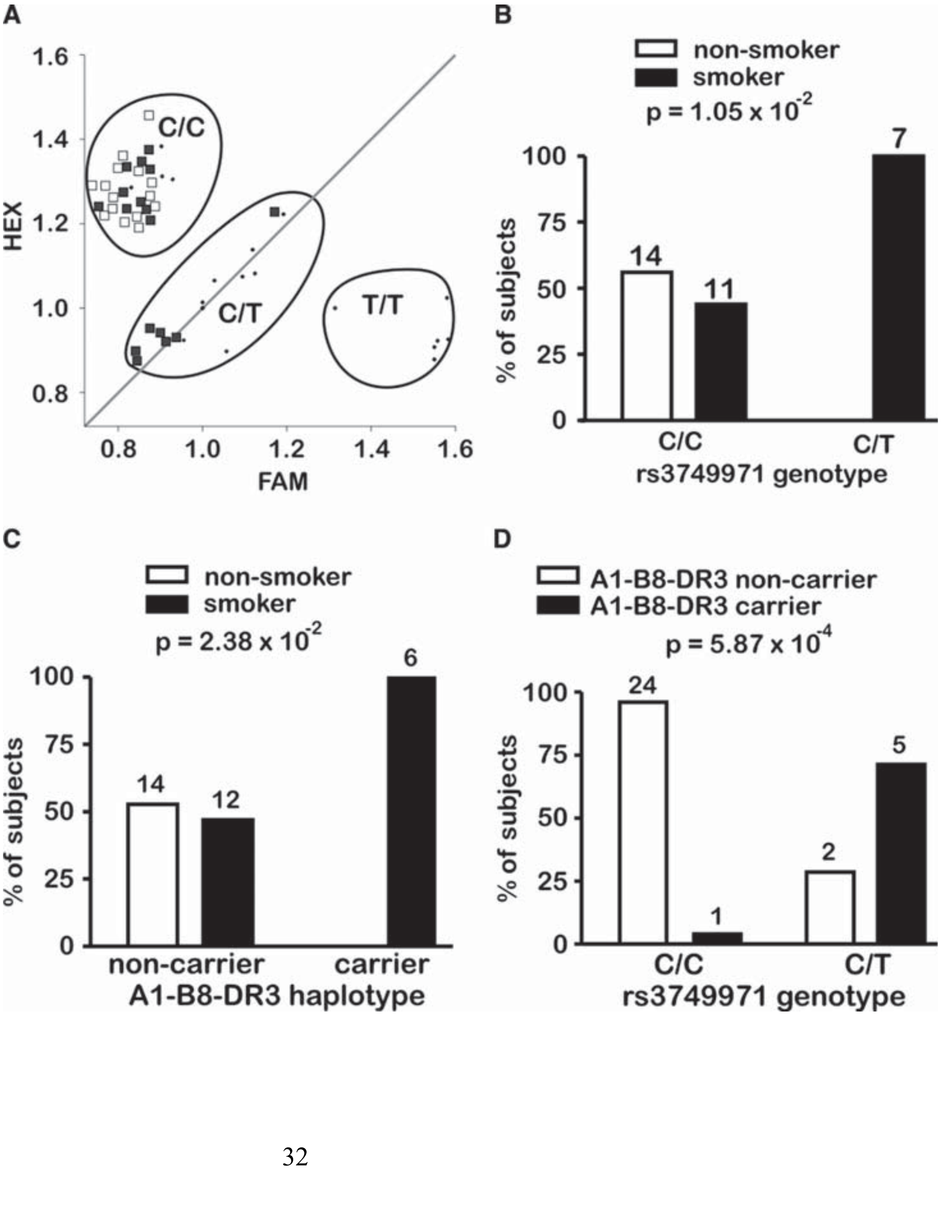


FIG. 3. Association of smoking behavior with the *rs3749971* genotype and the A1-B8-DR3 haplotype. (**A**) HEX and FAM are primer fluorophores marking the *rs3749971* T and the C allele primers, respectively. Values indicate the normalized relative number of PCR cycles necessary to reach the genotyping threshold for each allele. HEX values around 1 (or below) are indicative for the presence of the T allele, and the same is true for FAM values, indicating the presence of the C allele. Filled squares: 17 smokers; open squares: 15 nonsmokers; dots: controls. (**B**) Association of smoking behavior with *rs3749971* genotypes. (**C**) Association of smoking behavior with the carrier state of A1-B8-DR3. (**D**) Association of A1-B8-DR3 with *rs3749971* genotypes. (**B–D**) The number of individuals in each category is indicated above the corresponding bars, as well as the *p*-values, as determined by Fisher's exact test.

the ligand binding site. Keller *et al*. (2007) have recently demonstrated that a variant of an OR gene can substantially influence sensitivity (in both intensity and pleasantness) to specific odors in humans. They showed that a mutant allele of the *OR7D4* gene encoding an OR with two amino acid substitutions (residues 88 and 133) as compared to the most common allele causes functional impairment of the receptor *in vitro*. These substitutions also alter the perception of the smell of androstenone and androstadienone in a significant manner. Residues 88 and 133 are located within the first extracellular and the second intracellular loop. At least residue 88 is close to the beginning of the third transmembrane domain, suggesting that the region in the vicinity of residue 97 in the OR12D3 protein might indeed be involved in ligand binding. Further support for a role of amino acids close to the residue 97 in ligand binding is provided by the recently published crystal structure of the human β_2 adrenergic G-protein–coupled receptor (Rasmussen *et al.*, 2007).

Apart from social and psychological factors (Pomerleau *et al.*, 1992; Barman *et al.*, 2004; Lerer *et al.*, 2006), it has been shown that the individual genetic constitution plays a role in initiating and continuing tobacco consumption (Sullivan and Kendler, 1999; Li *et al.*, 2003). Genes within the xHLA contribute as well, as suggested by the finding of an HLA haplotype–dependent association of smoking in two independent European cohorts, with a clear-cut gender bias toward females (Füst *et al.*, 2004). The present study confirms these results by identifying an HLA-linked OR allele that is associated with tobacco abuse. In addition, our work implies that the rs3749971^T allele is involved in the smoking-induced aggravation of certain autoimmune diseases that can be observed in patients carrying A1-B8-DR3 (Hegediüs *et al.*, 2004; Holm *et al.*, 2005). Given the fact that the A1-B8-DR3 haplotype occurs with a frequency of 5–10% in Caucasians (Alper *et al.*, 2006), the *rs3749971* polymorphism should be suitable for large-scale screening tests, particularly among young people. Studies of the isolated OR12D3 protein will be indispensable to evaluate the consequences of the Thr97Ile exchange within this receptor on its ligand specificity.

Acknowledgments

This work was supported by the Berliner Krebsgesellschaft and by the Senate of Berlin (Ernst-von-Leyden and NaFöG fellowships to P.S.C.S.), by the Wellcome Trust (R.H., M.M. and S.B) as well as by the Monika Kutzner-Stiftung Berlin, and by the National Office for Research and Technology, Hungary.

Disclosure Statement

The authors declare that no competing interests exist.

References

Abaffy T, Malhotra A, Luetje CW (2007) The molecular basis for ligand specificity in a mouse olfactory receptor: a network of functionally important residues. J Biol Chem 282:1216–1224.

Alper CA, Larsen CE, Dubey DP, *et al.* (2006) The haplotype structure of the human major histocompatibility complex. Hum Immunol 67:73–84.

American Cancer Society (2003) 12th World Conference on Tobacco or Health. Tobacco Control Profiles. American Cancer Society, Atlanta.

American Cancer Society (2007) Cancer facts and figures 2007. American Cancer Society, Atlanta.

Baker RR, Pereira da Silva JR, Smith G (2004) The effect of tobacco ingredients on smoke chemistry. Part I: Flavorings and additives. Food Chem Toxicol 42:S3–S37.

Barman SK, Pulkkinen L, Kaprio J, Rose RJ (2004) Inattentiveness, parental smoking and adolescent smoking initiation. Addiction 99:1049–1061.

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265.

de Bakker PIW, McVean G, Sabeti PC, *et al.* (2006) A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet 38:1166–1172.

Dyment DA, Ebers GC, Sadovnick AD (2004) Genetics of multiple sclerosis. Lancet Neurol 3:104–110.

Ehlers A, Beck S, Forbes SA, *et al.* (2000) MHC-linked olfactory receptor loci exhibit polymorphism and contribute to extended HLA/OR-haplotypes. Genome Res 10:1968–1978.

Franklin TR, Wang Z, Wang J, *et al.* (2007) Limbic activation to cigarette smoking cues independent of nicotine withdrawal: a perfusion FMRI study. Neuropsychopharmacology 32:2301–2309.

Füst G, Arason GJ, Kramer J, *et al.* (2004) Genetic basis of tobacco smoking: strong association of a specific major histocompatibility complex haplotype on chromosome 6 with smoking behavior. Int Immunol 16:1507–1514.

Global Youth Tobacco Survey Collaborating Group (2003) Differences in worldwide tobacco use by gender: findings from the Global Youth Tobacco Survey. J Sch Health 73:207–215.

Hawkes CH (2007) Smoking is a risk factor for multiple sclerosis: a metaanalysis. Mult Scler 13:610–615.

Hegediüs L, Brix TH, Vestergaard P (2004) Relationship between cigarette smoking and Graves' ophthalmopathy. J Endocrinol Invest 27:265–271.

Herrera BM, Cader MZ, Dyment DA, *et al.* (2006) Follow-up investigation of 12 proposed linkage regions in multiple sclerosis. Genes Immun 7:366–371.

Holm IA, Manson JE, Michels KB, *et al.* (2005) Smoking and other lifestyle factors and the risk of Graves' hyperthyroidism. Arch Intern Med 165:1606–1611.

Horton R, Wilming L, Rand V, *et al.* (2004) Gene map of the extended human MHC. Nat Rev Genet 5:889–899.

Hunt PJ, Marshall SE, Weetman AP, *et al.* (2001) Histocompatibility leucocyte antigens and closely linked immunomodulatory genes in autoimmune thyroid disease. Clin Endocrinol 55:491–499.

Jarvis MJ (2004) Why people smoke. BMJ 328:277–279.

Jordens JZ, Lanham S, Pickett MA, *et al.* (2000) Amplification with molecular beacon primers and reverse line blotting for the detection and typing of human papillomaviruses. J Virol Meth 89:29–37.

Katada S, Hirokawa T, Oka Y, *et al.* (2005) Structural basis for a broad but selective ligand spectrum of a mouse olfactory receptor: mapping the odorant-binding site. J Neurosci 25:1806–1815.

Keller A, Zhuang H, Chi Q, *et al.* (2007) Genetic variation in a human odorant receptor alters odour perception. Nature 449: 468–472.

Klareskog L, Stolt P, Lundberg K, *et al.* (2006) A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination. Arthritis Rheum 54:38–46.

Koch M, van Harten A, Uyttenboogaart M, de Keyser J (2007) Cigarette smoking and progression in multiple sclerosis. Neurology 69:1515–1520.

Lerer E, Kanyas K, Karni O, *et al.* (2006) Why do young women smoke? II. Role of traumatic life experience, psychological characteristics and serotonergic genes. Mol Psychiatry 11:771–781.

Li MD, Cheng R, Ma JZ, Swan GE (2003) A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. Addiction 98:23–31.

Linn-Rasker SP, van der Helm-van Mil AHM, van Gaalen FA, *et al.* (2006) Smoking is a risk factor for anti-CCP antibodies only in rheumatoid arthritis patients who carry HLA-DRB1 shared epitope alleles. Ann Rheum Dis 65:366–371.

Man O, Gilad Y, Lancet D (2004) Prediction of the odorant binding site of olfactory receptor proteins by human-mouse comparisons. Protein Sci 13:240–254.

Milinski M, Wedekind C (2001) Evidence for MHC-correlated perfume preferences in humans. Behav Ecol 12:140–149.

Miretti MM, Walsh EC, Ke X, *et al.* (2005) A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. Am J Hum Genet 76:634–646.

Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press, New York.

Palczewski K, Kumasaka T, Hori T, *et al.* (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. Science 289:739–745.

Pomerleau CS, Pomerleau OF, Flessland KA, Basson SM (1992) Relationship of tridimensional personality questionnaire scores and smoking variables in female and male smokers. J Subst Abuse 4:143–154.

Rasmussen SGF, Choi H, Rosenbaum DM, *et al.* (2007) Crystal structure of the human β2 adrenergic G-protein-coupled receptor. Nature 450:383–387.

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175.

Schmiedeberg K, Shirokova E, Weber HP, *et al.* (2007) Structural determinants of odorant recognition by the human olfactory receptors OR1A1 and OR1A2. J Struct Biol 159:400–412.

Sullivan PF, Kendler KS (1999) The genetic epidemiology of smoking. Nicotine Tob Res 1:S51–S59.

Warren CW, Jones NR, Eriksen MP, Asma S (2006) Global Tobacco Surveillance System (GTSS) collaborative group: patterns of global tobacco use in young people and implications for future chronic disease burden in adults. Lancet 367: 749–753.

Weetman AP (2000) Graves' disease. N Engl J Med 343:1236–1248.

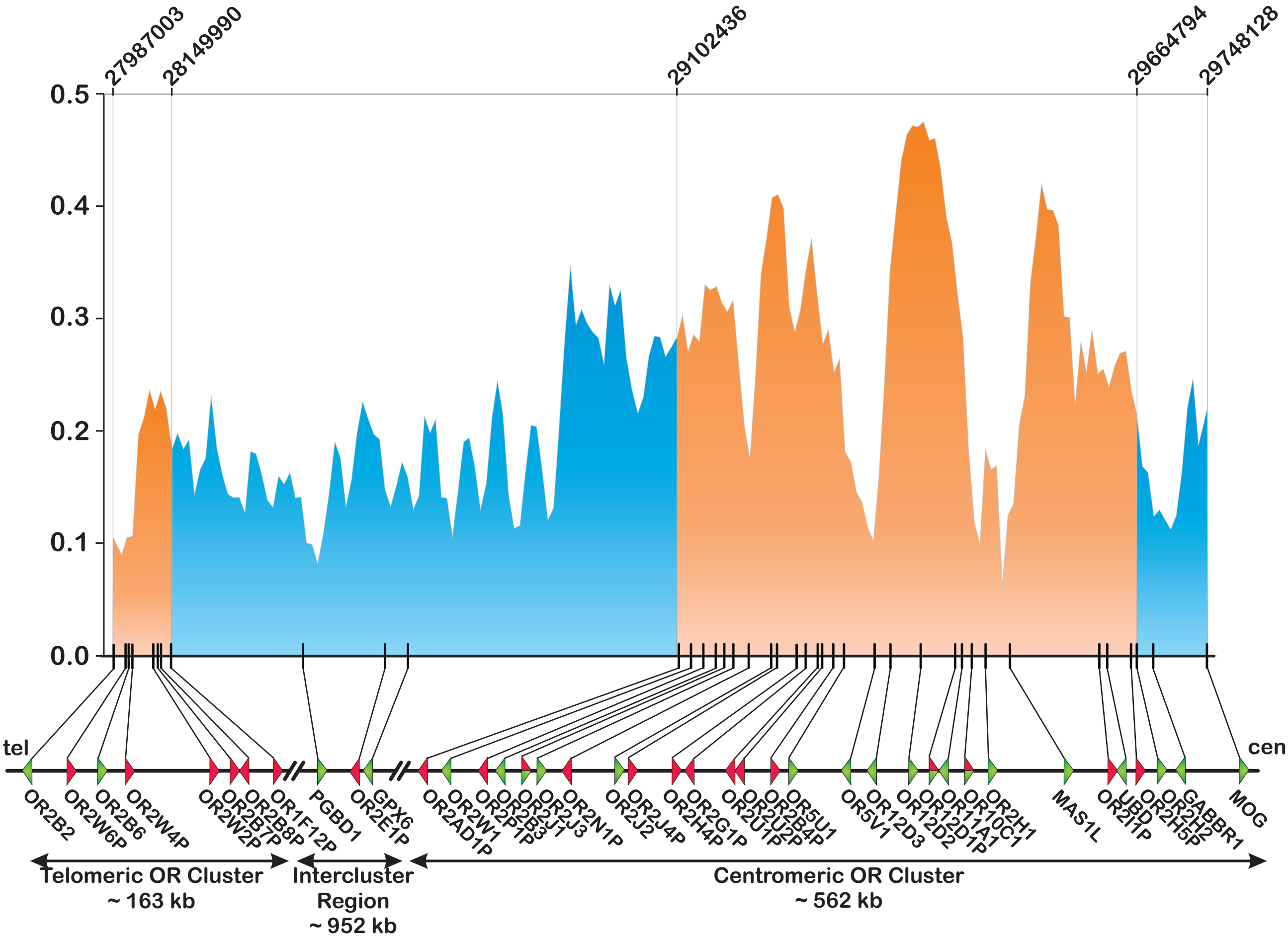
Younger RM, Amadou C, Bethel G, *et al.* (2001) Characterization of clustered MHC-linked olfactory receptor genes in human and mouse. Genome Res 11:519–530.

Zhao Z, Fu YX, Hewett-Emmett D, Boerwinkle E (2003) Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene 312:207–213.

Ziegler A, Uchańska-Ziegler B (2006) Rheumatoid arthritis etiology and HLA-linked odorant receptor gene polymorphisms: comment on the article by Klareskog *et al.* Arthritis Rheum 54:2705–2706.

Address reprint requests to:
Prof. Dr. Andreas Ziegler
Institut für Immungenetik
Charité-Universitätsmedizin Berlin
Thielallee 73
14195–Berlin
Germany

E-mail: andreas.ziegler@charite.de



Supplementary Figure 1: Diversity of 1170 SNPs in a region of the xHLA complex
This region encompasses the two odorant receptor (OR) clusters on chromosome 6p, between the genes *OR2B2* and *MOG*). The average number of nucleotide differences per site ( ) was determined based on sliding windows (18 bases wide, stepping 6 bases after each measurement). The regions of both OR clusters are plotted in orange, and blue is used for the regions outside them. The ticks under the graph mark the position of OR genes and pseudogenes in relation to the plot. Above the graph, the NCBI 36 coordinates (bp) of selected loci are depicted. The line underneath the graph shows OR genes and pseudogenes at their approximate chromosomal locations. Triangles indicate transcriptional orientation, and are filled in green (genes), red (pseudogenes) or in both colours (haplotype-dependent genes/pseudogenes).
tel: directon to the telomere; cen: direction to the centromere

Supplementary Table 1

Compilation of all 1170 SNPs assessed in the genomic region between *OR2B2* and *MOG*, with indication of ID at the dbSNP database, and genomic coordinates according to NCBI36. SNPs chosen as tagSNPs are indicated.

Order	dbSNP ID	Coordinate (NCBI 36)	tagSNP
1	*rs2394015*	27978337	
2	*rs200945*	27989883	
3	*rs201910*	27991536	
4	*rs13218875*	27991991	
5	*rs9468254*	27992895	
6	*rs1015075*	27995254	
7	*rs6927241*	27997313	
8	*rs6933825*	27998610	
9	*rs156737*	28003192	tagSNP #1
10	*rs9468256*	28003483	
11	*rs12663899*	28007558	
12	*rs7760871*	28009020	tagSNP #2
13	*rs7742529*	28010848	
14	*rs7742858*	28010971	
15	*rs10456357*	28011446	
16	*rs6910968*	28012160	
17	*rs2893937*	28018029	
18	*rs276366*	28035460	
19	*rs9295753*	28036424	tagSNP #3
20	*rs276364*	28039599	
21	*rs276363*	28041280	
22	*rs7769416*	28056553	
23	*rs9380031*	28062677	tagSNP #4
24	*rs7748445*	28069926	tagSNP #5
25	*rs1143887*	28072802	
26	*rs149901*	28073482	
27	*rs156743*	28075068	
28	*rs12333142*	28076983	
29	*rs149946*	28078010	
30	*rs6939591*	28087157	
31	*rs10484402*	28087604	tagSNP #6
32	*rs9368540*	28089652	
33	*rs149971*	28090131	
34	*rs149972*	28091206	
35	*rs149973*	28091592	
36	*rs149974*	28093075	
37	*rs6456804*	28093367	
38	*rs2840222*	28093541	
39	*rs2893947*	28094129	
40	*rs149975*	28094319	
41	*rs9368545*	28107023	
42	*rs1529749*	28108340	
43	*rs3926997*	28108508	
44	*rs149941*	28109012	
45	*rs149942*	28109589	
46	*rs149943*	28110367	
47	*rs185741*	28111250	
48	*rs149897*	28114629	tagSNP #7
49	*rs149896*	28116503	
50	*rs175954*	28119564	
51	*rs149900*	28122576	
52	*rs149962*	28123897	
53	*rs149965*	28126668	
54	*rs9393879*	28126923	
55	*rs203888*	28129568	
56	*rs9468274*	28134056	
57	*rs6903723*	28135789	
58	*rs13198131*	28137003	
59	*rs9380046*	28138478	
60	*rs7741570*	28141484	
61	*rs149952*	28141864	
62	*rs149957*	28144999	
63	*rs9380047*	28145872	
64	*rs203881*	28146096	
65	*rs4713135*	28147565	
66	*rs5021186*	28149070	
67	*rs6913038*	28152735	tagSNP #8
68	*rs3956922*	28153995	
69	*rs172164*	28154457	
70	*rs203876*	28154652	
71	*rs12211199*	28155049	
72	*rs203877*	28156603	
73	*rs9393885*	28157988	
74	*rs427348*	28166344	tagSNP #9
75	*rs1853097*	28166614	
76	*rs175955*	28167713	
77	*rs13197574*	28168218	
78	*rs3823180*	28169723	
79	*rs203893*	28170045	
80	*rs9380051*	28172070	
81	*rs169433*	28172715	
82	*rs203891*	28175777	
83	*rs10946952*	28181443	
84	*rs203885*	28184865	
85	*rs203884*	28185353	
86	*rs6924324*	28190345	
87	*rs9368554*	28190690	
88	*rs1654775*	28191404	
89	*rs4713137*	28191500	
90	*rs6456807*	28194055	
91	*rs1770132*	28194644	
92	*rs478616*	28196359	
93	*rs7747772*	28198641	
94	*rs539474*	28200285	
95	*rs6922063*	28202345	
96	*rs1770129*	28203125	
97	*rs4713140*	28205172	tagSNP #10
98	*rs1383394*	28205474	
99	*rs6914592*	28208091	
100	*rs4713141*	28209657	
101	*rs6941992*	28214120	
102	*rs4713142*	28214326	tagSNP #11
103	*rs4713145*	28214806	
104	*rs3757188*	28215336	
105	*rs3757186*	28215641	
106	*rs1904841*	28216064	
107	*rs1904840*	28216211	
108	*rs2791331*	28216691	
109	*rs2277103*	28217612	
110	*rs868987*	28218127	
111	*rs1890809*	28218457	
112	*rs2791333*	28219093	
113	*rs1225715*	28221352	
114	*rs4713149*	28227777	
115	*rs2622318*	28227865	
116	*rs6908414*	28229807	
117	*rs6932313*	28230328	
118	*rs1225604*	28230544	
119	*rs1150666*	28231907	
120	*rs1150667*	28232042	
121	*rs6928131*	28233428	
122	*rs1225603*	28233606	
123	*rs9393897*	28235689	
124	*rs1225618*	28237692	
125	*rs1150671*	28239022	
126	*rs1564442*	28240405	tagSNP #12
127	*rs1225713*	28241091	
128	*rs9283884*	28243639	
129	*rs1144707*	28243978	
130	*rs3173443*	28245006	
131	*rs4713152*	28245433	
132	*rs9393902*	28245709	tagSNP #13
133	*rs1150675*	28245996	
134	*rs1150677*	28248113	
135	*rs1150678*	28250255	
136	*rs2840214*	28250520	
137	*rs6904045*	28252208	
138	*rs6904277*	28252423	
139	*rs1233667*	28254653	
140	*rs9295761*	28255966	
141	*rs1225591*	28256731	
142	*rs1225593*	28258498	
143	*rs1150683*	28263293	
144	*rs12205680*	28265083	
145	*rs1237873*	28266933	
146	*rs1233704*	28274902	
147	*rs1233702*	28275869	
148	*rs1233701*	28276705	
149	*rs768484*	28277094	
150	*rs1233698*	28277602	
151	*rs735765*	28278276	
152	*rs1233708*	28281198	tagSNP #14
153	*rs1150696*	28283528	
154	*rs1150697*	28283615	
155	*rs1233666*	28284394	
156	*rs1150700*	28290349	
157	*rs1150701*	28291865	
158	*rs1233664*	28293705	tagSNP #15
159	*rs1736894*	28294533	
160	*rs1736892*	28294957	
161	*rs1736891*	28295080	
162	*rs11967622*	28296593	
163	*rs1736890*	28297368	
164	*rs12211332*	28298142	
165	*rs3757183*	28299838	
166	*rs9461443*	28302608	
167	*rs11751492*	28303299	
168	*rs1150706*	28304593	
169	*rs2299031*	28304697	
170	*rs1150707*	28305584	
171	*rs1679723*	28307579	
172	*rs1233663*	28308211	
173	*rs7206*	28309117	
174	*rs6918043*	28309447	
175	*rs11752073*	28312772	tagSNP #16
176	*rs7769054*	28317176	
177	*rs1736889*	28317269	
178	*rs967005*	28318667	
179	*rs7757215*	28320468	
180	*rs1150712*	28321218	
181	*rs13200462*	28326178	
182	*rs1679732*	28329243	
183	*rs10456362*	28329795	
184	*rs2185955*	28330959	
185	*rs2394048*	28331170	
186	*rs7750106*	28331659	tagSNP #17
187	*rs12000*	28335415	
188	*rs1635*	28335583	
189	*rs1679705*	28337175	
190	*rs1778508*	28337860	
191	*rs1778507*	28338696	
192	*rs9468322*	28339222	
193	*rs2799076*	28340257	
194	*rs4713154*	28340648	
195	*rs4713155*	28342381	
196	*rs2799077*	28342576	
197	*rs1778484*	28348777	
198	*rs1419183*	28350773	
199	*rs1150725*	28351679	
200	*rs1150724*	28358215	tagSNP #18
201	*rs2142731*	28358892	
202	*rs11756111*	28360838	tagSNP #19
203	*rs1150722*	28361511	
204	*rs12526248*	28362667	
205	*rs13211507*	28365356	
206	*rs2039813*	28373455	
207	*rs2142730*	28374128	
208	*rs2223287*	28375287	
209	*rs2281043*	28376476	
210	*rs6456811*	28378026	tagSNP #20
211	*rs1062169*	28378230	
212	*rs1339899*	28379581	
213	*rs4711166*	28388516	
214	*rs6901990*	28389327	
215	*rs7759855*	28390842	
216	*rs13195487*	28395454	
217	*rs742107*	28398693	
218	*rs707907*	28399219	
219	*rs4711167*	28402867	tagSNP #21
220	*rs6902583*	28403512	
221	*rs9468343*	28407130	
222	*rs6456812*	28407861	
223	*rs6910120*	28408295	
224	*rs7752721*	28409146	
225	*rs4713161*	28410483	
226	*rs1416920*	28410763	
227	*rs1416919*	28410863	
228	*rs723476*	28413083	
229	*rs6908459*	28413472	
230	*rs213244*	28414887	tagSNP #22
231	*rs2108926*	28416726	tagSNP #23
232	*rs1119211*	28417115	tagSNP #24
233	*rs6918631*	28420435	
234	*rs6929449*	28421781	
235	*rs213240*	28423854	
236	*rs9468346*	28425693	
237	*rs6921919*	28433180	
238	*rs213233*	28433654	
239	*rs1555047*	28436662	
240	*rs1555046*	28437016	
241	*rs213230*	28438243	
242	*rs213229*	28438666	
243	*rs213228*	28439231	tagSNP #25
244	*rs7773018*	28439310	
245	*rs11751928*	28443357	
246	*rs7773051*	28448295	

Order	dbSNP ID	Coordinate (NCBI 36)	tagSNP
247	*rs7449675*	28448851	
248	*rs6916243*	28450900	
249	*rs6938371*	28451712	
250	*rs7769304*	28453760	
251	*rs7774981*	28454889	
252	*rs2072319*	28456639	
253	*rs7765637*	28456876	
254	*rs1556957*	28457905	
255	*rs3799500*	28460636	tagSNP #26
256	*rs7763846*	28461760	
257	*rs2859379*	28466174	
258	*rs2232428*	28467679	
259	*rs2232426*	28468638	tagSNP #27
260	*rs2232425*	28468817	
261	*rs2041230*	28473494	
262	*rs1005125*	28475334	
263	*rs6907950*	28478225	
264	*rs6908137*	28478372	
265	*rs4254981*	28479381	
266	*rs2531828*	28481769	
267	*rs2859361*	28483586	
268	*rs4382246*	28488227	
269	*rs6899389*	28489119	
270	*rs3757181*	28489875	
271	*rs2394098*	28490731	
272	*rs6922169*	28490911	
273	*rs10484540*	28495422	
274	*rs7451680*	28497230	
275	*rs1024629*	28498379	
276	*rs1015690*	28498505	tagSNP #28
277	*rs3800328*	28498822	
278	*rs2859365*	28499444	
279	*rs7740351*	28507391	
280	*rs7765989*	28508274	
281	*rs2859366*	28508416	
282	*rs6930745*	28510082	
283	*rs11752919*	28511582	
284	*rs2531801*	28512676	
285	*rs11750984*	28519426	
286	*rs2859367*	28522145	
287	*rs6939966*	28523864	tagSNP #29
288	*rs2531808*	28528763	
289	*rs2531810*	28529744	
290	*rs2531812*	28532025	
291	*rs7382749*	28533734	
292	*rs1041926*	28534275	tagSNP #30
293	*rs2021745*	28534965	
294	*rs2859372*	28535109	tagSNP #31
295	*rs2859374*	28536124	
296	*rs9468379*	28539163	tagSNP #32
297	*rs2859376*	28541699	
298	*rs2531816*	28550085	tagSNP #33
299	*rs12526477*	28550942	
300	*rs1029328*	28555894	
301	*rs1015811*	28556065	
302	*rs2531817*	28556873	
303	*rs2859350*	28561132	
304	*rs6927023*	28562200	
305	*rs993998*	28566341	
306	*rs13195077*	28570852	
307	*rs6913125*	28572692	
308	*rs2859356*	28573334	tagSNP #34
309	*rs12661782*	28574421	
310	*rs2191037*	28574967	
311	*rs2531822*	28576280	
312	*rs6937042*	28576408	
313	*rs2071966*	28581304	tagSNP #35
314	*rs11755403*	28581907	
315	*rs2108925*	28586203	
316	*rs6922986*	28586574	
317	*rs434112*	28588812	
318	*rs406113*	28591461	
319	*rs11757000*	28592848	
320	*rs403774*	28593258	
321	*rs423118*	28593659	
322	*rs9366720*	28595070	
323	*rs2215220*	28596853	
324	*rs377514*	28599929	
325	*rs6924526*	28604293	
326	*rs13202071*	28605748	
327	*rs769189*	28608214	
328	*rs448450*	28608985	
329	*rs1159276*	28609411	
330	*rs426922*	28610167	
331	*rs451774*	28610529	tagSNP #36
332	*rs413488*	28611336	
333	*rs4533983*	28613922	
334	*rs389118*	28615358	
335	*rs6940168*	28618957	
336	*rs454182*	28619143	
337	*rs2394102*	28619996	
338	*rs384399*	28621822	
339	*rs4711173*	28622562	
340	*rs393414*	28629295	tagSNP #37
341	*rs6456825*	28630674	
342	*rs414745*	28634634	
343	*rs422089*	28635385	
344	*rs418092*	28641925	
345	*rs454746*	28645230	
346	*rs12193707*	28650834	
347	*rs420463*	28664033	
348	*rs7738979*	28665291	tagSNP #38
349	*rs418914*	28676507	
350	*rs380914*	28677928	
351	*rs442439*	28680985	tagSNP #39
352	*rs7773645*	28682285	
353	*rs911178*	28682394	
354	*rs7381750*	28683907	
355	*rs10223644*	28686951	
356	*rs9501180*	28687450	
357	*rs13218450*	28689302	
358	*rs4711176*	28707007	
359	*rs7742658*	28708471	
360	*rs6925972*	28709913	
361	*rs6925044*	28713410	
362	*rs6933018*	28718340	tagSNP #40
363	*rs963937*	28718713	
364	*rs7773193*	28719313	
365	*rs7743296*	28722589	
366	*rs7382309*	28728176	
367	*rs6910105*	28731989	
368	*rs12210398*	28734376	
369	*rs9295775*	28735448	
370	*rs3051146*	28735513	
371	*rs9885928*	28758332	tagSNP #41
372	*rs4333411*	28759712	
373	*rs2394123*	28766679	
374	*rs1319076*	28770369	
375	*rs4143771*	28772610	
376	*rs12111360*	28775427	
377	*rs7383248*	28776459	
378	*rs6908726*	28779322	
379	*rs2893974*	28779877	
380	*rs6901325*	28783609	
381	*rs9368571*	28785737	
382	*rs7775835*	28786336	tagSNP #42
383	*rs1539586*	28787847	
384	*rs6937342*	28789489	
385	*rs9366725*	28793960	
386	*rs5029693*	28795521	
387	*rs6909960*	28796900	
388	*rs1419094*	28802371	
389	*rs2226184*	28802875	
390	*rs1954407*	28803128	
391	*rs7772682*	28803647	
392	*rs9393929*	28804042	
393	*rs12176003*	28805701	
394	*rs7755641*	28807348	
395	*rs10498733*	28807670	
396	*rs6456834*	28808331	
397	*rs7745385*	28808857	
398	*rs6456835*	28809471	
399	*rs7766599*	28811411	
400	*rs2394130*	28812754	tagSNP #43
401	*rs6905768*	28813784	
402	*rs1233573*	28814079	
403	*rs3869043*	28814513	
404	*rs880638*	28814890	
405	*rs3905240*	28819312	
406	*rs7764322*	28823296	
407	*rs2394128*	28826851	
408	*rs1233590*	28828699	
409	*rs1233591*	28828969	
410	*rs1233596*	28835668	
411	*rs1233597*	28836324	
412	*rs1233599*	28839178	
413	*rs6923801*	28839751	
414	*rs3852213*	28840354	
415	*rs1233603*	28842021	
416	*rs1233606*	28843503	
417	*rs1233610*	28844978	
418	*rs7741972*	28846980	
419	*rs1233616*	28847874	
420	*rs7771281*	28849271	
421	*rs4398727*	28851341	
422	*rs3888926*	28853266	
423	*rs1233619*	28853442	
424	*rs3891338*	28854516	
425	*rs4248135*	28854839	
426	*rs1233621*	28855641	
427	*rs7758798*	28857010	tagSNP #44
428	*rs7751425*	28860781	
429	*rs7751451*	28860880	
430	*rs1742753*	28861643	
431	*rs4324813*	28863574	
432	*rs12665150*	28867700	
433	*rs878587*	28869844	
434	*rs7745088*	28872512	
435	*rs7454923*	28874097	
436	*rs2765222*	28875063	
437	*rs2765219*	28878336	
438	*rs2754766*	28878742	
439	*rs1967743*	28879041	
440	*rs9368576*	28880337	
441	*rs11758303*	28880678	
442	*rs2754767*	28883228	
443	*rs3131343*	28883551	
444	*rs4324798*	28884104	
445	*rs417919*	28887410	
446	*rs6928657*	28888296	
447	*rs398795*	28892018	tagSNP #45
448	*rs386628*	28892383	
449	*rs377392*	28892728	tagSNP #46
450	*rs7762991*	28893425	
451	*rs172330*	28894871	
452	*rs6902254*	28896033	
453	*rs10484543*	28899189	
454	*rs209184*	28899276	
455	*rs209181*	28900470	
456	*rs880157*	28901772	
457	*rs11966708*	28902266	
458	*rs12660128*	28908373	
459	*rs3118357*	28910142	
460	*rs6928422*	28910791	
461	*rs9257248*	28911285	
462	*rs209176*	28912180	
463	*rs422331*	28918811	
464	*rs3131335*	28922440	
465	*rs12180795*	28930730	
466	*rs9501112*	28932952	
467	*rs9380095*	28933814	
468	*rs176461*	28934399	tagSNP #47
469	*rs9461483*	28935352	
470	*rs3132388*	28935873	
471	*rs209166*	28937066	
472	*rs6456858*	28938016	
473	*rs3132389*	28939015	
474	*rs3131336*	28939605	
475	*rs2187805*	28940474	
476	*rs3118370*	28941095	
477	*rs7775657*	28943246	
478	*rs209164*	28944450	
479	*rs3131093*	28945431	
480	*rs3132392*	28946623	
481	*rs209160*	28947873	
482	*rs7763661*	28951426	tagSNP #48
483	*rs209153*	28954243	
484	*rs209152*	28954456	
485	*rs209151*	28954718	
486	*rs9501154*	28956347	
487	*rs172326*	28957784	
488	*rs2006758*	28958376	
489	*rs1016472*	28959361	
490	*rs429369*	28959902	tagSNP #49
491	*rs6930087*	28960419	
492	*rs7750338*	28961151	
493	*rs2032500*	28962263	
494	*rs3135309*	28963799	
495	*rs2148008*	28963825	
496	*rs169679*	28964566	
497	*rs209148*	28965263	
498	*rs209147*	28966160	tagSNP #50
499	*rs209146*	28966844	
500	*rs7749527*	28967913	
501	*rs6456867*	28968493	
502	*rs381808*	28970163	
503	*rs6942070*	28971416	
504	*rs3130838*	28974518	
505	*rs209131*	28975745	
506	*rs209130*	28975790	
507	*rs1536215*	28975905	
508	*rs209128*	28977163	
509	*rs1056032*	28978965	
510	*rs209126*	28980436	
511	*rs209125*	28980620	
512	*rs209122*	28983492	
513	*rs2269553*	28984499	
514	*rs3135293*	28985237	
515	*rs1794588*	28985763	
516	*rs916284*	28986621	
517	*rs9295781*	28987764	
518	*rs209121*	28989650	
519	*rs3132377*	28993966	

Order	dbSNP ID	Coordinate (NCBI 36)	tagSNP
520	rs9468444	28995698	
521	rs6906867	28997289	
522	rs9295782	28998211	
523	rs2230683	28999168	
524	rs1061625	28999450	
525	rs2765229	29000911	
526	rs2894066	29001919	
527	rs3763338	29002303	
528	rs1237485	29002336	
529	rs12193226	29003331	
530	rs2015436	29004906	
531	rs1233508	29005625	
532	rs3118361	29006279	
533	rs2182230	29007688	
534	rs932776	29009026	
535	rs2032502	29009557	
536	rs1004062	29010125	
537	rs6912843	29012154	
538	rs3135329	29013096	
539	rs3130895	29013783	
540	rs3131070	29014523	
541	rs3130843	29016206	
542	rs4713186	29017457	
543	rs7762289	29018227	
544	rs6931968	29018868	
545	rs2071790	29019794	
546	rs2071789	29020068	
547	rs2071788	29020299	
548	rs763009	29023100	
549	rs4947339	29024244	
550	rs6456876	29026928	
551	rs9357074	29028080	
552	rs3131073	29028964	
553	rs4947256	29029932	
554	rs1476016	29030711	
555	rs3130845	29031359	tagSNP #51
556	rs9380100	29031972	tagSNP #52
557	rs3131075	29032618	
558	rs12386522	29036159	
559	rs12110866	29046345	tagSNP #53
560	rs4713190	29050823	
561	rs4713191	29051859	
562	rs4713192	29051870	
563	rs3135322	29054910	
564	rs3130837	29056082	
565	rs9257445	29057196	
566	rs968722	29057894	
567	rs9380102	29059004	
568	rs6904975	29060025	
569	rs12527531	29060379	
570	rs3129791	29062282	
571	rs3135321	29063515	
572	rs3129792	29064405	
573	rs6934470	29064740	
574	rs6934947	29064784	
575	rs10946966	29065572	
576	rs7776164	29066447	
577	rs6456879	29070593	
578	rs2269555	29072836	
579	rs6908206	29072861	
580	rs9468459	29073680	
581	rs6920392	29075027	
582	rs6906909	29076135	
583	rs9257453	29076919	tagSNP #54
584	rs6933976	29077511	
585	rs6456881	29078089	
586	rs3129793	29078165	
587	rs6456883	29078358	
588	rs6916645	29082925	
589	rs6923005	29084060	
590	rs3130891	29085206	
591	rs9468461	29087035	
592	rs3130893	29088695	
593	rs6930903	29089232	
594	rs9357075	29090275	
595	rs6916161	29091834	
596	rs6916923	29092151	
597	rs9468464	29092500	tagSNP #55
598	rs9380103	29093193	
599	rs6909302	29094363	
600	rs6941946	29096571	
601	rs7768299	29098183	
602	rs1543796	29098606	
603	rs2394512	29100172	
604	rs2394513	29100496	
605	rs9461497	29102115	
606	rs7741520	29102558	
607	rs2269554	29102977	
608	rs3135320	29104537	
609	rs4713197	29105573	
610	rs6927986	29106722	
611	rs7755402	29110050	
612	rs9468471	29112079	
613	rs9468473	29114238	
614	rs7743837	29114832	
615	rs7749435	29115835	tagSNP #56
616	rs6919044	29117556	
617	rs6924824	29118141	
618	rs6456886	29118575	
619	rs6456889	29118812	
620	rs7341218	29120104	tagSNP #57
621	rs9468474	29121121	
622	rs2143574	29122014	
623	rs6456890	29123114	
624	rs9380106	29123969	
625	rs6456891	29124575	
626	rs4713199	29125543	
627	rs7747023	29133668	
628	rs12665818	29140172	
629	rs6904527	29141041	
630	rs2179648	29144516	
631	rs6918175	29147218	
632	rs3131083	29147965	
633	rs4713200	29148774	
634	rs999265	29149791	
635	rs2050231	29150377	
636	rs3130758	29150925	
637	rs1159519	29151568	tagSNP #58
638	rs3131085	29152689	
639	rs9393941	29153620	
640	rs2064365	29154408	
641	rs9348821	29155191	
642	rs3129787	29156677	
643	rs11758255	29158698	
644	rs7740805	29159858	
645	rs6903771	29160556	tagSNP #59
646	rs7752270	29161982	tagSNP #60
647	rs6456901	29163426	
648	rs9380109	29165955	
649	rs2013972	29166981	
650	rs12665108	29170178	
651	rs6917293	29171042	
652	rs4538737	29172756	
653	rs6907184	29173378	
654	rs4713201	29175247	
655	rs6905729	29176613	
656	rs2394518	29177328	
657	rs9295790	29178419	
658	rs3131090	29180277	
659	rs3131091	29181083	
660	rs6456908	29183061	
661	rs11754944	29183651	
662	rs3130762	29185776	
663	rs3116838	29186140	
664	rs6456909	29187145	
665	rs3749977	29188333	
666	rs3130764	29188338	
667	rs3129106	29189150	
668	rs3129109	29192219	
669	rs3129110	29193215	
670	rs3130766	29194816	
671	rs2064162	29196802	
672	rs7740128	29202683	
673	rs6456913	29204982	
674	rs7760364	29205923	
675	rs6930603	29206817	
676	rs3130778	29207566	
677	rs7738990	29209108	
678	rs3129119	29210847	
679	rs12529022	29211556	
680	rs2223363	29212656	
681	rs9295794	29212919	
682	rs7775256	29213780	
683	rs2267632	29214452	
684	rs9468487	29215036	
685	rs9468488	29215616	
686	rs9393945	29216275	
687	rs6917665	29217585	
688	rs3116846	29222605	
689	rs3130718	29222610	
690	rs3130720	29223737	
691	rs3130724	29226518	
692	rs6902241	29227147	
693	rs3129126	29229623	
694	rs7750858	29234658	
695	rs3130732	29235461	
696	rs2142906	29237975	
697	rs3130737	29243200	
698	rs3918428	29244102	
699	rs2394520	29245212	
700	rs3129151	29246307	tagSNP #61
701	rs3129154	29247778	
702	rs9468491	29248729	
703	rs3116855	29249624	
704	rs3129157	29249735	
705	rs3129158	29249764	
706	rs3116856	29249841	
707	rs3129159	29250046	
708	rs3130743	29250056	
709	rs2206042	29250468	
710	rs3116857	29252119	
711	rs3130745	29256611	
712	rs3116817	29257553	
713	rs2006752	29259304	
714	rs3116820	29259663	
715	rs3129171	29263745	
716	rs3129173	29267640	
717	rs6904810	29269451	
718	rs1977074	29271850	tagSNP #62
719	rs3129093	29278281	
720	rs760804	29279059	
721	rs6906270	29279872	tagSNP #63
722	rs2394546	29283619	
723	rs736466	29284316	
724	rs1883329	29285120	tagSNP #64
725	rs2206040	29286532	
726	rs2206041	29286624	
727	rs3130817	29287255	
728	rs3129095	29288303	
729	rs719746	29289617	
730	rs2894083	29290325	
731	rs3129096	29291365	
732	rs11758554	29291888	
733	rs3116837	29292679	
734	rs6933230	29293306	tagSNP #65
735	rs3130801	29296198	tagSNP #66
736	rs3130803	29297174	
737	rs714470	29299430	
738	rs7757500	29300165	
739	rs3130811	29301281	
740	rs3130812	29301804	
741	rs3130813	29303030	
742	rs6901837	29304156	
743	rs3130814	29304200	
744	rs2207337	29308264	
745	rs10456369	29309134	
746	rs3117345	29310484	
747	rs3130816	29311930	
748	rs7742939	29313352	
749	rs7770244	29320306	
750	rs3129105	29321059	
751	rs7747464	29322936	
752	rs3117330	29333788	
753	rs3117329	29335636	tagSNP #67
754	rs7761697	29338093	
755	rs3117328	29338570	
756	rs3130827	29338676	
757	rs10484545	29342503	
758	rs6904130	29343186	
759	rs3130829	29343893	
760	rs6942318	29346010	
761	rs3117327	29347145	
762	rs3117326	29348372	
763	rs2394550	29356453	
764	rs5875188	29356556	tagSNP #68
765	rs6456942	29359392	
766	rs9468508	29361748	
767	rs6901923	29361879	
768	rs7741086	29362358	
769	rs1884123	29364431	tagSNP #69
770	rs1033569	29365783	
771	rs1033568	29365978	
772	rs3130835	29366874	
773	rs4446587	29367634	
774	rs3117425	29368442	
775	rs4398768	29369046	
776	rs7383161	29372666	
777	rs6899932	29374102	tagSNP #70
778	rs7754926	29377077	
779	rs3117424	29377294	
780	rs4594993	29378508	
781	rs3117426	29380022	
782	rs7774255	29382366	
783	rs9257694	29382496	
784	rs1474859	29383595	
785	rs9257697	29384190	
786	rs6910451	29388164	
787	rs9393954	29390362	
788	rs9257711	29391455	
789	rs1535151	29392363	
790	rs720831	29392530	
791	rs9257712	29393201	tagSNP #71
792	rs4713209	29395301	

Order	dbSNP ID	Coordinate (NCBI 36)	tagSNP
793	*rs9257718*	29399739	
794	*rs9468515*	29400272	
795	*rs9257723*	29403613	
796	*rs12660111*	29404160	tagSNP #72
797	*rs9257726*	29404935	
798	*rs9357078*	29405678	
799	*rs9257731*	29407383	
800	*rs6930435*	29409234	tagSNP #73
801	*rs6903989*	29410042	
802	*rs6917520*	29412559	
803	*rs6927977*	29413965	
804	*rs3130805*	29414468	
805	*rs9348827*	29416405	
806	*rs9257748*	29417780	
807	*rs1014258*	29418692	
808	*rs9257751*	29419450	
809	*rs3117435*	29420773	
810	*rs7745768*	29422975	tagSNP #74
811	*rs9257756*	29424211	
812	*rs4321865*	29424902	
813	*rs9257768*	29430097	tagSNP #75
814	*rs7356951*	29430664	
815	*rs9257769*	29431087	
816	*rs3117438*	29431270	
817	*rs9257770*	29431849	
818	*rs6930033*	29431916	
819	*rs7357041*	29433110	
820	*rs4713210*	29433444	
821	*rs3129685*	29433613	
822	*rs9257777*	29435169	
823	*rs7754402*	29436845	
824	*rs16718*	29443137	
825	*rs9380120*	29443563	
826	*rs4713211*	29444080	
827	*rs9295804*	29444173	
828	*rs6934993*	29445526	
829	*rs4713213*	29446988	
830	*rs7772982*	29449033	
831	*rs7753474*	29449334	tagSNP #76
832	*rs7773534*	29449390	
833	*rs9380122*	29450262	
834	*rs3749971*	29450801	
835	*rs3749970*	29450851	
836	*rs5003264*	29452015	
837	*rs2144425*	29452965	
838	*rs2144426*	29453055	
839	*rs238884*	29453703	
840	*rs238880*	29455070	
841	*rs1419640*	29458879	tagSNP #77
842	*rs238872*	29459898	
843	*rs3117444*	29460965	
844	*rs3094555*	29461678	
845	*rs3129681*	29462403	
846	*rs3094549*	29463168	
847	*rs1419638*	29463539	
848	*rs1419637*	29463578	
849	*rs1419635*	29463938	
850	*rs4711185*	29464066	
851	*rs442694*	29464707	
852	*rs9257813*	29465382	
853	*rs4713214*	29466016	
854	*rs7452887*	29466696	
855	*rs4452630*	29466953	
856	*rs9257819*	29468203	
857	*rs1362065*	29468697	
858	*rs2022077*	29469144	
859	*rs9257823*	29469755	
860	*rs1362063*	29470313	
861	*rs9257827*	29470923	tagSNP #78
862	*rs4713216*	29472024	
863	*rs4713217*	29472202	
864	*rs9257834*	29472492	
865	*rs3128853*	29472664	
866	*rs2073154*	29472692	
867	*rs2073151*	29472828	
868	*rs2073150*	29473118	
869	*rs2073149*	29473300	
870	*rs4713218*	29473851	
871	*rs2158279*	29474142	
872	*rs1024470*	29474254	
873	*rs1028411*	29475276	
874	*rs4713220*	29475375	
875	*rs4711187*	29475467	
876	*rs4713221*	29475619	
877	*rs1819784*	29476152	
878	*rs9405124*	29476682	
879	*rs2394604*	29477135	
880	*rs2394605*	29477187	
881	*rs2394606*	29477219	
882	*rs2394607*	29477386	
883	*rs1419634*	29477539	
884	*rs2894099*	29478724	
885	*rs1362061*	29479032	
886	*rs1362060*	29479282	
887	*rs429479*	29480190	
888	*rs994321*	29480223	
889	*rs1419633*	29481080	
890	*rs1419632*	29481116	
891	*rs7453752*	29484933	
892	*rs2097771*	29486306	
893	*rs12192194*	29487758	
894	*rs1544403*	29489773	tagSNP #79
895	*rs720497*	29490300	tagSNP #80
896	*rs1362074*	29491126	
897	*rs1362073*	29491187	
898	*rs3029487*	29491488	
899	*rs10807055*	29491998	
900	*rs3131024*	29493104	
901	*rs1011985*	29493639	
902	*rs12207410*	29494473	
903	*rs6456947*	29496173	
904	*rs2523421*	29498307	
905	*rs1419643*	29499801	
906	*rs7754054*	29500368	
907	*rs2074470*	29502292	
908	*rs7770592*	29503612	tagSNP #81
909	*rs7757269*	29504082	
910	*rs10946990*	29504511	
911	*rs6904456*	29505417	
912	*rs2394609*	29505865	
913	*rs6910208*	29505948	
914	*rs3131025*	29506680	tagSNP #82
915	*rs7349863*	29506724	
916	*rs11754009*	29514561	
917	*rs9468532*	29515243	
918	*rs2074469*	29515835	
919	*rs2074466*	29516178	
920	*rs9468533*	29516740	
921	*rs7765791*	29517958	tagSNP #83
922	*rs11752013*	29518624	
923	*rs6937718*	29519614	
924	*rs9368601*	29520288	
925	*rs7738722*	29522151	
926	*rs7738742*	29522211	
927	*rs7739243*	29522478	
928	*rs6929603*	29523517	
929	*rs6935708*	29524519	
930	*rs2523442*	29525684	
931	*rs3094574*	29529959	
932	*rs1419647*	29530551	
933	*rs9257863*	29533453	
934	*rs6456951*	29534072	
935	*rs6903755*	29534858	
936	*rs7768854*	29536699	
937	*rs2021729*	29537260	
938	*rs2073148*	29538038	
939	*rs3128854*	29539566	
940	*rs1345228*	29540258	tagSNP #84
941	*rs2107191*	29542163	
942	*rs2746149*	29543223	
943	*rs1419646*	29544310	
944	*rs6937864*	29545795	
945	*rs11961170*	29546973	
946	*rs11966831*	29548230	
947	*rs7751705*	29549486	
948	*rs2746150*	29550569	
949	*rs1233495*	29551989	
950	*rs7756110*	29553363	
951	*rs6925408*	29554438	
952	*rs3094573*	29554500	tagSNP #85
953	*rs6925744*	29554557	
954	*rs7771335*	29555906	
955	*rs2107189*	29559036	
956	*rs11756628*	29559867	
957	*rs10484547*	29560642	
958	*rs9501675*	29561168	
959	*rs11965733*	29562018	
960	*rs2066951*	29562357	
961	*rs9348832*	29563889	
962	*rs11756938*	29564522	
963	*rs1233493*	29566109	
964	*rs1233492*	29566345	
965	*rs3128844*	29567128	
966	*rs1233491*	29569598	
967	*rs1233490*	29569782	
968	*rs6915084*	29570049	
969	*rs1233489*	29570911	
970	*rs6926506*	29571836	
971	*rs6932526*	29572837	
972	*rs1345227*	29574824	
973	*rs7766082*	29575874	
974	*rs1233487*	29576677	tagSNP #86
975	*rs9393967*	29577357	
976	*rs1233486*	29577520	
977	*rs4711192*	29578654	
978	*rs3131019*	29578853	
979	*rs11961013*	29579802	
980	*rs757256*	29580784	
981	*rs3130858*	29581369	tagSNP #87
982	*rs3130860*	29582253	
983	*rs734960*	29583010	
984	*rs1002187*	29583335	
985	*rs1233482*	29583550	
986	*rs3131020*	29583770	
987	*rs3131021*	29584010	
988	*rs3131022*	29584552	
989	*rs1557820*	29585108	
990	*rs2158281*	29585651	
991	*rs1233478*	29585689	
992	*rs2285791*	29586169	
993	*rs1010408*	29587401	tagSNP #88
994	*rs12527641*	29588293	
995	*rs3094572*	29588861	
996	*rs1362075*	29589996	
997	*rs2523433*	29590300	
998	*rs1592410*	29591842	
999	*rs1015869*	29592541	
1000	*rs1015868*	29593248	
1001	*rs3094564*	29594220	
1002	*rs9257890*	29594783	
1003	*rs2745400*	29596124	
1004	*rs724078*	29596927	
1005	*rs7752486*	29600442	
1006	*rs2745403*	29601050	
1007	*rs7450360*	29601862	
1008	*rs969931*	29602775	
1009	*rs1233427*	29604542	
1010	*rs9468549*	29605447	
1011	*rs1233426*	29605980	
1012	*rs6908631*	29606301	
1013	*rs1233425*	29606375	
1014	*rs362546*	29607294	
1015	*rs909967*	29607415	
1016	*rs886381*	29608103	
1017	*rs362544*	29608165	
1018	*rs1233422*	29608290	
1019	*rs9968*	29608915	
1020	*rs2294748*	29609060	
1021	*rs1233421*	29609392	tagSNP #89
1022	*rs1233420*	29611201	
1023	*rs6930217*	29611660	
1024	*rs6911709*	29611821	
1025	*rs6911894*	29611890	
1026	*rs6912673*	29612037	
1027	*rs6935418*	29612135	
1028	*rs1233418*	29613022	
1029	*rs419957*	29613647	
1030	*rs414282*	29614138	
1031	*rs422241*	29614178	
1032	*rs362514*	29615188	
1033	*rs6936699*	29616108	
1034	*rs10484548*	29616151	
1035	*rs414390*	29617752	
1036	*rs362541*	29618128	
1037	*rs362540*	29618203	
1038	*rs3025657*	29619094	
1039	*rs2534795*	29619219	
1040	*rs6935895*	29619639	
1041	*rs2534794*	29619728	tagSNP #90
1042	*rs407161*	29621698	
1043	*rs3094576*	29624135	
1044	*rs376681*	29624986	
1045	*rs365488*	29625851	
1046	*rs2745411*	29626338	
1047	*rs417374*	29626368	
1048	*rs7763501*	29626830	
1049	*rs5875195*	29627328	
1050	*rs3052069*	29627329	
1051	*rs3906305*	29627455	
1052	*rs453658*	29627744	
1053	*rs446145*	29628683	
1054	*rs11724*	29629215	
1055	*rs389419*	29629341	
1056	*rs7771629*	29629845	
1057	*rs3215532*	29630014	tagSNP #91
1058	*rs2294745*	29630048	
1059	*rs6915177*	29630484	
1060	*rs2534791*	29630608	
1061	*rs7739536*	29631368	
1062	*rs444013*	29631542	
1063	*rs8337*	29631572	
1064	*rs404240*	29631853	tagSNP #92
1065	*rs2076484*	29631899	

Order	dbSNP ID	Coordinate (NCBI 36)	tagSNP
1066	*rs2534790*	29632064	
1067	*rs12665228*	29633487	
1068	*rs362536*	29634836	
1069	*rs64036*	29635163	
1070	*rs2272991*	29635495	
1071	*rs362513*	29636214	
1072	*rs995185*	29637033	
1073	*rs1233405*	29637650	
1074	*rs362531*	29638351	
1075	*rs362512*	29638524	
1076	*rs422878*	29638840	
1077	*rs362527*	29639714	
1078	*rs2523447*	29640631	tagSNP #93
1079	*rs388234*	29641187	
1080	*rs3094577*	29641762	
1081	*rs1119080*	29644577	
1082	*rs1235162*	29645116	
1083	*rs1003582*	29646295	
1084	*rs12527115*	29647093	
1085	*rs1233399*	29647374	
1086	*rs1003581*	29648096	tagSNP #94
1087	*rs362509*	29648753	tagSNP #95
1088	*rs6912437*	29649286	
1089	*rs362525*	29651538	
1090	*rs362524*	29651665	
1091	*rs362522*	29653100	tagSNP #96
1092	*rs1233397*	29653607	tagSNP #97
1093	*rs4568477*	29654804	
1094	*rs1233391*	29658510	
1095	*rs1233389*	29660375	
1096	*rs1233388*	29661619	
1097	*rs3129034*	29663701	
1098	*rs362521*	29664651	tagSNP #98
1099	*rs1233386*	29666082	
1100	*rs1233384*	29667187	tagSNP #99
1101	*rs6912589*	29669742	
1102	*rs3095273*	29674261	
1103	*rs3025644*	29675815	
1104	*rs3025643*	29677847	tagSNP #100
1105	*rs2267633*	29678733	
1106	*rs2076483*	29679437	tagSNP #101
1107	*rs10946999*	29679501	
1108	*rs3025638*	29680813	
1109	*rs2076482*	29681146	
1110	*rs12193396*	29682289	
1111	*rs5875197*	29682488	
1112	*rs740884*	29682523	
1113	*rs740882*	29683348	
1114	*rs881284*	29683939	
1115	*rs29230*	29684285	tagSNP #102
1116	*rs2076489*	29684329	
1117	*rs29263*	29684482	
1118	*rs29262*	29684484	
1119	*rs29258*	29685509	
1120	*rs29257*	29685922	
1121	*rs2076488*	29686421	
1122	*rs29255*	29687436	
1123	*rs29253*	29688328	
1124	*rs29227*	29688480	
1125	*rs29226*	29688903	tagSNP #103
1126	*rs29225*	29688933	
1127	*rs29251*	29689649	
1128	*rs6927867*	29690417	
1129	*rs6919973*	29695767	
1130	*rs29223*	29696437	
1131	*rs25629*	29697294	
1132	*rs29220*	29697559	
1133	*rs3828923*	29698608	
1134	*rs6938190*	29698760	
1135	*rs6938734*	29698952	
1136	*rs3025626*	29699757	
1137	*rs2267635*	29700324	
1138	*rs715044*	29701681	
1139	*rs2021749*	29706012	
1140	*rs29243*	29706995	
1141	*rs9257924*	29711023	
1142	*rs9461540*	29712152	
1143	*rs444189*	29713823	
1144	*rs3095267*	29714937	tagSNP #104
1145	*rs3025623*	29715638	
1146	*rs1233374*	29716587	
1147	*rs29236*	29717102	
1148	*rs3117289*	29717814	
1149	*rs29273*	29718882	tagSNP #105
1150	*rs29233*	29719122	
1151	*rs29232*	29719324	
1152	*rs3131854*	29719778	tagSNP #106
1153	*rs3129073*	29723708	
1154	*rs439812*	29724493	tagSNP #107
1155	*rs29269*	29725632	
1156	*rs29272*	29726251	
1157	*rs29228*	29731622	
1158	*rs29234*	29731995	tagSNP #108
1159	*rs3130250*	29732884	
1160	*rs2535267*	29733641	
1161	*rs2252711*	29734203	
1162	*rs2535260*	29736767	
1163	*rs2256266*	29740203	
1164	*rs6905408*	29741707	tagSNP #109
1165	*rs2071652*	29743249	
1166	*rs2071653*	29743860	
1167	*rs2273192*	29746111	
1168	*rs9357083*	29747576	
1169	*rs2535242*	29748614	tagSNP #110
1170	*rs9461544*	29749235	

Supplementary Table 2

List of SNPs captured by the 12 tagSNPs significantly associated with the haplotype A1-B8-DR3. Their number in supplementary Table 1 and their genomic position in relation to known or predicted genes are indicated.

tagSNP #10 - Representative for 9 SNPs:
97: intronic region of *ZSCAN16*
104: between *ZSCAN16* and *ZNF192*
151: between the pseudogenes *LOC222701* and *LOC222699*
181: intronic region of *ZSCAN4*
182: between *ZSCAN4* and the predicted gene *C6orf194*
183: between *ZSCAN4* and the predicted gene *C6orf194*
190: between the predicted gene *C6orf194* and *ZNF187*
196: telomeric, within 2 kb of a mRNA transcript for *ZNF187*
198: intronic region of *ZNF187*
tagSNP #23 - Representative for 2 SNPs:
226: intronic region of *ZNF323*
231: intronic region of *ZNF323*
tagSNP #24 - Representative for 3 SNPs:
232: intronic region of *ZNF323*
237: within the 5’ UTR of *ZSCAN3*
280: between *ZSCAN12* and the pseudogene *COX11P*
tagSNP #27 - Representative for 6 SNPs:
259: intronic region of *ZSCAN12*
205: intronic region of *PGBD1*
77: centromeric, within 2 kb of a mRNA transcript for *ZNF165*
46: in the pseudogene *OR2W2P*
27: between the pseudogene *OR2W4P* and *LOC340192*
4: centromeric, within 2 kb of a mRNA transcript for *OR2B2*
tagSNP #29 - Representative for 2 SNPs:
283: between *ZSCAN12* and *COX11P*
287: centromeric, within 2 kb of a mRNA transcript for *COX11P*
tagSNP #42 - Representative for 2 SNPs:
382: between *LOC646160* and the pseudogene *LOC442181*
396: in the pseudogene *LOC442181*
tagSNP #51 - Representative for 34 SNPs:
378: between *LOC646160* and *LOC442181*
392: between *LOC646160* and *LOC442181*
398: between *LOC442181* and *LOC646192*
417: between *LOC442181* and *LOC646192*
423: between *LOC442181* and *LOC646192*
430: intronic region of *LOC646192*
443: between *LOC646192* and hypothetical gene *LOC401242*
444: between *LOC646192* and hypothetical gene *LOC401242*
459: between *LOC646192* and hypothetical gene *LOC401242*
461: between *LOC646192* and hypothetical gene *LOC401242*
464: between *LOC646192* and hypothetical gene *LOC401242*
470: in the hypothetical gene *LOC401242*
473: in the hypothetical gene *LOC401242*
474: centromeric, within 2 kb of a mRNA transcript for the hypothetical gene *LOC401242*
476: between the hypothetical gene *LOC401242* and *TRIM27*
479: between the hypothetical gene *LOC401242* and *TRIM27*
480: between the hypothetical gene *LOC401242* and *TRIM27*
494: between the hypothetical gene *LOC401242* and *TRIM27*

504: between the hypothetical gene *LOC401242* and *TRIM27*
514: intronic region of *TRIM27*
523: coding region of *TRIM27*: synonymous substitution
532: between *TRIM27* and *KRT18P1*
539: between *TRIM27* and *KRT18P1*
552: between *TRIM27* and *KRT18P1*
555: between *TRIM27* and *KRT18P1*
564: between *KRT18P1* and *ZNF311*
570: between *KRT18P1* and *ZNF311*
590: between *ZNF311* and *OR2AD1P*
592: between *ZNF311* and *OR2AD1P*
638: in *LOC646260* (*SAR1P1*)
716: between *OR2J4P* and *OR2H4P*
762: between *OR2U2P* and *OR2B4P*
774: between *OR2B4P* and *OR5U1*
834: coding region of *OR12D3*: non synonymous substitution in residue 97 (Thr > Ile)
tagSNP #73 - Representative for 3 SNPs:
800: between *DDX6P* and *OR5V1*
801: between *DDX6P* and *OR5V1*
807: between *DDX6P* and *OR5V1*
tagSNP #74 - Representative for 4 SNPs:
445: between *LOC646192* and *LOC401242*
455: between *LOC646192* and *LOC401242*
804: between *DDX6P* and *OR5V1*
810: between *DDX6P* and *OR5V1*
tagSNP #86 - Representative for 2 SNPs:
949: centromeric, within 2 kb of a mRNA transcript for *LOC646366*
974: between *LOC442195* and *GPR53P*
tagSNP #92 - Representative for 10 SNPs:
851: intronic region of *OR5V1* and *OR12D3*
887: intronic region of *OR5V1* and *OR12D3*
942: in pseudogene *UBDP1*
948: telomeric, within 2 kb of a mRNA transcript for *LOC646366*
963: centromeric to *RPS17P1*, in the hypothetical transcript for *LOC442195*
966: between *RPS17P1* and *GPR53P*
969: between *RPS17P1* and *GPR53P*
1055: in pseudogene *OR2I1P*
1064: coding region of *UBD*: synonymous substitution
1082: intronic region of *GABBR1*
tagSNP #104 - Representative for 4 SNPs:
1144: between *MOG* and *GABBR1*
1148: between *MOG* and *GABBR1*
1153: between *MOG* and *GABBR1*
1157: telomeric, within 2 kb of a mRNA transcript for *GABBR1*

3. Variation and Linkage Disequilibrium within Olfactory Receptor Gene Clusters Linked to the Human Major Histocompatibility Complex

3.1. Summary

The study published by Ehlers and co-workers [Ehlers et al., 2000] established, using a group of 10 human cell lines, a milestone for the MHC-linked OR research by sequencing 13 of its 34 loci and assessing their polymorphism. The following study aimed at completing the initiated investigation by sequencing all three remaining functional loci, in addition to further nine genes known so far only as pseudogenes. Moreover, integration with publicly available genomic information was performed through the inclusion of sequence data from the recently completed human MHC haplotype project [Horton et al., 2008], for which eight further (homozygous) human cell lines were sequenced. The gene *OR2B8P*, previously known only as a pseudogene, was found to have functional alleles, while *OR1F12*, previously known as a potentially functional gene, was found to present a truncated allele.

Questions about LD dynamics within the OR polymorphisms, as well as between the latter and the HLA complex, could be answered and discussed in this study. Moreover, based on phylogenetic and comparative analyses of the 18 cell lines, a comprehensive picture of the genotypic and phenotypic variability of both MHC-linked OR gene clusters was generated. Beyond its population genetics relevance, these results build the basis for the functional assessment of the participation of HLA-linked OR loci in reproduction.

3.2. Manuscript

Santos PSC, Uehara CJS, Ziegler A, Uchanska-Ziegler B, Bicalho MG: Variation and linkage disequilibrium within odorant receptor gene clusters linked to the human major histocompatibility complex. Hum Immunol. 2010, 71(9):843-50.

DOI:10.1016/j.humimm.2010.06.011

Pages 44-64

4. Comparative Genomic Analysis of MHC-linked Olfactory Receptor Repertoires among 14 Vertebrate Species

Comparative genomic analyses belong to the most powerful tools in evolutionary genetics. The recent sequencing of the genomes of different organisms has shown that almost all vertebrates studied in detail so far have one or more clusters of genes encoding olfactory receptors (OR) in close linkage to the MHC. A systematic comparison of the sequences of these receptors from different organisms has been, however, limited to humans and rodents. Additionally, a comparative analysis focusing on genomic morphology (i.e. linkage, position and transcriptional orientation of loci or gene clusters) of MHC-linked OR genes is, to the best of my knowledge, still absent. In this chapter, I present the so far most comprehensive comparison of protein sequence and genomic morphology of MHC-linked OR genes performed to date, among 14 vertebrates for which enough sequence information has been generated by different sequencing projects, and made available through internet-based databanks. 464 peptide sequences (Table 1) from human, chimpanzee, orangutan, rhesus macaque, dog, cat, cow, pig, horse, mouse, rat, opossum, frog and zebra fish were assessed.

Table 4.1: List of organisms analyzed for their MHC-linked OR genes, with indication of number of protein sequences assessed, the respective chromosome, as well as the approximate total length of the OR clusters

Organism	Short	# Protein Seqs	Chr.	Length (Mb)
Homo sapiens (Human)	Hsa	34	6p	2,3
Pan troglodytes (Chimpanzee)	Ptr	13	6p	3,0
Pongo pygmaeus (Orangutan)	Ppy	13	6	3,0
Macaca mulatta (Rhesus macaque)	Mamu	17	4	3,0
Canis familiaris (Dog)	Cfa	10	35	1,5
Felis catus (Cat)	Fca	8	B2	1,5
Bos taurus (Cow)	Bta	57	23	1,5
Sus scrofa (Pig)	Ssc	77	7	3,4
Equus caballus (Horse)	Eca	71	20	4,0
Mus musculus (Mouse)	Mumu	44	17	1,2
Rattus norvegicus (Rat)	Rno	59	20p	1,6
Monodelphis domestica (Opossum)	Mdo	38	2	9,0
Xenopus tropicalis (Frog)	Xtr	3	Scf_396	0,02
Danio rerio (Zebrafish)	Dre	20	15	0,3
Total		**464**		

Phylogenetic analyses demonstrate that most OR families present in the two human MHC-linked clusters are also present in the other species, although some are specifically expanded in certain taxa (Fig. 4.1). A map of the respective genomic regions was generated for each species, through which several morphological features such as the number of loci, the gene order and localization, as well as the presence of particular OR gene families could be identified as typical for each of the taxa assessed: primates, ungulates, carnivores, rodents, marsupial, amphibian and teleost (Fig. 4.2).

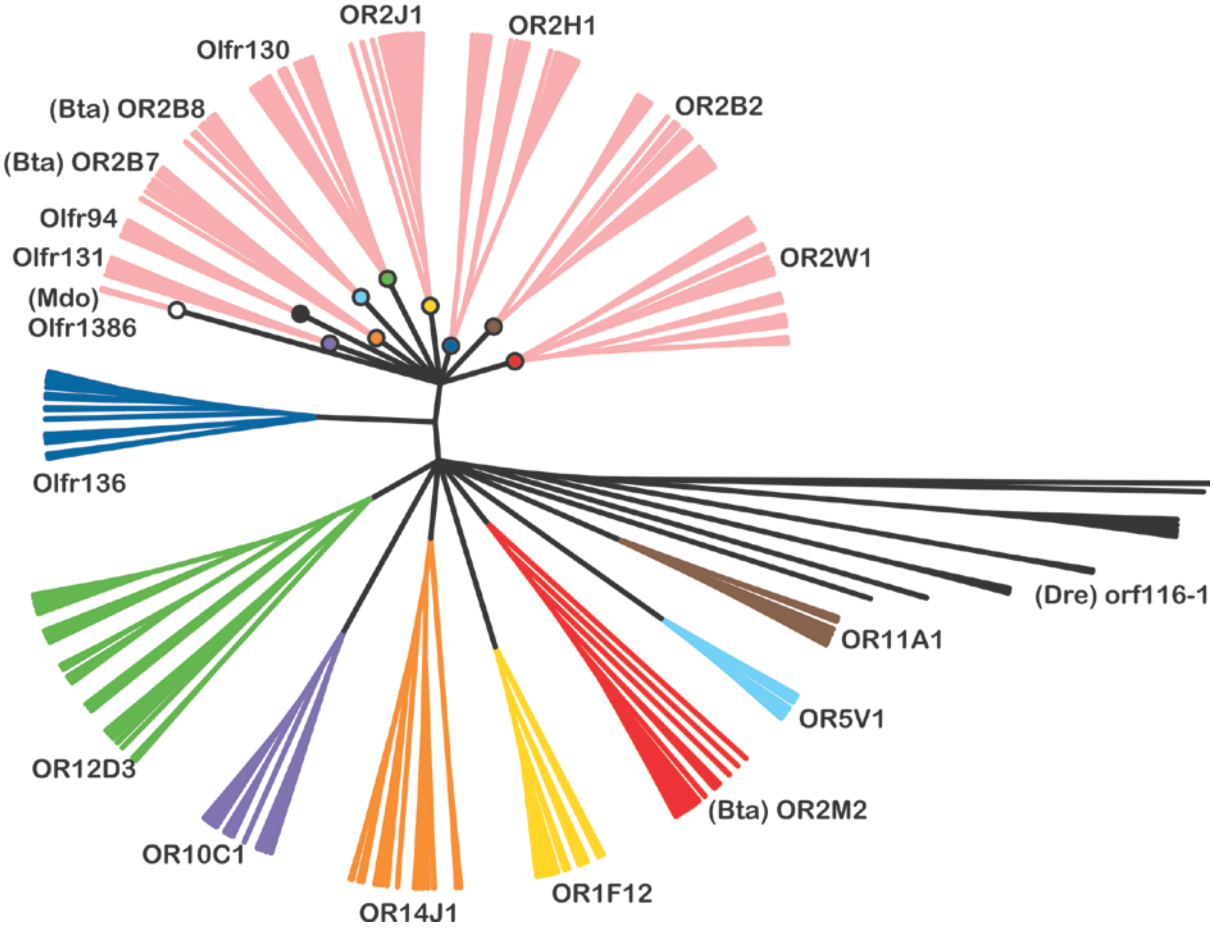


Figure 4.1: Phylogenetic tree depicting the evolutionary relationships of 464 MHC-linked OR amino acid sequences from 14 vertebrates. The colours used to identify each branch of sequences on the tree are the same used in the corresponding loci of all panels of Fig. 4.2. For each branch the name of one locus is given as representative of the group. The tree was inferred using the neighbour-joining method.

The genomes of humans, mouse, rat and zebra fish are the best established vertebrate genome assemblies available. Genome browsers like ENSEMBL (www.ensembl.org/) and VEGA (http://vega.sanger.ac.uk/) gather information from different genome projects worldwide and make them available with official gene nomenclatures and coordinates. Obtaining their sequences involved finding the MHC through the literature or an electronic BLAST search against the whole genome, scanning its vicinity for OR genes, and downloading them manually.

In contrast, most genes of the remaining ten vertebrate genome assemblies still lack a nomenclature for OR genes, and the MHC cannot easily be identified through the respective genome browsers. Data mining in these cases involved, besides of literature research, several runs of BLAST searches for unequivocal localization of the MHC within a chromosome, with following BLAST searches for OR loci in a range of 10 Mb around the MHC. Additionally, InterPro (http://www.ebi.ac.uk/interpro/) protein signatures were used to locate OR genes. InterPro is an online database that integrates data from several institutions aiming at the classification and automatic annotation of proteins and genomes [Hunter et al., 2009]. It uses information from known amino acid sequence motifs in order to recognize the structure and predict the function or a large number of proteins. After testing the strategy of sequence capturing based on InterPro signatures on four well established genomes (human, mouse, rat and zebra fish) with 100% concordance, I used InterPro in order to obtain sequences from the remaining vertebrates. Signatures used were “IPR000725” (olfactory receptor), and in some cases “IPR000276” (seven transmembrane g-protein-coupled receptor, rhodopsin-like). Nucleotide or amino acid sequences were obtained from the web servers of the following institutions: NCBI (www.ncbi.nlm.nih.gov, pig and horse), LGD (http://home.ncifcrf.gov/ccr/lgd/, cat), JGI (http://genome.jgi-psf.org/, frog) and ENSEMBL (www.ensembl.org, all other organisms). In most cases, the electronic tool BioMart (www.biomart.org) could be used in order to identify InterPro OR signatures within the sequences.

Primates (Fig. 4.2, panel a) exhibit two OR clusters telomeric to the MHC, with a large histone cluster flanking the distal OR cluster, and well conserved sub-organization of both clusters. The proximal OR cluster includes the framework genes gamma-aminobutyric acid (GABA) B receptor 1 (*GABBR1*), myelin oligodendrocyte glycoprotein (*MOG*) and ubiquitin D (*UBD*). As discussed in the article shown in chapter 4, apparently functional orthologs of the human gene *OR2W6P* are present within the MHC distal OR cluster, in the same position and transcriptional orientation. In the case of the macaque, the *OR2W6P* ortholog is the only locus that remained functional in that cluster. Orangutan is the only primate that seems to have "lost", through pseudogenization, all members of the OR12D family (discussed in chapters 2 and 3). Although the human loci *OR2B8P* and *OR12D1P* are widely considered as pseudogenes, they were included in Fig. 4.2 due to segregating functional alleles found in the study presented in the chapter 3.

Rodents (mouse and rat, Fig. 4.2, panel b) exhibit a single OR cluster telomeric to the MHC, homologous to the centromeric MHC-linked OR cluster of primates. In both organisms, the homologous histone clusters are located on other chromosomes, which possess a syntenic region harbouring OR loci that are evolutionarily related to those of the telomeric OR cluster on human Chr6p. The basic organization is remarkably conserved between the two murine species, which have, as compared to the other organisms, the number of loci being homologous to the human families OR2N1, OR14J, OR1F and OR10C strongly expanded. The apparently obligatory framework genes are also present, close to the MHC.

In comparison to primates, ungulates (horse, cow and pig, Fig. 4.2, panel c) were found to have a third cluster of MHC-linked OR, with all loci being similar to human gene *OR2M3* (which is harboured within another large OR cluster on human Chr. 1q44). The histone genes of ungulates are split in three relatively short clusters, which have conserved positions between horse and cow, but not in pig. Another peculiarity of *Sus scrofa* refers to the MHC-proximal OR cluster: horse and cow have the three framework genes close to the MHC, whereas the pig has them neighbouring the telomeric side of the same OR cluster, in opposite transcriptional direction. This fact, as well as the positions and transcriptional orientations of the pig orthologs of *OR2J3* and neighbours, are suggestive of a chromosomal inversion that has probably taken place after the divergence of pigs from the ancestral ungulate.

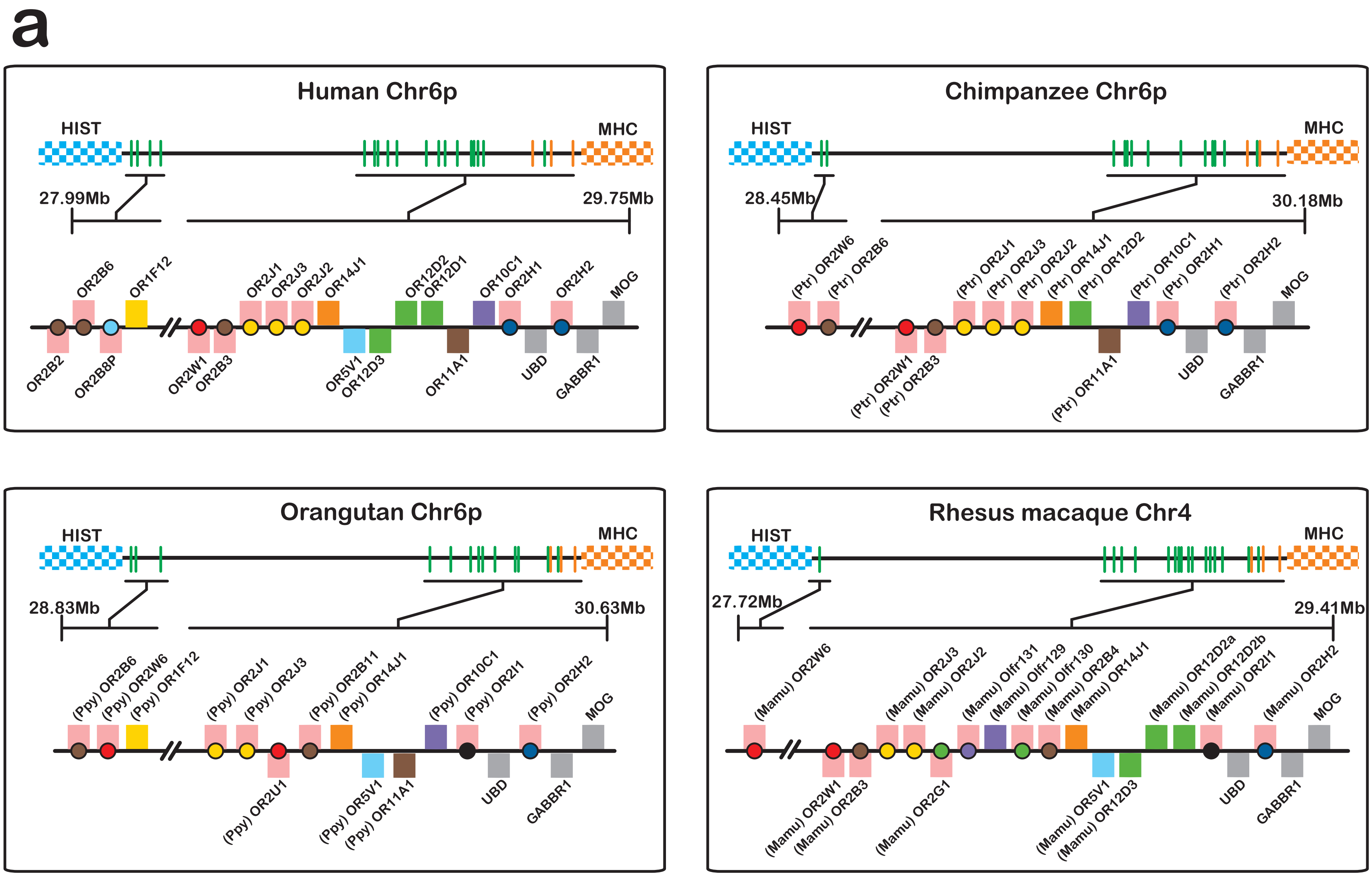
Carnivores (cat and dog, Fig. 4.2, panel d) were the only vertebrates in this analysis not found to have OR linked to the MHC, but only to the MHC framework genes *GABBR1* and *MOG*. As discussed elsewhere [Yuhki et al., 2007], a chromosomal break between the MHC and OR seems to have taken place before the evolutionary split of canine and feline groups. In both cases, the OR clusters are very close to the telomeres of ChrB2 (cat) and Chr35 (dog), and a large histone cluster is present linked to the OR clusters. Apart from chicken, these carnivores are so far the only vertebrates known to lack MHC-linked OR genes.

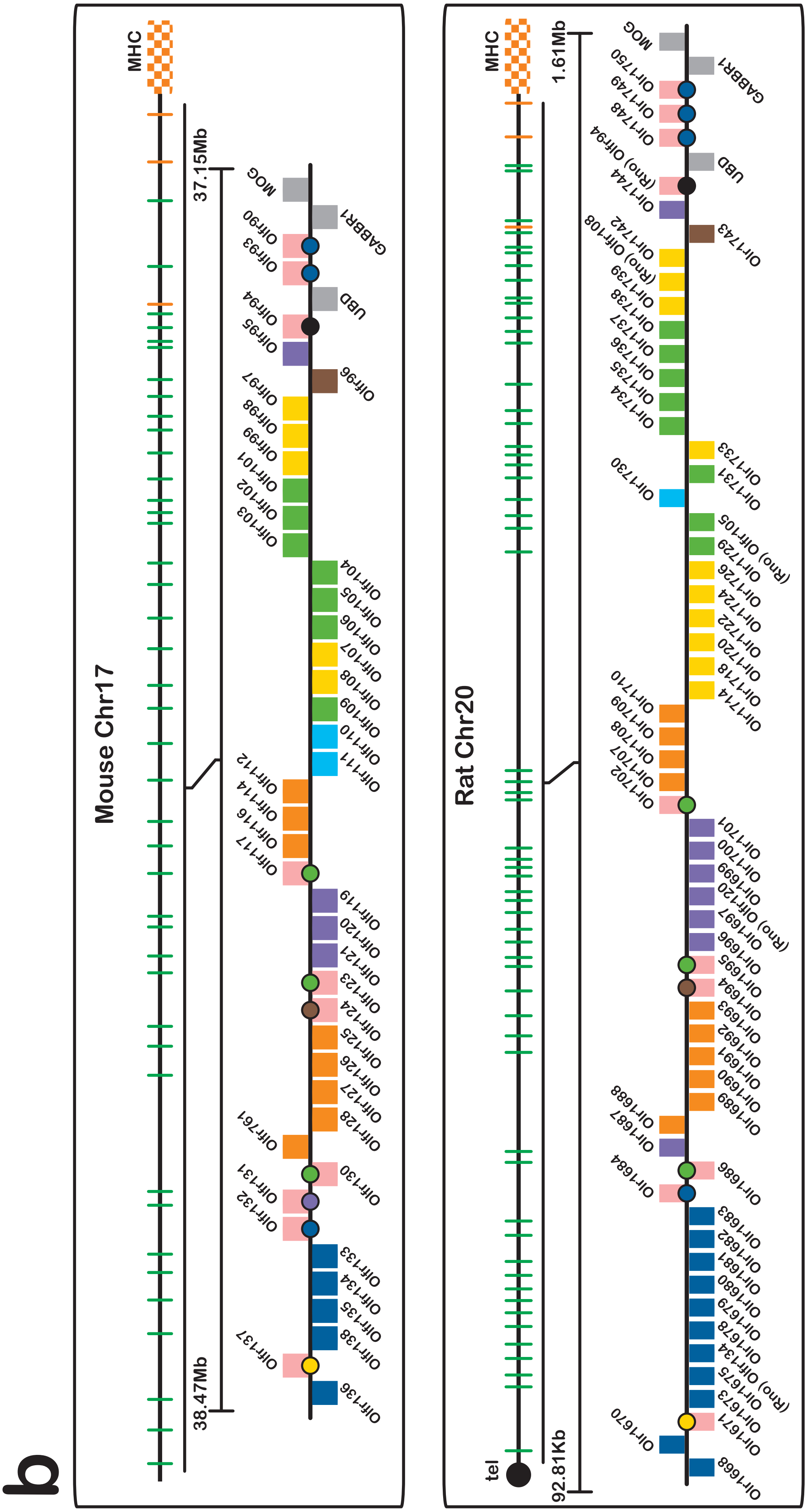
Opossum, the only marsupial that could be studied (Fig. 4.2, panel e), presents two relatively distant MHC-linked OR clusters, with the MHC-distal cluster harbouring an interspersed histone cluster. Similarly to ungulates, the opossum possesses also OR that cluster with members of the OR2M subfamily (Fig. 4.1) within the MHC-distal cluster. The framework genes are present in the MHC-proximal OR cluster. This species lacks loci of some gene families present in almost all other mammals (as OR5V and OR14J). On the other hand, the opossum is so far the only species in which homologs of the mouse Olfr1386 are MHC-linked. In all other organisms, including the mouse, this gene is not linked to the MHC.

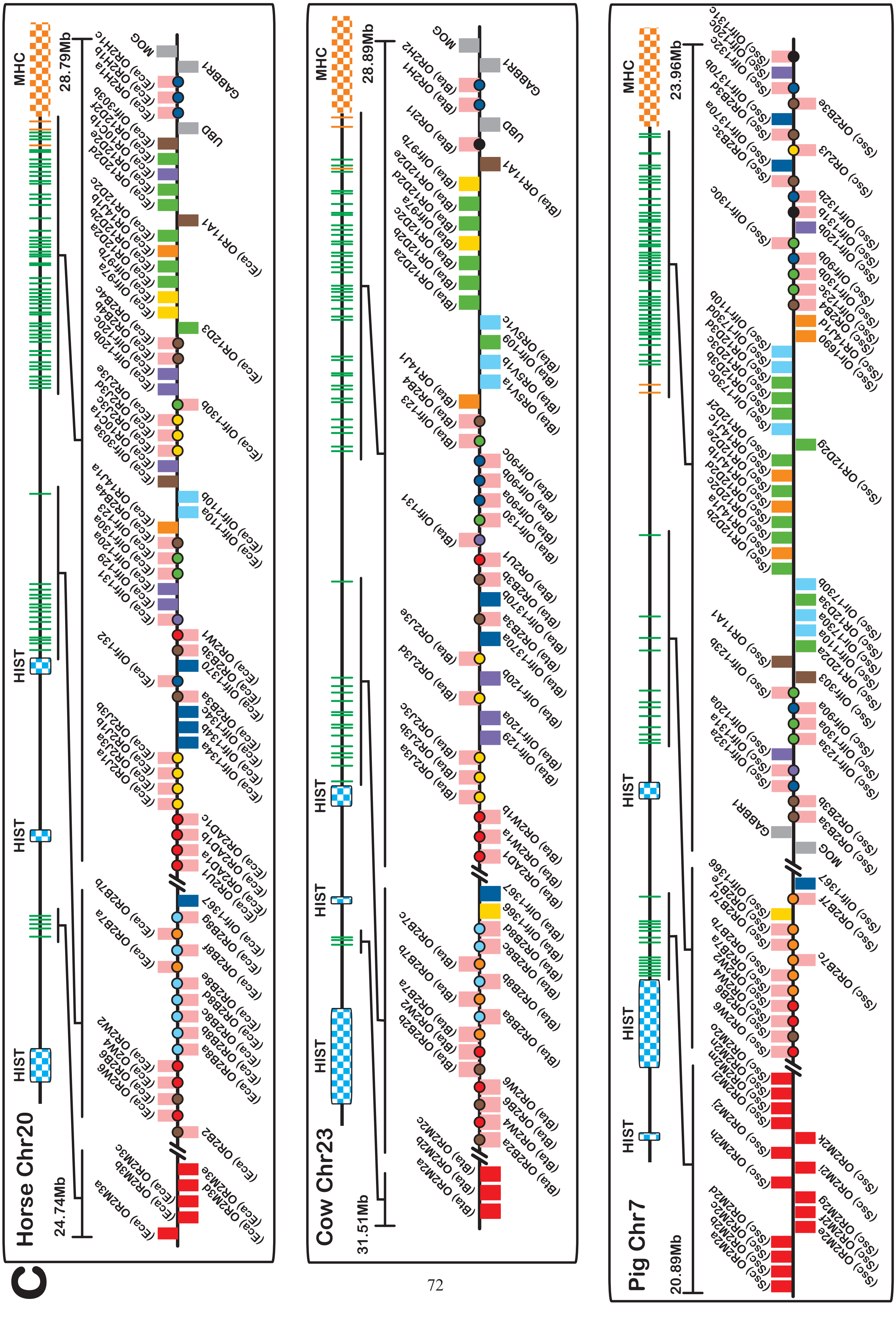
The genomic assembly of the frog (Fig. 4.2, panel f) is still incomplete for the MHC region and its vicinities. However, at least part of the MHC class II, three OR genes with sequence similarity to human *OR1F12*, and the framework gene *GABBR1* were confirmed to be linked so far, in a segment provisionally called “scaffold 396”. It remains unclear whether further OR, framework or MHC loci are linked to this region.

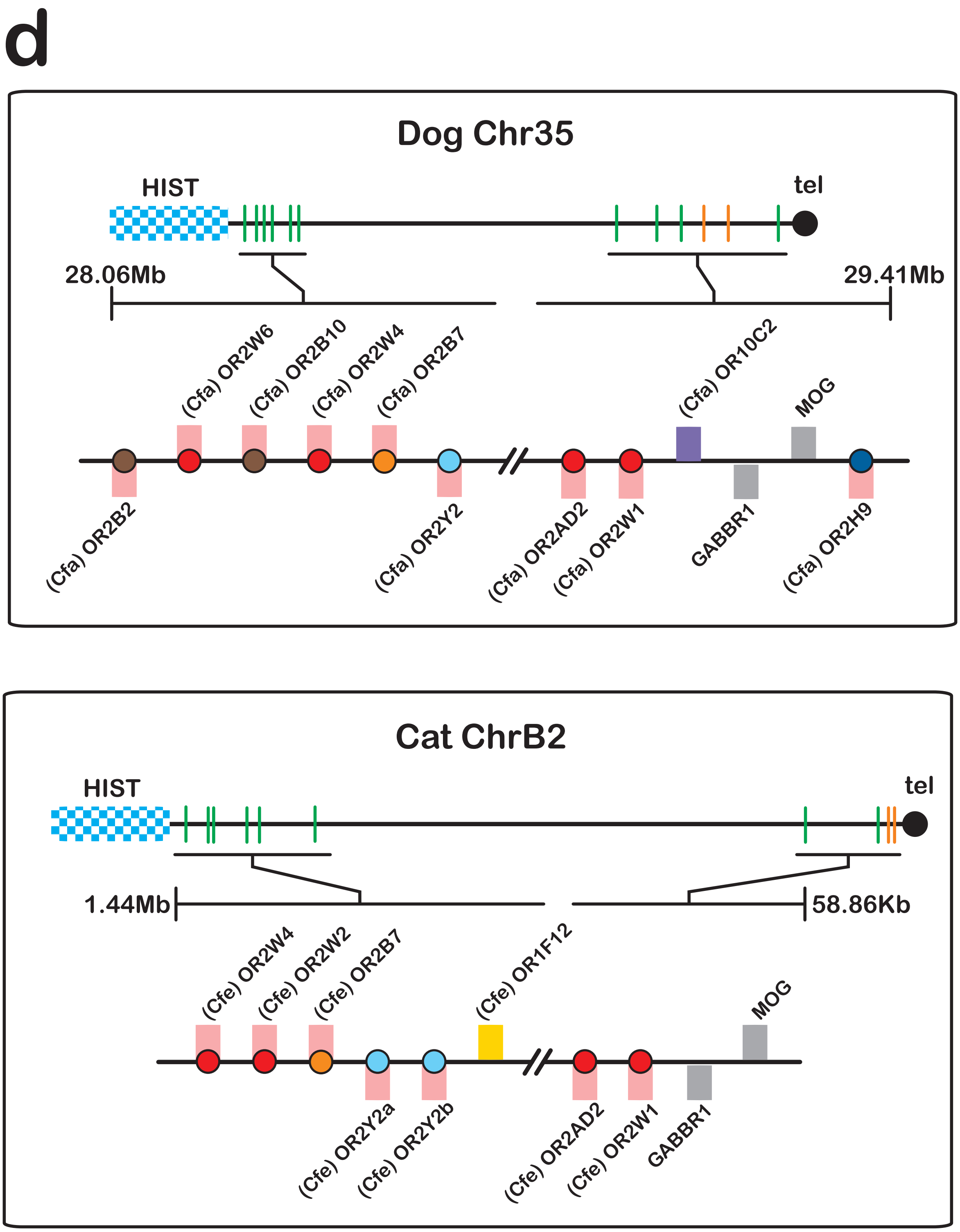
Although all terrestrial vertebrates analyzed are characterized by MHC-linked OR loci belonging to a defined subgroup of families (Fig. 4.2, panels a to f), this is different in the only teleost included in the analysis: in the zebra fish, a distinct group of OR are linked to genes that in mammals are part of the MHC class I and III regions (Fig. 4.2, panel g).

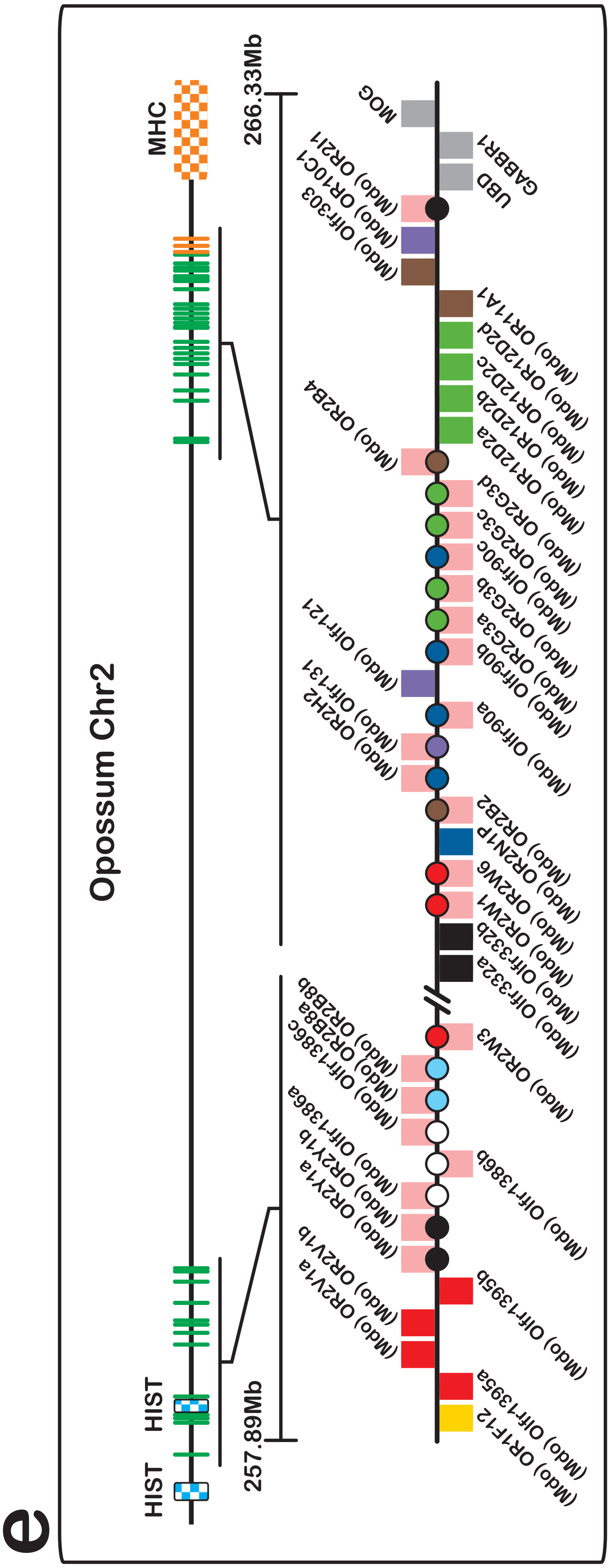
Figure 4.2 (next six pages): Gene maps of the MHC-linked OR genes in 14 vertebrates grouped in seven taxa: primates (a), rodentia (b), ungulata (c), carnivora (d), marsupialia (e), amphibia (f) and teleostei (g). Within each species map, an overview of the region is given above and a detailed view is given below. In the overview, the MHC complex and histone clusters are shown when present (in orange and blue checked boxes, respectively), as well as the relative positions of each OR gene (green vertical ticks) and selected framework genes (orange vertical ticks). The names of each OR locus are depicted in the detailed view below, where only human, mouse, rat and zebra fish have widely accepted nomenclatures for these loci, and have the official gene symbols shown. For all other species, I assigned a name for each locus based on sequence similarity to known genes, having the human and the mouse genomes as references. A gene was considered homolog of the human or mouse locus with which its amino acid sequence had the highest identity, considering 75% identity as a minimum threshold. In these cases, the gene designation of the locus is preceded by the species code, and has an additional letter in cases of multiple homologs (“Bta_OR2M2a” for instance, meaning the *Bos taurus* homolog of human *OR2M2*, fist copy). The colours of the boxes correspond to the clustering of the respective amino acid sequence in the phylogenetic tree (Fig. 4.1), and these are placed above or below the base line in order to indicate transcriptional orientation to the right or to the left, respectively. MHC framework genes are depicted in grey.

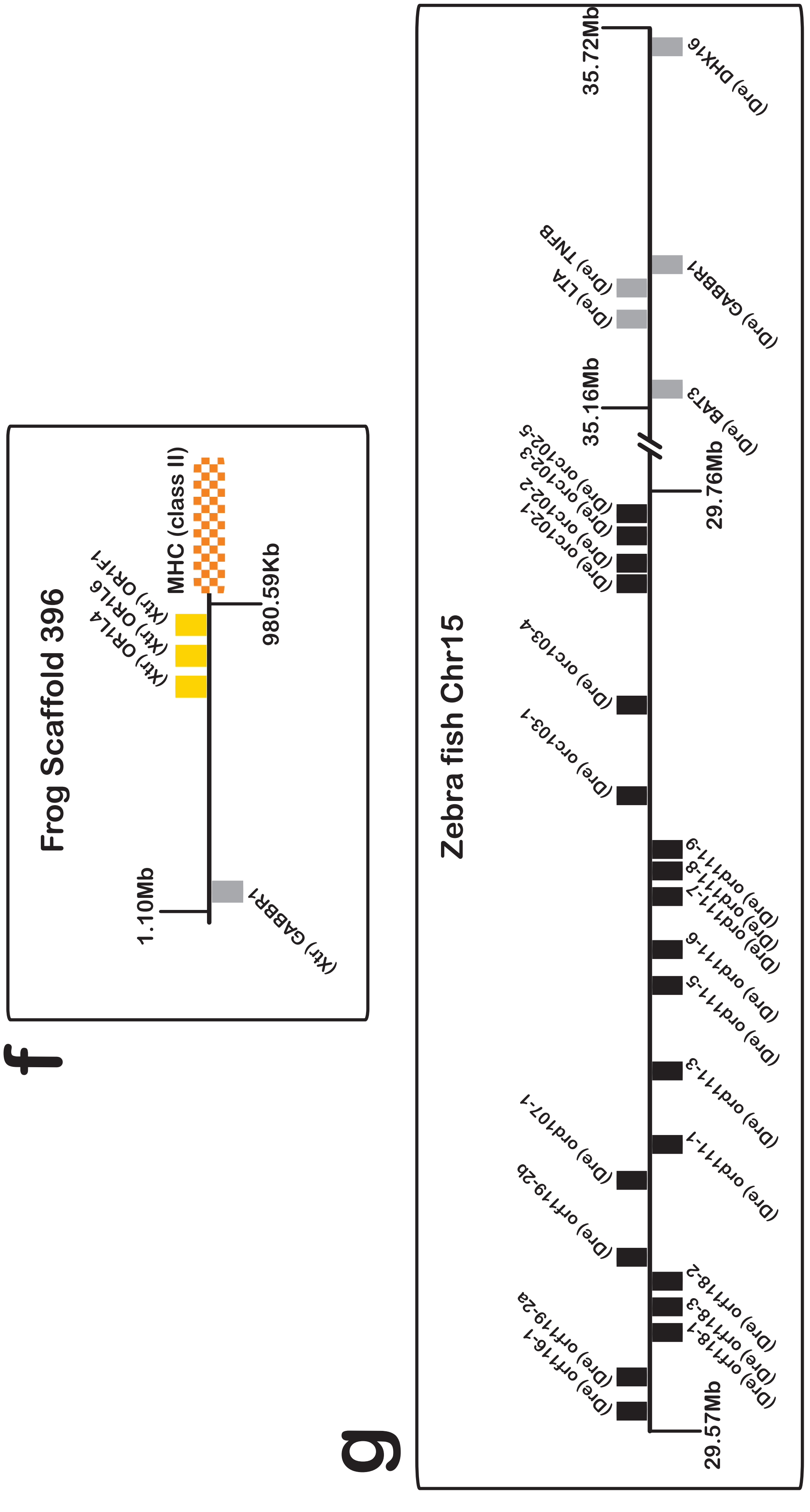

a
Human Chr6p
HIST
MHC
27.99Mb
29.75Mb
OR2B6
OR1F12
OR2J1
OR2J3
OR2J2
OR14J1
OR12D2
OR12D1
OR10C1
OR2H1
OR2H2
MOG
OR2B2
OR2B8P
OR2W1
OR2B3
OR5V1
OR12D3
OR11A1
UBD
GABBR1
Chimpanzee Chr6p
HIST
MHC
28.45Mb
30.18Mb
(Ptr) OR2W6
(Ptr) OR2B6
(Ptr) OR2J1
(Ptr) OR2J3
(Ptr) OR2J2
(Ptr) OR14J1
(Ptr) OR12D2
(Ptr) OR10C1
(Ptr) OR2H1
(Ptr) OR2H2
MOG
(Ptr) OR2W1
(Ptr) OR2B3
(Ptr) OR11A1
UBD
GABBR1
Orangutan Chr6p
HIST
MHC
28.83Mb
30.63Mb
(Ppy) OR2B6
(Ppy) OR2W6
(Ppy) OR1F12
(Ppy) OR2J1
(Ppy) OR2J3
(Ppy) OR2B11
(Ppy) OR14J1
(Ppy) OR10C1
(Ppy) OR2I1
(Ppy) OR2H2
MOG
(Ppy) OR2U1
(Ppy) OR5V1
(Ppy) OR11A1
UBD
GABBR1
Rhesus macaque Chr4
HIST
MHC
27.72Mb
29.41Mb
(Mamu) OR2W6
(Mamu) OR2J3
(Mamu) OR2J2
(Mamu) Olfr131
(Mamu) Olfr129
(Mamu) Olfr130
(Mamu) OR2B4
(Mamu) OR14J1
(Mamu) OR12D2a
(Mamu) OR12D2b
(Mamu) OR2I1
(Mamu) OR2H2
MOG
(Mamu) OR2W1
(Mamu) OR2B3
(Mamu) OR2G1
(Mamu) OR5V1
(Mamu) OR12D3
UBD
GABBR1


b
Mouse Chr17
MHC
37.15Mb
38.47Mb
MOG
GABBR1
Olfr90
Olfr93
UBD
Olfr94
Olfr95
Olfr96
Olfr97
Olfr98
Olfr99
Olfr101
Olfr102
Olfr103
Olfr104
Olfr105
Olfr106
Olfr107
Olfr108
Olfr109
Olfr110
Olfr111
Olfr112
Olfr114
Olfr116
Olfr117
Olfr119
Olfr120
Olfr121
Olfr123
Olfr124
Olfr125
Olfr126
Olfr127
Olfr128
Olfr761
Olfr130
Olfr131
Olfr132
Olfr133
Olfr134
Olfr135
Olfr138
Olfr137
Olfr136
Rat Chr20
MHC
1.61Mb
tel
92.81Kb
MOG
GABBR1
Olr1750
Olr1749
Olr1748
(Rno) Olfr94
UBD
Olr1744
Olr1743
Olr1742
(Rno) Olfr108
Olr1739
Olr1738
Olr1737
Olr1736
Olr1735
Olr1734
Olr1733
Olr1731
Olr1730
Olr1729
(Rno) Olfr105
Olr1726
Olr1724
Olr1722
Olr1720
Olr1718
Olr1714
Olr1710
Olr1709
Olr1708
Olr1707
Olr1702
Olr1701
Olr1700
Olr1699
(Rno) Olfr120
Olr1697
Olr1696
Olr1695
Olr1694
Olr1693
Olr1692
Olr1691
Olr1690
Olr1689
Olr1688
Olr1687
Olr1686
Olr1684
Olr1683
Olr1682
Olr1681
Olr1680
Olr1679
Olr1678
(Rno) Olfr134
Olr1675
Olr1673
Olr1671
Olr1670
Olr1668


C
Horse Chr20
MHC
HIST
HIST
HIST
28.79Mb
24.74Mb
MOG
GABBR1
UBD
(Eca) OR2H1c
(Eca) OR2H1b
(Eca) OR2H1a
(Eca) Olfr303b
(Eca) OR12D1b
(Eca) OR10C1b
(Eca) OR12D2e
(Eca) OR12D2d
(Eca) OR12D2c
(Eca) OR14J1b
(Eca) OR12D2b
(Eca) OR12D2a
(Eca) Olfr97b
(Eca) Olfr97a
(Eca) OR2B4c
(Eca) OR2B4b
(Eca) Olfr130c
(Eca) Olfr120b
(Eca) OR2J3e
(Eca) OR2J3d
(Eca) OR2J3c
(Eca) OR10C1a
(Eca) Olfr303a
(Eca) OR14J1a
(Eca) OR2B4a
(Eca) Olfr123
(Eca) Olfr130a
(Eca) Olfr120a
(Eca) Olfr129
(Eca) Olfr131
(Eca) Olfr132
(Eca) OR2J3b
(Eca) OR2J1b
(Eca) OR2J3a
(Eca) OR2J1a
(Eca) OR2B7b
(Eca) OR2B7a
(Eca) OR2W6
(Eca) OR2B6
(Eca) OR2W4
(Eca) OR2W2
(Eca) OR2M3c
(Eca) OR2M3b
(Eca) OR2M3a
(Eca) OR11A1
(Eca) OR12D3
(Eca) Olfr130b
(Eca) Olfr110b
(Eca) Olfr110a
(Eca) OR2W1
(Eca) OR2B3b
(Eca) Olfr1370
(Eca) Olfr134c
(Eca) OR2B3a
(Eca) Olfr134b
(Eca) Olfr134a
(Eca) OR2AD1c
(Eca) OR2AD1b
(Eca) OR2AD1a
(Eca) OR2U1
(Eca) Olfr1367
(Eca) OR2B8g
(Eca) OR2B8f
(Eca) OR2B8e
(Eca) OR2B8d
(Eca) OR2B8c
(Eca) OR2B8b
(Eca) OR2B8a
(Eca) OR2B2
(Eca) OR2M3e
(Eca) OR2M3d
Cow Chr23
MHC
HIST
HIST
HIST
28.89Mb
31.51Mb
MOG
GABBR1
UBD
(Bta) OR2H2
(Bta) OR2H1
(Bta) OR2I1
(Bta) Olfr97b
(Bta) OR12D2e
(Bta) OR12D2d
(Bta) Olfr97a
(Bta) OR12D2c
(Bta) OR12D2b
(Bta) OR12D2a
(Bta) OR14J1
(Bta) OR2B4
(Bta) Olfr123
(Bta) Olfr131
(Bta) OR2J3e
(Bta) OR2J3d
(Bta) OR2J3c
(Bta) OR2J3b
(Bta) OR2J3a
(Bta) OR2B7c
(Bta) OR2B7b
(Bta) OR2B7a
(Bta) OR2W2
(Bta) OR2B2b
(Bta) OR2M2c
(Bta) OR2M2b
(Bta) OR2M2a
(Bta) OR11A1
(Bta) OR5V1c
(Bta) Olfr109
(Bta) OR5V1b
(Bta) OR5V1a
(Bta) Olfr90c
(Bta) Olfr90b
(Bta) Olfr90a
(Bta) Olfr130
(Bta) OR2U1
(Bta) OR2B3b
(Bta) Olfr1370b
(Bta) OR2B3a
(Bta) Olfr1370a
(Bta) Olfr120b
(Bta) Olfr120a
(Bta) Olfr129
(Bta) OR2W1b
(Bta) OR2W1a
(Bta) OR2AD1
(Bta) Olfr1367
(Bta) Olfr1366
(Bta) OR2B8d
(Bta) OR2B8c
(Bta) OR2B8b
(Bta) OR2B8a
(Bta) OR2W6
(Bta) OR2B6
(Bta) OR2W4
(Bta) OR2B2a
Pig Chr7
MHC
HIST
HIST
HIST
23.96Mb
20.89Mb
GABBR1
(Ssc) Olfr131c
(Ssc) Olfr120c
(Ssc) Olfr132c
(Ssc) Olfr1370b
(Ssc) OR2B3d
(Ssc) Olfr1370a
(Ssc) OR2B3c
(Ssc) Olfr130c
(Ssc) Olfr110b
(Ssc) Olfr130d
(Ssc) OR12D3d
(Ssc) OR12D3c
(Ssc) OR12D3b
(Ssc) Olfr1730c
(Ssc) OR12D2f
(Ssc) OR14J1c
(Ssc) OR12D2e
(Ssc) OR14J1b
(Ssc) OR12D2d
(Ssc) OR12D2c
(Ssc) OR14J1a
(Ssc) OR12D2b
(Ssc) OR11A1
(Ssc) Olfr123b
(Ssc) Olfr120a
(Ssc) Olfr131a
(Ssc) Olfr132a
(Ssc) Olfr1366
(Ssc) OR2B7e
(Ssc) OR2B7d
(Ssc) OR2B7b
(Ssc) OR2B7a
(Ssc) OR2W7
(Ssc) OR2W4
(Ssc) OR2B6
(Ssc) OR2W6
(Ssc) OR2M2o
(Ssc) OR2M2n
(Ssc) OR2M2m
(Ssc) OR2M2l
(Ssc) OR2M2j
(Ssc) OR2M2h
(Ssc) OR2M2d
(Ssc) OR2M2c
(Ssc) OR2M2b
(Ssc) OR2M2a
(Ssc) OR2B3e
(Ssc) OR2J3
(Ssc) Olfr132b
(Ssc) Olfr131b
(Ssc) Olfr120b
(Ssc) Olfr90b
(Ssc) Olfr130b
(Ssc) Olfr123a
(Ssc) OR2B4
(Ssc) OR14J1d
(Ssc) Olfr169
(Ssc) OR12D2g
(Ssc) Olfr1730b
(Ssc) OR12D3a
(Ssc) Olfr1730a
(Ssc) Olfr110a
(Ssc) OR12D2a
(Ssc) Olfr303
(Ssc) Olfr90a
(Ssc) Olfr130a
(Ssc) Olfr133a
(Ssc) OR2B3b
(Ssc) OR2B3a
(Ssc) MOG
(Ssc) Olfr1367
(Ssc) OR2B7f
(Ssc) OR2B7c
(Ssc) OR2M2k
(Ssc) OR2M2i
(Ssc) OR2M2g
(Ssc) OR2M2f
(Ssc) OR2M2e


d
Dog Chr35
HIST
tel
28.06Mb
29.41Mb
(Cfa) OR2W6
(Cfa) OR2B10
(Cfa) OR2W4
(Cfa) OR2B7
(Cfa) OR10C2
MOG
(Cfa) OR2B2
(Cfa) OR2Y2
(Cfa) OR2AD2
(Cfa) OR2W1
GABBR1
(Cfa) OR2H9
Cat ChrB2
HIST
tel
1.44Mb
58.86Kb
(Cfe) OR2W4
(Cfe) OR2W2
(Cfe) OR2B7
(Cfe) OR1F12
MOG
(Cfe) OR2Y2a
(Cfe) OR2Y2b
(Cfe) OR2AD2
(Cfe) OR2W1
GABBR1


e
Opossum Chr2
MHC
HIST
HIST
266.33Mb
257.89Mb
MOG
GABBR1
UBD
(Mdo) OR2I1
(Mdo) OR10C1
(Mdo) Olfr303
(Mdo) OR11A1
(Mdo) OR12D2d
(Mdo) OR12D2c
(Mdo) OR12D2b
(Mdo) OR12D2a
(Mdo) OR2B4
(Mdo) OR2G3d
(Mdo) OR2G3c
(Mdo) Olfr90c
(Mdo) OR2G3b
(Mdo) OR2G3a
(Mdo) Olfr90b
(Mdo) Olfr121
(Mdo) Olfr131
(Mdo) Olfr90a
(Mdo) OR2H2
(Mdo) OR2B2
(Mdo) OR2N1P
(Mdo) OR2W6
(Mdo) OR2W1
(Mdo) Olfr332b
(Mdo) Olfr332a
(Mdo) OR2W3
(Mdo) OR2B8b
(Mdo) OR2B8a
(Mdo) Olfr1386c
(Mdo) Olfr1386a
(Mdo) Olfr1386b
(Mdo) OR2Y1b
(Mdo) OR2Y1a
(Mdo) OR2V1b
(Mdo) OR2V1a
(Mdo) Olfr1395b
(Mdo) Olfr1395a
(Mdo) OR1F12


f
Frog Scaffold 396
1.10Mb
(Xtr) GABBR1
(Xtr) OR1L4
(Xtr) OR1L6
(Xtr) OR1F1
MHC (class II)
980.59Kb
g
Zebra fish Chr15
29.57Mb
(Dre) orf116-1
(Dre) orf119-2a
(Dre) orf118-1
(Dre) orf118-3
(Dre) orf118-2
(Dre) orf119-2b
(Dre) ord107-1
(Dre) ord111-1
(Dre) ord111-3
(Dre) ord111-5
(Dre) ord111-6
(Dre) ord111-7
(Dre) ord111-8
(Dre) ord111-9
(Dre) orc103-1
(Dre) orc103-4
(Dre) orc102-1
(Dre) orc102-2
(Dre) orc102-3
(Dre) orc102-5
29.76Mb
35.16Mb
(Dre) BAT3
(Dre) LTA
(Dre) TNFB
(Dre) GABBR1
35.72Mb
(Dre) DHX16

5. Transmission Distortion in the Human Genome

In this chapter, I will present three manuscripts that describe the work performed focusing the phenomenon of allelic transmission distortion in humans. While two of them (5.1 and 5.2) address the TD problematic in phenomenon itself both in publicly available and in independently sampled populations, the third manuscript (5.3) addresses a technical problem associated with one of the most important online resources for human sequence variation today: the international HapMap project.

5.1. Transmission Distortion in the Human Chromosome 6p

5.1.1. Summary

The fact that the mouse t haplotypes are under extreme transmission distortion (TD, a departure from Mendel's law of independent segregation of alleles in a given heterozygous locus or region) affect a region syntenic to large parts of the short arm human chromosome 6 (Chr6p) – including the HLA complex and linked olfactory receptors – led us to investigate the same phenomenon in the populations genotyped through the international haplotype mapping consortium (HapMap). Beyond the 180 founder haplotypes of the CEU population assessed for the study described on chapter 2, the HapMap had, at that time, SNP data from 60 family trios (father-mother-adult child) available – 30 from the CEU and 30 from the YRI population. The study presented in the following article aimed at testing, *in silico*, the possibility that alleles of SNPs on Chr6p could be under TD. Both OR clusters and the MHC were found to segregate conforming to Mendelian expectation. However, a centromeric region of Chr6p harbouring, among others, the transcription-factor encoding locus *RUNX2* was found to be under strong TD. Besides of discussing the importance of TD studies in healthy cohorts, and of giving examples of studies that could be biased because TD was not considered, the article describes a way to use linkage disequilibrium data in order to choose subsets of SNPs carrying haplotypic information and thus circumvent the statistical problem of multiple testing.

One additional result of this work concerns the data resource property of the international HapMap project. Scientists world-wide use the HapMap as a guide for study design and interpretation, as an overview of genomic diversity can be interactively browsed by the user through the internet. The results of the following work were compiled in a code file (Supplemental File S3) that can be uploaded by readers into the HapMap genome browser and viewed within the genomic context. This is a first step towards a future genome-wide TD map, which should ideally be available for browsing by HapMap users investigating a genomic region. As we point out in the paper, TD is a feature that has to be taken into account during interpretation of results from phenotype association studies. In this sense, the ability to check, through an internet browser, if one given region of interest shows evidence of TD or not, would be extremely helpful.

5.1.2. Publication

Santos PS, Höhne J, Schlattmann P, König IR, Ziegler A, Uchanska-Ziegler B, Ziegler A (2009): Assessment of transmission distortion on chromosome 6p in healthy individuals using tagSNPs. European Journal of Human Genetics 17:1182-1189

DOI:10.1038/ejhg.2009.16

Pages 78-90

5.2. Transmission Distortion in Southern Brazilian Families

5.2.1. Summary

The work described in this chapter aimed at verifying, with a larger and independent cohort, the results obtained with the TD assessment of Chr6p using the HapMap families from the CEU and YRI populations (see 5.1). Even though results of that study were statistically significant, the fact that TD was observed among CEU but not among YRI trios left several questions open. Is TD an ethnicity-dependent feature? Is the YRI cohort an exception to the rule? Or is the TD observed just an artefact resulting from sample stratification within the CEU sample of families and inflated by linkage disequilibrium in the S-M-R region? The study shown in this section aimed at answering these questions. Genomic DNA from 239 individuals from 141 family trios from Curitiba, Southern Brazil, were obtained through a cooperation with a Brazilian immunogenetics laboratory that recruits volunteer bone marrow donors, which, in some cases, concur in participating in population genetics studies like the one presented here. All individuals were genotyped for SNPs covering the S-M-R region. An additional genome-wide search for SNPs showing evidence of TD, based on criteria from the International HapMap project was performed as well. The results of the first study (Section 5.1) were confirmed in Southern Brazilian trios, and TD could be established as a general feature of the S-M-R region among Caucasian populations. However, TD was not found for other SNPs in the genome, although suggestive results were obtained for the follistatin-like 1 (*FSTL1*) locus on chromosome 2.

5.2.2. Manuscript

Santos PS, Höhne J, Schlattmann P, Poerner F, Bicalho MG, Ziegler A, Uchanska-Ziegler B: Presence of Transmission Distortion on Human Chromosome 6p Revealed by SNP Genotyping of Southern Brazilian Families. Submitted.

Pages 92-112

Presence of Transmission Distortion on Human Chromosome 6p Revealed by SNP Genotyping of Southern Brazilian Families

Pablo Sandro Carvalho Santos[1,4], Johannes Höhne[2,4], Peter Schlattmann[2], Fabiana Poerner[3], Maria da Graça Bicalho[3], Andreas Ziegler[1], and Barbara Uchanska-Ziegler[1]

[1]Institut für Immungenetik and [2]Institut für Biometrie und klinische Epidemiologie, Charité–Universitätsmedizin Berlin, Freie Universität Berlin, Berlin, Germany; [3]LIGH/UFPR: Laboratório de Imunogenética e Histocompatibilidade, Departamento de Genética da Universidade Federal do Paraná, Curitiba, Brazil

[4]These authors contributed equally to this work.

E-mail addresses:
PSCS: pablo.santos@charite.de, JH: johannes.hoehne@charite.de,
PS: peter.schlattmann@charite.de, FP: fpoerner11@gmail.com, MGB: ligh@ufpr.br,
AZ: andreas.ziegler@charite.de, BU-Z: bziegler@charite.de

Corresponding author:
Andreas Ziegler, Institut für Immungenetik, Charité-Universitätsmedizin Berlin, Campus Benjamin Franklin, Freie Universität Berlin, Thielallee 73, 14195 Berlin, Germany.
Tel: +49-30-450 564731, Fax:+49-30-450 564920,
E-mail: andreas.ziegler@charite.de

Running Head: Transmission Distortion in Chr6p among Southern Brazilian Families

Abstract

Background. Transmission distortion (TD) is a statistically significant violation of Mendel's expected 1:1 allelic transmission ratio from heterozygous parents to their offspring. We have previously reported evidence for TD of single nucleotide polymorphisms (SNPs) within human chromosome 6p around the loci *SUPT3H, MIR586* and *RUNX2* (*S-M-R* region) among 30 HapMap family trios of European ancestry through a systematic *in silico* investigation. Several questions however, regarding the relatively small sample size, the differences between investigated populations, the haplotype phasing and the possibility of genotyping errors remained unanswered.

Results. We now report a search for TD in a fully independent cohort of 141 Southern Brazilian family trios of predominantly European ancestry by SNP genotyping. We primarily focused on the *S-M-R* region, but performed, in addition, a genome-wide search within the 30 CEU trios for markers for which paternal and maternal TD were reciprocally compensated within each family trio. The over four-fold increased sample size allowed us to dispense with the use of software-based haplotype phasing, and thus consider undisputable transmissions only, without losing statistical power. Strong evidence of TD however, was found only for the *S-M-R* region.

Conclusions. Although it still remains enigmatic why TD is present, and how a considerable degree of heterozygosity is retained in the corresponding loci, TD of the *S-M-R* region is clearly a phenomenon present in the general, "unaffected" population. As a consequence, TD has to be taken into account when assessing the outcome of linkage studies, especially in the case of the intensively investigated *RUNX2* locus.

Background

Transmission Distortion (TD) refers to the contravention of Mendel's law of independent assortment of characteristics. The best known example of TD, first described over 70 years ago, concerns the mouse t haplotypes, which are an extreme – yet not completely understood – example of distorted segregation of large segments of mouse chromosome 17. [1] In the present study, we will refer to TD in a broad sense, denoting a statistically significant deviation from the generally expected 1:1 transmission ratio of alleles from heterozygous parents to their offspring, without regard to its molecular causes or mechanisms.

TD has already been described for different loci in many species, [1–4] including humans. [5–7] In humans, TD has traditionally been investigated among families selected for different diseases or phenotypic traits, but the focus of these studies has recently been extended also to families without a disease phenotype. [5,8–14] TD is not only interesting as a biological phenomenon, but the detailed identification of loci under TD in the general, healthy population, is a requirement for the correct interpretation of linkage and association studies based on allelic transmission within families. If ignored, the presence of TD can introduce serious bias into such studies. [5,9,12,15,16]

We recently reported evidence of TD in a region of the short arm of human chromosome 6 (Chr6p) harboring three genes: the microRNA locus *MIR586* and the two transcription factor-encoding *SUPT3H* and *RUNX2*. [14] The close physical linkage of these two transcription factor loci is apparently an extreme case of phylogenetic conservation, which can be observed not only in all mammals, birds and reptiles studied so far, [17] but also amphibians, [18] fish, [19] and even in the demosponge *Amphimedon queenslandica*, indicating the existence of linkage between *SUPT3H* and *RUNX2* for at least 700 million years. [20] Whereas little information is currently available for *MIR586* and *SUPT3H*, this is not the case for *RUNX2*. The protein encoded by this gene is a master transcription factor crucially related to bone, cartilage and tooth morphogenesis. [21–24] It was also shown to be expressed in non-skeletal tissues, [25] and to be associated with many central processes such as cell growth, determination of cell fate, epigenetic regulation, [26,27] and hematopoiesis. [28,29] A recent query to the STRING database yielded 119 human proteins as functional partners of Runx2, with high confidence prediction. [30] Additionally, its role on clinical conditions as osteoporosis, [31,32] cancers of breast, [33] prostate, [34] bone, [35] and lymphoid tissues [36,37] makes Runx2 a promising target also for therapeutic interventions. [38] It seems therefore essential to assess the possibility that the linkage disequilibrium (LD) block around *RUNX2*, *MIR586* and *SUPT3H* (the "*S-M-R* region") is under TD in the general, healthy population. Likewise, since TD in this chromosomal segment appears to be an ethnicity-dependent phenomenon, [14] an investigation of trios belonging to further ethnic groups is desirable.

The main limitation of many investigations addressing TD – including our own previous analysis [14] – was the small number of trios, and consequently the reduced statistical power. In the present investigation we analyzed the occurrence of TD in a sample of family trios over four times as large as previously. These trios belong to a completely independent population which is, however, ethnically related to the CEU sample. Additionally, we investigated whether it would be possible to detect TD among single nucleotide polymorphisms (SNPs) from the International HapMap Project using a group of criteria that considered not only a sex-compensatory version of the transmission/disequilibrium test (TDT), [39] but also the LD shape of the respective region, and the number of SNPs involved. Following this rationale, we performed a genome-wide search for markers fulfilling those criteria, and genotyped a second group of SNPs in the larger, independent cohort mentioned above, looking for evidence of TD.

Results and Discussion

The analysis of SNPs from the *S-M-R* region allowed us to assess the existence of TD for that region in the Brazilian population. Considering only Euro-Brazilian trios (Fig. 1), TD was observed for rs12530016 and rs2038765 (parent-unspecific p-value = 0.0052 and 0.0086 respectively). While mothers and fathers seem to equally contribute to the effect observed for rs2038765, this is not the case for rs12530016. For this marker, mothers are driving the observed distortion (mother-specific p-value = 0.0236). The same analysis, considering all analyzed trios, is depicted in figure 2. TD is still observed for rs2038765 (parent-unspecific p-value = 0.0024), but is only marginal for rs12530016 (parent-unspecific p-value = 0.0347).

Regarding the sex-compensatory analysis, the only locus with a marker showing relatively strong TD was rs4676781, in the follistatin-like 1 (*FSTL1*) locus. When only Euro-Brazilians were considered, the parent-unspecific p-value was 0.0072, and the mother specific p-value was 0.0186, whereas for all analyzed trios the parent-unspecific p-value was 0.0085, and the mother specific p-value was 0.0133. However, these results did not withstand correction for multiple comparisons (Bonferroni).

Although all markers were genotyped together, the present study addressed two independent problems: (i) Is TD present in the *S-M-R* region of an unrelated population? and (ii) Are loci picked out of a genome-wide search from the HapMap CEU population according to a series of criteria including TD, sex-compensatory effects and LD, under TD in an independent population? While we could demonstrate the presence of TD in the S-M-R region also among Brazilian trios, we did not find unequivocal evidence for TD with regard to further loci.

The six markers analyzed for the *S-M-R* region were chosen as tagSNPs for a genomic segment spanning ~ 1.7 Mb, for which we had previously obtained evidence supporting the existence of TD in the CEU HapMap population. [14] Two of these SNPs showed evidence of TD: one telomeric (rs12530016), in high LD with markers mapping to the *SUPT3H* and *MIR586* loci, and one central SNP (rs2038765), in complete LD with at least 62 markers distributed over *SUPT3H* and the telomeric half of *RUNX2*, which is consistent with our previous findings. A list of SNPs tagged by rs12530016 and rs2038765, as well as a visual representation of their relative positions, are given in figure 3 and table 2. The sexes of the parents responsible for the observed TD seem to imply a difference between our present findings and those from the HapMap: while TD in the CEU population was largely a paternal effect, the distortion observed in the Brazilian population is an added effect from both sexes, even with mothers of European ancestry dominating in the case of rs12530016 (Fig. 2). The only explanation for this discrepancy which we regard as likely is the lack of statistical power in the previous HapMap analyses, [14] in which only thirty trios had been tested. Moreover, the relatively high number of trios available for the present study allowed us to dispense software-based phasing of haplotypes, which was necessary in the HapMap trios and is a possible source of bias. This means that we considered only undisputable transmissions (e.g. in the case of a heterozygous father and a homozygous mother) yet keeping high statistical power.

With the second analysis, we addressed a problem emerging from a genome-wide search for loci in the HapMap CEU population possessing several features (see methods) that can be taken as evidence of TD. Although some SNPs seemed to show evidence of TD (in particular, the loci *FSTL1*, neuroligin 1 (*NLGN1)*, huntingtin (*HTT*) and the zinc finger protein 667 (*ZNF667*) deserve attention in follow-up studies), none withstood correction for multiple testing. Having genotyped 48 markers, TD in any given SNP would have to yield a p-value lower than 10^{-3} in order to "survive" correction. Therefore, we cannot exclude the possibility that results are due to chance alone, even in suggestive cases as *FSTL1*: as in the case of the product of the *RUNX2* gene, the FSTL1 protein is involved with carcinogenesis, [45] and is predicted to interact with the bone morphogenetic protein 2 and the noggin precursor, [30] both essential for several developmental processes. Additionally, the *FSTL1* locus is associated with two sharp LD blocks in all populations (www.hapmap.org).

Whereas for most families of German, Polish, Ukrainian and Japanese ancestries the immigration to Southern Brazil is a relatively recent event (around 100 years ago), this is different for families with African, Portuguese, Italian and, of course, Amerindian ancestries. [41] A certain degree of ethnic admixture is therefore expected to be present in many families assessed here (Euro-Brazilians or not). Even so, genotyping of the *S-M-R* region still revealed evidence of TD for rs2038765 among all Brazilian families (Fig. 1, 75% of Euro-Brazilians). As expected, the restriction of the sample only to Euro-Brazilians (Fig. 2) yielded a picture more similar to the HapMap CEU population, with the emergence of one additional TD signal due to rs12530016. The fact that we are still able to detect TD, despite the assumed genetic admixture, reinforces its presence.

After performance of the genome-wide search for SNPs under TD that supported our sex-compensatory analysis, the HapMap entered a new phase with the release of HapMap3. This was an extreme improvement compared to HapMap2, as seven new populations were added to the project, and a new set of markers was genotyped, making inter-population comparisons easier to perform. However, it seems that many markers under TD have been preferentially excluded from the genotyping plates of HapMap3, possibly because of irregularities in Hardy-Weinberg equilibrium. As a result, TD is hardly detectable in HapMap3. Even so, mothers of a Maasai (MKK) population still show evidence of TD in the *S-M-R* region, while (HapMap3) CEU mothers exhibit strong evidence of TD on the rs4676781 marker within the *FSTL1* locus (results not shown).

The present study further substantiates the evidence that the *S-M-R* region is under TD not only in disease families, [7] but also in healthy individuals. [14] It remains unclear, however, what forces keep this region in a heterozygous state, and also what the biological significance of this phenomenon might be. As TD has sometimes been shown to relate to genotypes of maternal or paternal grandparents, [8,10] we believe that this kind of assessment is desirable for future studies that address TD within the *S-M-R* region. The 1000 genomes project (www.1000genomes.org), once concluded for family trios, is expected to shed new light on these questions. Whereas a whole-genome search for loci under TD is expected to face strong statistical obstacles due to the principal problem of multiple testing (see Santos et *al.* [14] for an extended discussion), the *S-M-R* region undoubtedly represents a set of genetic markers that are suitable to investigate not only the relationship between TD and LD, but also the reasons for continued TD despite the apparent absence of sex-compensatory effects.

Subjects and Methods

Study population

Our cohort was composed of 141 Southern Brazilian Family trios, from 49 unrelated families (49 mothers, 49 fathers and 141 children, totaling 239 individuals) with an average of 2.88 children per family. All subjects belonged to families recruited in the context of searching for compatible bone marrow donors for patients with a transplant indication. All family members were aged 18 or above, signed an informed consent allowing population studies and were residents of the city (or surroundings) of Curitiba, State of Paraná, Southern Brazil. Curitiba's ethnic composition is a result of immigration waves in the nineteenth Century from Europe (mainly from areas corresponding to today's Germany, Poland, Ukraine and Italy), as well as migration movements within Brazil (through which people of Portuguese, West African, Lebanese and Japanese ancestry settled in the region). [40-42] All individuals were asked to reveal information about their ethnic origin and, accordingly, 36 families were of European origin (Portuguese, Italian, German or Polish backgrounds). The other families were of African (3), Japanese (1), Amerindian (3), or unknown (6) ancestry. There was no conflicting self-classification within family members. We excluded subjects for which the genotyping of more than 50% of markers failed, as well as those showing more than 1 Mendelian error (pointing at the possibility of an individual not being a biological child of both parents in the trio), which led to the exclusion of 19 individuals (11 for the former and 8 for the latter reason). DNA was extracted from 250 µl of fresh buffy coat by a salting out method. [43]

Candidate SNPs

In order to assess the Brazilian sample for TD in the *S-M-R* region, we selected 10 SNPs from this segment for genotyping. Markers were selected as tagSNPs, and covered the area reported to be under TD [14]: the entire reading frames of *SUPT3H* and *MIR586*, as well as the telomeric half of *RUNX2*. The tagging procedure was performed as previously described. [14] Additionally, we selected a second group of SNPs from throughout the genome, in order to shed light on a further problem: if a locus is under TD, one would expect it to be within a region of strong and well defined LD (strong LD block with sharp borders), and the reciprocal compensation of paternal and maternal TD could be an explanation why loci under TD remain heterozygous in a population. With this in mind, we performed a genome-wide search for markers under TD in the HapMap CEU population (results not shown) fulfilling the mentioned criteria, and selected fifty SNPs, from which twelve were coding for a non-synonymous amino acid exchange, and five had been reported as transcript-specific. [44]

All candidate SNPs were genotyped using the SNPlex® Genotyping System (experiments outsourced to Geneservice™, www.geneservice.co.uk). After genotyping, SNPs had to fulfill the following criteria in order to be kept for further analysis: less than two Mendelian errors, minor allele frequency of at least 10%, lack of complete LD with a neighboring SNP (in order to avoid redundancy), and successful genotyping in at least 50% of individuals. 49 SNPs (six from the *S-M-R* region and 43 from the genome-wide approach) passed all inclusion criteria and were considered further. Since only six SNPs (that do not segregate independently from each other) were tested in the first part of the study, p-values were not corrected for multiple testing. The second set of SNPs consisted of 43 SNPs, mostly independent from each other, and these tests underwent therefore statistical multiple testing correction. A list of the 49 SNP IDs with other details is given in table 1.

Statistical analyses

To investigate the parental specific TD in the *S-M-R* region, we computed the standard TDT [39] for each SNP 3 times: We tested the father-specific, the mother-specific and the parent-unspecific components. Under the null hypothesis (Mendelian transmission, $\theta = 0.5$) the TDT statistic follows a χ_1^2 distribution. The test statistic is computed in analogy to the McNemar test with $(b-c)^2 / (b+c)$, where b and c are given in the contingency table described by Spielman and colleagues. [39] A modification of this test was used to quantify sex-compensatory effects in the genome-wide approach of this study. Sex-compensatory TD occurs if allelic transmission from both the father and the mother are non-Mendelian and compensate each other. The test statistic is computed with

$$\frac{\left((b_F + c_M) - \ (c_F + \ b_M)\right)^2}{b_F + c_F + b_M + c_M}$$

Thereby we test if $\theta_F < 0.5 < \theta_M$ or $\theta_M < 0.5 < \theta_F$. Indices stand for father ('F') and mother ('M'). The derivation of this test is shown in the appendix.

List of abbreviations

CEU: Utah residents with ancestry from Northern and Western Europe (HapMap population), LD: linkage disequilibrium, MKK: Maasai in Kinyawa, Kenya (HapMap population), nsyn: non synonymous, *S-M-R* region: region on chromosome 6p including the loci *SUPT3H*, *MIR586* and *RUNX2*, SNP: single nucleotide polymorphism, TD: transmission distortion, TDT: transmission/disequilibrium test, trans: transcript specific.

Acknowledgements

We are grateful to C. Backhaus for excellent technical assistance. The Berliner Krebsgesellschaft e.V. supported this study (*Ernst-von-Leyden* fellowship to PSCS and a grant to BU-Z).

Conflict of interest

The authors declare that no competing interests exist.

Appendix

For a biallelic marker with alleles X and Y, heterozygote fathers transmit X more likely and heterozygous mothers transmit Y more likely – or vice versa. According to the contingency table described by Spielman and colleagues, [39] b_F specifies the number of heterozygous fathers that transmit X, and c_F heterozygous fathers transmit Y. The same holds for mothers with b_M and c_M.

In terms of the Chi-square test, we have two observations:

$$O_1 : b_F + c_M$$
$$O_2 : c_F + b_M.$$

If there is no sex-compensatory effect, we expect to observe

$$E_1 = E_2 = \frac{b_F + c_F + b_M + c_M}{2}$$

We can apply the χ^2 test

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{\left(b_F + c_M - \frac{b_F + c_F + b_M + c_M}{2}\right)^2}{\frac{b_F + c_F + b_M + c_M}{2}} + \frac{\left(c_F + b_M - \frac{b_F + c_F + b_M + c_M}{2}\right)^2}{\frac{b_F + c_F + b_M + c_M}{2}}$$

$$= \frac{\left((b_F + c_M) - (c_F + b_M)\right)^2}{b_F + c_F + b_M + c_M}$$

.

Since we have two observations, the test statistic follows a χ^2 distribution with one degree of freedom.

References

1. Lyon MF (2003). Transmission ratio distortion in mice. Annu Rev Genet 37: 393-408.
2. Lyttle TW (1993). Cheaters sometimes prosper: distortion of Mendelian segregation by meiotic drive. Trends Genet. 9(6):205-210.
3. Wu G, Hao L, Han Z, Gao S, Latham KE, de Villena FP, Sapienza C (2005). Maternal transmission ratio distortion at the mouse Om locus results from meiotic drive at the second meiotic division. Genetics. 170(1):327-334.
4. Aparicio JM, Ortego J, Calabuig G, Cordero PJ (2009). Evidence of subtle departures from Mendelian segregation in a wild lesser kestrel (*Falco naumanni*) population. Heredity. Article in press.
5. Eaves IA, Bennett ST, Forster P, Ferber KM, Ehrmann D, Wilson AJ, Bhattacharyya S, Ziegler AG, Brinkmann B, Todd JA (1999). Transmission ratio distortion at the INS-IGF2 VNTR. Nat Genet. 22(4):324-325.
6. Becker T, Jansen S, Tamm S, Wienker TF, Tümmler B, Stanke F (2007). Transmission ratio distortion and maternal effects confound the analysis of modulators of cystic fibrosis disease severity on 19q13. Eur J Hum Genet. 15(7):774-778.
7. Sull JW, Liang KY, Hetmanski JB, Fallin MD, Ingersoll RG, Park J, Wu-Chou YH, Chen PK, Chong SS, Cheah F, Yeow V, Park BY, Jee SH, Jabs EW, Redett R, Jung E, Ruczinski I, Scott AF, Beaty TH (2008). Differential parental transmission of markers in RUNX2 among cleft case-parent trios from four populations. Genet Epidemiol. 32(6):505-512.
8. Naumova AK, Leppert M, Barker DF, Morgan K, Sapienza C (1998). Parental origin-dependent, male offspring-specific transmission-ratio distortion at loci on the human X chromosome. Am J Hum Genet. 62(6):1493-1499.
9. Paterson AD, Petronis A (1999). Transmission ratio distortion in females on chromosome 10p11-p15. Am J Med Genet. 88(6):657-661.
10. Naumova AK, Greenwood CM, Morgan K (2001). Imprinting and deviation from Mendelian transmission ratios. Genome. 44(3):311-320.
11. Zöllner S, Wen X, Hanchard NA, Herbert MA, Ober C, Pritchard JK (2004). Evidence for extensive transmission distortion in the human genome. Am J Hum Genet. 74(1):62-72.
12. Friedrichs F, Brescianini S, Annese V, Latiano A, Berger K, Kugathasan S, Broeckel U, Nikolaus S, Daly MJ, Schreiber S, Rioux JD, Stoll M (2006). Evidence of transmission

ratio distortion of DLG5 R30Q variant in general and implication of an association with Crohn disease in men. Hum Genet. 119(3):305-311.

13. Hanchard N, Rockett K, Udalova I, Wilson J, Keating B, Koch O, Nijnik A, Diakite M, Herbert M, Kwiatkowski D (2006). An investigation of transmission ratio distortion in the central region of the human MHC. Genes Immun. 7(1):51-58.
14. Santos PS, Höhne J, Schlattmann P, König IR, Ziegler A, Uchanska-Ziegler B, Ziegler A (2009). Assessment of transmission distortion on chromosome 6p in healthy individuals using tagSNPs. Eur J Hum Genet, 17(9):1182-1189.
15. Greenwood CM, Morgan K (2000). The impact of transmission-ratio distortion on allele sharing in affected sibling pairs. Am J Hum Genet. 66(6):2001-2004.
16. Evans DM, Morris AP, Cardon LR, Sham PC (2006). A note on the power to detect transmission distortion in parent-child trios via the transmission disequilibrium test. Behav Genet. 36(6):947-950.
17. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P (2009). Ensembl 2009. Nucleic Acids Res. 37(Database issue):D690-7.
18. Bowes JB, Snyder KA, Segerdell E, Gibb R, Jarabek C, Noumen E, Pollet N, Vize PD (2008). Xenbase: a Xenopus biology and genomics resource. Nucleic Acids Res. 36(Database issue):D761-7.
19. Glusman G, Kaur A, Hood L, Rowen L (2004). An enigmatic fourth runt domain gene in the fugu genome: ancestral gene loss versus accelerated evolution. BMC Evol Biol. 4:43.
20. Robertson AJ, Larroux C, Degnan BM, Coffman JA (2009). The evolution of Runx genes II. The C-terminal Groucho recruitment motif is present in both eumetazoans and homoscleromorphs but absent in a haplosclerid demosponge. BMC Res Notes;2:59.
21. Yoshida CA, Furuichi T, Fujita T, Fukuyama R, Kanatani N, Kobayashi S, Satake M, Takada K, Komori T (2002). Core-binding factor beta interacts with Runx2 and is required for skeletal development. Nat Genet. 32(4):633-638.

22. Stein GS, Lian JB, van Wijnen AJ, Stein JL, Montecino M, Javed A, Zaidi SK, Young DW, Choi JY, Pockwinse SM (2004). Runx2 control of organization, assembly and activity of the regulatory machinery for skeletal gene expression. Oncogene 23(24):4315-4329.
23. Chen S, Gluhak-Heinrich J, Wang YH, Wu YM, Chuang HH, Chen L, Yuan GH, Dong J, Gay I, MacDougall M (2009). Runx2, osx, and dspp in tooth development. J Dent Res. 88(10):904-909.
24. Javed A, Afzal F, Bae JS, Gutierrez S, Zaidi K, Pratap J, van Wijnen AJ, Stein JL, Stein GS, Lian JB (2009). Specific residues of RUNX2 are obligatory for formation of BMP2-induced RUNX2-SMAD complex to promote osteoblast differentiation. Cells Tissues Organs. 189(1-4):133-137.
25. Jeong JH, Jin JS, Kim HN, Kang SM, Liu JC, Lengner CJ, Otto F, Mundlos S, Stein JL, van Wijnen AJ, Lian JB, Stein GS, Choi JY (2008). Expression of Runx2 transcription factor in non-skeletal tissues, sperm and brain. J Cell Physiol. 217(2):511-517.
26. Young DW, Hassan MQ, Pratap J, Galindo M, Zaidi SK, Lee SH, Yang X, Xie R, Javed A, Underwood JM, Furcinitti P, Imbalzano AN, Penman S, Nickerson JA, Montecino MA, Lian JB, Stein JL, van Wijnen AJ, Stein GS (2007). Mitotic occupancy and lineage-specific transcriptional control of rRNA genes by Runx2. Nature. 445(7126):442-446.
27. Young DW, Hassan MQ, Yang XQ, Galindo M, Javed A, Zaidi SK, Furcinitti P, Lapointe D, Montecino M, Lian JB, Stein JL, van Wijnen AJ, Stein GS (2007). Mitotic retention of gene expression patterns by the cell fate-determining transcription factor Runx2. Proc Natl Acad Sci USA. 104(9):3189-3194.
28. de Bruijn MF, Speck NA (2004). Core-binding factors in hematopoiesis and immune function. Oncogene. 24;23(24):4238-4248.
29. Okumura AJ, Peterson LF, Lo MC, Zhang DE (2007). Expression of AML/Runx and ETO/MTG family members during hematopoietic differentiation of embryonic stem cells. Exp Hematol. 35(6):978-988.
30. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C (2009). STRING 8 – a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 37(Database issue):D412-6.
31. Doecke JD, Day CJ, Stephens AS, Carter SL, van Daal A, Kotowicz MA, Nicholson GC, Morrison NA (2006). Association of functionally different RUNX2 P2 promoter alleles with BMD. J Bone Miner Res. (2):265-273.

32. Lee HJ, Koh JM, Hwang JY, Choi KY, Lee SH, Park EK, Kim TH, Han BG, Kim GS, Kim SY, Lee JY (2009). Association of a RUNX2 promoter polymorphism with bone mineral density in postmenopausal Korean women. Calcif Tissue Int. 84(6):439-445.
33. Barnes GL, Hebert KE, Kamal M, Javed A, Einhorn TA, Lian JB, Stein GS, Gerstenfeld LC (2004). Fidelity of Runx2 activity in breast cancer cells is required for the generation of metastases-associated osteolytic disease. Cancer Res. 64(13):4506-4513.
34. Akech J, Wixted JJ, Bedard K, van der Deen M, Hussain S, Guise TA, van Wijnen AJ, Stein JL, Languino LR, Altieri DC, Pratap J, Keller E, Stein GS, Lian JB (2009). Runx2 association with progression of prostate cancer in patients: mechanisms mediating bone osteolysis and osteoblastic metastatic lesions. Oncogene. Article in press.
35. Sadikovic B, Yoshimoto M, Chilton-MacNeill S, Thorner P, Squire JA, Zielenska M (2009). Identification of interactive networks of gene expression associated with osteosarcoma oncogenesis by integrated molecular profiling. Hum Mol Genet. 18(11):1962-1975.
36. Blyth K, Vaillant F, Hanlon L, Mackay N, Bell M, Jenkins A, Neil JC, Cameron ER (2006). Runx2 and MYC collaborate in lymphoma development by suppressing apoptotic and growth arrest pathways in vivo. Cancer Res. 66(4):2195-2201.
37. Kuo YH, Zaidi SK, Gornostaeva S, Komori T, Stein GS, Castilla LH (2009). Runx2 induces acute myeloid leukemia in cooperation with Cbfbeta-SMMHC in mice. Blood. 113(14):3323-3332.
38. Chua CW, Chiu YT, Yuen HF, Chan KW, Man K, Wang X, Ling MT, Wong YC (2009). Suppression of androgen-independent prostate cancer cell aggressiveness by FTY720: validating Runx2 as a potential antimetastatic drug screening platform. Clin Cancer Res. 15(13):4322-4335.
39. Spielman RS, McGinnis RE, Ewens WJ (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 52(3):506-516.
40. Probst CM, Bompeixe EP, Pereira NF, de O Dalalio MM, Visentainer JE, Tsuneto LT, Petzl-Erler ML (2000). HLA polymorphism and evaluation of European, African and Amerindian contribution to the White and Mulatto populations from Parana, Brazil. Human Biology, 72:597.
41. Wachowicz, R. 2002. A História do Paraná. 10th edition. Imprensa Oficial do Paraná, Curitiba.

42. Sens-Abuázar C, Santos PS, Bicalho MG, Petzl-Erler ML, Sperandio-Roxo V (2009). MHC microsatellites in a Southern Brazilian population. Int J Immunogenet, 36(5):269-274.

43. Lahiri DK and Nurnberger JI Jr (1991). A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies. Nucleic Acids Res. 19(19):5444.

44. Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J (2008). Genome-wide analysis of transcript isoform variation in humans. Nat Genet, 40(2):225-231.

45. Chan QK, Ngan HY, Ip PP, Liu VW, Xue WC, Cheung AN (2009). Tumor suppressor effect of follistatin-like 1 in ovarian and endometrial carcinogenesis: a differential expression and functional analysis. Carcinogenesis. 30(1):114-121.

Table 1. List of the 49 analyzed SNPs, the loci or region they belong to, and the sums of transmission events of both alleles from heterozygous parents to their offspring. nsyn: non synonymous coding SNPs; trans: transcript specific SNPs. Region 1 includes loci *ARHGAP29* and *ABCD3*; region 2 includes loci *GHR* and *SEPP1*; region 3 includes loci *GDF9*, *UQCRQ*, *LEAP2*, and *AFF4*; region 4 includes loci *PLAC9* and *ANXA11*; and region 5 includes the loci *KIF21A* and *ABCD2*.

***S-M-R* analysis**					
SNP ID	**Locus**	**All Brazilians**		**Euro-Brazilians**	
		Fathers	**Mothers**	**Fathers**	**Mothers**
rs12530016	*S-M-R*	12:19	19:31	10:19	10:23
rs13215618	*S-M-R*	21:24	20:24	19:21	19:17
rs2038765	*S-M-R*	11:23	20:37	7:16	17:30
rs6458419	*S-M-R*	17:20	14:21	15:17	13:15
rs1555681	*S-M-R*	23:19	20:11	20:18	15:10
rs6901786	*S-M-R*	19:25	15:20	14:17	15:17
Genome-wide approach					
SNP ID	**Locus**	**All Brazilians**		**Euro-Brazilians**	
		Fathers	**Mothers**	**Fathers**	**Mothers**
rs6426724	*USP48*	23:18	15:14	13:15	11:9
rs16825896 (nsyn)	*USP48*	23:26	14:12	19:16	9:10
rs6667223	*USP48*	26:27	19:13	22:16	12:11
rs11749 (nsyn)	*DNALI1*	19:23	19:17	11:17	15:12
rs12126162	*NEGR1*	19:14	22:20	15:12	16:16
rs6699841	*NEGR1*	16:22	24:22	14:16	23:19
rs1929132	Region 1	27:24	13:14	25:21	9:9
rs11165122	Region 1	24:23	10:7	20:19	9:7
rs12057415	*ABCD3*	24:26	10:9	21:22	10:8
rs1356424	*TSGA10*	20:14	25:23	16:11	21:18
rs1581249	*TSGA10*	21:17	26:20	13:11	23:17
rs12712041	*TSGA10*	16:20	25:21	13:13	22:19
rs2053724 (nsyn)	*DPP10*	22:20	21:20	16:15	15:17
rs1259294	*FSTL1*	18:26	23:21	14:23	17:19
rs4676781	*FSTL1*	15:23	9:23	13:22	7:19
rs11705889	*NLGN1*	25:24	16:21	24:23	12:16
rs11713253	*NLGN1*	24:32	18:28	21:24	14:22
rs1983060	*NLGN1*	28:19	30:20	25:16	26:14
rs362331 (nsyn)	*HTT*	28:17	21:11	20:14	18:10
rs2910864	Region 2	23:29	23:15	17:20	19:11
rs230819	Region 2	24:26	25:15	17:18	22:11
rs24705	Region 3	32:28	26:16	22:13	24:15
rs4705874	Region 3	28:26	28:19	19:12	26:18

rs2074506 (nsyn)	*VARS2*	5:3	4:12	4:3	4:10
rs9449444 (nsyn)	*IBTK*	18:21	19:19	11:14	14:13
rs10869500 (trans)	*OSTF1*	17:25	21:19	15:14	15:9
rs1054402	*PAPPA*	14:18	21:10	12:16	18:8
rs17302884	*PAPPA*	24:16	21:24	19:12	14:17
rs4837520	*PAPPA*	18:25	18:12	13:20	15:10
rs7071579	Region 4	20:16	37:21	15:11	24:15
rs1556897	Region 4	20:18	28:15	16:12	23:15
rs10769716 (nsyn)	*GVIN1*	24:21	21:9	16:12	18:9
rs17121881 (nsyn)	*AMICA1*	12:16	23:22	11:15	21:20
rs7968837	Region 5	22:13	20:14	18:10	19:11
rs7301705 (nsyn)	*OR6C74*	20:21	18:19	17:18	12:12
rs10083789 (nsyn)	*USP31*	17:15	13:16	14:11	11:16
rs9898390	*SMG6*	10:18	16:21	5:14	10:13
rs8074850	*SMG6*	28:28	14:15	23:24	9:8
rs4986764 (nsyn)	*BRIP1*	24:22	14:17	14:13	13:16
rs6505780 (trans)	*CEP192*	20:21	27:22	16:12	21:20
rs527839 (trans)	*CEP192*	24:20	24:22	18:13	20:20
rs3760849 (nsyn)	*ZNF667*	20:19	21:28	11:16	12:25
rs17738540 (nsyn)	*SEC14L4*	5:12	6:9	5:12	4:9

Table 2. List of markers tagged by each of the two SNPs found to be under TD among Southern Brazilian family trios (*rs2038765* and *rs12530016*).

SNPs tagged by rs2038765		**SNPs tagged by rs12530016**	
Marker ID	**Coordinate**	**Marker ID**	**Coordinate**
rs2023311	44786284	rs12530016	44974300
rs12205657	44788593	rs13206526	44924159
rs6924185	45125206	rs16869119	44922870
rs11961316	45137576	rs3799972	44931251
rs11970412	45145494	rs3799974	44932926
rs2038765	45184472	rs3823252	44918396
rs12198982	45185892	rs9296450	45061764
rs12193720	45211340	rs9472414	45054484
rs12209161	45212299		
rs10456542	45215557		
rs1324536	45224239		
rs12193812	45248739		
rs11965706	45249419		
rs17209636	45250495		
rs12206568	45260725		
rs12213735	45260826		
rs12198376	45261659		
rs10948212	45264911		
rs12205860	45265301		

rs10948213	45266493
rs6919813	45267155
rs6919998	45267267
rs6919873	45267314
rs6920046	45267409
rs12191566	45267980
rs12211519	45268508
rs10456122	45271035
rs17209678	45271579
rs12212745	45274236
rs10948214	45278444
rs12193030	45283188
rs2093900	45284661
rs10456543	45286885
rs12192890	45290412
rs12206561	45290787
rs17288250	45300000
rs17288257	45301054
rs6927213	45314261
rs10456549	45333739
rs12191262	45334994
rs4443508	45343930
rs10948220	45344632
rs4400216	45345933
rs4479922	45356349
rs11964690	45366046
rs10948223	45381899
rs12191751	45390089
rs17209741	45400416
rs12201555	45402378
rs12205523	45404596
rs12201899	45406182
rs10948226	45407152
rs12199256	45411471
rs17209769	45429183
rs11966878	45436243
rs12194628	45437354
rs12203466	45438241
rs10807321	45438952
rs17288320	45446681
rs12216308	45450283
rs12210230	45452832
rs17288327	45457855

Figure Titles and Legends:

Figure 1. TD plot of the 49 genotyped SNPs among all Brazilian family trios analyzed in this study (N = 122). For each marker, the pink, green and blue bars contain information regarding the corresponding loci, rs number and chromosome, respectively. Yellow vertical bars are used to indicate those for which TD was found with a p-value below 10^{-2}, while two horizontal lines indicate p-value thresholds of 10^{-2} and $10^{-2.5}$. Green circles correspond to TD measured among fathers, red for mothers, while black circles indicate parent-unspecific TD. Regions 1 to 5 contain more than one locus, and their codes are given in table 1.

Figure 2. TD plot of the 49 genotyped SNPs among Euro-Brazilian family trios (N = 98). For each marker, the pink, green and blue bars contain information regarding the corresponding loci, rs number and chromosome, respectively. Yellow vertical bars are used to indicate those for which TD was found with a p-value below 10^{-2}, while two horizontal lines indicate p-value thresholds of 10^{-2} and $10^{-2.5}$. Green circles correspond to TD measured among fathers, red for mothers, while black circles indicate parent-unspecific TD. Regions 1 to 5 contain more than one locus, and their codes are given in table 1.

Figure 3. Screen shot of the HapMap genome browser displaying the positions of each group of tagged SNPs within the *S-M-R* region. Red ticks are used for SNPs tagged by rs2038765 and blue for those tagged by rs12530016. The reading frames of *SUPT3H* and *RUNX2* can be seen under the tagSNP tracks. Relative exon and intron positions of alternative transcripts, as well as transcriptional orientations are indicated.

Fig 1

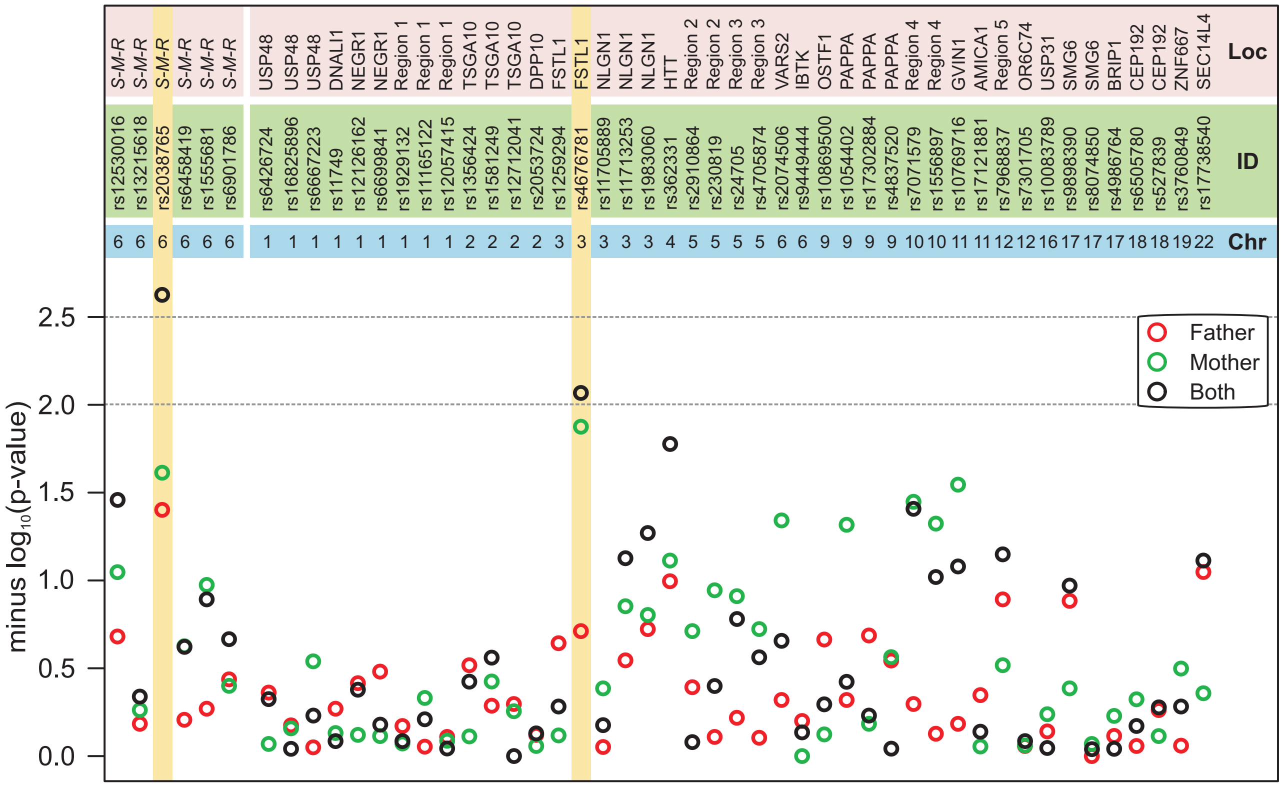

Loc
S-M-R S-M-R S-M-R S-M-R S-M-R S-M-R USP48 USP48 USP48 DNALI1 NEGR1 NEGR1 Region 1 Region 1 Region 1 TSGA10 TSGA10 TSGA10 DPP10 FSTL1 FSTL1 NLGN1 NLGN1 NLGN1 HTT Region 2 Region 2 Region 3 Region 3 VARS2 IBTK OSTF1 PAPPA PAPPA PAPPA Region 4 Region 4 GVIN1 AMICA1 Region 5 OR6C74 USP31 SMG6 SMG6 BRIP1 CEP192 CEP192 ZNF667 SEC14L4
ID
rs12530016 rs13215618 rs2038765 rs6458419 rs1555681 rs6901786 rs6426724 rs16825896 rs6667223 rs11749 rs12126162 rs6699841 rs1929132 rs11165122 rs12057415 rs1356424 rs1581249 rs12712041 rs2053724 rs1259294 rs4676781 rs11705889 rs11713253 rs1983060 rs362331 rs2910864 rs230819 rs24705 rs4705874 rs2074506 rs9449444 rs10869500 rs1054402 rs17302884 rs4837520 rs7071579 rs1556897 rs10769716 rs17121881 rs7968837 rs7301705 rs10083789 rs9898390 rs8074850 rs4986764 rs6505780 rs527839 rs3760849 rs17738540
Chr
6 6 6 6 6 6 1 1 1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 3 4 5 5 5 5 6 6 9 9 9 9 10 10 11 11 12 12 16 17 17 17 18 18 19 22
minus log10(p-value)
2.5
2.0
1.5
1.0
0.5
0.0
Father
Mother
Both

Fig 2

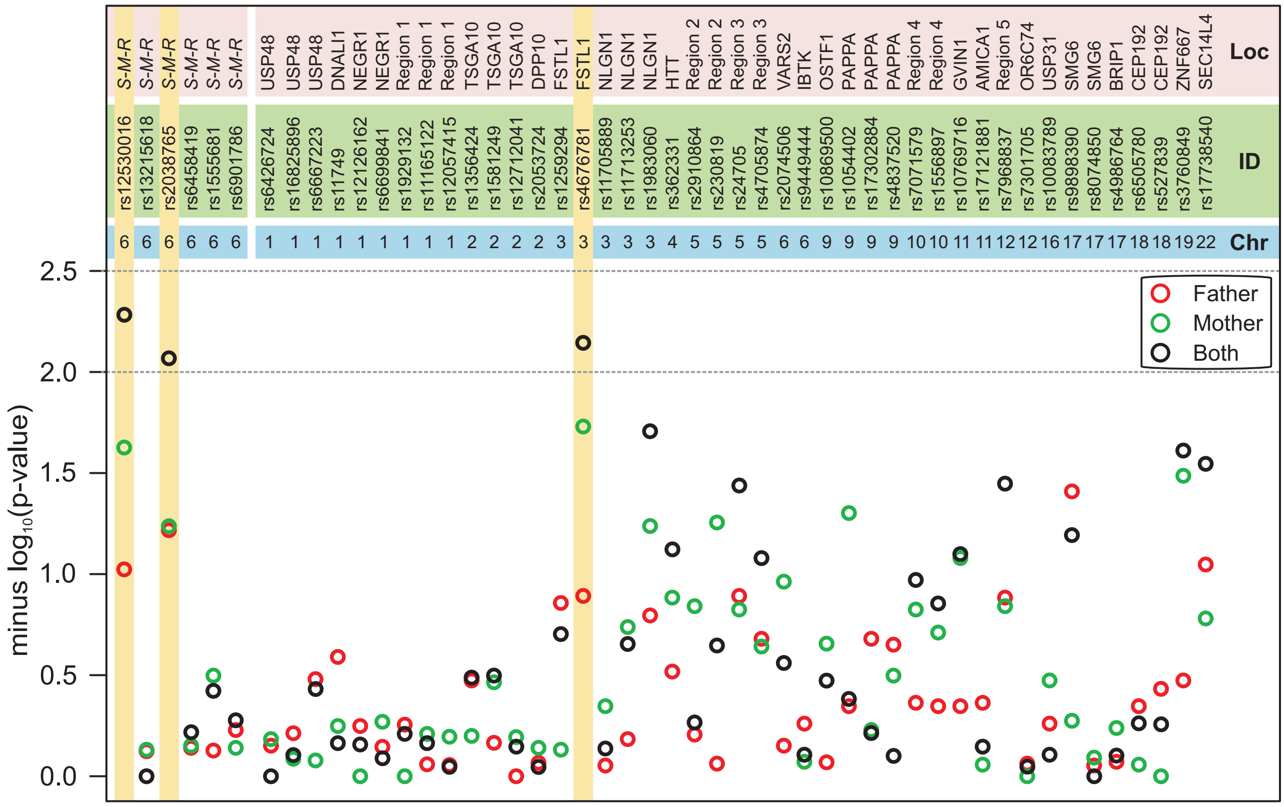
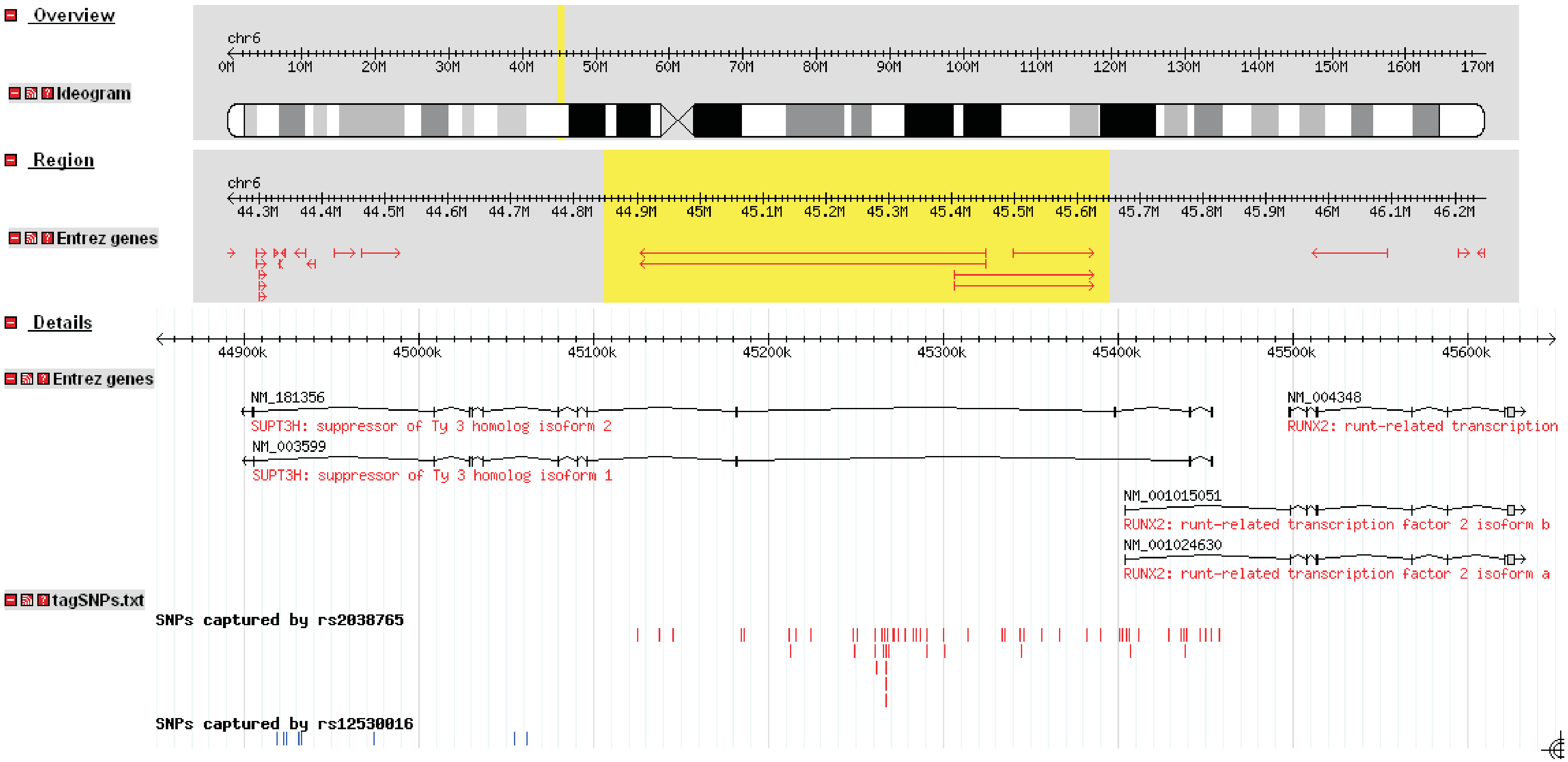

Loc
S-M-R S-M-R S-M-R S-M-R S-M-R S-M-R USP48 USP48 USP48 DNALI1 NEGR1 NEGR1 Region 1 Region 1 Region 1 TSGA10 TSGA10 TSGA10 DPP10 FSTL1 FSTL1 NLGN1 NLGN1 NLGN1 HTT Region 2 Region 2 Region 3 Region 3 VARS2 IBTK OSTF1 PAPPA PAPPA PAPPA Region 4 Region 4 GVIN1 AMICA1 Region 5 OR6C74 USP31 SMG6 SMG6 BRIP1 CEP192 CEP192 ZNF667 SEC14L4
ID
rs12530016 rs13215618 rs2038765 rs6458419 rs1555681 rs6901786 rs6426724 rs16825896 rs6667223 rs11749 rs12126162 rs6699841 rs1929132 rs11165122 rs12057415 rs1356424 rs1581249 rs12712041 rs2053724 rs1259294 rs4676781 rs11705889 rs11713253 rs1983060 rs362331 rs2910864 rs230819 rs24705 rs4705874 rs2074506 rs9449444 rs10869500 rs1054402 rs17302884 rs4837520 rs7071579 rs1556897 rs10769716 rs17121881 rs7968837 rs7301705 rs10083789 rs9898390 rs8074850 rs4986764 rs6505780 rs527839 rs3760849 rs17738540
Chr
6 6 6 6 6 6 1 1 1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 3 4 5 5 5 5 6 6 9 9 9 9 10 10 11 11 12 12 16 17 17 17 18 18 19 22
minus log10(p-value)
2.5
2.0
1.5
1.0
0.5
0.0
Father
Mother
Both

Fig 3


Overview
chr6
0M 10M 20M 30M 40M 50M 60M 70M 80M 90M 100M 110M 120M 130M 140M 150M 160M 170M
Ideogram
Region
chr6
44.3M 44.4M 44.5M 44.6M 44.7M 44.8M 44.9M 45M 45.1M 45.2M 45.3M 45.4M 45.5M 45.6M 45.7M 45.8M 45.9M 46M 46.1M 46.2M
Entrez genes
Details
44900k 45000k 45100k 45200k 45300k 45400k 45500k 45600k
Entrez genes
NM_181356
SUPT3H: suppressor of Ty 3 homolog isoform 2
NM_003599
SUPT3H: suppressor of Ty 3 homolog isoform 1
NM_004348
RUNX2: runt-related transcription
NM_001015051
RUNX2: runt-related transcription factor 2 isoform b
NM_001024630
RUNX2: runt-related transcription factor 2 isoform a
tagSNPs.txt
SNPs captured by rs2038765
SNPs captured by rs12530016

6. Linkage Disequilibrium between Microsatellites and MHC loci

6.1. Summary

The HLA complex of a large number of individuals is routinely genotyped in forensic institutes and for tissue transplantation purposes. In these cases, a detailed knowledge of the LD patterns of the MHC in different human populations has not only a scientific or medical, yet also economic relevance. Because microsatellite markers are relatively easy (and also cheaper) to genotype than entire HLA alleles, a detailed knowledge of the LD profile between HLA alleles and microsatellites of the HLA complex seems very opportune. The study shown in the following article aimed at describing the LD patterns between three generally genotyped HLA loci (HLA-B, HLA-DQB1 and HLA-DRB1) and four known microsatellite markers harboured in the HLA complex. The results point at high confidence LD values for some allele combinations. While microsatellite genotyping cannot substitute direct MHC genotyping, it can clearly be used as an alternative to test repetitions in the case of incomplete genotyping or as a guide for ambiguity solving.

6.2. Publication

Sens-Abuázar C, Santos PS, Bicalho MG, Petzl-Erler ML, Sperandio-Roxo V (2009): MHC microsatellites in a Southern Brazilian population. International Journal of Immunogenetics 36:269-274.

DOI:10.1111/j.1744-313X.2009.00864.x

Pages 114-119

7. Final Discussion and Conclusions

Apart from the results presented in chapter 4, all other analyses are part of manuscripts which have either been published or are currently under review. For this reason, the results of the analyses provided in chapters 2, 3, 5 and 6 have already been discussed within the manuscripts and will not be repeated here. In this chapter, I aim to outline the implications of the results from this thesis, and discuss those from chapter 4.

7.1. MHC-linked Olfactory Receptor Genes

All three studies addressing MHC-linked OR genes in this thesis (Chapters 2, 3 and 4) take advantage of online resources in order to complete the analyzed data and serve their argumentation. While the comparison of MHC-linked OR genes among different species (Chapter 4) was performed exclusively based on publicly available data, the two other studies focusing on these genes in humans were based on a combination of the use of electronic data (HapMap genotypings for chapter 2 and MHCHP resequencing for chapter 3) with laboratory-generated data (genotyping of Hungarian cohort in chapter 2, and sequencing of ten cell lines in chapter 3).

The main result of the study shown in chapter 2 was the linkage of a polymorphism in an OR gene to smoking, which is, to my knowledge, the first time a human behaviour was associated to an OR gene. One amino acid exchange on the *OR12D3* gene – found to be in strong LD with the ancient HLA haplotype HLA A1-B8-DR3 – was found to correlate with smoking habits in Hungarian women. This finding has possible implications for the early identification of individuals with increased risk to become smokers. However, the ethical aspects of such testing, and the risk that people carrying the relevant allele become subject of genetic discrimination (in cases of health or life insurances for instance), are apparently still unresolved [Mould et al., 2003; Hall et al., 2008; Aymé et al., 2008]. In addition, this study also provided a panel of 110 tagSNPs covering both MHC-linked OR gene clusters (Chapter 2, supplementary tables 1 and 2). These markers capture almost all of the variation present in the region for the HapMap CEU population with high confidence ($r^2 \geq 0.8$), and can be utilized as a resource in future studies. As shown before [Mueller et al., 2005; Xing et al., 2008], tagSNPs are expected to be transferrable among cohorts of related ancestry.

The analyses from chapters 2 and 3 intersect each other at the LD assessment of the region encompassing both MHC-linked OR gene clusters. While the first study (Chapter 2) builds on the high levels of LD in order to define tagSNPs and polymorphisms typical for different HLA haplotypes, the results of the second study (Chapter 3) indicate the presence of recombination between OR clusters and the MHC, as well as limited LD among functional OR gene polymorphisms. However, these results are not contradictory, since the HapMap cohort (ninety unrelated persons were included in the first study) is widely considered a population sample, with the corresponding high frequencies of common HLA haplotypes (eleven times A1-B8-DR3 and 5 times A3-B7-DR15, for instance), while this is radically different for the 18 cell lines assessed in the second study. Both, the ten Berlin cell lines as well as the eight MHCHP cell lines, were specifically chosen based on their MHC diversity [Ziegler et al., 1985; Volz et al., 1992; Ehlers et al., 2000; Horton et al., 2008].

Another intersection among analyses from chapters 2 and 3 regards the *OR12D3* polymorphism found to be in high LD with the haplotype HLA-A1-B8-DR3 and associated with smoking habits in the first study. Assessing the cell lines, the mentioned polymorphism was shown, for the first time, to be also linked to a different HLA haplotype: A32-B44-DR4, from the Caucasian cell line SSTO [Horton et al., 2008], indicating one probable recombination event between *OR12D3* and the HLA complex. Also here, this finding has no implications for population genetics, as the sample in which it was found is especially prone to unexpected combinations.

The work presented in chapter 4 was dedicated to the description and phylogenetic analysis of OR gene families linked to the MHC in all vertebrate species that have been sufficiently sequenced to date, and from which genomic data is available through internet-based databanks. The comparison of the structure and sequences from the MHC-linked OR gene clusters among fourteen vertebrates described here is, so far, the most comprehensive study regarding this region among different organisms.

The phylogenetic tree (Chapter 4, Fig. 4.1) suggests a common ancestry of all these loci among terrestrial vertebrates, as most gene families (branches with different colours in the tree) contain at least one gene from almost each species. This is different only for zebra fish, for which the products of these OR genes present sequences that are more similar to each other than to any other MHC-linked OR sequence from any other vertebrates, and therefore cluster in the phylogenetic analysis (Chapter 4, Fig. 4.1, black branches), suggesting an independent origin. If this is correct, at least two independent evolutionary events must have led to the linkage between OR genes and MHC genes, a fact that reinforces the biological relevance of the linkage between MHC and OR genes that is seen in all vertebrates so far, except dog and cat.

The apparently obligatory linkage of certain MHC framework genes with MHC-linked OR genes is a remarkable finding of this investigation. Even in the case of carnivores, the genes *GABBR1* and *MOG* "kept" linked to the OR genes after the presumed chromosomal segmentation that split these from the MHC [Yuhki et al., 2007]. In this context, the fact that the split between MHC and OR clusters coincides with a strong reduction in the number of functional OR loci, may be seen as suggestive of a functional relationship that may have lost importance after the split. Moreover, possible effects of domestication and high levels of inbreeding of these two species on the genomic "architecture" of the MHC and surrounding regions still need to be studied. This aspect is also relevant for the other domesticated animals that were assessed here: cow, pig and horse.

The framework gene *GABBR1* encodes the subunit 1 of the gamma-aminobutyric acid (GABA) B receptor (GABA$_B$), a seven transmembrane receptor of GABA, which is one of the main inhibitory neurotransmitters in the vertebrate central nervous system [Grifa et al., 1998; Goei et al., 1998]. This tissue is a probable background of interaction between the products of OR genes and *GABBR1*: an influence of GABA$_B$ in the behaviour of olfactory receptor neurons of zebra fish [Tabor et al., 2008], turtle [Wachowiak et al., 1999], frog [Duchamp-Viret et al., 2000], rat [Panzanelli et al., 2004], and mouse [Vucinić et al., 2006] has been reported.

Apart from being an inhibitory neurotransmitter, GABA (also through its receptor GABA$_B$) plays a central role in the sperm acrosome reaction [Hu et al., 2002; Burrello et al., 2004]. The known sperm specific expression of OR genes (discussed in chapter 3) indicates an additional probable background for interaction between OR genes and *GABBR1* products.

It must remain an open question whether the interaction of gene products, common regulation or epistatic effects of either OR gene or framework genes can explain the apparently obligatory genomic linkage observed between these loci. Interestingly, a search within the Drosophila genome browser (http://flybase.org/) revealed that even in this arthropod, which diverged from vertebrates approximately 990 million years ago [Blair Hedges & Kumar, 2003], the ortholog of human *GABBR1* (named "GABA-B-R1") is linked to at least four OR genes (Or33a, Or33b, Or33c and Or35a) on Drosophila chromosome 2L. Since OR genes from Drosophila and vertebrates are evolutionarily unrelated [Bargmann, 2006; Nozawa and Nei, 2007], apart from the fact that both are GPCR, this linkage is additionally suggestive of an important functional relationship between OR and the GABA receptor genes, which, for some as yet unknown reason, requires these genes to be physically close to each other. Considering that there are only ~ 60 OR genes in the Drosophila genome [Robertson et al., 2003], it is rather improbable that the linkage discussed here is due to chance alone.

The phylogenetic tree of the MHC-linked OR genes (Chapter 4, Fig. 4.1) indicates a common ancestry of all these loci, at least in terrestrial vertebrates, while the genomic structure depicted on the gene maps (Chapter 4, Fig. 4.2) reflects the close relationship among the different species within each assessed taxon. The case of the horse is, in this context, of special interest: Although they lack members of most mammal orders, the gene maps presented here corroborate the common understanding that horse, pig and cow (ungulates) are closely related, building a group that is apart from the one including cat and dog (Carnivora). In fact, the phylogenetic relationships between the orders Cetartiodactila (even-toed ungulates, including pigs, cows and whales), Perissodactila (odd-toed ungulates, including horses, tapirs and rhinoceroses) and Carnivora (including cats, dogs and hyenas) have long been subject of debate with contradictory results [Novacek, 1992; Graur et al., 1997; Cao et al., 2000; Murphy et al., 2001; Nishihara et al., 2006; Kitazoe et al., 2007], and have apparently not yet been resolved, although the monophyly of each of these three taxa is undisputed [Murphy et al., 2004].

According to one view [Murphy et al., 2001; Nishihara et al., 2006], Perissodactila and Carnivora are closely related to each other, building a sister group of Cetartiodactila. A phylogenetic tree based on these results is given in Fig. 7.1, panel a. In line with this notion, one recent study based on sequence analysis of genomic retroposon insertions suggests the name “Pegasoferae” for designating the alleged monophyletic clade that incorporates odd-toed ungulates, Carnivora and Chiroptera (bats) [Nishihara et al., 2006].

The competing view [Kitazoe et al., 2007] coincides with the classical “family trees” of mammals, which are based on the morphology of fossils and extant species, and suggests that Perissodactila and Cetartiodactila belong to one monophyletic taxon (here called Ungulata), a clade apart from Carnivora [Archibald, 1998; Shoshani and McKenna, 1998]. Fig. 7.1, panel b presents a phylogenetic tree based on these results. A recent molecular analysis reinforcing this notion [Kitazoe et al., 2007] considered abrupt changes of the rate of molecular evolution that are likely to have taken place at the time of the divergence of these three orders, around the Cretaceous-Paleogene extinction event [Rohde and Muller, 2005; Kitazoe et al., 2007]. This geological "moment", also known as the K-T (Cretaceous to Tertiary) boundary, is associated with abrupt climate changes and the mass extinction of plants and animals, with the sudden extinction of land dinosaurs as its best known feature [Labandeira et al., 2002; Rohde and Muller, 2005; Krug et al., 2009; Keller et al., 2009]. The Cretaceous-Paleogene extinction event is also associated with an accelerated rate of molecular evolution that had, so far, not been taken into consideration by other phylogenetic reconstructions based on molecular data [Kitazoe et al., 2007].

The recent studies supporting the two competing models for the phylogenetic history among the clades Cetartiodactila, Perissodactila, and Carnivora are based on diverging methodologies, but have the fact in common that they are built on sequence comparisons (nuclear DNA, mitochondrial DNA, amino acid, or retroposons). In contrast, the analysis presented here (Chapter 4) uses an alternative approach in order to provide strong support for the second view (horse, pig and cow belonging to a group apart of cat and dog): instead of considering many individual sequences of the different organisms, I compare the "genomic anatomy" of relatively long, syntenic genomic regions, considering the presence or absence of gene classes (OR, histones, MHC, framework), their positions relative to each other, as well as their transcriptional orientations. The presence of three orthologous histone clusters interspersed by three orthologous OR clusters in which the families, positions and transcriptional orientations are widely conserved, provides evidence for the stronger similarity of horses to Cetartiodactila than to Carnivora (Fig. 4.2, panels c and d, and Fig. 7.1, panel c).

Phylogenetic inferences are known to be more prone to bias related to convergent evolution when they are based on classical anatomic analyses, as compared to studies based on nucleic acid or amino acid sequences comparisons [Shoshani and McKenna, 1998]. On the other side, sequences are also susceptible to convergent evolution, as recently shown specifically for odorant receptor amino acid sequences [Hayden et al., 2010]. In that investigation, Hayden and collaborators analyzed around 50 thousand individual OR gene sequences in the genomes of fifty aquatic, semi-aquatic, terrestrial and flying mammals, and observed that sequence similarities between OR genes tended to depend more on the habitat and other ecological traits of the assessed animals than on the phylogenetic relationships between the species.

Since there is currently no genomic assembly available for most of the vertebrates assessed in that study, no assumption can be made about the role of MHC-linked OR genes. It is additionally interesting to observe that Hayden and co-workers [Hayden et al., 2010] assumes the notion that groups Perissodactila, Carnivora and Chiroptera under one single clade, apart from Cetartiodactila, as proposed by Nishihara and colleagues [Nishihara et al., 2006], to be part of the "consensus" phylogenetic tree of mammals.

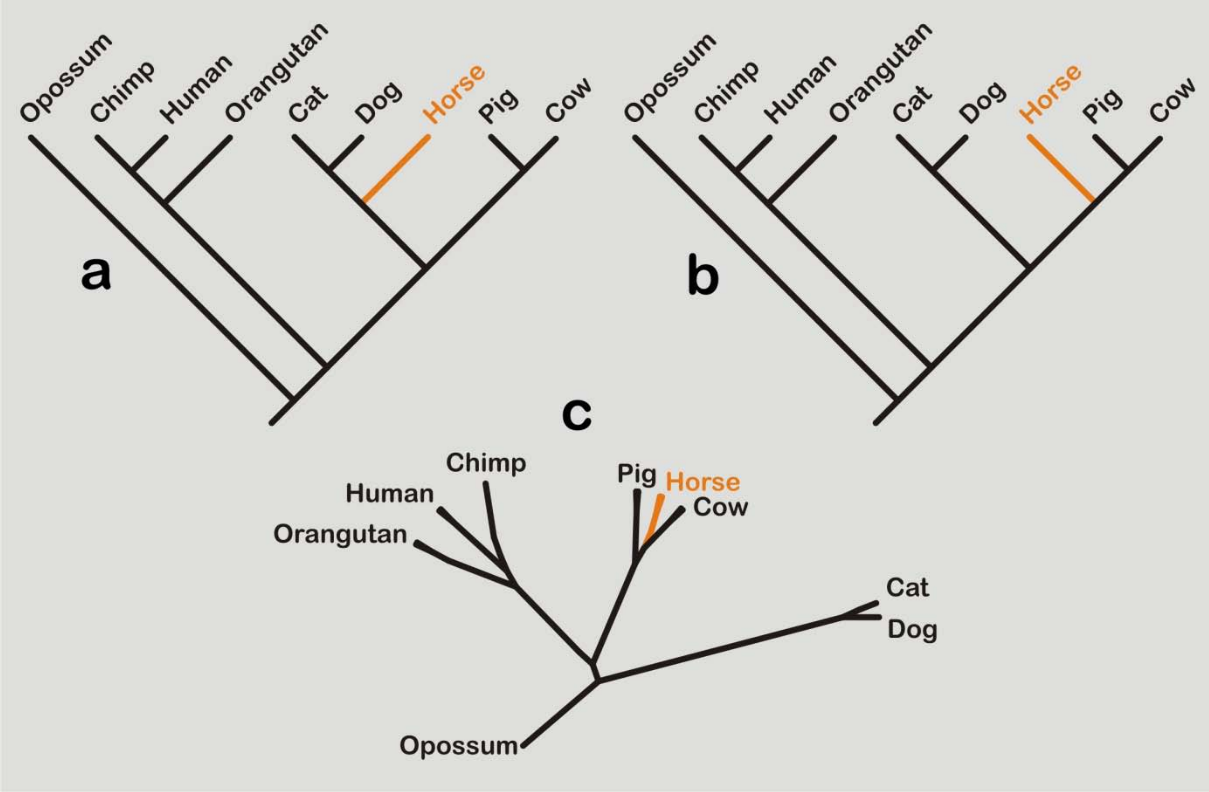


Fig. 7.1: Three phylogenetic trees of nine organisms which had their MHC-linked OR genes assessed (see Chapter 4). The tree on panel a represents the phylogenetic history of these clades as found by Murphy, Nishihara and colleagues [Murphy et al., 2001; Nishihara et al., 2006]. The tree on panel b represents the phylogenetic history of these clades as found by Kitazoe and colleagues [Kitazoe et al., 2007]. The tree on panel c represents the phylogenetic history of these clades as it can be inferred from the results regarding the genomic structure of organisms from Chapter 4. Since no assumptions of relationships between the four groups could be made, an unrooted tree was generated. The horse branches are highlighted, as they are subject to debate (see text).

In contrast both to classical anatomy and to sequence analysis, the approach described here represents a view into the genome from an intermediate distance. The size of the genomic region analyzed here, as well as the number of assessed species are certainly small, in order to provide steady evidence for resolving the debate around the evolutionary history of eutherians. For example, the series of chromosomal inversions as well as the expansion of the genes belonging to the OR2M family, which seem to have taken place specifically in the pig genome, make horse and cow seem more similar to each other than to pig (Fig. 4.2, panels c and d, and Fig. 7.1, panel c). This artificial similarity is probably related to the insufficient number of species and genomic regions assessed. It remains an interesting question whether the genomic architecture of segments as the one assessed here is also prone to convergent evolution. As for today, there is, to the best of my knowledge, no reason to believe so.

7.2. Linkage Disequilibrium and Transmission Distortion

In human genetics, LD is generally understood from a genealogical point of view: ancestral haplotypes have been broken and shuffled by means of recombination events throughout evolutionary history [Ardlie et al., 2002; Slatkin, 2008]. Because of population constraints such as bottlenecks and genomic drift, the linkage of some segments was maintained and, according to this view, also these will be shuffled through meiotic recombination with time.

Nevertheless, the assessment of TD described here (Chapter 5), can shed new light on the traditional understanding of LD creation and maintenance. Besides its possible biological implications that are discussed in the manuscripts, TD can be interpreted as one of the evolutionary forces shaping the LD landscape of the human genome, although it has apparently been widely ignored. If present in a given genomic region, TD will not only protect LD blocks from the “erosive” work of recombination by conserving LD blocks through the prevention of recombination events, but also by actively generating such LD blocks. If TD is an ethnically-specific phenomenon, as described in the two manuscripts of chapter 5, it should be expected to be an additional force responsible for the distinct LD profiles observed in different populations. Moreover, recombination hotspots are expected to be found flanking the LD block (which could be, in this case, called a TD block). The observed correspondence between LD and TD from the results of this thesis (see section 5.1, Fig. 1), as well as from other reports (see section 5.1, supplementary Fig. S5) support this point of view. This proposition remains a hypothesis which needs extensive further testing in order to be proven, but this idea has, to the best of my knowledge, not been suggested before.

Based on HapMap data, several studies found a high correspondence between the LD profiles among ethnically related populations [Mueller et al., 2005; Xing et al., 2008; Hu et al., 2008].

A detailed knowledge of LD data has practical consequences that include haplotype tagging (shown in chapter 2), lowering the number of markers in order to avoid statistical burdens linked to multiple testing (chapter 5), and increasing the efficiency of genotyping (chapter 6). The work with microsatellites presented in chapter 6 describes an approach that takes advantage of the LD present between relevant HLA alleles and neighbouring microsatellite loci within the MHC in order optimize the genotyping of HLA alleles (for example through resolving ambiguities and avoiding repetition of experiments). According to studies suggesting the transferability of tagSNPs among related populations [Mueller et al., 2005; Xing et al., 2008; Hu et al., 2008], the results shown in this article (Chapter 6) should also be valid for other populations of European ancestry.

Finally, taking the existence of several new internet-based resources into account, and the increased relevance that these databases and tools for data mining have reached for life sciences research within the last years [Buckingham, 2004; Krallinger et al., 2008; Hubbard et al., 2009; Sayers et al., 2010], the implementation of a web-based visual databank for integrating results from different studies relevant to TD on the human genome seems very appropriate. The works presented in the chapter 5 are initial steps towards this goal. In fact, one recently published study [Deng et al., 2009] builds on the published suggestion of the first of the two LD investigations presented here (Section 5.1), describing a first large scale map of transmission distortion (for all human autosomes), based on HapMap data. The article by Deng and colleagues [Deng et al., 2009] cites the paper presented in section 5.1 several times, and reports the use of a similar strategy to detect SNPs with evidence of TD.

The studies presented in the chapter 5 had two of the HapMap populations as a reference – CEU and YRI – because these consisted, in contrast to the others, of family trios and were therefore suitable also for investigating TD (see also sections 1.4.2 and 1.5). As briefly mentioned in section 1.4.2, the international HapMap project has recently reached a historical milestone with the release of genotyping data from the Phase III samples in November 2008. With this release (Public Release #26), the HapMap increased data richness through the inclusion of seven new populations to the original four genotyping panels, totalling eleven populations that can now be compared to each other in the context of genotypic and haplotypic variation. Three of the new populations (ASW, MEX and MKK) are structured in the form of family trios, as originally only the CEU and YRI panels. This allows scientists to perform a very reliable phasing of the genotypings, as well as to assess allelic and haplotypic transmission within families in a way that has never been possible before. Moreover, a series of new quality control (QC) checks for both samples and markers were introduced.

Nevertheless, when the investigation reported in section 5.1 was repeated using data from the latest HapMap release, we observed that TD was now completely absent in Chr6p, and generally lower in the rest of the genome. At first glance, this would suggest that the section 5.1 analyses are flawed, disappearing with the new HapMap data. However, three important facts suggest an alternative explanation:

1. Most SNPs exhibiting TD in the article (section 5.1) were not included in the Phase III genotyping panels. While around 50% of the SNPs from phase II "survived" QC and were kept in the Phase III release (taking the whole chromosome 6 as an example), this was the case for only ~25% of the SNPs from genomic areas with evidence of TD. As discussed in both TD studies (sections 5.1 and 5.2), TD seems to be an ethnicity-related property. Therefore, the stringency of QC may have led, unintentionally, to the systematic exclusion of markers that show skewed allele segregation rates, as they are less likely to fulfil Hardy-Weinberg expectancy thresholds at least in one of the eleven populations analyzed. As currently given in the HapMap page from the Sanger Centre (http://www.sanger.ac.uk/humgen/hapmap3/), SNPs had to pass QC apparently in all populations in order to be included in the Phase III. Another recent report [Fardo et al., 2009] discusses and reinforces the fact that the exclusion of SNPs not in Hardy-Weinberg equilibrium might be counterproductive (or unnecessary) in the context of disease association studies.

2. TD for markers residing in the *SUPT3H* area could undoubtedly be confirmed, in an independent population, through the results from the second TD study (section 5.2). Although all twelve *SUPT3H* markers assessed in that study had been taken from the HapMap Phase II SNP set, only four of these were kept and genotyped in the Phase III release.

3. The possibility that the reason for exclusion of markers shown to be under TD could have been duplicated identification numbers, mapping problems or other inconsistencies was checked and found not to be the case for any of the investigated markers, at least on Chr6p. The possibility that the fathers responsible for TD observed using the Phase II data were those excluded due to QC proceedings was also investigated, and again this

was not the case: TD can still be observed around *SUPT3H* among the CEU fathers belonging to family trios kept in the Phase III release.

It seems therefore that the criteria for SNP inclusion in the latest HapMap Phase III data set were too stringent, as SNPs with deviations from the expected transmission ratio like the one we reported (section 5.1) were preferentially excluded from the data set. As a consequence, it becomes very difficult to compare data using the Phase II release with those from Phase III in the context of allelic segregation distortion, since investigations focusing on TD with the Phase III data are expected to have their results artificially distorted. The current status of the HapMap website (http://hapmap.ncbi.nlm.nih.gov/) lists HapMap3 as the newest update of the HapMap resources, and unadvised users have no reason to prefer the earlier release, which is still available.

7.3. Concluding Remarks

The results from the analyses focusing on the MHC-linked OR genes carried out as part of this work contribute to the understanding of human variation regarding this genomic region, and have implications for the conservation of genomic structure among vertebrates. The region is characterized by high LD with the MHC, a feature used in the other studies as the basis for the selection of tagSNPs, genotyping optimization and the first TD analysis.

While most results obtained and discussed represent the descriptive answers to a the series of questions raised throughout this work, two new ideas (both originated from unexpected “side effects” of the analyses) that were suggested here remained without further testing, and will be subject of future work: the use of genomic anatomy for phylogenetic inferences, and the role of TD as a force shaping the LD landscape of the human genome.

8. Bibliography

1. Aloni R, Olender T, Lancet D. Ancient genomic architecture for mammalian olfactory receptor clusters. Genome Biol. 2006, 7(10):R88.
2. Alper CA, Awdeh ZL, Yunis EJ. Complotypes and extended haplotypes in laboratory medicine. Complement Inflamm. 1989, 6(1):8-18.
3. Alper CA, Awdeh Z, Yunis EJ. Conserved, extended MHC haplotypes. Exp Clin Immunogenet. 1992, 9(2):58-71.
4. Alper CA, Larsen CE, Dubey DP, Awdeh ZL, Fici DA, Yunis EJ. The haplotype structure of the human major histocompatibility complex. Hum Immunol. 2006, 67(1-2):73-84.
5. Amadou C, Younger RM, Sims S, Matthews LH, Rogers J, Kumanovics A, Ziegler A, Beck S, Lindahl KF. Co-duplication of olfactory receptor and MHC class I genes in the mouse major histocompatibility complex. Hum Mol Genet. 2003, 12(22):3025-40.
6. Aparicio JM, Ortego J, Calabuig G, Cordero PJ. Evidence of subtle departures from Mendelian segregation in a wild lesser kestrel (*Falco naumanni*) population. Heredity. 2010, article in press.
7. Archibald JD. Fossil Evidence for a Late Cretaceous Origin of "Hoofed" Mammals. Science. 1996, 272(5265):1150-3.
8. Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nat Rev Genet. 2002, 3(4):299-309.
9. Axelsson E, Albrechtsen A, van AP, Li L, Megens HJ, Vereijken AL, Crooijmans RP, Groenen MA, Ellegren H, Willerslev E, Nielsen R. Segregation distortion in chicken and the evolutionary consequences of female meiotic drive in birds. Heredity. 2010, article in press.
10. Aymé S, Matthijs G, Anastasiadou V, Fatmahan A, Braga S, Burn J, Cassiman JJ, Cornel M, Coviello D, Evers-Kiebooms G, Gorry P, Hodgson S, Kääriäinen H, Kosztolányi G, Kristoffersson U, Macek M Jr, Patch C, Schmidtke J, Sequeiros J, Stoppa-Lyonnet D, Tranebjaerg L, Heyningen V, van Ommen GJ. Patenting and licensing in genetic testing: recommendations of the European Society of Human Genetics. Eur J Hum Genet. 2008, 16 Suppl 1:S10-9.
11. Bargmann CI. Comparative chemosensation from receptors to ecology. Nature. 2006, 444(7117):295-301.
12. Bennett D. The T-locus of the mouse. Cell. 1975, 6:441-54.

13. Blair Hedges S, Kumar S. Genomic clocks and evolutionary timescales. Trends Genet. 2003, 19(4):200-6.

14. Blomhoff A, Olsson M, Johansson S, Akselsen HE, Pociot F, Nerup J, Kockum I, Cambon-Thomsen A, Thorsby E, Undlien DE, Lie BA. Linkage disequilibrium and haplotype blocks in the MHC vary in an HLA haplotype specific manner assessed mainly by DRB1*03 and DRB1*04 haplotypes. Genes Immun. 2006, 7(2):130-40.

15. Bosch E, Laayouni H, Morcillo-Suarez C, Casals F, Moreno-Estrada A, Ferrer-Admetlla A, Gardner M, Rosa A, Navarro A, Comas D, Graffelman J, Calafell F, Bertranpetit J. Decay of linkage disequilibrium within genes across HGDP-CEPH human samples: most population isolates do not show increased LD. BMC Genomics. 2009, 10:338.

16. Brown RE, Roser B, Singh PB. Class I and class II regions of the major histocompatability complex both contribute to individual odors in congenic inbred strains of rats. Behav Genet. 1989, 19:659-674.

17. Buck L, Axel R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell. 1991, 65(1):175-87.

18. Buckingham S. Bioinformatics: data's future shock. Nature. 2004, 428(6984):774-7.

19. Burrello N, Vicari E, D'Amico L, Satta A, D'Agata R, Calogero AE. Human follicular fluid stimulates the sperm acrosome reaction by interacting with the gamma-aminobutyric acid receptors. Fertil Steril. 2004, 82 Suppl 3:1086-90.

20. Cao Y, Fujiwara M, Nikaido M, Okada N, Hasegawa M. Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data. Gene. 2000, 259(1-2):149-58.

21. Carroll LS, Meagher S, Morrison L, Penn DJ, Potts WK. Fitness effects of a selfish gene (the Mus t complex) are revealed in an ecological context. Evolution. 2004, 58(6):1318-28.

22. Chakraborty R, Stivers DN, Deka R, Yu LM, Shriver MD, Ferrell RE. Segregation distortion of the CTG repeats at the myotonic dystrophy locus. Am J Hum Genet. 1996, 59(1):109-18.

23. Chesley P, Dunn LC. The Inheritance of Taillessness (Anury) in the House Mouse. Genetics. 1936, 21(5):525-36.

24. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. Nat Genet. 2001, 29(2):229-32.

25. Dausset J. The major histocompatibility complex in man. Science. 1981, 213:1469-74.

26. de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, Morrison J, Richardson A, Walsh EC, Gao X, Galver L, Hart J, Hafler DA, Pericak-Vance M, Todd JA, Daly MJ, Trowsdale J, Wijmenga C, Vyse TJ, Beck S, Murray SS, Carrington M, Gregory S, Deloukas P, Rioux JD. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet. 2006, 38(10):1166-72.

27. Dean NL, Loredo-Osti JC, Fujiwara TM, Morgan K, Tan SL, Naumova AK, Ao A. Transmission ratio distortion in the myotonic dystrophy locus in human preimplantation embryos. Eur J Hum Genet. 2006, 14(3):299-306.

28. Deng L, Zhang D, Richards E, Tang X, Fang J, Long F, Wang Y. Constructing an initial map of transmission distortion based on high density HapMap SNPs across the human autosomes. J Genet Genomics. 2009, 36(12):703-9.

29. Duchamp-Viret P, Delaleu JC, Duchamp A. GABA(B)-mediated action in the frog olfactory bulb makes odor responses more salient. Neuroscience. 2000, 97(4):771-7.

30. Eaves IA, Bennett ST, Forster P, Ferber KM, Ehrmann D, Wilson AJ, Bhattacharyya S, Ziegler AG, Brinkmann B, Todd JA. Transmission ratio distortion at the INS-IGF2 VNTR. Nat Genet. 1999, 22(4):324-5.

31. Ehlers A, Beck S, Forbes SA, Trowsdale J, Volz A, Younger R, Ziegler A. MHC-linked olfactory receptor loci exhibit polymorphism and contribute to extended HLA/OR-haplotypes. Genome Res. 2000, 10(12):1968-78.

32. Eizaguirre C, Yeates SE, Lenz TL, Kalbe M, Milinski M. MHC-based mate choice combines good genes and maintenance of MHC polymorphism. Mol Ecol. 2009, 18(15):3316-29.

33. Eklund AC, Belchak MM, Lapidos K, Raha-Chowdhury R, Ober C. Polymorphisms in the HLA-linked olfactory receptor genes in the Hutterites. Hum Immunol. 2000, 61(7):711-7.

34. Evans K, Fryer A, Inglehearn C, Duvall-Young J, Whittaker JL, Gregory CY, Butler R, Ebenezer N, Hunt DM, Bhattacharya S. Genetic linkage of cone-rod retinal dystrophy to chromosome 19q and evidence for segregation distortion. Nat Genet. 1994, 6(2):210-3.

35. Fardo DW, Becker KD, Bertram L, Tanzi RE, Lange C: Recovering unused information in genome-wide association studies: the benefit of analyzing SNPs out of Hardy-Weinberg equilibrium. Eur J Hum Genet. 2009, 17:1676-82.

36. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, Domingo R Jr, Ellis MC, Fullan A, Hinton LM, Jones NL, Kimmel BE, Kronmal GS,

Lauer P, Lee VK, Loeb DB, Mapa FA, McClelland E, Meyer NC, Mintier GA, Moeller N, Moore T, Morikang E, Prass CE, Quintana L, Starnes SM, Schatzman RC, Brunke KJ, Drayna DT, Risch NJ, Bacon BR, Wolff RK. A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. 1996, Nat Genet. 1996, 13(4):399-408.

37. Fernando MM, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, Vyse TJ, Rioux JD. Defining the role of the MHC in autoimmunity: a review and pooled analysis. PLoS Genet. 2008, 4(4):e1000024.

38. Fishman L, Aagaard J, Tuthill JC. Toward the evolutionary genomics of gametophytic divergence: patterns of transmission ratio distortion in monkeyflower (Mimulus) hybrids reveal a complex genetic basis for conspecific pollen precedence. Evolution. 2008, 62(12):2958-70.

39. Fukuda N, Yomogida K, Okabe M, Touhara K. Functional characterization of a mouse testicular olfactory receptor and its role in chemosensing and in regulation of sperm motility. J Cell Sci. 2004, 117(24):5835-45.

40. Füst G, Arason GJ, Kramer J, Szalai C, Duba J, Yang Y, Chung EK, Zhou B, Blanchong CA, Lokki ML, Bödvarsson S, Prohászka Z, Karádi I, Vatay A, Kovács M, Romics L, Thorgeirsson G, Yu CY. Genetic basis of tobacco smoking: strong association of a specific major histocompatibility complex haplotype on chromosome 6 with smoking behavior. Int Immunol. 2004, 16(10):1507-14.

41. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. Science. 2002, 296(5576):2225-9.

42. Ganetzky B. On the components of segregation distortion in Drosophila melanogaster. Genetics. 1977, 86(2 Pt. 1):321-55.

43. Girirajan S, Elsea SH. Distorted Mendelian transmission as a function of genetic background in Rai1-haploinsufficient mice. Eur J Med Genet. 2009, 52(4):224-8.

44. Glusman G, Yanai I, Rubin I, Lancet D. The complete human olfactory subgenome. Genome Res. 2001, 11:685-702.

45. Goei VL, Choi J, Ahn J, Bowlus CL, Raha-Chowdhury R, Gruen JR. Human gamma-aminobutyric acid B receptor gene: complementary DNA cloning, expression, chromosomal location, and genomic organization. Biol Psychiatry. 1998, 44(8):659-66.

46. Graur D, Gouy M, Duret L. Evolutionary affinities of the order Perissodactyla and the phylogenetic status of the superordinal taxa Ungulata and Altungulata. Mol Phylogenet Evol. 1997, 7(2):195-200.

47. Greenwood CM, Morgan K. The impact of transmission-ratio distortion on allele sharing in affected sibling pairs. Am J Hum Genet. 2000, 66(6):2001-4.

48. Grifa A, Totaro A, Rommens JM, Carella M, Roetto A, Borgato L, Zelante L, Gasparini P. GABA (gamma-amino-butyric acid) neurotransmission: identification and fine mapping of the human GABAB receptor gene. Biochem Biophys Res Commun. 1998, 250(2):240-5.

49. Gu CC, Yu K, Rao DC. Characterization of LD structures and the utility of HapMap in genetic association studies. Adv Genet. 2008, 60:407-35.

50. Hall WD, Gartner CE, Carter A. The genetics of nicotine addiction liability: ethical and social policy implications. Addiction. 2008, 103(3):350-9.

51. Hanchard N, Rockett K, Udalova I, Wilson J, Keating B, Koch O, Nijnik A, Diakite M, Herbert M, Kwiatkowski D. An investigation of transmission ratio distortion in the central region of the human MHC. Genes Immun. 2006, 7(1):51-8.

52. Haston CK, Humes DG, Lafleur M. X chromosome transmission ratio distortion in Cftr +/- intercross-derived mice. BMC Genet. 2007, 8:23.

53. Hayden S, Bekaert M, Crider TA, Mariani S, Murphy WJ, Teeling EC. Ecological adaptation determines functional mammalian olfactory subgenomes. Genome Res. 2010, 20(1):1-9.

54. Herrmann BG, Koschorz B, Wertz K, McLaughlin KJ, Kispert A. A protein kinase encoded by the t complex responder gene causes non-mendelian inheritance. Nature. 1999, 402(6758):141-6.

55. Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 1968, 38:226:31.

56. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC Jr, Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S. Gene map of the extended human MHC. Nat Rev Genet. 2004, 5(12):889-99.

57. Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, Forbes S, Gilbert JG, Halls K, Harrow JL, Hart E, Howe K, Jackson DK, Palmer S, Roberts AN, Sims S, Stewart CA, Traherne JA, Trevanion S, Wilming L, Rogers J, de Jong PJ, Elliott JF, Sawcer S, Todd JA, Trowsdale J, Beck S. Variation analysis and gene annotation of

eight MHC haplotypes: the MHC Haplotype Project. Immunogenetics. 2008, 60(1):1-18.

58. Hu C, Jia W, Zhang W, Wang C, Zhang R, Wang J, Ma X; International Type 2 Diabetes 1q Consortium, Xiang K. An evaluation of the performance of HapMap SNP data in a Shanghai Chinese population: analyses of allele frequency, linkage disequilibrium pattern and tagging SNPs transferability on chromosome 1q21-q25. BMC Genet. 2008, 9:19.

59. Hu JH, He XB, Wu Q, Yan YC, Koide SS. Biphasic effect of GABA on rat sperm acrosome reaction: involvement of GABA(A) and GABA(B) receptors. Arch Androl. 2002, 48(5):369-78.

60. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P (2009). Ensembl 2009. Nucleic Acids Res. 37(Database issue):D690-7.

61. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009, 37(Database issue):D211-5.

62. IMGT/HLA Database, 2010. Available from http://hla.alleles.org/

63. International HapMap Consortium. The International HapMap Project. Nature. 2003, 426: 789-796.

64. International HapMap Consortium. A haplotype map of the human genome. Nature. 2005, 437:1299-1320.

65. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007, 449(7164):851-61.

66. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2004, 409:860-921.

67. Jacob S, McClintock MK, Zelano B, Ober C. Paternally inherited HLA alleles are associated with women's choice of male odor. Nat Genet. 2002, 30(2):175-9.

68. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. Haplotype tagging for the identification of common disease genes. Nat Genet. 2001, 29(2):233-7.

69. Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris AP, Bentley D, Cardon LR, Deloukas P. The impact of SNP density on fine-scale patterns of linkage disequilibrium. Hum Mol Genet. 2004, 13(6):577-88.

70. Keller G, Sahni A, Bajpai S. Deccan volcanism, the KT mass extinction and dinosaurs. J Biosci. 2009, 34(5):709-28.

71. Kitazoe Y, Kishino H, Waddell PJ, Nakajima N, Okabayashi T, Watabe T, Okuhara Y. Robust time estimation reconciles views of the antiquity of placental mammals. PLoS ONE. 2007, 2:e384.

72. Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol. 2008, 9 Suppl 2:S8.

73. Kriz V, Mares J, Wentzel P, Funa NS, Calounova G, Zhang XQ, Forsberg-Nilsson K, Forsberg M, Welsh M. Shb null allele is inherited with a transmission ratio distortion and causes reduced viability in utero. Dev Dyn. 2007, 236(9):2485-92.

74. Krug AZ, Jablonski D, Valentine JW. Signature of the end-Cretaceous mass extinction in the modern biota. Science. 2009, 323(5915):767-71.

75. Labandeira CC, Johnson KR, Wilf P. Impact of the terminal Cretaceous event on plant-insect associations. Proc Natl Acad Sci U S A. 2002, 99(4):2061-6.

76. Lancet D, Pace U. The molecular basis of odor recognition. Trends Biochem Sci. 1987, 12:63.

77. Lane RP, Cutforth T, Young J, Athanasiou M, Friedman C, Rowen L, Evans G, Axel R, Hood L, Trask B. Genomic analysis of orthologous mouse and human olfactory receptor loci. Proc. Natl Acad. Sci. USA. 2001, 98:7390-7395.

78. Lemire M, Roslin NM, Laprise C, Hudson TJ, Morgan K. Transmission-ratio distortion and allele sharing in affected sib pairs: a new linkage statistic with reduced bias, with application to chromosome 6q25.3. Am J Hum Genet. 2004, 75(4):571-86.

79. Lewontin RC, Kojiana K. The evolutionary dynamics of complex polymorphisms. Evolution. 1960, 14:458-72.

80. Lewontin RC. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics. 1964, 49(1):49-67.

81. Lie BA, Thorsby E. Several genes in the extended human MHC contribute to predisposition to autoimmune diseases. Curr Opin Immunol. 2005, 17:526-31

82. Lyon MF, Zenthon J, Evans EP, Burtenshaw MD, Willison KR. Extent of the mouse t-complex and its inversions shown by in situ hybridization. Immunogenetics. 1988, 27:375-82.

83. Lyon MF. Transmission ratio distortion in mice. Annu Rev Genet. 2003, 37:393-408.

84. Lyttle TW. Cheaters sometimes prosper: distortion of Mendelian segregation by meiotic drive. Trends Genet. 1993, 9(6):205-10.

85. Mägi R, Pfeufer A, Nelis M, Montpetit A, Metspalu A, Remm M. Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation. BMC Genomics. 2007, 8:159.

86. Malnic B, Godfrey PA, Buck LB. The human olfactory receptor gene family. Proc Natl Acad Sci U S A. 2004, 101(8):2584-9.

87. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest. 2008, 118(5):1590-605.

88. Milinski M. The Major Histocompatibility Complex, Sexual Selection, and Mate Choice. Annual Review of Ecology, Evolution, and Systematics. 2006, 37(1): 159-186.

89. Miretti MM, Walsh EC, Ke X, Delgado M, Griffiths M, Hunt S, Morrison J, Whittaker P, Lander ES, Cardon LR, Bentley DR, Rioux JD, Beck S, Deloukas P. A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. Am J Hum Genet. 2005, 76(4):634-46.

90. Mombaerts P. Seven-transmembrane proteins as odorant and chemosensory receptors. Science. 1999, 286(5440):707-11.

91. Mould A. Implications of genetic testing: discrimination in life insurance and future directions. J Law Med. 2003, 10(4):470-87.

92. Mueller JC, Lõhmussaar E, Mägi R, Remm M, Bettecken T, Lichtner P, Biskup S, Illig T, Pfeufer A, Luedemann J, Schreiber S, Pramstaller P, Pichler I, Romeo G, Gaddi A, Testa A, Wichmann HE, Metspalu A, Meitinger T. Linkage disequilibrium patterns and tagSNP transferability among European populations. Am J Hum Genet. 2005, 76(3):387-98.

93. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, Springer MS. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science. 2001, 294(5550):2348-51.

94. Murphy WJ, Pevzner PA, O'Brien SJ. Mammalian phylogenomics comes of age. Trends Genet. 2004, 20(12):631-9.

95. Murphy AM, Meade KG, Hayes PA, Park SD, Evans AC, Lonergan P, MacHugh DE. Transmission ratio distortion at the growth hormone gene (GH1) in bovine preimplantation embryos: An in vitro culture-induced phenomenon? Mol Reprod Dev. 2008, 75(5):715-22.

96. Naumova AK, Leppert M, Barker DF, Morgan K, Sapienza C. Parental origin-dependent, male offspring-specific transmission-ratio distortion at loci on the human X chromosome. Am J Hum Genet. 1998, 62(6):1493-9.

97. Naumova AK, Greenwood CM, Morgan K. Imprinting and deviation from Mendelian transmission ratios. Genome. 2001, 44(3):311-20.

98. Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet. 2008, 9(12):951-63.

99. Ngai J, Dowling MM, Buck L, Axel R, Chess A. The family of genes encoding odorant receptors in the channel catfish. Cell. 1993, 72(5):657-66.

100. Nishihara H, Hasegawa M, Okada N. Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. Proc Natl Acad Sci USA. 2006, 103(26):9929-34.

101. Nordborg M, Tavaré S. Linkage disequilibrium: what history has to tell us. Trends Genet. 2002, 18(2):83-90.

102. Novacek MJ. Mammalian phylogeny: shaking the tree. Nature. 1992, 356(6365):121-5.

103. Nozawa M, Nei M. Evolutionary dynamics of olfactory receptor genes in Drosophila species. Proc Natl Acad Sci U S A. 2007, 104(17):7122-7.

104. Ober C, Weitkamp LR, Cox N, Dytch H, Kostyu D, Elias S. HLA and mate choice in humans. Am J Hum Genet. 1997, 61:497-504.

105. Olsson M, Madsen T, Nordby J, Wapstra E, Ujvari B, Wittsell H. Major histocompatibility complex and mate choice in sand lizards. Proc Biol Sci. 2003, 270 Suppl 2:S254-6.

106. Panzanelli P, López-Bendito G, Luján R, Sassoé-Pognetto M. Localization and developmental expression of GABA(B) receptors in the rat olfactory bulb. J Neurocytol. 2004, 33(1):87-99.

107. Pardo-Manuel de Villena F, Sapienza C. Nonrandom segregation during meiosis: the unfairness of females. Mamm Genome. 2001, 12(5):331-9.

108. Parmentier M, Libert F, Schurmans S, Schiffmann S, Lefort A, Eggerickx D, Ledent C, Mollereau C, Gérard C, Perret J, Grootegoed A, Vassart G. Expression of members of the putative olfactory receptor gene family in mammalian germ cells. Nature. 1992, 355(6359):453-5.

109. Paterson AD, Sun L, Liu XQ; Framingham Heart Study. Transmission ratio distortion in families from the Framingham Heart Study. BMC Genet. 2003, 4 Suppl 1:S48.

110. Paterson AD, Waggott D, Schillert A, Infante-Rivard C, Bull SB, Yoo YJ, Pinnaduwage D. Transmission-ratio distortion in the Framingham Heart Study. BMC Proc. 2009, 3 Suppl 7:S51.

111. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science. 2001, 294(5547):1719-23.

112. Purushothaman D, Elliott RW, Ruvinsky A. A search for transmission ratio distortions in offspring from crosses between inbred mice. J Genet. 2008, 87(2):127-31.

113. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. Linkage disequilibrium in the human genome. Nature. 2001, 411(6834):199-204.

114. Reusch TB, Häberli MA, Aeschlimann PB, Milinski M. Female sticklebacks count alleles in a strategy of sexual selection explaining MHC polymorphism. Nature. 2001, 414(6861):300-2.

115. Robertson HM, Warr CG, Carlson JR. Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. Proc Natl Acad Sci U S A. 2003, 100 Suppl 2:14537-42.

116. Rohde RA, Muller RA. Cycles in fossil diversity. Nature. 2005, 434(7030):208-10.

117. Ryder LP, Svejgaard A, Dausset J. Genetics of HLA disease association. Annu Rev Genet. 1981, 15:169-87.

118. Santos PS, Schinemann JA, Gabardo J, Bicalho Mda G. New evidence that the MHC influences odor perception in humans: a study with 58 Southern Brazilian students. Horm Behav. 2005, 47(4):384-8.

119. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, John Wilbur W, Yaschenko E, Ye J. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2010, 38(Database issue):D5-16.

120. Schulze TG, Zhang K, Chen YS, Akula N, Sun F, McMahon FJ. Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. Hum Mol Genet. 2004, 13(3):335-42.

121. Schwensow N, Eberle M, Sommer S. Compatibility counts: MHC-associated mate choice in a wild promiscuous primate. Proc Biol Sci. 2008, 275(1634):555-64.

122. Shoshani J, McKenna MC. Higher taxonomic relationships among extant mammals based on morphology, with selected comparisons of results from molecular data. Mol Phylogenet Evol. 1998, 9(3):572-84.

123. Singh PB, Brown RE, Roser B. MHC antigens in urine as olfactory recognition cues. Nature. 1987, 14-20;327(6118):161-4.

124. Skelding KA, Gerhard GS, Simari RD, Holmes DR Jr. The effect of HapMap on cardiovascular research and clinical practice. Nat Clin Pract Cardiovasc Med. 2007, 4(3):136-42.

125. Slatkin M. Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008, 9(6):477-85.

126. Snell GD. Methods for the study of histocompatibility genes. J Genet. 1948, 49:87-108.

127. Sommer S. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. Front Zool. 2005, 2:16.

128. Spehr M, Gisselmann G, Poplawski A, Riffell JA, Wetzel CH, Zimmer RK, Hatt H. Identification of a testicular odorant receptor mediating human sperm chemotaxis. Science. 2003, 299(5615):2054-8.

129. Tabor R, Yaksi E, Friedrich RW. Multiple functions of GABA A and GABA B receptors during pattern processing in the zebrafish olfactory bulb. Eur J Neurosci. 2008, 28(1):117-27.

130. Tatsura H, Nagao H, Tamada A, Sasaki S, Kohri K, Mori K. Developing germ cells in mouse testis express pheromone receptors. FEBS Lett. 2001, 488(3):139-44.

131. Thompson EE, Haller G, Pinto JM, Sun Y, Zelano B, Jacob S, McClintock MK, Nicolae DL, Ober C. Sequence variations at the human leukocyte antigen-linked olfactory receptor cluster do not influence female preferences for male odors. Hum Immunol. 2010, 71(1):100-3.

132. Trowsdale J. HLA genomics in the third millennium. Curr Opin Immunol. 2005, 17:498-504.

133. Vanderhaeghen P, Schurmans S, Vassart G, Parmentier M. Olfactory receptors are displayed on dog mature sperm cells. J Cell Biol. 1993, 123(6):1441-52.

134. Vanderhaeghen P, Schurmans S, Vassart G, Parmentier M. Molecular cloning and chromosomal mapping of olfactory receptor genes expressed in the male germ line: evidence for their wide distribution in the human genome. Biochem Biophys Res Commun. 1997a, 237(2):283-7.

135. Vanderhaeghen P, Schurmans S, Vassart G, Parmentier M. Specific repertoire of olfactory receptor genes in the male germ cells of several mammalian species. Genomics. 1997b, 39(3):239-46.

136. Vandiedonck C, Knight JC. The human Major Histocompatibility Complex as a paradigm in genomics research. Brief Funct Genomic Proteomic. 2009, 8(5):379-94.

137. VanLiere JM, Rosenberg NA. Mathematical properties of the r2 measure of linkage disequilibrium. Theor Popul Biol. 2008, 74(1):130-7.

138. Véron N, Bauer H, Weisse AY, Lüder G, Werber M, Herrmann BG. Retention of gene products in syncytial spermatids promotes non-Mendelian inheritance as revealed by the t complex responder. Genes Dev. 2009, 23(23):2705-10.

139. Volz A, Fonatsch C, Ziegler A. Regional mapping of the gene for autosomal dominant spinocerebellar ataxia (SCA1) by localizing the closely linked D6S89 locus to 6p24.2-p23.05. Cytogenet Cell Genet. 1992, 60(1):37-9.

140. Volz A, Ehlers A, Younger R, Forbes S, Trowsdale J, Schnorr D, Beck S, Ziegler A. Complex transcription and splicing of odorant receptor genes. J Biol Chem. 2003, 278(22):19691-701.

141. Vucinić D, Cohen LB, Kosmidis EK. Interglomerular center-surround inhibition shapes odorant-evoked input to the mouse olfactory bulb in vivo. J Neurophysiol. 2006, 95(3):1881-7.

142. Wachowiak M, Cohen LB. Presynaptic inhibition of primary olfactory afferents mediated by different mechanisms in lobster and turtle. J Neurosci. 1999, 19(20):8808-17.

143. Walensky LD, Ruat M, Bakin RE, Blackshaw S, Ronnett GV, Snyder SH. Two novel odorant receptor families expressed in spermatids undergo 5'-splicing. J Biol Chem. 1998, 273(16):9378-87.

144. Wedekind C, Seebeck T, Bettens F, Paepke AJ. MHC-dependent mate preferences in humans. Proc Biol Sci. 1995, 260(1359):245-9.

145. Willison KR, Lyon MF. A UK-centric history of studies on the mouse t-complex. Int. J. Dev. Biol. 2000, 44:57-63.

146. Woelfing B, Traulsen A, Milinski M, Boehm T. Does intra-individual major histocompatibility complex diversity keep a golden mean? Philos Trans R Soc Lond B Biol Sci. 2009, 364(1513):117-28.

147. Wu G, Hao L, Han Z, Gao S, Latham KE, de Villena FP, Sapienza C. Maternal transmission ratio distortion at the mouse Om locus results from meiotic drive at the second meiotic division. Genetics. 2005, 170(1):327-34.

148. Xing J, Witherspoon DJ, Watkins WS, Zhang Y, Tolpinrud W, Jorde LB. HapMap tagSNP transferability in multiple populations: general guidelines. Genomics. 2008, 92(1):41-51.

149. Yamazaki K, Boyse EA, Miké V, Thaler HT, Mathieson BJ, Abbott J, Boyse J, Zayas ZA, Thomas L. Control of mating preferences in mice by genes in the major histocompatibility complex. J Exp Med. 1976, 144(5):1324-35.

150. Young JM, Friedman C, Williams EM, Ross JA, Tonnes-Priddy L, Trask BJ. Different evolutionary processes shaped the mouse and human olfactory receptor gene families. Hum Mol Genet. 2002 Mar 1;11(5):535-46.

151. Younger RM, Amadou C, Bethel G, Ehlers A, Lindahl KF, Forbes S, Horton R, Milne S, Mungall AJ, Trowsdale J, Volz A, Ziegler A, Beck S. Characterization of clustered MHC-linked olfactory receptor genes in human and mouse. Genome Res. 2001, 11(4):519-30.

152. Yuhki N, Beck T, Stephens R, Neelam B, O'Brien SJ. Comparative genomic structure of human, dog, and cat MHC: HLA, DLA, and FLA. J Hered. 2007, 98(5):390-9.

153. Yunis EJ, Larsen CE, Fernandez-Viña M, Awdeh ZL, Romero T, Hansen JA, Alper CA. Inheritable variable sizes of DNA stretches in the human MHC: conserved extended haplotypes and their fragments or blocks. Tissue Antigens. 2003, 62(1):1-20.

154. Zaykin DV, Pudovkin A, Weir BS. Correlation-based inference for linkage disequilibrium with multiple alleles. Genetics. 2008, 180(1):533-45.

155. Zhang K, Qin ZS, Liu JS, Chen T, Waterman MS, Sun F. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. Genome Res. 2004, 14(5):908-16.

156. Zhang X, Firestein S. The olfactory receptor gene superfamily of the mouse. Nat Neurosci. 2002, 5(2):124-33.

157. Ziegler A, Müller C, Heinig J, Radka SF, Kömpf J, Fonatsch C. Monosomy 6 in a human lymphoma line induced by selection with a monoclonal antibody. Immunobiology. 1985, 169(5):455-60.

158. Ziegler A. Biology of chromosome 6. DNA Seq. 1997, 8(3):189-201.

159. Ziegler A EA, Forbes S, Trowsdale J, Uchanska-Ziegler B, Volz A, Younger R, Beck S. Polymorphic olfactory receptor genes and HLA loci constitute extended haplotypes. In: Kasahara M (ed.) *major histocompatibility complex – evolution, structure, and function.* Springer Verlag: Tokyo, 2000a, pp 110-130.

160. Ziegler A, Ehlers A, Forbes S, Trowsdale J, Volz A, Younger R, Beck S. Polymorphisms in olfactory receptor genes: a cautionary note. Hum Immunol. 2000b, 61(12):1281-4.

161. Ziegler A, Dohr G, Uchanska-Ziegler B. Possible roles for products of polymorphic MHC and linked olfactory receptor genes during selection processes in reproduction. Am J Reprod Immunol. 2002, 48(1):34-42. 19.

162. Ziegler A, Kentenich H, Uchanska-Ziegler B. Female choice and the MHC. Trends Immunol. 2005, 26(9):496-502.

163. Zöllner S, Wen X, Hanchard NA, Herbert MA, Ober C, Pritchard JK. Evidence for extensive transmission distortion in the human genome. Am J Hum Genet. 2004, 74(1):62-72.

164. Zozulya S, Echeverri F, Nguyen T. The human olfactory receptor repertoire. Genome Biol. 2001, 2:0018.1-0018.12.

9. Oral Communications and Posters

1. **Santos PSC**, Beck S, Füst G, Horton R, Miretti M, Uchanska-Ziegler B, Ziegler A. Analysis of the Genomic Variation at the two HLA-linked Odorant Receptor Clusters in different extended HLA Haplotypes. 8th International Meeting on Human Genome Variation and Complex Genome Analysis. Hong Kong, China, Sep 2006 (Poster).

2. **Santos, PSC.** A Map of Transmission Distortion in the Human extended MHC (xMHC). 2nd Göttingen Workshop on Immunogenetics. Göttingen, Germany, Feb 2007 (Talk).

3. **Santos PSC**, Uchanska-Ziegler B, Ziegler A. A Map of Transmission Ratio Distortion on chromosome 6p. 12th Human Genome Organisation Meeting. Montreal, Canada, May 2007 (Talk and Poster).

4. **Santos, PSC.** Assessment of Transmission Distortion on Human Chromosome 6p in healthy individuals using tagSNPs. 16. Jahrestagung der Deutschen Gesellschaft für Immungenetik. Essen, Germany, Sep 2008 (Talk).

5. **Santos PSC**, Höhne J, Schlattmann P, König IR, Ziegler A, Uchanska-Ziegler B, Ziegler A. Assessment of transmission distortion on chromosome 6p in healthy individuals using tagSNPs. 10th Human Genome Variation Meeting 2008. Toronto, Canada, October 2008 (Poster).

6. **Santos PSC**, Kellermann T, Uchanska-Ziegler B, Ziegler A. Comparison of MHC-linked odorant receptor repertoires among 14 vertebrates. 23rd European Immunogenetics and Histocompatibility Conference (EFI). Ulm, Germany, May 2009 (Poster)

7. **Santos PSC.** Confirmation of Transmission Distortion in the Human Genome through SNP Genotyping. 11th Human Genome Variation Meeting 2009. Tallinn, Estonia, Sep 2009 (Talk)