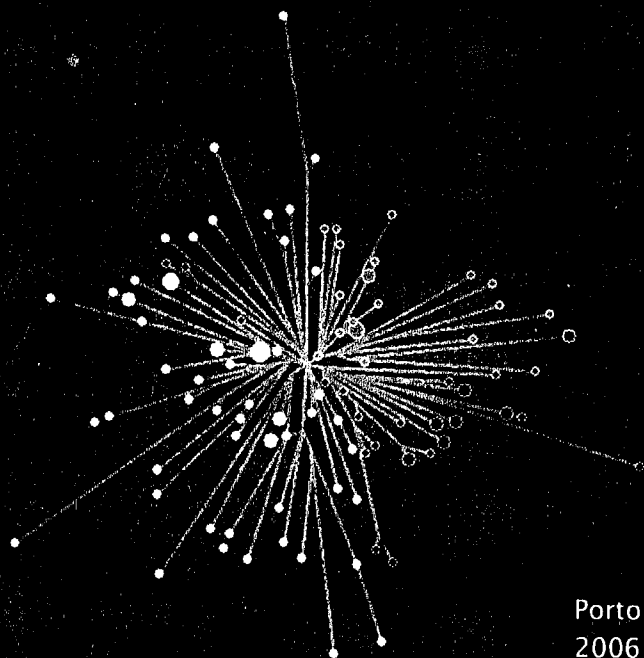


Gene evolution and regulation within the Xq21.3/Yp11.2 hominid-specific homology block

Alexandra Manuel Ferreira Lopes



Porto
2006



FC

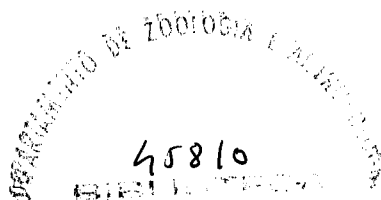
Biblioteca
Faculdade de Ciências
Universidade do Porto



0000055054

100-0/27

Alexandra Manuel Ferreira Lopes



**GENE EVOLUTION AND REGULATION WITHIN
THE Xq21.3/Yp11.2 HOMINID-SPECIFIC
HOMOLOGY BLOCK**



Porto

2006

Dissertação apresentada à Faculdade de
Ciências da Universidade do Porto para a
obtenção do grau de Doutor em Biologia

A thesis submitted for the degree of Doctor
of Philosophy in the Faculty of Science,
University of Porto

ACKNOWLEDGMENTS / AGRADECIMENTOS

Muitos foram os que, de múltiplas formas, contribuíram para a realização deste trabalho. Aproveito este espaço para deixar uma merecida nota de gratidão.

Ao Professor Doutor António Amorim quero agradecer a discussão de ideias, bem como as críticas e sugestões, que foram as linhas orientadoras deste trabalho. A forma como incentivou a que formulasse as minhas próprias questões e a confiança demonstrada na delineação dos trabalhos foram decisivas na minha formação ao nível científico.

Ao Professor Doutor Sobrinho Simões agradeço a oportunidade de integrar uma equipa de investigadores notável como a que encontrei no IPATIMUP e a forma como incute em cada jovem investigador a vontade de fazer sempre mais e melhor.

I would like to thank Professor Timothy Crow for kindly receiving me in his research group and for sharing such a great enthusiasm on human evolution. I also thank Dr. Norman Ross for his remarkable contribution to a collaborative research plan, as well as the whole group, namely, Maria, Adam, Tom, Rekha, James and Nic for sharing ideas and results and for having received me in such a friendly way.

I wish to thank Dr. Carole Sargent for her contribution to the work presented and for a critical reading of the first chapter of this dissertation.

Special thanks to Patricia Blanco-Arias for sharing results and ideas and for all words of support.

I would like to acknowledge all of the co-authors of the publications presented in this thesis for their contributions. Special thanks to Francesc Calafell for all his help.

Gostaria de incluir uma referência à Fundação para a Ciência e Tecnologia pela concessão da bolsa de doutoramento (SFRH/BD/7006/2001) sem a qual não poderia ter realizado este trabalho.

À Sandra Beleza, Sandra Martins, Luísa Azevedo e Susana Seixas um imenso obrigada pela forma carinhosa como me apoiaram desde que integrei o grupo. O

debate de ideias, a partilha de experiências e os (muitos!) momentos de bom humor foram alguns dos melhores momentos que passei durante o doutoramento.

Aos restantes colegas de bancada do laboratório da Genética Populacional, nomeadamente Cíntia, Luísa Pereira, Sandra Alves, Bárbara, Filipe, Pedro, Ana, Iva, Elisabete, Rita, Sofia, Raquel, Mafalda, Gil, Margarida e Susana Pereira, e também à Ângela e à Marta, quero agradecer a boa disposição e a agradável convivência nas longas horas passadas no laboratório. Um agradecimento especial para a Leonor, que me “iniciou” na genética molecular.

Ao Luís, um Muito Obrigada pelas inúmeras sugestões, conselhos e críticas nas longas conversas sobre “vida, universo e tudo o mais”. O teu entusiasmo e a tua forma despretensiosa de fazer ciência foram um exemplo para mim.

Ao Gil, à Ana Sofia e ao Rui Reis agradeço a troca de ideias e o apoio. E tenho também que agradecer ao Rui a preciosa revisão do resumo em francês!

Quero agradecer aos restantes colegas do IPATIMUP pela partilha de experiência e de conhecimento mas também pela forma calorosa como me trataram ao longo dos últimos cinco anos. Obrigada ao Rui e ao Wagner pela ajuda na formatação da tese.

Ao João e aos meus amigos de sempre, em especial à Maria João, à Ana e à Marisa, agradeço o apoio incondicional e as palavras de ânimo nos momentos mais difíceis.

Dedico este trabalho aos meus familiares mais próximos que estiveram sempre presentes e o tornaram possível.

Aos meus pais e às minhas irmãs, Susana e Célia, por me terem sempre incentivado a aprender mais mas, acima de tudo, pelo carinho e pelo apoio, que significam muito para mim.

Ao meu avô, pelo exemplo de tenacidade e de boa disposição e por demonstrar tanto orgulho em ter uma neta Bióloga.

"Seen in the light of evolution, biology is, perhaps, intellectually the most satisfying and inspiring science. Without that light it becomes a pile of sundry facts some of them interesting or curious but making no meaningful picture as a whole."

Theodosius Dobzhansky

SUMMARY

The Xq21.3/Yp11.2 homology block arose by a duplicative translocation approximately 5 Mya, after the divergence of human (*Homo sapiens*) and chimpanzee (*Pan troglodytes*) lineages. This region is thus present on both sex chromosomes in humans but it is restricted to the X chromosome in non-human primates and other mammals. The existence of a hominid-specific block of homology between the sex chromosomes poses interesting questions in evolutionary terms, regarding gene regulation through X-inactivation, and could have contributed to gene dosage differences between humans and chimpanzees, due to gene duplication in this region. Moreover, it offers a unique opportunity to the study of genetic variability in two genomic environments with high sequence identity but distinct exposure to recombination.

The present dissertation is focused on the evolution and regulation of *Protocadherin X/Y* (*PCDHX/Y* or *PCDH11X/Y*) gene pair, which is located within the referred region of homology and is mainly expressed in brain. The issues addressed refer to different evolutionary scales and can be summarized as following:

I. *PCDH11X* evolution and regulation in hominoids.

Assuming that in non human primates and other mammals the strict X-linkage of *Pcdh11* results in X-inactivation of the gene in females, subsequently to the human-specific gene duplication, an adjustment of gene dosage between sexes might have occurred, compensating the gain of an extra copy of the gene on the Y chromosome in human males. If in fact the inactivation status of the gene has changed after the transposition event, this might have contributed to the existence of differences in *PCDH11X* expression levels in brain between humans and chimpanzees, since humans would have two active copies of the gene (on both X chromosomes in females and on the X and Y in males), while in chimpanzees only one *Pcdh11X* gene copy would be transcribed in both sexes. To determine the inactivation status of this gene in humans and in great apes (chimpanzee and gorilla) promoter methylation and *PCDH11X* expression levels in brain cortex of humans and chimpanzees were analysed. Absence of CpG island methylation on both the active and the inactive X chromosomes in humans and an up to twofold excess in the abundance of *PCDH11X* transcripts in females compared to males are strong indications that *PCDH11X* escapes inactivation. While methylation levels in the putative promoter region of *PCDH11X* revealed no differences between humans and great apes, expression was found to be approximately 2-fold higher in human compared to chimpanzee brain. The

inter-species comparisons performed more easily fit a model of *PCDH11X* escaping inactivation only in humans, although the nature of the differences in gene expression levels between humans and chimpanzees remains elusive. The analysis of histone methylation and acetylation, epigenetic modifications also related to X chromosome inactivation, in human and gorilla (*Gorilla gorilla*) lymphoblastoid cell lines was inconclusive, due to artefacts inherent to cell immortalization and culture.

The differences in *PCDH11X* expression levels in males and females could relate to sexual dimorphism in brain. The impact of these findings on differences in brain physiology and/or structure in both normal individuals (between sexes) and in individuals with sex chromosomes aneuploidies (e.g. Klinefelter's syndrome) can only be a matter of speculation at the present time, urging for functional studies on *PCDH11X/Y* protein products.

To assess the evolutionary change occurring at the protein level in this gene pair after duplication and transposition, the relative rate of molecular evolution of *PCDH11X* and *PCDH11Y* was analysed, revealing acceleration in the rate of Y-linked aminoacid substitutions. This gene pair does not seem however to be evolving free from selective constraints, which could indicate a process of pseudogeneization. It is possible that *PCDH11Y* is shifting in function, since it has accumulated not only sequence differences but also a few structural rearrangements.

II. *PCDH11X/Y* evolutionary patterns in human populations.

To better understand the evolutionary dynamics of repetitive sequences in human sex chromosomes seven new X/Y homologous microsatellites located within *PCDH11X/Y* were analysed in samples from Portugal and Mozambique. Higher diversity was found in X-linked microsatellites than in their Y chromosome counterparts; when comparing populations, Mozambicans showed more allele diversity for the X chromosome, but the contrary was true for the Y chromosome microsatellites. Sharp differences were observed on X/Y allele distributions, concerning both the presence of private alleles and a different modal repeat length for X and Y-linked markers. Evolutionary patterns predicted for the Xq21.3/Yp11.2 region under study, relying on five intragenic single nucleotide polymorphisms within *PCDH11X* and two within *PCDH11Y* also revealed distinct scenarios for X and Y chromosomes. Greater microsatellite diversity was displayed by African X chromosomes within the most common haplotypes shared by both populations, whilst higher microsatellite diversity was found in Portugal for the ancestral Y chromosome haplotype.

Overall, information drawn from Yp11.2 polymorphisms is in accordance with the diversity patterns observed for other Y-linked loci, with a remarkable differentiation

between populations observed for Y chromosome haplotypes and thus greater divergence among Y chromosomes in human populations. The analysed Y-linked binary markers are phylogenetically equivalent to previously described mutations and will not contribute to a refinement of Y chromosome clades.

The diversity patterns of Xq21.3, inferred from haplotypes within *PCDH11X* in populations from Africa, Asia and Europe are compatible with the general assumptions of the "Out of Africa" model of human evolution. The contrasting patterns of evolution within the homologous region on Yp do not necessarily defy the referred model, if the specificities inherent to Y-linked markers are properly accounted for.

RESUMO

O bloco de homologia entre Xq21.3/Yp11.2 surgiu em consequência de uma translocação duplicativa, há aproximadamente 5 milhões de anos, após divergência da linhagem que deu origem à espécie humana (*Homo sapiens*) da que originou o chimpanzé (*Pan troglodytes*). Esta região está, portanto, presente em ambos os cromossomas sexuais humanos mas restrita ao cromossoma X quer em primatas não humanos quer em outros mamíferos. A existência de um bloco de homologia que é especificamente humano suscita questões interessantes em termos evolutivos no que respeita à regulação génica por inactivação no cromossoma X, podendo ter contribuído para uma alteração na dose de transcritos entre humanos e chimpanzés, devido à ocorrência de duplicação de genes nesta região. Para além das questões referidas, este bloco proporciona uma oportunidade única para o estudo da variabilidade genética em dois ambientes genómicos com elevada identidade de sequência mas expostos de forma distinta à recombinação. A presente dissertação foca a evolução e regulação do par de genes *Protocaderina XY* (*PCDHXY* ou *PCDH11XY*), que está localizado nesta região de homologia e que é expresso maioritariamente no cérebro. As questões estudadas referem-se a diferentes escalas evolutivas e podem ser sumariadas do seguinte modo:

I. Regulação e evolução do gene *PCDH11X* em hominóides.

Assumindo que em primatas não humanos e outros mamíferos a ligação do gene *Pcdh11* estritamente ao cromossoma X resulta na sua inactivação em fêmeas, posteriormente à duplicação, que é específica de humanos, surge a hipótese de ter ocorrido um ajustamento da dose génica, de forma a compensar o ganho de uma cópia extra do gene no cromossoma Y em humanos. Se o estado de inactivação do gene de facto se alterou após a transposição, pode ter contribuído para a existência de diferenças entre humanos e chimpanzés no nível de expressão da *PCDH11X* no cérebro, já que humanos passariam a ter duas cópias activas do gene (em ambos os cromossomas X no sexo feminino e em ambos os cromossomas sexuais no sexo masculino), enquanto que em chimpanzés apenas uma cópia de *Pcdh11X* seria transcrita em ambos os sexos. De forma a determinar o estado de inactivação deste gene em humanos e em primatas superiores (chimpanzé e gorila), analisaram-se os níveis de metilação na região promotora e os níveis de expressão de *PCDH11X* em córtex cerebral de humanos e chimpanzés. A ausência de metilação das ilhas CpG em ambos os cromossomas X (activo e inactivo) em humanos e um excesso de aproximadamente duas vezes na abundância de transcritos do gene *PCDH11X* em

indivíduos do sexo feminino, quando comparados com os do sexo oposto, são uma forte indicação de que este gene escapa à inactivação. Enquanto que os níveis de metilação na provável região promotora do gene não revelaram diferenças entre humanos e primatas superiores, a expressão do gene foi aproximadamente duas vezes superior em cérebro humano, quando comparado com o de chimpanzé. Os resultados obtidos nas comparações inter-espécies realizadas adequam-se mais facilmente a um modelo em que *PCDH11X* escapa inactivação apenas em humanos, apesar de a natureza das diferenças no nível de expressão entre humanos e chimpanzés ser ainda intangível. A análise de metilação e acetilação de histonas, algumas alterações epigenéticas também relacionadas com a inactivação do cromossoma X, em linhas celulares linfoblastóides de humano e de gorila (*Gorilla gorilla*) revelou-se inconclusiva, devido à existência de artefactos inerentes à cultura de células e à immortalização de linhas celulares.

As diferenças encontradas no nível de expressão da *PCDH11X* entre os sexos podem estar relacionadas de algo modo com a existência de dimorfismos sexuais no cérebro humano. O impacto destes achados na compreensão tanto de diferenças fisiológicas e/ou anatómicas entre sexos em indivíduos normais, como em características específicas em indivíduos com aneuploidias dos cromossomas sexuais (como por exemplo indivíduos com o síndrome de Klinefelter) apenas pode ser especulado, sendo necessários, no futuro, estudos funcionais nas proteínas codificadas pelos genes *PCDH11X/Y*.

De modo a avaliar as alterações que ocorreram a nível da proteína neste par de genes no decurso da evolução, a taxa relativa de evolução molecular de *PCDH11X* e *PCDH11Y* foi analisada, revelando uma aceleração na taxa de substituição de aminoácidos no gene no cromossoma Y. Este par de genes não parece estar de qualquer modo a evoluir livre de constrições selectivas, o que poderia indicar um processo de pseudogeneização. É possível que *PCDH11Y* esteja a sofrer uma alteração de função já que acumulou, para além de alterações de sequência, alguns rearranjos estruturais.

II. Padrões evolutivos dos genes *PCDH11X/Y* em populações humanas.

De modo a melhor compreender a dinâmica evolutiva das sequências repetitivas nos cromossomas sexuais humanos, foram analisados sete novos microssatélites homólogos, localizados dentro dos genes *PCDH11X/Y*, em amostras da população portuguesa e da população moçambicana. Os microssatélites do cromossoma X revelaram maior diversidade do que os correspondentes do cromossoma Y. Em termos de comparação populacional, os moçambicanos revelaram

maior diversidade no cromossoma X, enquanto que, para o cromossoma Y se verificou o oposto. Foram observadas diferenças significativas nas distribuições de microssatélites do cromossoma X e do cromossoma Y, em relação à presença de alelos exclusivos e a diferentes modais para o tamanho da sequência repetitiva em cada um dos cromossomas sexuais.

Os padrões evolutivos previstos para a região em Xq21.3/Yp11.2 sob estudo, com base em cinco polimorfismos bialélicos dentro de *PCDH11X* e em dois em *PCDH11Y*, revelaram cenários distintos para os cromossomas X e Y: a população africana apresenta maior diversidade de microssatélites para os haplótipos partilhados, a europeia revela maior diversidade para o cromossoma Y. De um modo global, a informação deduzida a partir dos polimorfismos em Yp11.2 está em concordância com os padrões de diversidade observados para outros loci no cromossoma Y, em que é observada uma diferenciação notável nas frequências de haplótipos entre populações, verificando-se portanto em populações humanas uma maior divergência entre cromossomas Y. Os polimorfismos bialélicos analisados são filogeneticamente equivalentes a mutações previamente descritas no cromossoma Y e portanto não irão contribuir para o refinamento dos clados já definidos.

Os padrões de diversidade em Xq21.3, inferidos a partir de haplótipos de marcadores bialélicos dentro do gene *PCDH11X* analisados em populações provenientes do continente africano, asiático e europeu são compatíveis com os pressupostos gerais do modelo "Out of Africa" para a evolução humana. Os padrões contrastantes de evolução dentro da região homóloga em Yp não desafiam necessariamente o modelo referido se as especificidades inerentes a marcadores do cromossoma Y forem devidamente consideradas.

RESUMÉ

Le bloc d'homologie dans Xq21.3/Yp11.2 est apparu en conséquence d'une translocation duplicative, survenue il y a environ 5 millions d'années, après la divergence entre les lignages de l'homme (*Homo sapiens*) et du chimpanzé (*Pan troglodytes*). Cette région est donc présente dans tous les deux chromosomes sexuels humains mais est seulement dans le chromosome X des primates non humains et d'autres mammifères. L'existence d'un bloc d'homologie qui est spécifiquement humain suscite des questions intéressantes du point de vue évolutif à l'égard de la régulation génique par l'inactivation du chromosome X et son rôle dans l'établissement des différences dans la dose de transcripts entre humains et chimpanzés, dû à la présence de la duplication des gènes dans cette région. En outre, ce bloc fournit une occasion unique pour étudier la variabilité génétique dans deux environnements génomiques d'identité de séquences élevée mais exposés de forme distincte à la recombinaison.

La présente dissertation est centrée sur la régulation et l'évolution de la paire de gènes *Protocadherin X/Y* (*PCDHX/Y* ou *PCDH11X/Y*), localisée dans ce bloc d'homologie et qui est exprimée essentiellement dans le cerveau. Les questions étudiées font allusion aux différentes échelles évolutives et peuvent être résumées comme suit:

I. Règlement et évolution du gène *PCDH11X* chez les hominoides.

En supposant que chez les primates non humains et autres mammifères la liaison du gène *Pcdh11* strictement au chromosome X résulte de son inactivation chez les femelles, ultérieurement à la duplication, qui est spécifique des humains, apparaît l'hypothèse d'avoir existé un ajustement de la dose génique, de manière à compenser le profit d'une copie supplémentaire du gène dans le chromosome Y chez les humains. Si l'état d'inactivation du gène s'est de fait modifié après la transposition, il peut avoir contribué à l'existence de différences entre humains et chimpanzés du niveau d'expression de la *PCDH11X* dans le cerveau, depuis ils humains auraient alors par deux copies actives du gène (dans chacun des deux chromosomes X pour le sexe féminin et dans chacun des deux chromosomes sexuels pour le sexe masculin), tandis que chez les chimpanzés seulement une copie de *Pcdh11X* serait transcrite dans pour chacun des deux sexes.

De manière à déterminer l'état d'inactivation de ce gène chez les humains et chez les primates supérieurs (chimpanzé et gorille), les niveaux de méthylation de la région promotrice et les niveaux d'expression de *PCDH11X* dans cortex le cérébral d'humains et des chimpanzés ont été analysés. L'absence de méthylation des îles CpG

dans chacun des deux chromosomes X (actif et inactif) chez les humains et un excès d'approximativement deux fois dans l'abondance de transcripts du gène *PCDH11X* chez les individus du sexe féminin, alors que comparés avec ceux du sexe opposé, sont un important indice de que ce gène échappe à l'inactivation. Tandis que les niveaux de méthylation dans la probable région promotrice du gène n'ont pas révélé de différences entre des humains et des primats supérieurs, l'expression du gène a été approximativement deux fois supérieure dans le cerveau humain, alors que comparé avec ceux de chimpanzé. Les résultats obtenus avec les comparaisons inter-espèces réalisées s'ajustent plus facilement à un modèle dans lequel *PCDH11X* échappe à l'inactivation seulement chez les humains, malgré le fait que la nature des différences du niveau d'expression entre humains et chimpanzés soit encore intangible. L'analyse de méthylation et d'acétylation des histones, quelques modifications épigénétiques qui concernent aussi l'inactivation du chromosome X, dans des lignes cellulaires EBV-lymphoblastoïdes d'humain et de gorille (*Gorilla gorilla*) s'est révélée peu concluante, dû à l'existence d'artefacts inhérents à la culture et transformation de lignées cellulaires.

Les différences du niveau d'expression de la *PCDH11X* trouvées entre les sexes peuvent d'une certaine manière être en rapport avec l'existence de dimorphismes sexuels dans le cerveau humain. L'impact de ces découvertes dans la compréhension tant de différences physiologiques et/ou anatomiques entre les sexes chez des individus normaux, que dans des caractéristiques spécifiques chez les individus avec des aneuploidies des chromosomes sexuels (comme par exemple des personnes avec le syndrome de Klinefelter) ne peut en ce moment qu'être spéculé, étant nécessaire, à l'avenir, des études fonctionnelles des protéines codifiées par les gènes *PCDH11X/Y*.

Afin d'évaluer les modifications qui se sont produites au niveau de la protéine dans cette paire de gènes au cours de l'évolution, le taux relatif d'évolution moléculaire de *PCDH11X* et *PCDH11Y* a été analysé, révélant une accélération dans le taux de substitution d'acides aminés dans le gène dans le chromosome Y. Cette paire de gènes ne semble pas être à évoluer libre de contraintes sélectives, ce qui pourrait indiquer un processus de pseudogénéisation. Il est possible que *PCDH11Y* soit en train de souffrir une modification de fonction étant donné qu'il a accumulé, outre des modifications de séquence, quelques réarrangements structurels.

II. Normes évolutives des gènes *PCDH11X/Y* dans des populations humaines.

Afin de mieux comprendre la dynamique évolutive des séquences répétitives dans les chromosomes sexuels humains, sept nouveaux microsatellites homologues, localisés à l'intérieur des gènes *PCDH11X/Y*, dans des échantillons de la population

portugaise et de la population mozambicaine ont été analysés. Les microsatellites du chromosome X ont révélé plus grande diversité que les correspondants du chromosome Y. Dans des termes de comparaison de populations, les Mozambicains ont révélé plus grande diversité dans le chromosome X, tandis que pour le chromosome Y s'est vérifié l'opposé. Des différences significatives ont été observées dans les distributions de microsatellites du chromosome X et du chromosome Y, concernant la présence d'allèles exclusifs et les différents modaux pour la dimension de la séquence répétitive dans chacun des chromosomes sexuels.

Les normes évolutives prévues pour la région dans Xq21.3/Yp11.2 étudiées basées sur cinq polymorphismes bialléliques à l'intérieur de *PCDH11X* et sur deux dans *PCDH11Y*, ont révélé des scénarios distincts pour les chromosomes X et Y: La population africaine présente plus grande diversité pour le chromosome X, l'Européenne révèle plus grande diversité pour le chromosome Y. D'une manière globale, les informations déduites à partir des polymorphismes dans Yp11.2 est en accord avec les normes de diversité observées pour d'autres loci dans le chromosome Y, dans lequel est observée une différenciation remarquable dans les fréquences d'haplotypes entre populations, se vérifiant donc dans des populations humaines une plus grande divergence entre chromosomes Y. Les polymorphismes bialléliques analysés sont phylogénétiquement équivalents à des mutations préalablement décrites dans le chromosome Y et donc n'iront pas contribuer au raffinement des clades déjà définies.

Les normes de diversité dans Xq21.3, observées à partir d'haplotypes dans le gène *PCDH11X* analysées chez des populations provenant du continent africain, asiatique et européen sont compatibles avec les présupposés généraux du modèle "Out of Africa" pour l'évolution humaine. Les normes contrastantes d'évolution dans la région homologue de Yp ne défient pas nécessairement le modèle énoncé si les spécificités inhérentes aux marqueurs du chromosome Y sont dûment considérées.

INDEX

i. SUMMARY	11
ii. RESUMO	15
iii. RESUMÉ	19
iv. ABBREVIATIONS	27

GENERAL INTRODUCTION

1. Characterization of human X and Y chromosomes	31
2. Evolution of mammalian sex chromosomes	33
2.1. Present X-Y homology	36
2.1.1. The Xq21.3/Yp11.2 homology block	37
2.1.1.1. Gene structure and putative functions of <i>PCDH11X/Y</i>	38

SECTION I: *PCDH11X/Y* REGULATION AND EVOLUTION IN HOMINIDS

INTRODUCTION

1. Gene Dosage Compensation	47
1.1. Emergence of X-inactivation in mammals	47
1.1.1. Exceptions to the model	48
1.1.2. "Snap-shots" of dosage compensation evolution	49
1.2. Mechanisms of X-inactivation	51
1.2.1. Epigenetic modifications of the inactive X chromosome	52
1.2.2. Strategies for assessing the inactivation status of X-linked genes	53
2. The Xq21.3/Yp11.2 transposition and human-chimpanzee differences	56
AIMS	63

RESULTS

1. Inactivation status of *PCDH11X* in humans and in the chimpanzee

Article 1 - Lopes AM, Ross N, Close J, Dagnall A, Amorim A, Crow TJ. Inactivation status of <i>PCDH11X</i> : sexual dimorphisms in gene expression levels in brain. Hum Genet (in press)	67
---	----

2. <i>PCDH11X</i> histone modifications and CpG island methylation in human and gorilla female lymphoblastoid cell lines (AM Lopes and N Ross, unpublished results)	79
---	----

3. Methylation of *PCDH11X* in Klinefelter's syndrome (XXY)

Article 2 - Ross N, Wadekar R, Lopes A, Dagnall A, Close J, DeLisi LE, Crow TJ. Methylation of two Homo sapiens-specific X-Y homologous genes in Klinefelter's syndrome (XXY) Am J Med Genet B Neuropsychiatr Genet (submitted).....	85
---	----

DISCUSSION

1. Inactivation status and gene dosage of <i>PCDH11X</i>	105
2. Rates of molecular evolution of <i>PCDH11X</i> and its paralogue <i>PCDH11Y</i>	108

CONCLUDING REMARKS	115
---------------------------------	-----

REFERENCES	119
-------------------------	-----

SECTION II: *PCDH11X/Y* EVOLUTIONARY PATTERNS IN HUMAN POPULATIONS

INTRODUCTION

1. The significance of genetic variation.....	147
2. Evolutionary and population genetics frameworks	147
2.1. Genetic perspectives on human origins	147
2.2. The re-shaping of human diversity	149
3. The sex chromosomes in population genetics	150
3.1. Properties of the Y chromosome	150
3.2. Properties of the X chromosome	153
3.3. Properties of the Xq21.3/Yp11.2 regions.....	154

AIMS	157
-------------------	-----

RESULTS

1. Microsatellite diversity comparisons and evolutionary patterns of *PCDH11X/Y*

Article 3 - Lopes AM, Calafell F, Amorim A (2004) Microsatellite variation and evolutionary history of <i>PCDHX/Y</i> gene pair within the Xq21.3/Yp11.2 hominid-specific homology block. Mol Biol Evol 21(11):2092-2101.	161
---	-----

2. *PCDH11Y* biallelic markers in a large sample of well characterized Portuguese Y chromosome lineages

Article 4 - Beleza S, Gusmão L, Lopes A, Alves C, Gomes I, Giouzeli M, Calafell F, Carracedo A, Amorim A. Micro-phylogeographic and demographic history of Portuguese male lineages. *Ann Hum Genet* (in press)..... 173

DISCUSSION

1. Xq21.3/Yp11.2 microsatellite diversity in two human populations.....	191
1.1. The genetic background of the populations under study.....	191
1.1.1. Male and female lineages in Mozambique	191
1.1.2. Male and female lineages in Portugal	192
1.2. Comparison of X- and Y-linked microsatellite variation	193
1.3. Comparison of African versus European X/Y microsatellite diversity.....	196
2. <i>PCDH11Y</i> haplotypes within the Y chromosome phylogeny.....	197
3. Diversity patterns within <i>PCDH11X</i>	198

CONCLUDING REMARKS..... 208

REFERENCES..... 211

ABBREVIATIONS

bp	Base pair
CNR	Cadherin-Related Neuronal Receptor
CNS	Central Nervous System
CpG	Cytosin phospho Guanine
DHPLC	Denaturing High Performance Liquid Chromatography
DNA	Deoxyribonucleic Acid
Kb	Kilobases
Ky, Kya	Thousand years, thousand years ago
LD	Linkage Disequilibrium
LGM	Last Glacial Maximum
LINE	Long Interspersed Repeat Element
Mb	Megabases
MSY	Male Specific Region of the Y chromosome
mtDNA	Mitochondrial DNA
My, Mya	Million years, million years ago
NRY	Non-Recombining Region of the Y chromosome
ORF	Open-Reading Frame
PAR1, PAR2	Pseudoautosomal region 1, pseudoautosomal region 2
PCR, RT-PCR	Polymerase Chain Reaction, Reverse Transcription-PCR
RAO	Recent African Origin
RNA, mRNA	Ribonucleic Acid, messenger RNA
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
TMRCA	Time to Most Recent Common Ancestor
TSS	Transcriptional Start Site
UTR	Untranslated Region
XCR	Conserved Region of the X Chromosome
Xp, Yp	Y chromosome's short arm, X chromosome's short arm
Xpter	Terminal region of X chromosome's short arm
Xq, Yq	Y chromosome's long arm, X chromosome's long arm
YCC	Y Chromosome Consortium
YCR	Conserved Region of the Y Chromosome

GENERAL INTRODUCTION

1. Characterization of human X and Y chromosomes

Chromosomal systems of sex determination are diverse and similar systems can be found in evolutionarily distant taxa. In our species, females present a pair of similar X chromosomes (46, XX karyotype), while males, the heterogametic sex, carry a pair of dimorphic X and Y chromosomes (46, XY karyotype). The Y chromosome is male determining since testis development and thus male sexual differentiation are triggered in the early embryo by the gene product of one Y-linked locus, *SRY* (Sex Determining Region on the Y; Sinclair *et al.*, 1990). The XX/XY system is almost universal across placental mammals (Eutheria), with few exceptional cases noteworthy by the presence of composite or multiple sex chromosome systems or, more strikingly, by the absence of a Y chromosome (Fredga, 1983 and 1988).

Although sharing regions of homology, sex chromosomes are strikingly different in size, morphology and gene content. Recently, the nearly complete sequence of both human X and Y chromosomes has been published (Skaletsky *et al.*, 2003; Ross *et al.*, 2005), providing a detailed picture of their differences and similarities and thus allowing a better understanding of sex chromosomes biology. One of the most striking differences between sex chromosomes is the enormous size disparity, since the X chromosome comprises roughly 5% of euchromatin of a diploid cell, while its homologue represents approximately only 1% (Vallender *et al.*, 2005), resulting in a manifold difference in euchromatic material. This difference is also reflected in the number of genes found on the X chromosome (approximately 1000; Ross *et al.*, 2005) and on the Y chromosome (156; Skaletsky *et al.*, 2003). Moreover, the Y is highly enriched in heterochromatin, which constitutes the bulk (roughly 40Mb) of the distal long arm of this chromosome (Figure 1) and largely supersedes the total of euchromatic DNA sequence (23 Mb). The different size and morphology of sex chromosomes allow their distinction at the cytogenetic level.

Recombination between the X and the Y is restricted to regions of homology in the extremities of the short and long arms of the chromosomes, referred to as pseudoautosomal regions or PARs (Cooke *et al.*, 1985; Freije *et al.*, 1992). These regions are highly recombinogenic, with the occurrence of an obligate X-Y crossing over in PAR1 in male meiosis (Burgoyne, 1982) and a 6-fold higher rate of recombination in PAR2, compared with X-specific sequences (Kvaloy *et al.*, 1994; Li and Hamer, 1995). Outside the pairing regions, the X chromosome recombines only in female meiosis, while the Y chromosome remains unpaired throughout generations, contributing to the existence of X- and Y-specific sequences with different degrees of divergence (Lhan and Page, 1999a; Skaletsky *et al.*, 2003).



Figure 1 – Simplified scheme of human X and Y chromosomes. The pseudoautosomal regions are in violet and the heterochromatic region of the Y chromosome is in grey. Green regions represent the most ancient regions of sex chromosomes (XCR and YCR) while in red are represented regions that have been added to the X and Y less than 130 Mya (see also Figure 3). On the Y chromosome both red and green regions are non-recombining. From Graves (2002).

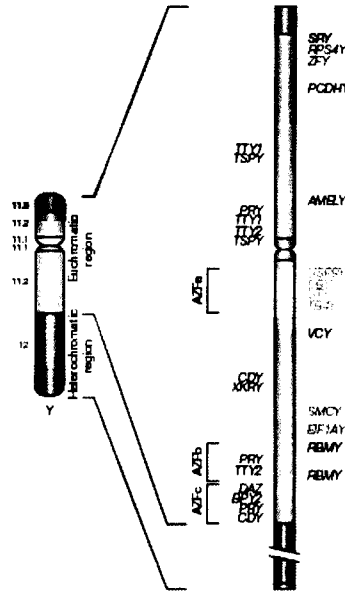


Figure 2 – Active genes on the Y chromosome. The non-recombining region (NRY) is in yellow, the grey bar is the centromere and the pseudoautosomal regions are in red. Genes named to the right on the scheme have active homologues on the X chromosome, while genes on the left do not. Genes in red are widely expressed housekeeping genes; genes in black are expressed in testis only; genes in green do not fall in any of the patterns of expression previously mentioned: *Amelogenin Y* (*AMELY*) is expressed in tooth buds and *Protocadherin Y* (*PCDHY*) is expressed mainly in brain. From Lahn *et al.* (2001).

The human X chromosome harbours approximately 1000 genes involved in several different functions (Ross *et al.*, 2005). In contrast, a distinctive feature of the Y chromosome is the paucity of diversity in gene function (Figure 2) with the presence of few functional classes of Y-linked genes, regarding their expression pattern (Lahn and Page, 1997; Lahn *et al.*, 2001; Skaletsky *et al.*, 2003). Notably, genes widely expressed in the body are found in single copy and have an X-linked homologue, while genes expressed predominantly or exclusively in the testes are present in multi-copy

and fail to identify homologous sequences on the X chromosome. The latter belong to some of the well known protein-coding gene families involved in spermatogenesis, such as *DAZ* (Deleted in AZoospermia), and localize to ampliconic regions, where several copies of the same transcript are found organized in palindromes (Skaletsky *et al.*, 2003). Interestingly, frequent gene conversion between palindrome arms in human and chimpanzee lineages is thought to account for the high sequence identity between gene copies (~99.9% identity), and might be contributing to the conservation of the integrity of testis-specific genes in the absence of X-Y crossing over (Rozen *et al.*, 2003). An enrichment of genes involved in sex and reproduction, in comparison to autosomes, was also reported on the X chromosome (Saifi and Chandra, 1999; Hurst, 2001), as well as of genes expressed in brain and associated with general cognitive ability (Qiu *et al.*, 2002; Zechner *et al.*, 2001). Even so, the functional specialization of the Y chromosome is far more dramatic, with a remarkable accumulation of Y-specific spermatogenic gene families by differentiation from an X homologue, as in the case of *RBMY* (RNA Binding Motif Protein Y-linked; Delbridge *et al.*, 1999; Mazeyrat *et al.*, 1999), by autosomal transposition, such as *DAZ* (Saxena *et al.*, 1996) or retrotransposition, such as *CDY* (Chromodomain Protein Y-linked; Lahn and Page, 1999b), followed by intense gene amplification.

Enriched in heterochromatin and repeats, and lacking most of the genes present on its counterpart, the Y chromosome has been considered a degenerate copy of the X chromosome and this is attributed to its abstinence from crossing over. The deleterious effects of the lack of recombination are illustrated by the presence of many pseudogenes on the Y chromosome, which are homologous to X-linked genes (Skaletsky *et al.*, 2003). Notably, the Y chromosome is permissive to an alternative way of sequence exchange that, coupled with a safety-copy strategy, might be rescuing certain families of male-specific genes from degeneration.

2. Evolution of mammalian sex chromosomes

The peculiar characteristics of human X and Y chromosomes can only be understood in the light of their evolution.

In spite of being relatively widespread among different species and of sharing a striking resemblance by convergent evolution to a similar function, sex chromosomes from different lineages may not be phylogenetically related. Susumu Ohno proposed in 1967 that the mammalian X and Y chromosomes evolved from an autosome pair, following their recruitment into a chromosomal system for sex determination and several comparative studies have supported this view (Graves, 1995; Charlesworth,

1996). The major structural and genetic differences presently displayed by human sex chromosomes are thus considered derived characteristics that were acquired since X and Y chromosomes started diverging. The fact that birds Z and W sex chromosomes are not homologous to those of mammals (Ohno, 1967; Graves, 1995), places the starting point of X-Y differentiation after the split of birds and mammals from a common reptilian ancestor, approximately 300 Mya (Hedges and Kumar, 1998, 2004).

The acquisition of a dominant sex determining function by *SRY* on the “proto” Y chromosome, relatively to its homologue *SOX3* on the X chromosome (Stevanovic *et al.*, 1993; Foster and Graves, 1994), was the initial step in this differentiation (approximately 300 Mya; Skaletsky *et al.*, 2003), and the isolation of the sex determining region by suppression of X-Y recombination was the major driving force shaping sex chromosomes throughout evolution (Figure 3). Initially encompassing the sex-determining region and subsequently extended to genes with male-specific function (Charlesworth, 1991) suppression of recombination between sex chromosomes has probably been a discontinuous process. In fact, although gradual divergence might also have contributed to some extent to stop X-Y crossing over (Charlesworth *et al.*, 2005), successive large-scale chromosomal inversions are predicted to have occurred on the Y chromosome, disrupting gene order and progressively preventing alignment and exchange of most X-Y sequences in male meiosis (Lhan and Page, 1999a; Skaletsky *et al.*, 2003; Ross *et al.*, 2005). Exempt from recombination with the X chromosome along most of its length, the Y is predicted to have been exposed to accelerated genetic decay, rapidly accumulating deleterious mutations by drift (Muller’s ratchet; Muller, 1964; Rice 1994; Charlesworth, 1996), enhanced by a reduced effective population size (N_e), and by genetic hitchhiking with new favourable mutations (Rice, 1987). Moreover, recessive deleterious mutations on the Y can be easily kept from selective purge in males, by the protective effect of an intact gene copy on the X chromosome. This process of Y degeneration through large rearrangements and accumulation of mutations has been referred to as “attrition” (Graves *et al.*, 1998). In fact, many genes on the Y chromosome presenting a homologue on the X chromosome have become non-functional pseudogenes or have been completely lost during evolution (Skaletsky *et al.*, 2003; Ross *et al.*, 2005), leading to the presence of limited regions of homology between human sex chromosomes.

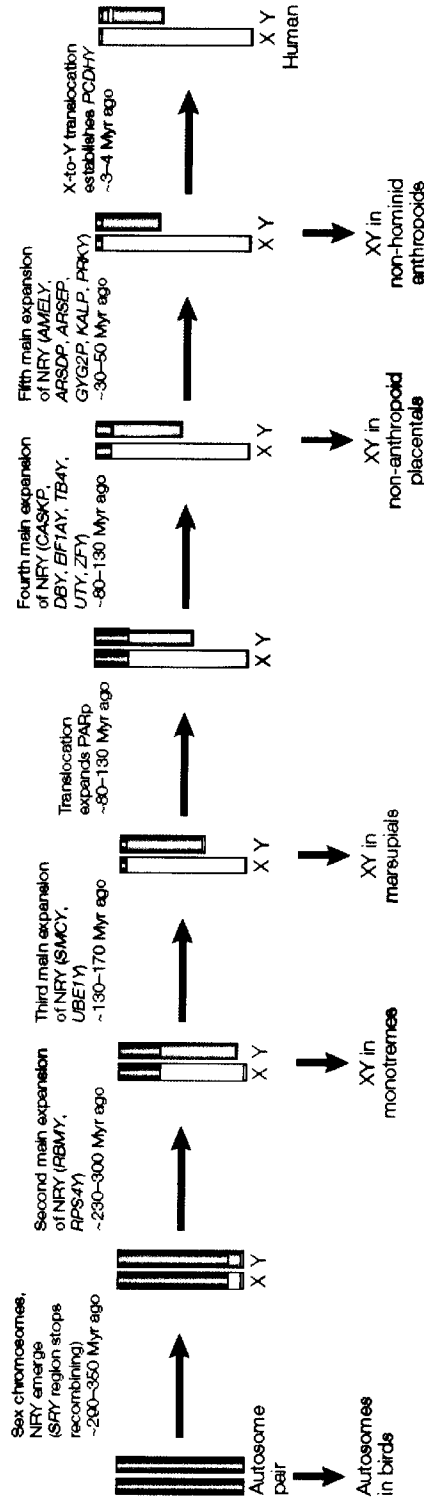


Figure 3 – Main events in the evolution of human sex chromosomes including translocations from the autosomes to sex chromosomes and an X-to-Y transposition, paralleled by the blockwise expansion of the NRY, probably mediated by a series of large scale inversions as posited by Lahn and Page (1999), and overall Y chromosome shrinkage. Blue regions are freely recombining; yellow regions are X chromosome specific; red regions are Y chromosome specific (NRY); green bar represents the X-transposed region on the Y. The diagram is not drawn to scale and centromeres are omitted. PARp, short arm pseudoautosomal region. From Lahn *et al.* (2001).

As a result of this unequal exposure do genetic degeneration, the X chromosome displays a remarkable conservation in structure, gene content and gene order across distantly related species (Ohno, 1967; Graves, 1998; Kohn, 2004) whereas the Y chromosome is characterized by structural rearrangements and gene degeneration to different extents in different groups of mammals (Graves 1995; Skaletsky *et al.*, 2003), being completely absent in two rodent species of the genus *Ellobius* (Fredga, 1983).

The presence of a non random subset of genes on sex chromosomes, compared to autosomes, is also thought to represent the outcome of a complex interaction of selective forces (Vallender and Lahn, 2004). Sexual antagonism (Rice, 1992) under constant selection is predicted to have contributed to a masculinization of the Y chromosome and conversely to a feminization of the X chromosome, due to a preferential representation of genes advantageous to one sex but detrimental to the other sex in the chromosomes that are more often present in the male or female context, respectively. Paradoxically, the hemizygous exposure in males of genes on the X chromosome would have favoured the allocation of genes with recessive male beneficial alleles on this chromosome, which would acquire male-specific expression as allele frequency, and thus the probability of homozygous females in the population, increased (Rice, 1984). This concept might explain the overrepresentation on the mammalian X chromosome of genes involved in spermatogenesis (Wang *et al.*, 2001; Hurst, 2001), most notably those expressed in early stages of spermatogenesis, contrasting with an impoverishment in genes implicated in the late stages of sperm production (Kill *et al.*, 2004). These observations have been related by some authors (Wu and Xu, 2003; Kill *et al.*, 2005) to the phenomenon of MSC1 (Meiotic Sex Chromosome Inactivation; Solari, 1974), which has been recently demonstrated to result in the transcriptional repression of all sex-linked germ-cell-specific genes during meiotic stages of spermatogenesis (Wang *et al.*, 2005). The enrichment of sex chromosomes in genes with sex biased functions has occurred by specialization of pre-existing genes with more general functions but also by gene movements such as transposition (Saxena *et al.*, 1996) and retrotransposition (Lahn and Page, 1999b; Emerson *et al.*, 2004; Khil *et al.*, 2005).

2.1. Present X-Y homology

In spite of the notorious differences displayed by sex chromosomes, the existence of X-Y homologous gene pairs bears witness to their common ancestry. Notably, apart from a few X-Y "fossil" gene pairs retaining ancestral homology (7 X-Y

orthologues; Ross *et al.*, 2005), the majority of genes shared by both sex chromosomes are located in homologous regions of more recent origin. These comprise three regions of high sequence identity between X and Y chromosomes: the previously characterized pseudoautosomal regions (PAR1 and PAR2) at the tips of both X and Y chromosomes, where recombination is maintained, and the Xq21.3/Yp11.2 homology block, which is embedded in the non-recombining region of sex chromosomes.

The 2.7 Mb PAR1 was added to the sex chromosomes of Eutheria (placental mammals) before their radiation approximately 105 My ago, in a larger block that included most of the short arm of the X chromosome (Xp11.3-pter), which is referred to as the “X-added region” or XAR (Figure 1 and Figure 3). Most of X-Y homologous genes locate to PAR1 and are freely exchanged between X and Y chromosomes.

The Xq/Yq 330 Kb PAR2 and the Xq21.3/Yp11.2 homology block were created by separate events involving the duplication and transposition of genetic material from the X to the Y chromosome, after the divergence of human and chimpanzee lineages (Freije *et al.*, 1992; Page *et al.*, 1984), by illegitimate recombination between LINE elements (Kvaloy *et al.*, 1994; Schwartz *et al.*, 1998).

The four genes within PAR2 (*HSPRY3*, *SYBL1*, *IL9R*, *CXYorf*) have identical copies on the X and on the Y chromosomes and undergo recombination both in female and in male meiosis, similarly to PAR1 genes.

2.1.1. The Xq21.3/Yp11.2 homology block

The Xq21.3/Yp11.2 region, also referred as “X-transposed region” on the Y chromosome or XTR (Skaletsky *et al.*, 2003), represents the largest block of homology between human sex chromosomes, spanning 3.91 Mb on the X and 3.38 Mb on the Y chromosome. Although no recombination occurs within this region in male meiosis, there is a high level of X-Y sequence identity (~99%), which can be reconciled with less than 5 million years of divergence (Schwartz *et al.*, 1998; Ross *et al.*, 2005). Nevertheless, several rearrangements and deletions account for structural differences between sex chromosomes within this homology block, namely its bipartite organization on Yp (Figure 4), compared to the continuum block it forms on Xq (Sargent *et al.*, 1996; Mumm *et al.*, 1997; Schwartz *et al.*, 1998; Skaletsky *et al.*, 2003), and the absence of four large regions on the Y chromosome, that have been deleted after the transposition (Vacca *et al.*, 1999; Sargent *et al.*, 2001; Blanco *et al.*, 2001; Blanco-Arias *et al.*, 2004a; Ross *et al.*, 2005). Only three protein-coding genes have been characterized in this region, two of which are X-Y shared, *PCDH11X/Y* or

PCDHX/Y (Yoshida and Sugano, 1999; Blanco *et al.*, 2000) and *TGIF2LX/Y* (Blanco-Arias *et al.*, 2002), while the third gene, *PABPC5* (Blanco *et al.*, 2001), is only present on the X chromosome, since it has been lost from the human Y chromosome (Vacca *et al.*, 1999; Sargent *et al.*, 2001; Blanco *et al.*, 2001).

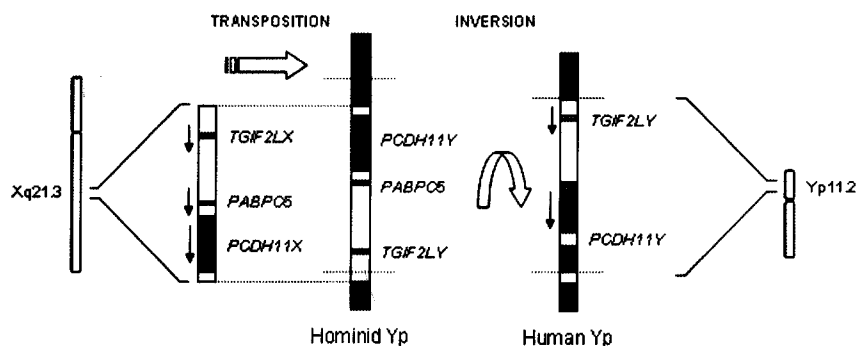


Figure 4 – Schematic representation of the Xq21.3 to Yp11.2 transposition and subsequent paracentric inversion on Yp which created a bipartite block on the Y chromosome. The breakpoints of insertion and inversion are symbolized by dashed lines. In black is Y-specific sequence on Yp, contrasting with X-Y homologous regions, with no background. Shadings indicate transcribed units within the homology block with black arrows indicating the direction of transcription (5'-3') in human X and Y chromosomes. The region of homology spans 3.91 Mb on the X and 3.38 Mb on the Y chromosome since several deletions have occurred on Yp, including the gene *PABPC5* which is not present on the human Y chromosome.

2.1.1.1 Gene structure and putative functions of *PCDH11X/Y*

PCDH11X/Y genes encode protocadherin molecules and are mainly expressed in brain (Blanco *et al.*, 2000). Cadherins are transmembrane proteins that mediate cell-cell adhesion through calcium-dependent homophilic interactions and are involved in embryonic and neural morphogenesis (reviewed by Gumbiner, 1996; Vleminckx and Kemler, 1999; Tepass *et al.*, 2000). Protocadherins are the largest subfamily of cadherins and comprise the alpha, beta and gamma gene clusters at human 5q31 (mouse chromosome 18c), as well as other cadherin-like molecules with six or seven extracellular domains and a cytoplasmic tail divergent from those of classic cadherins (Sano *et al.*, 1993; reviewed by Nollet *et al.*, 2000; Yagi and Takeichi, 2000; Frank and Kemler, 2002). In mammals, protocadherins are mainly expressed in the central nervous system (CNS) and a particularly interesting subgroup of these proteins, the CNRs (Cadherin-related Neuronal Receptors), which are represented in humans by the

Protocadherin α cluster, localize to the synaptic junction and are able to bind tyrosine kinase Fyn through their cytoplasmic domain (Kohmura *et al.*, 1998) and to participate in heterophilic interactions with reelin through their first extracellular domain (EC1) (Seenzaki *et al.*, 1999). Both of these molecules have important functions in the CNS: Fyn tyrosine kinase is known to be crucial for normal brain organization and function (Yagi and Takeichi, 2000) and reelin is a large extracellular matrix protein that functions in the cortical layering and positioning of neurons (D'Arcangelo *et al.*, 1995). Taken together, these observations suggest that CNRs may be involved in the establishment of neuronal connections or in signal transduction at the synaptic membrane (Wu and Maniatis, 1999; Yagi and Takeichi, 2000) and their differential expression patterns within local brain areas might be involved in the differentiation of neural subpopulations within the cortex, through interaction with reelin (Kohmura, 1998; Seenzaki *et al.*, 1999). Furthermore, an outstanding diversity of alternative transcripts can be formed from the *Pcdh* α and γ gene clusters due to an immunoglobulin-like organization, with multiple variable exons and a set of constant exons (Wu and Maniatis, 1999), that are driven by different promoters and combined by alternatively *cis*-splicing (Wang *et al.*, 2002; Tasic *et al.*, 2002). Additionally, each allele within the *Pcdh* α gene cluster is randomly expressed in individual cells, giving rise to combinations of different molecules being expressed in each neuron, a possible mechanism defining cell-specific identity within the CNS (Esumi *et al.*, 2005).

Sequence comparisons have led Blanco *et al.* (2000) and Blanco-Arias *et al.* (2004a) to propose that *PCDH11X/Y* genes belong to the CNR subgroup since they resemble these proteins in some aspects: 1) presence of a lysine-rich segment in the cytoplasmic region; 2) presence of a variant (TGD) RGD motif, a putative integrin binding domain; 3) presence of two copies of a SH2-SH3 Fyn tyrosine kinase binding domain (PXXP motif). Recently, these genes have been assigned to a new subfamily of protocadherin genes present both in man and in the mouse (δ - Protocadherins; Vanhalst *et al.*, 2005; Redies *et al.*, 2005). The adhesive properties of these molecules are apparently weaker than those of classic cadherins, as observed for other protocadherins, and their preference for homophilic interactions, which results in cell sorting both in cell culture and *in vivo*, may be important for specific recognition in cell-cell signalling (reviewed in Redies *et al.*, 2005). Together with *PCDH1*, *PCDH7* and *PCDH9* *PCDH11X/Y* constitute the protocadherin - $\delta 1$ subgroup, characterized by the presence of some distinctive features such as seven extracellular cadherin domains and the conservation of a short sequence motif (RRVTF) in their cytoplasmic domain

referred to as CM3 motif, which mediates their interaction with PP1 α (Vanhalst *et al.*, 2005), a protein implicated in regulating synaptic plasticity (Munton *et al.*, 2004). In mouse brain the expression patterns of the three $\delta 1$ – protocadherins are distinct and regionally restricted to specific anatomical and functional structures of the developing brain, being widely expressed in neuroanatomical structures of the olfactory and limbic systems (Vanhalst *et al.*, 2005). Thus, although the function of *PCDH11X/Y* is still not known, these observations suggest that, like other cadherins, they are involved in the regionalization and functional differentiation of brain grey matter. Moreover, Protocadherin11X protein has been detected by immunohistochemistry in several structures of the developing and adult human brain (Priddle *et al.*, 2002 and 2004).

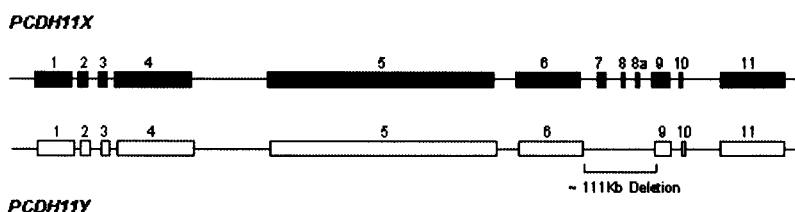


Figure 5 – Simplified gene structure of human *PCDH11X* and corresponding structure of *PCDH11Y*. Boxes represent exons (black, *PCDH11X*; white, *PCDH11Y*) and intervening line is intronic sequence. Exons 4-11 are coding; introns are not drawn to scale. Alternative 5'UTR exons described by Blanco-Arias *et al.* (2004a) were omitted for simplicity. The gene spans approximately 844 Kb on the X chromosome while on the Y chromosome is 111 Kb shorter, lacking exons 7, 8 and 8a.

Since the initial report of the X-linked protocadherin (*PCDH11*) by Yoshida and Sugano (1999) the genomic structure of *PCDH11X/Y* has been repeatedly revised (Blanco *et al.*, 2000; Blanco Arias *et al.*, 2004a) and it is presently established as consisting of at least 17 exons, 9 of which are coding (Figures 5 and 6). Exons 4 and 5 comprise all seven cadherin domains, the transmembrane domain and part of the cytoplasmic domain and can be considered the core of the ORF; exons 1 to 3, and part of exon 4, constitute the 5'UTR; exons 6 to 11 encode alternative cytoplasmic domains and 3' UTRs. Both genes exhibit a complex pattern of alternative splicing, potentially giving rise to variants with different 5' and 3' UTRs and/or different cytoplasmic domains. Moreover, transcription is apparently initiated from two different promoters that determine the existence of short and long isoforms, typically including different

sets of exons as well as different 5' and 3' UTRs (Blanco-Arias *et al.*, 2004a; Vanhalst *et al.*, 2005).

An overall 98.1% nucleotide and 98.3% amino acid identity was described for the short transcripts derived from each sex chromosome (Blanco *et al.*, 2000), a figure that remains largely unaltered when comparing the longest transcripts described by the same authors in 2004 (99.1% nucleotide and 98.3% amino acid identity). In spite of this relatively high similarity along alignable regions, there are some differences between X and Y transcripts that are noteworthy. The coding sequence of the Y chromosome does not comprise exons 7 and 8, which have been deleted in a larger block of approximately 111 Kb of genomic sequence (Blanco-Arias *et al.*, 2004a). Additionally a 13 base pair deletion within exon 4 has removed the initiator methionine of the Y transcripts and the first amino acid (aspartic acid) of the putative signal peptide, while it has brought into frame two alternative initiators, one upstream, giving rise to an enlarged signal peptide, and a downstream one that would lead to the absence of a signal peptide in the resulting protein (Blanco *et al.*, 2000). Evidence for the existence of the latter protein within the cytoplasm of apoptosis-resistant cells derived from a prostate cancer cell line has been reported (Chen *et al.*, 2002; Yang *et al.*, 2005). Overall, the Y transcripts can be derived from four different initiators, while transcription from the X chromosome is apparently initiated from a single position (Blanco-Arias *et al.*, 2004a). Finally, a stop codon in exon 6 of *PCDH11Y* gives rise to a splice variant with a cytoplasmic tail 9 amino acids shorter (Blanco *et al.*, 2000). Additional variation is introduced by alternative splicing that potentially gives rise to 8 variant cytoplasmic domains for *PCDH11X* and 4 for *PCDH11Y* (Blanco-Arias *et al.*, 2004a). The impact of these differences on the function of the *PCDH11* proteins encoded by different sex chromosomes is not known.

Although *Protocadherin11Y* is human-specific, its homologue is present on the X chromosome of different species of primates and other mammals (Blanco *et al.*, 2000; Blanco-Arias *et al.*, 2004b; Vanhalst *et al.*, 2005), and closely matches the partial sequence of an avian cadherin gene (Blanco *et al.*, 2000). In a comparative study of human, pig and mouse *PCDH11X* genes Blanco-Arias and co-workers (2004b) found that most of the coding exons were shared between human and mouse (exons 4,5,7,8,9,10 and 11). On the other hand, some exons previously identified in human transcripts were not detected in mouse and there was no evidence for alternative splicing in this species. These observations, together with the fact that *PCDH11X* is expressed ubiquitously in both mouse and pig tissues, in contrast with the restricted expression of most human isoforms to the CNS, led the authors to suggest that the *PCDH11X* gene family is undergoing species-specific evolution.

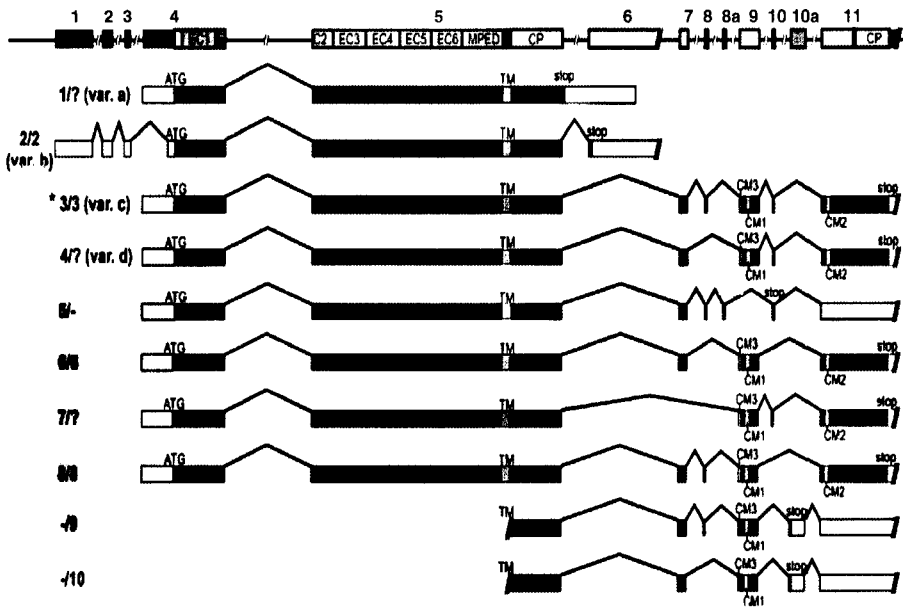


Figure 6 – Gene structure and alternative transcripts of human *PCDH11X* (isoform numbers before slash signs) and mouse *Pcdh11X* (isoform numbers after slash signs). Grey boxes represent exons; an additional exon is present in mouse-specific isoforms (10a). Black boxes represent coding sequences in the transcripts, white boxes represent 5' or 3' UTRs. EC, sequences coding the extracellular cadherin repeats; TM, sequence encoding the transmembrane domain; CM1, CM2 and CM3, conserved motifs. A minus sign symbolizes absence of a particular isoform in that species; question marks denote presumable transcripts that need to be validated experimentally in one of the two species. Variant annotations between parentheses refer to the literature on human *PCDH11X*. The variant with an asterisk is the “standard” long transcript. In variant (b) transcription is initiated from a promoter upstream exon 1 while in the remaining isoforms transcription begins upstream exon 4.1. In variant (a) exon 5 is elongated by intronic read-through. Adapted from Vanhalst *et al.* (2005). **Note:** *PCDH11Y* isoforms corresponding to variants (a), (b), (c) and (d), as well as additional human transcript variation (on the 5'UTR and signal peptide region and on the cytoplasmic domain) were characterized by Blanco *et al.* (2000) and Blanco-Arias *et al.* (2004a).

More recently, several splice variants of *PCDH11X* were detected in mouse brain (Figure 6), some of which are highly similar to human isoforms (Vanhalst *et al.*, 2005), suggesting a higher degree of conservation of *PCDH11X* regulation across species than previously thought.

***PCDH11X/Y* REGULATION AND EVOLUTION IN HOMINIDS**

INTRODUCTION

1. Gene Dosage Compensation

The accumulation of genes related to male reproduction on the Y chromosome, coupled with the extensive divergence of most Y genes from their X-homologues, have had a major impact on the regulation of gene dosage between the sexes. This is evident, as a substantial number of genes on the human X chromosome do not have a Y-linked homologue (1044 annotated genes in the nearly complete sequence of the X chromosome; Ross *et al.*, 2005) and the opposite is also true (approximately 15 protein-coding genes; Ross *et al.*, 2005). While genes only represented on the Y chromosome are expected to be non essential to females, many housekeeping genes, whose expression levels are crucial for both sexes, are restricted to the X chromosome. This imbalance is adjusted, in the particular case of mammals, by silencing almost all genes in one of the X chromosomes of females, early in development, through X-inactivation (Lyon, 1961) and thus equalizing gene dosage between males and females. This is a very effective mechanism, extending to the whole chromosome and to which only a small percentage of genes seem to completely escape (approximately 15% of X-linked genes; Carrel and Willard, 2005).

1.1. Emergence of X-inactivation in mammals

The widely accepted model for the establishment of dosage compensation by X-inactivation in sex chromosomes evolution states that it emerged by a response to the functional decay of Y-linked genes (Lyon, 1961; Ohno 1967; Charlesworth, 1996; Jegalian and Page, 1998). According to this view, reduced or restricted expression of genes on the Y chromosome initially led to an upregulation of X-linked gene expression in both sexes, followed by silencing of genes in one X chromosome of females. Evidence for this upregulation was found for one X-linked gene in mice (Adler, 1997) and recently it has been demonstrated both in mouse and in human somatic tissues (Nguyen and Disteche, 2006). The finding of several evolutionary strata on the X chromosome, created by a stepwise suppression of X-Y recombination, supported a more regional or even gene-by-gene establishment of X-inactivation (Lahn and Page, 1999a) than initially proposed (Lyon, 1961; Ohno 1967). This hypothesis has been strengthened by recent findings of Carrel and Willard (2005) who reported that escape from inactivation is related with the evolutionary strata on the X chromosome, with a higher number of genes escaping inactivation in the younger stratum, where homology of X-Y gene pairs is higher. The establishment of such stratification was only possible due to the conserved gene order observed across species on the eutherian X

chromosome, mainly attributed to the tight regulation to which this chromosome is subjected through X-inactivation, and often referred to as "Ohno's Law", honouring the author who first proposed it.

The model of "Y-driven" evolution described predicts that genes with homologues on the Y chromosome will escape inactivation, the only exception being intermediates which, while having lost their Y-linked counterparts, have not yet become X inactivated (Lahn and Page, 1999a). In agreement with this hypothesis, Carrel and Willard (2005) have recently described that the vast majority of human X-linked genes with Y homology escape inactivation, this proportion being higher when considering only those genes with functional Y orthologues. Concordantly, most of the X linked genes with no homology on the Y chromosome are inactivated.

1.1.1. Exceptions to the model

A considerable number of exceptional genes have been found in humans and in the mouse, some corresponding to the intermediates predicted by Jegalian and Page (1998), which escape inactivation in spite of the absence of a Y homologue (Brown *et al.*, 1995; Carrel *et al.*, 1996; Carrel and Willard, 2005), while others are X inactivated although presenting Y homology (Adler *et al.*, 1991; Carrel and Willard, 2005). According to data from Carrel and Willard data (2005), the former class of genes are mainly located in Xp, suggesting a relatively recent loss of their Y counterparts and supporting their status as intermediates. Additionally, the latter are mainly genes: 1) whose homologues on the Y are pseudogenes and thus probably non functional (e.g. *CASK*, *LOC139231*; Carrel and Willard, 2005); 2) genes belonging to the ampliconic class of X-Y homologous genes (Skaletsky *et al.*, 2003) that might be under less stringent dosage controls (human *TSPYL2* and *RBMX*; Carrel and Willard, 2005); or 3) genes with ubiquitous expression on the X chromosome but testis-specific on the Y (e.g. *Zfx* in the mouse; Adler *et al.*, 1991). Therefore, the great majority of genes defying the dosage compensation model also seem to be exceptional in terms of regulation.

A remarkable insurgent, *UBE1X*, was found to be X inactivated in mouse but it escapes X-inactivation in humans (Carrel *et al.*, 1996). Although in marsupials, in the mouse and other eutherians a Y-linked homologue of *UBE1X* can be detected, *UBE1Y* has been lost in the primate lineage, approximately 40 Mya (Mitchell *et al.*, 1998). The fact that *Ube1x* is widely expressed while *Ube1y* expression is restricted to testis (Mitchell *et al.*, 1991; Odorisio *et al.*, 1996) might have contributed to its susceptibility to X-inactivation in the mouse, but the escape from inactivation of *UBE1X* in humans is

difficult to explain in the light of the generally accepted model for dosage compensation evolution. It can be argued that this gene is still eluding X-inactivation due to a relatively recent loss of its Y homologue, thus being in an intermediate state to attain dosage compensation. Nevertheless this cannot be easily reconciled with *UBE1X* ancestral inactivation status in the mouse, unless if, before being lost in the primate lineage, *UBE1Y* was characterized by a more ubiquitous pattern of expression, as it was described in marsupials (Watson *et al.*, 2000), and *UBE1X* was correspondingly not X inactivated. In this view the present escape from inactivation in humans could be regarded as an example of inertia of dosage compensation mechanisms, but there are many uncertainties.

1.1.2. "Snap-shots" of dosage compensation evolution

The existence of genes that do not seem to be dosage compensated, irrespective of their presence on both X and Y chromosomes, highlights the importance of studying the regulation of each individual X-Y gene pair, in order to gain a better understanding of the global mechanisms. Of special interest are X-linked genes for which the sudden gain or loss of a Y homologue, through gene duplication or deletion, could be detected and placed in a relatively tight evolutionary time-frame, allowing to circumvent the confounding effects of gradual change in X-Y gene dosage throughout sex chromosomes evolution. This is the case for a small number of human genes, namely, *UBE1*, *HSPRY3*, *SYBL1*, *IL9R*, *CXYorf1*, *TGIF2L* and *PCDH11*. The particular evolutionary events they have been involved in, recently in evolution, make them valuable models to better understand the relationship between sex chromosome rearrangements and adjustment of gene dosage by X-inactivation.

The particular case of *UBE1X*, previously discussed, can be regarded as an example of a considerable asynchrony between the loss of a Y-linked gene and X-inactivation establishment. Nevertheless, having in consideration the arguments previously exposed, caution should be taken in interpreting the facts.

All the latter genes have been recently transposed from the X to the Y chromosome, in two separate events, after the divergence of human and chimpanzee lineages, during the last few million years. *HSPRY3X*, *SYBL1X*, *IL9RX*, and *CXorf1* transposition created the present pseudoautosomal region in human Xq/Yq (Freije *et al.*, 1992; Ciccodicola *et al.* 2000; Charchar *et al.*, 2003); *TGIF2L* and *PCDH11* were transposed from Xq21.3 to Yp11.2 and belong to this block of X-Y homology (Page *et al.*, 1984, Yoshida and Sugano, 1999; Blanco *et al.*, 2000; Blanco-Arias *et al.*, 2002).

Within PAR2, the most telomeric genes, *IL9RX* and *CXorf1*, escape inactivation, while the more proximal genes, *HSPRY3* and *SYBL1*, attain dosage compensation in a peculiar way, by silencing a copy of the gene on the inactive X chromosome in females and on the Y chromosome in males (D'Esposito *et al.*, 1996, 1997; Vermeesch *et al.*, 1997; Ciccodicola *et al.* 2000). These findings unravelled the existence of mechanisms of accommodation to the gain of genes on the Y chromosome, such as inactivation of Y-linked genes, and raised the possibility of a relatively quick integration of a gene into dosage compensation mechanisms, following sudden dosage imbalances. However, since the inactivation status of *IL9R* and *CXorf1* in great apes is not known, the latter hypothesis cannot be confirmed. Moreover, the different and complex path followed by each gene during mammalian evolution further obscures the mechanisms of regulation of each of these genes. For instance, while *HSPRY3*, *SYBL1* and *CXorf1* are all X-linked in eutherians (D'Esposito *et al.*, 1997; Ciccodicola *et al.* 2000; Charchar *et al.*, 2003) *IL9R* is autosomal in the mouse but X-linked in great apes (Vermeesch *et al.*, 1997). In fact, it has been proposed that the escape from inactivation of *IL9R* might reflect its recent introduction from an autosome to the X chromosome in great apes (D'Esposito *et al.*, 1997), with an initial absent Y homologue that was subsequently gain in the human lineage. On the contrary, the fact *SYBL1* is subject to X-inactivation has been attributed to its ancestral X-linkage in Eutherians (D'Esposito *et al.*, 1997), which in fact goes back to the marsupial lineage (Charchar *et al.*, 2003). Nevertheless, this argument is not forceful, considering that *CXorf1* is also X-linked in eutherians (Charchar *et al.*, 2003) but escapes inactivation and its Y homologue is also expressed. Thus, PAR2 genes experienced complex and independent evolutionary pathways and the patterns of expression and regulation now observed are more likely to be dependent on the specificities of their new genomic localization, namely in respect to their distance to Y chromosome heterochromatin (Charchar *et al.*, 2003). Although it would be interesting to compare their inactivation status in humans with that of great apes, the different history of X or autosomal linkage across lineages of each of these genes, allied to the fact that they have been involved in several rearrangements in Xq prior to the translocation to Yq (Charchar *et al.*, 2003), will definitely add controversy to possible findings concerning evolution of dosage compensation in this region.

Finally, two pairs of X-Y genes are remarkable in evolutionary terms, for the recent arrival of their Y-linked form on the Y chromosome. *TGIF2L* and *PCDH11* are represented both on the X and on the Y chromosome of humans but display X-linkage in other primates (Blanco *et al.*, 2000) and in the mouse (Blanco-Arias *et al.*, 2004b). The restricted expression of *TGIF2LX/Y* in testis (Blanco-Arias *et al.*, 2002) precludes

the need for its expression levels to be regulated in females. On the contrary, it is reasonable to admit that *PCDH11X/Y* expression, mainly in brain, is a probable target of dosage compensation between sexes. This gene pair thus seems to represent an extremely valuable model but, until this moment, there is only weak evidence of an escape from inactivation of *PCDH11X* in humans, from lymphoblastoid cell lines (Sudbrack *et al.*, 2001), and no studies of this kind have been performed in any other species.

1.2. Mechanisms of X-inactivation

Although dosage compensation can be found in very different animals, X-inactivation is restricted to placental mammals (Eutheria) and marsupials (Methateria), and it can be found in two different forms. Marsupials (Methateria) present an imprinted form of X-inactivation by which the paternal X chromosome is always inactivated (Sharman, 1971). A random form of X-inactivation is found in somatic lineages of Eutherian mammals, where the two X chromosomes - paternal and maternal - in each female cell have the same probability of becoming inactivated, resulting in a mosaic of different cell populations with alternate parental X chromosomes active. Eutheria (placental mammals) are believed to have diverged from Methateria (marsupials) 170 Mya (Hedges and Kumar, 1998) and thus imprinted X chromosome inactivation is considered an ancient form of inactivation that evolved to coexist with random X-inactivation (Cooper, 1971; Huynh and Lee, 2005), since it is also found in cells of the extra-embryonic tissues of certain eutherians, namely the mouse (Takagi and Sasaki, 1975) and the cow (Xue *et al.*, 2002).

Altered patterns of randomness in X-inactivation have been observed, due to a selective advantage of one cell population (e.g. in the presence of a structurally abnormal X chromosome), rare mutations altering the inactivation pattern, or by chance, which can be enhanced if there is a decreased number of precursor cells (reviewed by Brown and Robinson, 2000).

In humans, cell lineages of the embryo proper are randomly X inactivated early in development, in a multi-step process that involves counting the number of X chromosomes per diploid autosome set, choice of the X chromosome(s) to silence and finally silencing of all except one of the X chromosomes in each cell. It is well established that initiation of X-inactivation is controlled by one locus on Xq13, the X-inactivation centre (*XIC*) where the *XIST* (X-Inactive-Specific Transcript) localizes (Brockdorff *et al.*, 1992; Brown *et al.*, 1992). This large non coding RNA is transcribed exclusively from the inactive X chromosome and triggers silencing in *cis* (Penny *et al.*,

1996) by selectively covering the chromosome (Clemson *et al.*, 1996). The mechanisms by which counting and choice occur are still poorly understood (Heard, 2004) but in the mouse other elements of the *Xic* seem to be involved, namely *Tsix*, the antisense transcript of *Xist* and the intergenic *Xite* (Lee *et al.*, 1999; Ogawa and Lee, 2003; Lee, 2005), as well as a trans-acting factor, CTCF (Chao *et al.*, 2002). It has also been proposed that the inactive X chromosome might be inherited in a pre-inactivated state from the paternal germ line (Huynh and Lee, 2003), although this hypothesis is not consensual (Okamoto *et al.*, 2005).

1.2.1. Epigenetic modifications of the inactive X chromosome

X-inactivation is considered to be one of the most illustrative examples of epigenetic signalling, whereby two virtually identical X chromosomes are regulated differently. The transcriptionally silent state of the inactive X chromosome, attained by the action of *Xist* during embryogenesis, is maintained and inherited clonally in subsequent cell divisions. The remarkable stability of somatic random X-inactivation is attributed to various epigenetic modifications that characterize the inactive X chromosome. Most notably, the inactive X chromosome resembles constitutive heterochromatin, being condensed in interphase (Barr and Carr, 1962), hypoacetylated in histone H3 and H4 (Jeppesen and Turner, 1993; Boggs *et al.*, 1996) and replicating late in S phase (Priest *et al.*, 1967). Other histone modifications include H3 methylation in Lys 9 (H3K9) and Lys 27 (H3K27), hypomethylation of H3 in Lys 4 (H3K4) and enrichment in histone variant macroH2A1 (Contanzi and Pherson, 1998; Boggs *et al.*, 2002; Plath *et al.*, 2003). Additionally, CpG islands of housekeeping genes subject to X-inactivation are methylated in the inactive X (e.g. Pfeifer *et al.*, 1990; Norris *et al.*, 1991; Luoh *et al.*, 1995; Jegalian and Page 1998; Matarazzo *et al.*, 2002). That *Xist* RNA, DNA methylation and histone hypoacetylation can act synergistically to maintain X chromosome inactivation was demonstrated by the work of Csankovski and collaborators (2001). A correlation between expression and relative time of replication of individual genes has also been reported (Schmidt and Migeon 1990; Boggs and Chinault, 1994), with silenced genes on the inactive X chromosome replicating later than their active homologues. Hansen and colleagues (1996) proposed a model for regulation of X-inactivated genes involving both late replication and promoter methylation.

Genes subject to X-inactivation exhibit differential methylation of promoter associated CpG islands, with hypermethylation of the inactive allele while genes that escape inactivation do not (Goodfellow, 1988). This correlation is usually attributed to

housekeeping genes, which are more frequently associated with CpG islands (Bird, 1986), but has also been shown for a tissue-specific X-linked gene (Huber *et al.*, 1999). Concordantly, the reactivation of genes on the inactive X is observed in human-rodent somatic cell hybrids, following treatment with the demethylating agent 5-azacytidine (5-azadC) (Mohandas *et al.*, 1981; Graves, 1982). DNA methylation is generally assumed to be a late event in X-inactivation (Lock *et al.*, 1987; Bird, 2002) following the transcriptional repression previously imposed by *Xist*. Therefore, DNA methylation seems to be involved in the maintenance of the inactive state and, moreover, its semiconservative mode of inheritance has led to the proposal for a role as a conveyer of epigenetic cell "memory" (Riggs, 1975; Bird, 2002), shared with the also appealing "histone code" that is featured by different histone modifications (Strahl and Allis, 2000; Turner, 2002).

Although X-inactivation represents a very effective and stable gene silencing, a small percentage of human genes (~10%) show heterogeneous patterns of inactivation in human-rodent somatic cell hybrids and in cell lines (Carrel and Willard, 1999 and 2005) and one particular gene (*TIMP1*) has been shown to be expressed from the inactive X chromosome in some females but silenced on the inactive X of others (Anderson and Brown, 1999). The mechanisms underlying such heterogeneity are still not clearly defined (Anderson and Brown, 2002 and 2005).

1.2.2. Strategies for assessing the inactivation status of X-linked genes

The inactivation status of a large number of X-linked genes has been determined recurring to many different approaches. Some focus on the epigenetic modifications that differentiate active and inactive X chromosomes such as promoter DNA methylation and were performed initially by means of methylation sensitive restriction endonucleases (Carrel *et al.*, 1996; Jegalian and Page, 1998) and more recently taking advantage of techniques that allow the analysis of virtually all CpG sites within a DNA sequence (Frommer *et al.*, 1992; e.g. Matarazzo *et al.*, 2002). Differential methylation in active and inactive X chromosomes has also been used to assay X-inactivation patterns, namely to detect non random or skewed X-inactivation (e.g. Allen *et al.*, 1992; Hendriks *et al.*, 1992).

Alternatively, the analysis of expression from active and/or inactive X chromosomes has been performed, using Reverse Transcriptase - PCR (RT-PCR) mainly in: 1) cell lines from females heterozygous for transcribed polymorphisms and presenting non-random X-inactivation (e.g. Vermeesch *et al.*, 1997; Carrel and Willard 1999 and 2005); 2) material from individuals with different numbers of inactive X

chromosomes (e.g. Fisher *et al.*, 1990; Sudbrak *et al.*, 2001); and 3) human-rodent somatic cell hybrids retaining one inactive X (Xi) or one active X chromosome (Xa) (e.g. Brown *et al.*, 1997; Carrel *et al.* 1999; Carrel and Willard, 2005). RNA *in situ* hybridization can also be used to detect transcripts from the inactive X chromosome in interphase nuclei from cell lines expressing the gene of interest (Carrel *et al.*, 1996).

In recent reports, microarray technology was used to compare the expression of a high number of X-linked genes in female and in male lymphocytes (Craig *et al.*, 2004) and in lymphoblastoid cell lines from normal males, normal females and individuals with supernumerary X chromosomes (Sudbrak *et al.*, 2001). Using human-rodent somatic cell hybrids Carrel and Willard (2005) assembled a comprehensive inactivation profile of human X-linked genes, including essentially all genes that are expressed in fibroblast-based assay systems. In spite of these breakthroughs, a few genes remain to be analysed, mainly those with tissue-specific expression, whose transcripts are undetectable in the cell lines used.

For the vast majority of X-linked genes, a single approach was used to assess its inactivation status and only in a few instances different assays were performed for the same gene (e.g. Carrel *et al.*, 1996; D'Esposito *et al.*, 1996; Matarazzo *et al.*, 2002; Schneider-Gadicke *et al.*, 1989; Luoh *et al.*, 1995). Moreover, considerable heterogeneity was observed by Carrel and Willard (2005) in the levels of expression from the inactive X between different genes and between cell lines from different individuals. Since the basis for this heterogeneity is still unknown it would be desirable to analyse, for each gene, epigenetic modifications related to X-inactivation in parallel with expression levels in a panel of different individuals. Moreover, since the heterogeneity detected *in vitro* will only have biological meaning if also observed *in vivo*, this type of analysis should include native tissues.

In the particular case of *PCDH11X* this kind of methodology is not only desirable but also necessary, as some of the more classical approaches are not adequate. In fact, *PCDH11X* expression could not be detected from human-rodent somatic cell hybrids by RT-PCR (A. Dagnall, personal communication). Although *PCDH11X* is expressed at extremely low levels in lymphoblastoid and glioma cell lines (Blanco *et al.*, 2000) only two transcribed polymorphisms with rare variants were found in European populations (~2%; Giouzezi *et al.*, 2003).

Although *PCDH11X/Y* gene structure and splice variants have been analysed in detail, not much is known about their regulation, and the structure of the putative alternative promoters has not been characterized. *In silico* prediction algorithms retrieved only a marginal probability for the presence of a promoter upstream the

transcription start site (TSS) in exon 1 since no TATA-box was identified. Nevertheless, a CpG island, coincident with the TSS and the associated putative promoter, is conserved in human, pig and mouse, encompassing several promoter motifs, including two Sp1 sites (Blanco-Arias *et al.*, 2004b). An additional CpG island was found in the 5' end of *PCDH11X/Y*, 2 Kb downstream of exon 1 (Blanco-Arias *et al.*, 2004a), but it is not conserved in mouse.

The activation of a gene involves different combinations of ubiquitous and cell-specific regulatory proteins that include transcription factors and RNA Polymerase II (Pol II), which interact with different promoter elements (Carey and Smale, 2000). These promoter elements consist of multiple short sequences that usually lie within 1Kb region upstream and surrounding a transcription start site (TSS). The core promoter encompasses DNA sequences between approximately -40 and +50 relative to the TSS (Smale, 1994 *in* Carey and Smale, 2000) and is a positioning element, capable of supporting basal levels of expression *in vitro*. Two of these elements are sufficient to direct transcription when stimulated by an upstream activator (Smale, 1997 *in* Carey and Smale, 2000): the TATA-box motif, to which the TATA-binding protein subunit of the general transcription factor TFIID directly binds, and the initiator element that is recognized by Pol II and the TAFII factor. Although most genes present one of these elements in their core promoter, at least one class of promoters appears to lack both TATA and Inr elements but instead contain a high G/C content and multiple binding sites for the mammalian transcription factor Sp1 (Smale, 1994 *in* Carey and Smale, 2000). Sp1 protein can bind GC-box regulatory elements in the promoter and regulate expression, probably acting as a basal transcription factor, helping recruit the TATA-binding protein (Pugh and Tjian, 1991). Although Sp1 transcription factor is expressed ubiquitously in mouse embryos its expression levels change at different developmental stages and in different cell types, suggesting a role in specific developmental processes, in addition to the general role in cell growth (Saffer *et al.*, 1991). It has been shown to regulate expression of both housekeeping (Gidoni *et al.*, 1985; Giglioni *et al.*, 1989; Imataka *et al.*, 1992) and tissue-specific genes (Bonny *et al.*, 1998; Zhang *et al.*, 1999; Bunick *et al.*, 1990) namely specific neural genes involved in neuronal differentiation (Yan and Ziff, 1997; Okamoto *et al.*, 2002; Yo *et al.*, 2002). The transactivation activity of Sp1 can be suppressed by negative transcription regulators (Mukhopadhyay *et al.*, 1997) and DNA methylation of the Sp1 binding sites (Zhao *et al.*, 2005).

It is probable that *PCDH11X* proximal putative promoter is regulated by Sp1 elements, since no TATA-box or initiator is present in the sequences immediately

adjacent to the TSS. A role for DNA methylation is also plausible as these regulatory elements are embedded within a CpG island.

Contrarily to the putative promoter upstream of exon 1, a TATA-box 446 bp upstream the human TSS in exon 4.1, that is common to all *PCDH11X* long isoforms, has been detected (Promoterscan, <http://www.bimas.cit.nih.gov>; Prestridge, 1995), although no CpG island is present in this region (Blanco-Arias *et al.*, 2004a).

2. The Xq21.3/Yp11.2 transposition and human–chimpanzee differences

A major goal of human evolutionary genetics is to unravel the genetic changes that have contributed to new biological characteristics in the lineage leading to humans. To do so, the best strategy is to compare, at the molecular level, humans with other closely related primates, most notably to our sibling species, the common chimpanzee (*Pan troglodytes*) and the bonobo or pygmy chimpanzee (*Pan paniscus*). It is generally accepted that humans and chimpanzees diverged ~5.5 million years ago (Takahata and Satta, 1997; Chen *et al.*, 2001; Kumar and Hedges, 1998) and the genetic events that shaped the emergence of humans should be identifiable within the fixed differences accumulated between these species.

The high nucleotide identity between human and chimpanzee genomes, more precisely determined by the release of the draft genome sequence of *Pan troglodytes* as 98.94% (Chimpanzee Sequencing and Analysis Consortium, 2005), reflects the few million years elapsed since the two species last shared a common ancestor (5–8 My). Additionally the same data has confirmed the early notion that differences in aminoacid sequences of human and chimpanzee proteins are low (King and Wilson, 1975), establishing that orthologous proteins differ typically by only two amino acids, one per lineage. In spite of the high similarity, human-specific mutations have been found in a few genes (Chou *et al.*, 1998; Winter *et al.*, 2001; Enard *et al.*, 2002a; Gilad *et al.*, 2003; Stedman *et al.*, 2004). Most of these are inactivating mutations (Chou *et al.*, 1998; Winter *et al.*, 2001; Gilad *et al.*, 2003; Stedman *et al.*, 2004), in agreement with the concept that genetic loss in the human lineage should have caused some differences between the two species (Olson, 1999). In other instances accelerated evolution of a gene or a functional domain, indicated by an elevation of the ratio of nonsynonymous substitutions per nonsynonymous site (K_a) to the synonymous substitutions per synonymous site (K_s), has been detected in the human lineage (Zhang *et al.*, 2003; Birtle *et al.*, 2005) and taken as evidence of adaptive positive selection. In recent reports (Clark *et al.*, 2003; Nielsen *et al.*, 2005; Chimpanzee Sequencing and Analysis Consortium, 2005) a set of genes associated with diverse

biological functions have been indicated as under positive selection in human and chimpanzee lineages, although the results from different studies are not in complete agreement on the number of genes under a stronger selective regime in the lineage leading to humans.

Additionally, alternative hypotheses to account for the rapid evolutionary change occurred after the human and chimpanzee lineages split have been put forward. One of these hypotheses refers to differences in gene regulation (King and Wilson, 1975) that would cause alterations in levels or patterns of expression of certain genes. Expression levels of a considerable number of genes have been measured in humans and non human primates in a set of microarray experiments. The pioneer work of Enard *et al.* (2002b) provided a first insight on possible differences in expression levels between humans and chimpanzees in blood, liver and brain tissue and revealed a 5.5-fold acceleration of the rate of change in gene expression levels in brain on the lineage leading to humans, mainly caused by a set of genes with increased expression in human brain. Reanalysis of this data has partially (Hsieh *et al.*, 2003) or fully supported (Gu and Gu, 2003) the initial results, which have been confirmed in independent experiments (Cáceres *et al.*, 2003; Khaitovich *et al.*, 2005). A higher percentage of induced genes in the evolutionary branch leading to humans have also been reported in cultured fibroblasts (Karaman *et al.*, 2003). By contrast Udin *et al.* (2004) found similar levels of regulatory evolution in brain in human and in chimpanzee lineages.

Khaitovich *et al.* (2005) compared the patterns of evolution of both protein-coding sequences and expression levels of genes in brain, heart, liver and kidney between humans and chimpanzees and found evidence for co-evolution of regulatory and structural protein changes.

Changes in homeotic genes, regulating development, and alterations in transcription factors can be regarded as some of the most radical "regulatory mutations" first proposed by King and Wilson (1975). Sequence differences in proximal promoters are also predicted to alter gene expression but in a recent study aiming to relate differential promoter activity between humans and the chimpanzee to differences in expression levels, only a minority of genes showed a positive correlation (Heissig *et al.*, 2005). Epigenetic modifications such as DNA methylation are also crucial for gene regulation (Jaenisch and Bird, 2003) and differences in methylation at a few specific CpG sites have been found between human and chimpanzee brain tissue (Enard *et al.*, 2004), although the meaning of such differences is not clear. Interestingly, a gene conversion event involving the regulatory region of one sialic acid binding receptor

gene (*SIGLEC11*) has apparently contributed to the gain of human-specific expression in brain cortex microglia (Hayakawa *et al.*, 2005).

A possible role for genomic rearrangements in changes in gene expression is particularly appealing since these events can alter the position of genes relatively to regulatory sequences controlling transcription rate and patterns of expression, such as promoters and enhancers/repressors. In some instances such rearrangements may result in gene duplications, altering gene dosage. Cytogenetic studies have shown that human and chimpanzee karyotypes differ in the number of chromosomes, since human chromosome 2 results from the fusion of two small chromosomes seen in great apes (Yunis *et al.*, 1980). Additionally several inversions, insertions and deletions are exclusive of human or chimpanzee lineages (Cheng *et al.*, 2005; Newman *et al.*, 2005) and, most notably, there is a considerable number of lineage-specific duplications of genomic segments ranging from one to several Kb in length, referred to as segmental duplications (Bailey *et al.*, 2002; Fortna *et al.*, 2004; Cheng *et al.*, 2005). The finding of an association between human-specific duplications and upregulation of gene expression in humans (Khaitovich *et al.*, 2004; Cheng *et al.*, 2005), highlights the importance of genome plasticity in human evolution. It has also been proposed that aminoacid substitution would be faster in rearranged chromosomes due to the accumulation of positively selected alleles (Navarro and Barton, 2003) although this has failed to be fully confirmed, in a larger dataset (Chimpanzee Sequencing and Analysis Consortium, 2005).

Noteworthy are the large blocks of sequence transposed from the X to the Y chromosome in the human lineage, already characterized in previous sections - the Xq/Yq pseudoautosomal region or PAR2 and the Xq21.3 to Yp11.2 transposed region. Several genes within these regions have been duplicated only in humans - *HSPRY3*, *SYBL1*, *IL9R* and *CXYorf1* within PAR2; *TGIF2L* and *PCDH11* in Xq/Yp. However, due to sex-specific linkage, it is not straightforward to assume that such duplications resulted in altered gene dosage in the human lineage. The inactivation status of the X-linked genes in both species and the degree of allele sharing between sexes of the X- and Y-linked copies must be taken into account. In the case of PAR2 genes, there is no sex-specific linkage, X and Y isoforms are identical and could account for differences in expression levels between humans and chimpanzees. However, this does not seem to be the case for two of PAR2 genes, *SYBL1* and *HSPRY3*, since as previously referred, they are both silenced on the inactive X in females and on the Y chromosome in males in humans. Assuming that in chimpanzees *Syb11* is also subject to X-inactivation it is

reasonable to assume that gene dosage has been maintained after the transposition. *IL9R* and *CXYorf1* escape from X-inactivation in humans but, considering the arguments previously exposed, their status in the great apes is not easily predictable.

The human-specific Xq/Yp homology is particularly interesting, considering that the *PCDH11X/Y* gene pair is mainly expressed in brain (Blanco *et al.*, 2000; Blanco-Arias *et al.*, 2004a). A theory of the speciation of modern *Homo sapiens* has been put forward, where the Xq21.3/Yp11.2 transposition has a crucial role (Crow, 2002). According to this theory, *PCDH11X/Y* are the determinants of brain hemispheric asymmetry, a defining feature of the human brain and a putative correlate of language. Both changes in *PCDH11X/Y* sequence and expression levels between humans and chimpanzees could account for human uniqueness, relating to the referred characteristics. Consecutive phases of sexual selection, both by male and female mate choice were invoked to explain the establishment of such changes at the species level.

A detailed debate on the assumptions of the speciation thesis sustained by Crow is not under the scope of the present dissertation. Here it will be merely addressed the possible contribution of an alteration of gene dosage caused by the Xq to Yp duplicative transposition to differences in *PCDH11X/Y* expression levels between humans and chimpanzees. Assuming that in non human primates and other mammals the strict X-linkage of *Pcdh11X* results in X-inactivation of the gene in females, subsequently to the human-specific gene duplication, an adjustment of gene dosage between sexes might have occurred, compensating the gain of an extra copy of the gene on the Y chromosome in humans. If in fact the inactivation status of the gene has changed after the transposition event, this might have contributed to the existence of differences in *PCDH11X* expression levels in brain between humans and chimpanzees, since humans would have two active copies of the gene (on both X chromosomes in females and on the X and Y in males), while in chimpanzees only one *Pcdh11X* gene copy would be transcribed in both sexes. This hypothesis has the premise of functional equivalence of X- and Y-linked gene forms, even if transiently during hominid evolution, since otherwise the gain of a Y homologue in the human lineage would not contribute to increase the overall expression level of *PCDH11* transcripts and in this case a strict change of gene dosage could not be invoked.

The possible impact of the inactivation status of *Protocadherin11X* to sexual dimorphisms in normal individuals in humans as well as its relevance in sex chromosomes aneuploidies will also be considered. The discussion of the first part of the dissertation will be finalized with a brief glance on gene evolution following the transposition and the possible outcomes of the addition of new genes to a male-specific region.

AIMS

1. To establish the inactivation status of *PCDH11X* pre- and post-transposition by analysing:
 - a) CpG island methylation levels in human and in chimpanzee tissues (blood leukocytes and brain cortex);
 - b) expression levels in human and chimpanzee brain cortex.
2. To relate methylation levels in *PCDH11X* promoter region with histone modifications, in human and great ape (*Gorilla gorilla*) female lymphoblastoid cell lines.
3. To analyse the methylation levels of *PCDH11X* in sex chromosome aneuploidies: Klinefelter's syndrome (XXY).

RESULTS

1. Inactivation status of *PCDH11X* in humans and in the chimpanzee (*Pan troglodytes*)

Article 1

Lopes AM, Ross N, Close J, Dagnall A, Amorim A, Crow TJ. Inactivation status of *PCDH11X*: sexual dimorphisms in gene expression levels in brain. Hum Genet (in press)

Alexandra M. Lopes · Norman Ross · James Close
Adam Dagnall · António Amorim · Timothy J. Crow

Inactivation status of *PCDH11X*: sexual dimorphisms in gene expression levels in brain

Received: 21 September 2005 / Accepted: 23 November 2005
© Springer-Verlag 2006

Abstract Genes escaping X-inactivation are predicted to contribute to differences in gene dosage between the sexes and are the prime candidates for being involved in the phenotype observed in individuals with X chromosome aneuploidies. Of particular interest is *ProtocadherinX* (*PCDH11X* or *PCDHX*), a recently described gene expressed in brain. In humans, *PCDH11X* has a homologue on the Y chromosome and is predicted to escape from X-inactivation. Employing bisulphite sequencing analysis we found absence of CpG island methylation on both the active and the inactive X chromosomes, providing a strong indication that *PCDH11X* escapes inactivation in humans. Furthermore, a sexual dimorphism in levels of expression in brain tissue was observed by quantitative real-time PCR, with females presenting an up to 2-fold excess in the abundance of *PCDH11X* transcripts. We relate these findings to sexually dimorphic traits in the human brain. Interestingly, *PCDH11X/Y* gene pair is unique to *Homo sapiens*, since the X-linked gene was transposed to the Y chromosome after the human–chimpanzee lineages split. Although no differences in promoter methylation were found between humans and chimpanzees, evidence of an

upregulation of *PCDH11X* in humans deserves further investigation.

Introduction

The equalization of gene dosage between males and females in mammals is achieved by the silencing of most genes on one of the X chromosomes in females (X inactivation). This chromosome-wide gene silencing is accomplished through a cascade of events initiated by Xist (Penny et al. 1996) and involving a synergism with other epigenetic modifications that characterize the inactive X chromosome, such as DNA methylation, histone hypoacetylation and late replication (Lee and Jaenisch 1997; Csankovszki et al. 2001).

Assessment of the inactivation status of a large number of X-linked genes has been performed using many different approaches. Some rely on the analysis of expression from active and inactive X chromosomes mainly in (1) cell lines from heterozygous females presenting non-random inactivation (e.g. Vermeesch et al. 1997; Carrel and Willard 1999, 2005); (2) individuals with different numbers of inactive X chromosomes (e.g. Fisher et al. 1990; Sudbrak et al. 2001) and (3) rodent–human somatic cell hybrids retaining one Xa or one Xi (e.g. Brown et al. 1997; Carrel et al. 1999; Carrel and Willard 2005). Other more indirect approaches focus on the epigenetic modifications that differentiate active and inactive X chromosomes, namely promoter DNA methylation and histone modifications (Carrel et al. 1996; Jegalian and Page 1998; Matarazzo et al. 2002). In recent reports, microarray technology was used to compare the expression of a high number of X-linked genes in female and in male lymphocytes (Craig et al. 2004) and in lymphoblastoid cell lines from normal males, normal females and individuals with supernumerary X chromosomes (Sudbrak et al. 2001).

A comprehensive inactivation profile of human X-linked genes has also been published, including

Electronic Supplementary Material Supplementary material is available for this article at <http://dx.doi.org/10.1007/s00439-006-0134-0> and is accessible for authorized users.

A. M. Lopes (✉) · A. Amorim
IPATIMUP, Instituto de Patologia e Imunologia Molecular da
Universidade do Porto, R. Dr Roberto Frias, S/N,
4200-465 Porto, Portugal
E-mail: alopes@ipatimup.pt
Tel.: +351-2-25570700
Fax: +351-2-25570799

A. M. Lopes · A. Amorim
Faculdade de Ciências, Universidade do Porto,
4099-002 Porto, Portugal

N. Ross · J. Close · A. Dagnall · T. J. Crow
Department of Psychiatry, POWIC SANE Research Centre,
University of Oxford, Warneford Hospital,
OX3 7JX Oxford, UK

essentially all genes that are expressed in fibroblast-based assay systems (Carrel and Willard 2005). The results from this survey support the long-standing model of "Y driven" evolution of X inactivation (Lyon 1961; Ohno 1967; Charlesworth 1996; Jegalian and Page 1998), since the vast majority of human X-linked genes with functional homologues on the Y chromosome were found to escape inactivation, while most of the X-linked genes with no homology on the Y chromosome are inactivated. Nevertheless, the existence of a considerable number of exceptions to this model (Brown et al. 1995; Carrel et al. 1996; Carrel and Willard 2005) highlights the importance of studying the regulation of each individual X-Y gene pair.

Of particular interest is *Protocadherin11* X/Y (*PCDH11X/Y* or *PCDH11Y*), a human-specific gene pair, located in the Xq21.3/Yp11.2 homology block (Yoshida and Sugano 1999; Blanco et al. 2000), where the X-linked gene became transposed to the Y chromosome after the human-chimpanzee lineages split (Page et al. 1984; Schwartz et al. 1998; Skaletsky et al. 2003). The gain of a Y-linked homologue could have contributed to the escape from X-inactivation of *PCDH11X* in humans but to date there is only weak evidence for the escapee status of this gene, from experiments in lymphoblastoid cell lines (Sudbrak et al. 2001). Since *PCDH11X* is expressed mainly in brain (Blanco et al. 2000; Blanco-Arias et al. 2004a) additional analyses are needed.

Genes escaping X-inactivation are predicted to contribute to differences in gene dosage between the sexes (Brown and Greally 2003). Moreover, since *PCDH11X/Y* gene pair is unique to *Homo sapiens* we hypothesize that it should escape X-inactivation only in humans, in which case this gene duplication could have contributed to the existence of differences in gene expression levels in brain between species.

In the present study we aimed to establish the inactivation status of *PCDH11X*, analysing in humans and in the chimpanzee the methylation levels of two CpG islands in the 5' end of *PCDH11X/Y* genes, one of them associated with a putative promoter and reported to be highly conserved between human and mouse (Blanco-Arias et al. 2004b). Additionally, *PCDH11X* expression levels were analysed in brain tissue of female and male individuals and correlated with the observed levels of methylation.

Materials and methods

Methylation analysis by sodium bisulphite sequencing

Genomic DNA was extracted from human and non-human (*Pan troglodytes*) blood leucocytes by standard procedures and from brain tissue (grey matter) using Nucleon Soft Tissue DNA Extraction Kit, to be subjected to sodium bisulphite modification with CpG Genome DNA Modification Kit (Chemicon International) and used as template in PCR. A control sample

included in each bisulphite modification procedure was tested for a methylated region in the genome, to assure that the treatment was sensitive to methylation. Four fragments encompassing the whole *PCDH11X* CpG island 1 (~600 bp) and two fragments within CpG island 2 (~600 bp) were analysed using primers for the sense DNA strand (5'-3'). External and internal primers were designed in CpG-free regions or, when including a single CpG site, this position was synthesized as a degenerate base (C + T in forward primers and A + G in reverse primers), in order to avoid an amplification bias towards methylated or unmethylated alleles. Typically 40 ng of bisulphite-modified genomic DNA were amplified in a 25 µl reaction volume comprising Promega Master Mix (Promega), according to the manufacturer's protocol, and 0.3 µM of each primer. Reaction conditions for PCR were the following: 5-min pre-incubation step at 94°C, 35 cycles of denaturation at 94°C for 30 s, annealing for 1 min at the respective AT for each primer pair (see Electronic Supplementary Material) and extension at 72°C for 1 min and 30 s, followed by a final extension step at 72°C for 10 min. For 5' end CpG 1 fragments, an additional round of PCR with internal primers was necessary to obtain robust PCR products suitable for cloning. All PCR products were purified with QIAquick PCR purification kit (Qiagen), ligated into pGEM-T vector and transformed into JM109 *Escherichia coli* competent cells (Promega). Colonies were grown in LB-agar plates under ampicillin and blue/white selection. Insert-containing colonies were cultured and plasmid DNA was isolated, using GenElute Plasmid Miniprep Kit (Sigma), and used as template in a dideoxy cycle sequencing reaction using the Big Dye Terminator Cycle Kit (PE Applied Biosystems) and primers SP6/T7. Since better results were obtained when sequencing the reverse strand of inserts, the orientation of each fragment in the plasmid was determined previously to the sequencing reaction, using combinations of vector and insert primers. The products were purified using SigmaSpin Post-Reaction Clean-Up columns (Sigma) and run in an ABI 3100 sequencer (PE Applied Biosystems). The results were analysed using the Data Collection Software. For each sample a minimum of five and a maximum of ten clones were sequenced, to determine the overall level of methylation within *PCDH11X/Y* CpG islands in male and female samples. Although the primers used did not differentiate X and Y alleles, for some fragments it was possible to determine the chromosome of origin of each clone in male samples, allowing to assure that there was no detectable amplification or insertion bias.

Quantification of brain cDNAs by real-time PCR

Brain samples

Human brain tissues were obtained from the MRC London Neurodegenerative Diseases Brain Bank,

Department of Neuropathology/Neuroscience, Institute of Psychiatry. Brains were removed in autopsy, snap-frozen and sectioned by structural areas and stored at -80°C until further handling. For this study, grey matter was dissected from right superior temporal gyrus sections by briefly transferring the brain tissue to dry ice, to allow handling. Grey matter collected was immediately used for RNA extraction. All subjects used in this study (four females and five males) are from a pool of controls that have been established for future studies in which they will be compared with subjects with psychiatric disorders. These controls were selected to meet criteria such as brief agonal state (rapidity of death) and reduced post-mortem interval (period of time between death and tissue collection).

Pan troglodytes brain samples were kindly provided by Dr. Philipp Khaitovitch from Max-Planck Institute, Leipzig, Germany. Additionally, one female *Pan troglodytes* DNA sample was kindly provided by Carole Sargent from the Department of Pathology, University of Cambridge (UK).

RNA extraction and first-strand synthesis

Total RNA was extracted from brain samples using RNeasy Lipid Tissue Mini Kit (Qiagen) and quality verified by gel electrophoresis and optical densitometry. After DNase I treatment, $1\text{ }\mu\text{g}$ of total RNA was used as template for first-strand cDNA synthesis. RNAs were first denatured at 65°C for 5 min and cooled on ice. Reactions were incubated at 25°C for 5 min to allow primer annealing and at 60°C for 50 min in a final volume of $20\text{ }\mu\text{l}$ with reaction buffer (as supplied by the manufacturer), 5 mM DTT, 1 mM of each dNTP, 0.5 U of RNaseOUT, 2.5 μM of oligodT and 20 U of Super Script III RNase H-Reverse Transcriptase (Invitrogen). A final step of 15 min at 60°C was included to inactivate the enzyme.

Real-time PCR

Chromosome-specific primers were used for assessing the quantity of *PCDH11X* transcripts in males and females and primer specificity was confirmed by direct sequencing of real-time PCR products. Primers for *PCDH11X* were designed to amplify across exon 4 and exon 5, two exons that are common to all transcripts, in areas of conserved sequence between human and chimpanzee (determined by a Blat search available <http://www.genome.ucsc.edu/>). Quantitative real-time PCR reactions were carried out using $1\text{ }\mu\text{l}$ of the resulting cDNA in a reaction volume of $20\text{ }\mu\text{l}$ comprising Sybr Green PCR Master Mix (according to the manufacturer's protocol) and 0.3 μM of each primer in an Applied Biosystems 7000 detection system (ABI, Foster City), using standard PCR conditions and annealing/extension steps of chromosome-specific

primers were performed at 62°C , to ensure specificity. Each reaction was performed in triplicate and serial dilutions of human brain cDNA obtained from $1\text{ }\mu\text{g}$ of total RNA were used to generate a standard curve of cDNA quantity versus fluorescence signal, for both genes analysed. All primers were designed to span exon-exon boundaries, to prevent amplification of possible genomic DNA contaminants. Prior to setting up the quantification PCRs, the primers were used to verify the presence of a single band of the proper size on agarose gel and each product was sequenced. A dissociation protocol was always included in each real-time PCR protocol to ensure that no spurious PCR product was present in the final reaction product. *GAPDH* was used as an endogenous control for the amount of cDNA in each sample. All samples values subjected to statistical analysis (human and *Pan troglodytes*) have been determined in a single reaction, although partial independent replicates were also performed.

Statistical analysis

The equality of *PCDH11X* expression levels between sexes and between species was tested by means of the a non-parametric test (Mann-Whitney), since the assumption of a normal distribution for quantity of cDNA in the samples, as required for applying a two sample *t* test, could not be assured. As this is a conservative test, a complementary and more powerful analysis, which makes no assumptions regarding normality, was performed. This analysis relies on a Monte Carlo permutation test and was performed using the Stand-alone software (Simon 2000; available at <http://www.resampling.com>), where 1,000 randomization steps were implemented, in each comparison. For simplicity this method is referred to as Resampling.

Results

Methylation analysis

Based on the characterization of *PCDH11X/Y* mRNA transcripts, Blanco-Arias et al. (2004a) proposed the existence of two potential alternative promoters in humans, one in exon 1 and other in exon 4.1. Although in silico prediction algorithms retrieved only a marginal probability for the presence of a promoter upstream the TSS in exon 1, since no TATA-box was identified, a CpG island encompassing the transcriptional start site and several promoter motifs, including two Sp1 sites, was found to be conserved in human, pig and mouse (Blanco-Arias et al. 2004b). An additional CpG island was detected on the 5' end of the gene, 2 Kb downstream of exon 1 and overlapping exon 1a (Blanco-Arias et al. 2004a), which is poorly conserved in mouse (our observation by BLAST; <http://www.ncbi.nlm.nih.gov/BLAST>); no CpG

island was identified upstream of exon 4.1 (Blanco-Arias et al. 2004a). We analysed DNA methylation on both CpG islands on the 5' end of the gene (exon 1—hereafter referred as CpG island 1; downstream exon 1—referred as CpG island 2) in brain cortex (grey matter) of one male and one female in humans and in the chimpanzee. Additionally, methylation levels were also analysed on CpG island 1 in blood leucocytes in humans and in female chimpanzee.

Methylation levels of approximately 50% in female samples, i.e., half of the clones methylated, would be taken as evidence that the gene is subjected to X-inactivation in females, while absence of methylation would indicate an escape from inactivation. No methylation was expected in brain tissue in male samples since both X and Y transcripts have been detected in this tissue (Blanco et al. 2000; Blanco-Arias et al. 2004a). The analysis of methylation levels in human blood leucocytes was included for CpG island 1, to allow the comparison of brain with a non-expressing tissue.

For all samples and tissues analysed, the vast majority of CpG sites in both CpG islands was completely unmethylated and the few sites found to be methylated were, in most cases, in only approximately 10–20% of the clones analysed (Figs. 1, 2). In CpG island 1 all *Pan troglodytes* tissues analysed (male and female brain; female blood leucocytes) were completely devoid of methylated CpG sites (data not shown); the same was observed in human male brain in CpG island 2. Notably,

the region around exon 1 transcriptional start site was hypomethylated in all tissues in both species.

Quantification of brain *PCDH11X* cDNA by real-time PCR

The levels of *PCDH11X* expression in brain tissue (grey matter), normalized to *GAPDH*, were measured in humans (four females and five males) and in chimpanzees (two females and one male). Human females had higher levels of *PCDH11X* expression than males, resulting in female to male ratios of 2.6 and 2.5 (two experimental replicates; Table 1) and this difference was statistically significant ($P=0.014$ and $P=0.028$, respectively, Mann–Whitney test; $P=0.006$ for both comparisons, Resampling—see Materials and methods). It should be noted, however, that in both groups the mean standard deviation is considerable (Table 1), reflecting a high heterogeneity among samples, within sexes.

Although the mean expression levels of chimpanzee females were lower than the male chimpanzee (0.167 and 0.285, respectively), these comparisons are not informative, due to the reduced number of samples analysed (two females and one male). To test if *PCDH11X* escapes inactivation (h1) only in humans or (h2) in humans and in chimpanzees, several between-species comparisons were performed. Assuming hypothesis 1 (h1) male and female chimpanzees and human males

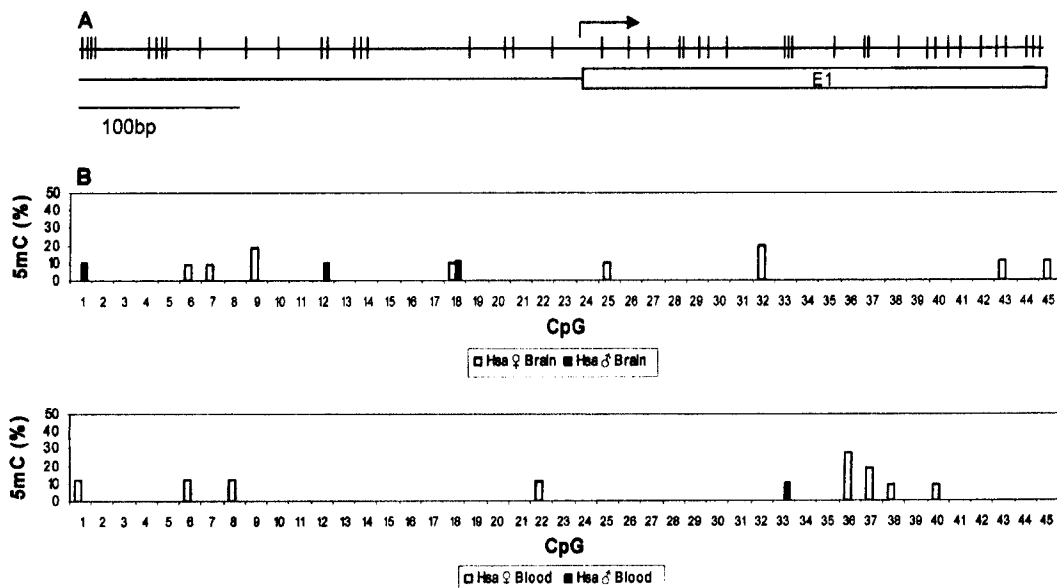


Fig. 1 Methylation levels in *PCDH11X* CpG island 1 in humans. **a** Schematic representation of human *PCDH11X* CpG island 1. Vertical lines in the scheme represent CpG sites; the transcriptional start site is symbolized by an arrow. **b** Percentage of cytosine

methylation at individual CpG sites (number of clones in which the CpG site was methylated) in males (*PCDH11X/Y*) and females (*PCDH11X/X*) in two different tissues (brain and blood leucocytes), in humans. The CpG sites are numbered 5' to 3'.

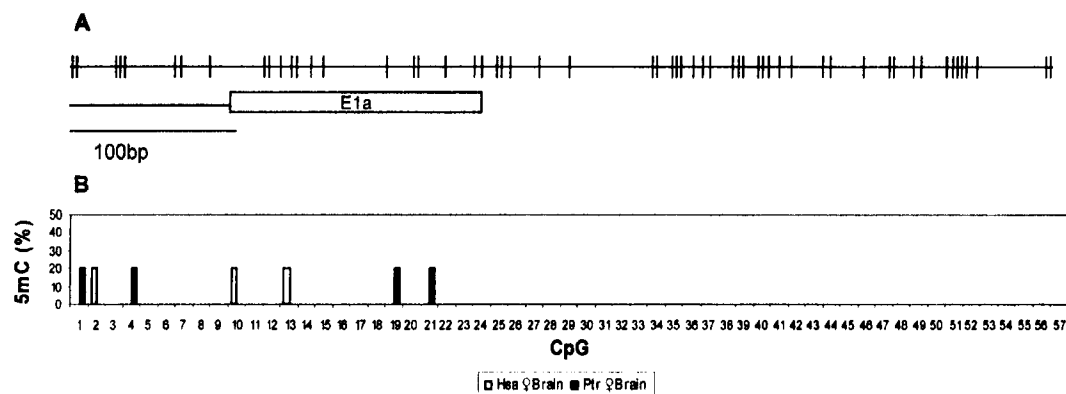


Fig. 2 Methylation levels in *PCDH11X* CpG island 2 in human and chimpanzee female brain. **a** Schematic representation of *PCDH11X* CpG island 2. Vertical lines in the scheme represent CpG sites. **b** Percentage of cytosine methylation at individual CpG

sites (number of clones in which the CpG site was methylated) in human and chimpanzee female brain. The CpG sites are numbered 5' to 3'; the correspondence between CpG sites in human and chimpanzee sequence is not complete

are expected to have one active copy of the gene, whereas human females would have two active copies of *PCDH11X* if the gene escapes from inactivation. When grouped together, expression levels of male and female chimpanzees were 4.6-fold (0.962/0.207) lower than human female, the difference being significant ($P=0.034$, Mann-Whitney; $P=0.008$, Resampling) but did not differ significantly from the human male expression levels ($P=0.456$, Mann-Whitney; $P=0.178$, Resampling).

However, assuming (h2), i.e. *PCDH11X* escapes from X-inactivation in both species, human and chimpanzee females would not differ in gene copy number. When these two groups were compared, the level of expression in human females was 5.8-fold (0.962/0.167) greater than in female chimpanzees indicating that expression in human females is greater than female chimpanzees. The difference did not reach statistical significance at the 0.05% level by the Mann-Whitney test ($P=0.064$), likely due to the lack of power of this test when comparing a reduced number of samples, but was highly significant by a Resampling analysis ($P=0.018$). These

results support (h1), *PCDH11X* escapes inactivation only in humans, although the upregulation observed for this gene in human females compared to chimpanzees was apparently much higher than the 2-fold difference expected.

In view of the manifold difference in *PCDH11X* expression levels between humans and chimpanzees we compared also *GAPDH* expression between species. Higher mean *GAPDH* expression levels were found in chimpanzees, compared to humans ($P=0.079$ by Mann-Whitney; $P=0.012$ by Resampling). This difference is thus contributing to the approximately 5-fold difference in *PCDH11X* expression levels between species since these expression levels are in fact ratios of the mean quantity of *PCDH11X* to the endogenous control. As a correction of the bias introduced by *GAPDH* differences into estimated *PCDH11X* expression levels, the mean quantity of *GAPDH* transcripts in human brain was used to normalize the levels of *Pcdh11X* expression in chimpanzee brain. As predicted by hypothesis (1) described above, the ratio of mean *Pcdh11X* (male and female chimpanzees grouped together) to mean human

Table 1 *PCDH11X* expression levels in human (*Hsa*) and chimpanzee (*Ptr*) brain

	Humans				Chimpanzees			
	F ($n=4$)		M ($n=5$)		F ($n=2$)		F and M ($n=3$)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
R#1	0.962 ^{a,b,d}	0.265	0.364 ^{c,d}	0.233	0.167 ^a	0.031	0.207 ^{b,c}	0.071
R#2	1.394 ^c	0.234	0.554 ^c	0.414	–	–	–	–

Measurements were performed by real-time PCR and normalized to *GAPDH*, in two experimental replicates (R#1 and R#2). The single male chimpanzee had a value of 0.285 and female chimpanzees had values of 0.190 and 0.145, respectively.

P values refer to Mann-Whitney test and Resampling analysis, respectively; superscripts show which group comparisons refer to. ^a $P=0.064$, ^b $P=0.018$; ^{b*} $P=0.034$, ^{*} $P=0.008$; ^c $P=0.456$, $P=0.178$; ^{d*} $P=0.014$, ^{*} $P=0.006$; ^{c*} $P=0.028$, ^{*} $P=0.006$

SD standard deviation, F female samples, M male samples, F and M combined male and female samples

GAPDH levels was similar to the *PCDH11X/GAPDH* ratio of human males (0.603 versus 0.554). However, the correspondent comparison of human female to chimpanzee female mean values (1.394 vs 0.758) and to the mean values obtained by considering chimpanzees as a group (1.394 vs 0.603) resulted in an approximately 2-fold difference (1.893 and 2.312 ratios, respectively). These results suggest that *PCDH11X* is upregulated in humans, even after correcting for between-species differences in *GAPDH* expression.

Discussion

It is generally accepted that the mammalian X and Y chromosomes were derived from an originally homologous pair of autosomes and that the Y has undergone progressive genetic decay (Ohno 1967). As genes on the Y decayed their X-linked homologues are assumed to have acquired dosage compensation through the mechanism of X-inactivation (Jegalian and Page 1998), although a few exceptions have also been identified (Brown et al. 1995; Carrel et al. 1996; Carrel and Willard 2005). Some X/Y homologies have arisen more recently in evolution as a result of duplicative translocations from the X to the Y chromosome. Two blocks of genes are known to have been so established in hominid evolution—the Xq/Yq pseudoautosomal region (PAR2) and the Xq21.3 to Yp translocation. Within PAR2 two genes, *SYBL1* and *HSPRY3*, are X-inactivated in females and undergo Y-specific inactivation in males, while the more telomeric genes, *IL9R* and *CXYorf1*, escape from X-inactivation (D'Esposito et al. 1996, 1997; Vermeesch et al. 1997; Ciccodicola et al. 2000).

Here we address the status with respect to X-inactivation of *PCDH11X*, considering that a Y-linked homologue arose through the Xq21.3/Yp duplicative translocation at 4 MYA in the hominid lineage. Our analysis focused on two indices of epigenetic modification, i.e. DNA methylation and X-linked levels of expression:

1. The absence of CpG island methylation on both active and inactive chromosomes in all tissues analysed is a strong indication that *PCDH11X* escapes inactivation in humans. Although the alternative promoter in exon 4.1 of *PCDH11X* lacks a CpG island and therefore is not amenable to methylation analysis, exon 4 is unlikely to be regulated independently of the exon 1 promoter (Pedone et al. 1994).
2. The findings from quantification of *PCDH11X* expression in human brain also suggest escape from X-inactivation, as a 2-fold excess of expression in females compared to males was detected.

These findings are in agreement with Sudbrak et al. (2001) who reported hybridization signals for *PCDH11X* with cDNA from lymphoblastoid cell lines, which were stronger in individuals with supernumerary X chromo-

somes. Our findings corroborate and extend these observations by quantitative real-time PCR in the relevant tissue, i.e. brain. By contrast Galfalvy et al. (2003) reported upregulation of *PCDH11X* expression in the prefrontal cortex in males using U133A oligonucleotide microarrays. However, the *PCDH11X* probe included in this array has a 100% sequence identity with the *PCDH11Y* homologous region, and the finding therefore probably reflects cross hybridization of the probe with *PCDH11Y* transcripts (Vawter et al. 2004). Thus the consensus of findings from these studies points to an imbalance of *PCDH11X/Y* gene products between the sexes, with *PCDH11X* transcripts present at a higher level in female, and *PCDH11Y* transcripts in male brains. *PCDH11Y* expression was detected in the same brain region here analysed for *PCDH11X* using Y-specific primers (data not shown), although the relative abundance of X- and Y-linked transcripts was not determined. *PCDH11X* dosage imbalance in female brain reported here is likely due to escape from X-inactivation and thus the sex differences found are expected to be shared by different cortical regions.

Protocadherin11X has 98.1–99.1% nucleotide identity (short and long isoforms, respectively) and 98.3% aminoacid identity with *PCDH11Y* (Blanco-Arias et al. 2004a), figures that reflect some differences between X and Y transcripts including the absence of exons 7 and 8 from *PCDH11Y*, the presence of a 13 bp deletion within exon 4 of *PCDH11Y* and a stop codon in exon 6 that gives rise to a splice variant with nine aminoacids missing from the cytoplasmic tail (Blanco et al. 2000). Either one of these differences or a difference in gene dosage could account for a sexual dimorphism, such as been reported for X/Y genes in mouse brain (Xu et al. 2002, 2005).

The function of *PCDH11X/Y* transcripts in brain is still not known but it seems probable that, like other cadherins expressed in the central nervous system these genes are involved in the regionalization and functional differentiation of brain grey matter. In fact, sequence comparisons have led Blanco et al. (2000) and Blanco-Arias et al. (2004a) to propose that *PCDH11X/Y* genes belong to the CNR (cadherin-related neuronal receptors) subgroup of protocadherins, which localize to the synaptic junction and may be involved in the establishment of neuronal connections or in signal transduction at the synaptic membrane (Wu and Maniatis 1999; Yagi and Takeichi 2000).

More recently, *PCDH11X/Y* have been assigned to a new subfamily of protocadherin genes present both in man and in the mouse (δ -Protocadherins; Vanhalst et al. 2005; reviewed in Redies et al. 2005), characterized by the presence of some distinctive features such as seven extracellular cadherin domains and the conservation of a short sequence motif (RRVTF) in their cytoplasmic domain referred to as CM3 motif, which mediates their interaction with PP1 α (Vanhalst et al. 2005), a protein implicated in regulating synaptic plasticity (Munton et al. 2004).

Most notably, in man the brain is sexually dimorphic with respect to asymmetry (the cerebral torque) from right frontal to left occipital across the anteroposterior axis (Barrick et al. 2004). Females are more strongly right-handed than males, acquire words more rapidly (Crow et al. 1998) and have faster brain growth (Kretschmann et al. 1979). By contrast, adult brain size is greater in males than females, cerebral asymmetry is more marked, particularly in the posterior part of the brain (Barrick et al. 2004) and there is a modest male superiority for spatial ability. Although most sexual dimorphisms are caused by male and female gonadal secretions, there is evidence for a direct role of sex chromosomes genes in brain sexual differentiation (Carruth et al. 2002; reviewed in Arnold 2004). In this context, an imbalance of *PCDH11X/Y* products between sexes might be relevant considering that these genes are expressed from early in development (Blanco et al. 2000).

Considerable variation in gene expression was noted, and has also been reported in microarray studies confined to male subjects (Enard et al. 2002; Khaitovich et al. 2004). X-linked gene expression has been found to vary extensively between individual females on account of incomplete X-inactivation of some genes (Carrel and Willard 2005). Allelic variation in expression between individuals has also been described in lymphoblastoid cell lines (Yan et al. 2002; Morley et al. 2004).

We found *PCDH11X* expression in human females to be approximately 2-fold higher than that of chimpanzees, when corrected for between-species *GAPDH* differences in expression. Given that we find no difference in DNA methylation, this raises the possibility that *PCDH11X* expression is controlled by a different mechanism. Alternatively, the upregulation of *PCDH11X* observed in human females could be due to a higher transcriptional rate, without differences in gene dosage, although the comparable *PCDH11X* expression levels presented by chimpanzees and human males make this possibility less forceful. Future experiments should focus on other epigenetic modifications, such as increased histone methylation and deacetylation, to explore alternative regulatory mechanisms of *PCDH11X* expression. It would be extremely valuable to extend these comparisons to a larger number of samples and to include a panel of different control genes.

PCDH11X expression levels have been estimated in human and in chimpanzee brain, in various microarray studies (Enard et al. 2002; Cáceres et al. 2003; Khaitovich et al. 2004; Uddin et al. 2004). The findings were somewhat discrepant between experiments, and this may be attributable to the small number of samples analysed in some studies and to the difficulty of distinguishing between hybridization to *PCDH11X* and *PCDH11Y*, as previously discussed.

In conclusion, we present evidence for the escape of *PCDH11X* from X-inactivation in humans and for the existence of a dosage imbalance in brain of X-linked

transcripts between males and females. Although *PCDH11X* has a homologue on the Y chromosome, which is expressed in the same brain region and thus could compensate for the observed imbalance, several differences distinguish *PCDH11X* from *PCDH11Y*, with possible functional impact on the encoded proteins. It is tempting to speculate on the possible consequences of the expression differences observed between sexes, namely how they could relate to sexually dimorphic traits. The findings of *PCDH11X* upregulation in the human brain deserve further investigation.

Acknowledgments The authors would like to acknowledge Patricia Blanco-Arias and Carole Sargent for their valuable contribution to this work. We also would like to thank Luis Teixeira da Costa and Raquel Seruca for scientific advice and Luis Cirnes and Marta Novais for technical support. We are grateful to Dr. P. Khaitovich, Dr. P. A. Morin and Dr. S. Paabo for kindly providing chimpanzee samples. We also thank the anonymous reviewers for their comments, which have contributed to the improvement of the manuscript. This work was partially supported by Fundação para a Ciência e Tecnologia (through grant SFRH/BD/7006/2001 and POCTI, Programa Operacional Ciência, Tecnologia e Inovação). Norman Ross, James Close, Adam Dagnall and Tim Crow are grateful to the UK Medical Research Council, TJ Crow Psychosis Research Trust, Stanley Foundation and SANE for support.

References

- Arnold AP (2004) Sex chromosomes and brain gender. *Nat Rev Neurosci* 5(9):701–708
- Barrick TR, Mackay CE, Prima S, Vandermeulen D, Crow TJ, Roberts N (2004) Automatic analysis of cerebral asymmetry: an exploratory study of the relationship between brain torque and planum temporale asymmetry. *Neuroimage* 24(3):678–691
- Blanco P, Sargent CA, Boucher CA, Mitchell M, Alfara NA (2000) Conservation of *PCDH11X* in mammals; expression of human X/Y genes predominantly in brain. *Mamm Genome* 11(10):906–914
- Blanco-Arias P, Sargent CA, Alfara NA (2004a) Protocadherin X (*PCDH11X*) and Y (*PCDH11Y*) genes; multiple mRNA isoforms encoding variant signal peptides and cytoplasmic domains. *Mamm Genome* 15(1):41–52
- Blanco-Arias P, Sargent CA, Alfara NA (2004b) A comparative analysis of the pig, mouse, and human *PCDH11X* genes. *Mamm Genome* 15(4):296–306
- Brown CJ, Greally JM (2003) A stain upon the silence: genes escaping X inactivation. *Trends Genet* 19(8):432–438
- Brown CJ, Miller AP, Carrel L, Rupert JL, Davies KE, Willard HF (1995) The DXS423E gene in Xp11.21 escapes X chromosome inactivation. *Hum Mol Genet* 4(2):251–255
- Brown CJ, Carrel L, Willard HF (1997) Expression of genes from the human active and inactive X chromosomes. *Am J Hum Genet* 60(6):1333–1343
- Cáceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, Geschwind DH, Lockhart DJ, Preuss TM, Barlow C (2003) Elevated gene expression levels distinguish human from non-human primate brains. *Proc Natl Acad Sci* 100(22):13030–13035
- Carrel L, Willard HF (1999) Heterogeneous gene expression from the inactive X chromosome: an X-linked gene that escapes X inactivation in some human cell lines but is inactivated in others. *Proc Natl Acad Sci* 96(13):7364–7369
- Carrel L, Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434(7031):400–404

- Carrel L, Clemson CM, Dunn JM, Miller AP, Hunt PA, Lawrence JB, Willard HF (1996) X inactivation analysis and DNA methylation studies of the ubiquitin activating enzyme E1 and PCTAIRE-1 genes in human and mouse. *Hum Mol Genet* 5(3):391-401
- Carrel L, Cottle AA, Goglin KC, Willard HF (1999) A first-generation X-inactivation profile of the human X chromosome. *Proc Natl Acad Sci* 96(25):14440-14444
- Carruth LL, Reisert I, Arnold AP (2002) Sex chromosome genes directly affect brain sexual differentiation. *Nat Neurosci* 5(10):933-934
- Charlesworth B (1996) The evolution of chromosomal sex determination and dosage compensation. *Curr Biol* 6(2):149-162
- Ciccodicola A, D'Esposito M, Esposito T, Gianfrancesco F, Migliaccio C, Miano MG, Matarazzo MR, Vacca M, Franze A, Cuccurese M, Cocchia M, Curci A, Terracciano A, Torino A, Cocchia S, Mercadante G, Pannone E, Archidiacono N, Rocchi M, Schlessinger D, D'Urso M (2000) Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. *Hum Mol Genet* 12:395-401
- Craig IW, Mill J, Craig GM, Loat C, Schalkwyk LC (2004) Application of microarrays to the analysis of the inactivation status of human X-linked genes expressed in lymphocytes. *Eur J Hum Genet* 12(8):639-646
- Crow TJ, Crow LR, Done DJ, Leask SJ (1998) Relative hand skill predicts academic ability: global deficits at the point of hemispheric indecision. *Neuropsychologia* 36(12):1275-1282
- Csankovszki G, Nagy A, Jaenisch R (2001) Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *J Cell Biol* 153(4):773-784
- D'Esposito M, Ciccodicola A, Gianfrancesco F, Esposito T, Flagiello L, Mazzarella R, Schlessinger D, D'Urso M (1996) A synaptobrevin-like gene in the Xq28 pseudoautosomal region undergoes X inactivation. *Nat Genet* 13(2):227-229
- D'Esposito M, Matarazzo MR, Ciccodicola A, Strazzullo M, Mazzarella R, Quaderi NA, Fujiwara H, Ko MS, Rowe LB, Ricco A, Archidiacono N, Rocchi M, Schlessinger D, D'Urso M (1997) Differential expression pattern of XqPAR-linked genes SYBL1 and IL9R correlates with the structure and evolution of the region. *Hum Mol Genet* 6:1917-1923
- Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Paabo S (2002) Intra- and interspecific variation in primate gene expression patterns. *Science* 296(5566):340-343
- Fisher EM, Beer-Romero P, Brown LG, Ridley A, McNeil JA, Lawrence JB, Willard HF, Bieber FR, Page DC (1990) Homologous ribosomal protein genes on the human X and Y chromosomes: escape from X inactivation and possible implications for Turner syndrome. *Cell* 63(6):1205-1218
- Galfalvy HC, Erraji-Benchekroun L, Smyrniotopoulos P, Pavlidis P, Ellis SP, Mann JJ, Sibille E, Arango V (2003) Sex genes for genomic analysis in human brain: internal controls for comparison of probe level data extraction. *BMC Bioinformatics* 4:37
- Jegalian K, Page DC (1998) A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* 394(6695):776-780
- Khaitovich P, Muetzel B, She X, Lachmann M, Hellmann I, Dietzsch J, Steigle S, Do HH, Weiss G, Enard W, Heissig F, Arendt T, Nieselt-Struwe K, Eichler EE, Paabo S (2004) Regional patterns of gene expression in human and chimpanzee brains. *Genome Res* 14(8):1462-1473
- Kretschmann HF, Schleicher A, Wingert F, Zilles K, Loeblich HJ (1979) Human brain growth in the 19th and 20th century. *J Neurol Sci* 40:169-188
- Lee JT, Jaenisch R (1997) The (epi)genetic control of mammalian X-chromosome inactivation. *Curr Opin Genet Dev* 7(2):274-280
- Lyon MF (1961) Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* 190:372-373
- Matarazzo MR, De Bonis ML, Gregory RI, Vacca M, Hansen RS, Mercadante G, D'Urso M, Feil R, D'Esposito M (2002) Allelic inactivation of the pseudoautosomal gene SYBL1 is controlled by epigenetic mechanisms common to the X and Y chromosomes. *Hum Mol Genet* 11(25):3191-3198
- Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* 430(7001):743-747
- Munton RP, Vizi S, Mansuy IM (2004) The role of protein phosphatase-1 in the modulation of synaptic and structural plasticity. *FEBS Lett* 567(1):121-128
- Ohno S (1967) Sex chromosomes and sex-linked genes. Springer, Berlin Heidelberg New York
- Page DC, Harper ME, Love J, Botstein D (1984) Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution. *Nature* 311(5982):119-123
- Pedone PV, Cosma MP, Ungaro P, Colantuoni V, Bruni CB, Zarrilli R, Riccio A (1994) Parental imprinting of rat insulin-like growth factor II gene promoters is coordinately regulated. *J Biol Chem* 269(39):23970-23975
- Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N (1996) Requirement for Xist in X chromosome inactivation. *Nature* 379(6561):131-137
- Redies C, Vanhalst K, Roy FV (2005) delta-Protocadherins: unique structures and functions. *Cell Mol Life Sci* 62(23):2840-2852
- Schwartz A, Chan DC, Brown LG, Alagappan R, Pettay D, Distche C, McGillivray B, de la Chapelle A, Page DC (1998) Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum Mol Genet* 7(1):1-11
- Simon JL, Resampling Stats. Inc. (2000) "Resampling Stats" Standalone software
- Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang SP, Waterston RH, Wilson RK, Rozen S, Page DC (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423(6942):825-837
- Sudbrak R, Wiczeorek G, Nuber UA, Mann W, Kirchner R, Erdogan F, Brown CJ, Wohrl D, Sterk P, Kalscheuer VM, Berger W, Lehrach H, Ropers HH (2001) X chromosome-specific cDNA arrays: identification of genes that escape from X-inactivation and other applications. *Hum Mol Genet* 10(1):77-83
- Uddin M, Wildman DE, Liu G, Xu W, Johnson RM, Hof PR, Kapatos G, Grossman LI, Goodman M (2004) Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. *Proc Natl Acad Sci* 101(9):2957-2962
- Vanhalst K, Kools P, Staes K, van Roy F, Redies C (2005) delta-Protocadherins: a gene family expressed differentially in the mouse brain. *Cell Mol Life Sci* 62(11):1247-1259
- Vawter MP, Evans S, Choudhary P, Tomita H, Meador-Woodruff J, Molnar M, Li J, Lopez JF, Myers R, Cox D, Watson SJ, Akil H, Jones EG, Bunney WE (2004) Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes. *Neuropsychopharmacology* 29(2):373-384
- Vermeech JR, Petit P, Kermouni A, Renaud JC, Van Den Berghe H, Marynen P (1997) The IL-9 receptor gene, located in the Xq/Yq pseudoautosomal region, has an autosomal origin, escapes X inactivation and is expressed from the Y. *Hum Mol Genet* 6:1-8

- Wu Q, Maniatis T (1999) A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* 97(6):779-790
- Yagi T, Takeichi M (2000) Cadherin superfamily genes: functions, genomic organization, and neurologic diversity. *Genes Dev* 14(10):1169-1180
- Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW (2002) Allelic variation in human gene expression. *Science* 297(5584):1143
- Yoshida K, Sugano S (1999) Identification of a novel protocadherin gene (PCDH11) on the human XY homology region in Xq21.3. *Genomics* 62(3):540-543
- Xu J, Burgoyne PS, Arnold AP (2002) Sex differences in sex chromosome gene expression in mouse brain. *Hum Mol Genet* 11(12):1409-1419
- Xu J, Taya S, Kaibuchi K, Arnold AP (2005) Sexually dimorphic expression of *Usp9x* is related to sex chromosome complement in adult mouse brain. *Eur J Neurosci* 21(11):3017-3022

2. *PCDH11X* histone modifications and CpG island methylation in human and in gorilla (*Gorilla gorilla*) female lymphoblastoid cell lines (AM Lopes and N Ross, unpublished results)

Histone modifications in the putative *PCDH11X* promoter within CpG island 1 were analysed in one human and one gorilla lymphoblastoid cell line by immunoprecipitation of chromatin (Orlando *et al.*, 1997). The results show that human alleles bind more histones associated with transcriptionally active genes - H3MeK4 (Boggs *et al.* 2002) and H4AcK5 (Davie *et al.*, 1978) - than gorilla alleles and the latter bind more histone associated with silenced genes (H3MeK9). This indicates that *PCDH11X* is expressed from both the active and the inactive X chromosome in humans while in gorilla *Pcdh11X* is only expressed from the active X chromosome.

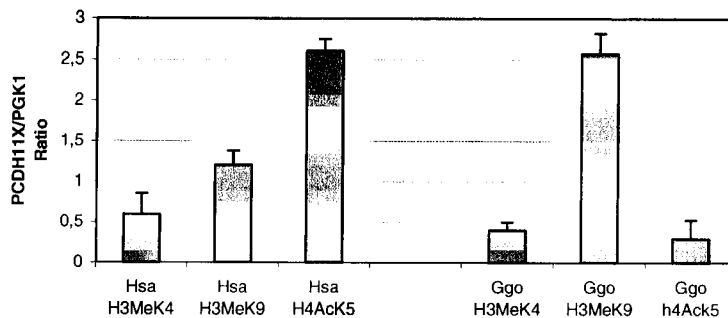


Figure I.1 - Results of two independent PCRs expressed as ratios of amounts of *PCDH11X* precipitated relative to *PGK1*. Chromatin obtained from transformed lymphocytes of human (*Hsa*) and gorilla (*Ggo*) females was treated with anti-trimethyl Ly4 of histone H3 (H3MeK4), anti-trimethyl Ly9 of histone H3 (H3MeK9) or anti-acetyl Lys5 of histone H4 (H4AcK5). Immunoprecipitates were analysed by semi-quantitative duplex PCR using fluorescent primers within the CpG islands of *Protocadherin11X* (*PCDH11X*) and *PGK1*.

To directly relate histone modifications with CpG island methylation levels, the status of selected CpG sites within CpG island 1 was determined in the same cell lines where chromatin immunoprecipitation was performed. This analysis was also performed in gorilla (*Gorilla gorilla*) uncultured blood leukocytes, for comparison.

An almost complete absence of methylation was observed in human and gorilla blood leukocytes (Figure I.2) in contrast to human and gorilla cell lines, which presented a high percentage of methylated clones, at several CpG sites (40%).

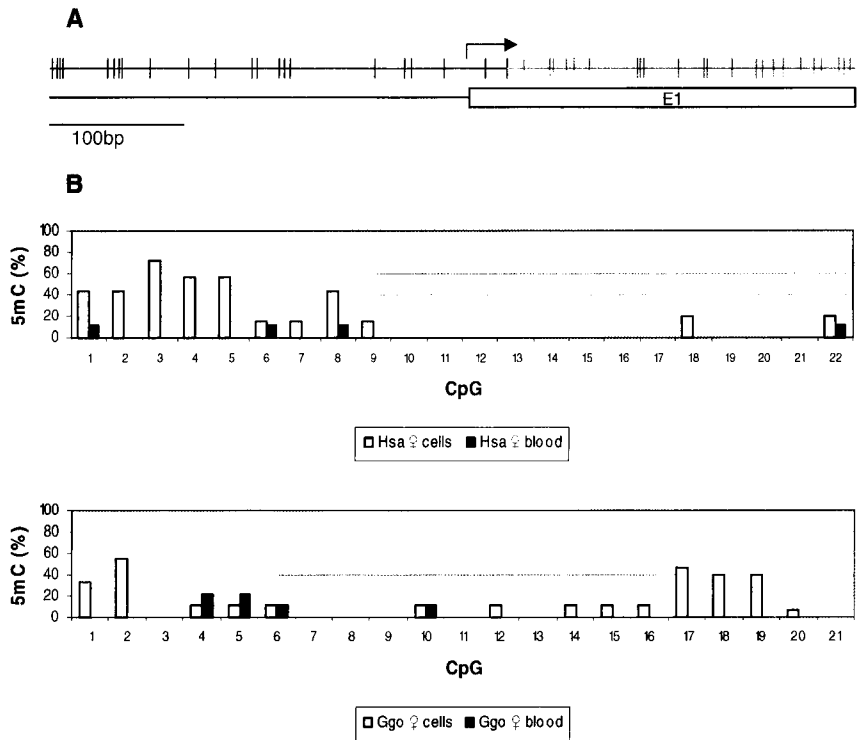


Figure I.2 - Methylation levels in the 5' region of *PCDH11X* CpG island 1 in lymphoblastoid cell lines and in blood leukocytes. (A) Schematic representation of *PCDH11X* CpG island 1, with analysed CpG sites represented as vertical black dashes. (B) Percentage of clones methylated at each CpG site, in human (*Hsa*) and gorilla (*Ggo*) female lymphoblastoid cell lines and blood leukocytes (for comparison). Correspondence between human and gorilla CpG sites is not complete.

Interestingly, the patterns of methylation were to some extent different between the two cell lines, with higher density of methylated CpGs around the transcriptional start site (which lies between CpG sites number 20 and 21) in the gorilla cell line, while in the human cell line all the sites with methylation levels higher than 20% were restricted to the 5' end of the CpG island.

It is difficult to reconcile the results obtained with different sources of biological material since in native tissues (human and gorilla blood leukocytes and human brain tissue) extremely low levels of methylation were found (Article 1 and present results) while cultured immortalized cell lines harboured higher levels of methylation.

An increase in methylation levels in human and mouse cell lines has been reported to be associated with cell line immortalization and to serial passages during cell culture (Shmookler Reis *et al.*, 1982; Jones *et al.*, 1990; Antequera *et al.*, 1990). Therefore, methylation levels detected *in vitro* might not reproduce the *in vivo* situation, as we have also demonstrated, invalidating the extrapolation from the patterns of histone modifications observed in these cell lines to the general biological status. Moreover, the fact that methylation patterns differ between the two cell lines further impairs this type of comparisons.

From the analysis of methylation in blood leukocytes it seems plausible that *Pcdh11X* escapes inactivation in gorilla since, as in humans, no methylation was found in the 5' end of CpG island 1. The aberrant levels of methylation found in gorilla cell lines, which are somewhat different from that of human cell lines, might explain the different patterns of histone modifications found between species.

3. Methylation of *PCDH11X* in Klinefelter's syndrome (XXY)

Article 2

Ross N, Wadekar R, **Lopes A**, Dagnall A, Close J, DeLisi LE, Crow TJ. Methylation of two *Homo sapiens*-specific X-Y homologous genes in Klinefelter's syndrome (XXY) Am J Med Genet B Neuropsychiatr Genet (submitted)

Methylation of two *Homo sapiens*- specific X-Y homologous genes in Klinefelter's syndrome (XXY)

Norman L J Ross^{1*}, Rekha Wadekar¹, Alexandra Lopes², Adam Dagnall¹, James Close¹,
Lynn E DeLisi³ and Timothy J Crow¹

1. SANE POWIC, University of Oxford, Warneford Hospital, Oxford OX3 7JX, United Kingdom.
2. IPATIMUP, Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Portugal.
3. The Department of Psychiatry New York University, New York, NY 10016.USA.

*Corresponding author: Norman Ross, SANE POWIC, University of Oxford, Warneford Hospital, Oxford OX3 7JX; Telephone:44 1865 455913; e-mail: norman.ross@psychiatry.ox.ac.uk

Short Title: DNA Methylation in Klinefelter's syndrome

Key Words: KS, DNA, *SYBL1*, *PCDH11XY*

ABSTRACT

An increased incidence of psychiatric and structural brain abnormalities in individuals with Klinefelter syndrome (KS, 47 XXY) could be due to the presence of extra copies of X-Y homologous genes that escape X-inactivation. Of particular interest are the two brain-expressed genes *Protocadherin11XY* (*PCDH11XY*) and the Synaptobrevin-like gene (*SYBL1*) which have been duplicated from the X chromosome to the Y chromosome to give X-Y homologous gene pairs that are specific to modern humans. We examined the DNA of KS individuals reported recently by DeLisi et al. [2005] and determined the parental origin of the X alleles, the degree of skewed X-inactivation and investigated the CpG island methylation status of *PCDH11XY* and *SYBL1* by bisulphite sequencing and quantification of methylated HpaII sites. We used a novel method for quantification of unmethylated CpGs with the restriction enzyme McrBC which cuts methylated but not unmethylated CpGs. The results showed that KS individuals have 2 methylated and one unmethylated *SYBL1* allele whereas *PCDH11XY* is unmethylated and escapes X-inactivation on the extra X chromosome. Overexpression of *PCDH11XY* in KS is probable and variable escape from inactivation of this *Homo sapiens*-specific gene could account for some abnormalities in KS. The origin of the parental alleles or their preferential X-inactivation was not associated with psychotic symptoms.

Several studies have reported verbal abnormalities and psychiatric conditions in KS males [reviewed in DeLisi et al. 1994, van Rijn S et al. 2005]. XXY men are found in about 0.8-1% of males with schizophrenia, approximately a 4-5 fold excess compared to normal male live births. Similarly approximately 1% of individuals with mental retardation have an XXY karyotype [Rosen et al. 1970]. In addition, structural abnormalities in the brain including enlarged ventricles and reversed or diminished cerebral asymmetries have been reported [Warwick et al. 1999, 2003, Rezaic et al. 2004].

Recently DeLisi et al. [2005] reported that KS individuals had abnormalities in both grey and white matter volumes of frontal and temporal lobes and white matter tracts leading to these areas. DeLisi et al. [2005] and others [Crow 1994; Geschwind et al. 1998] have postulated that excessive expression of one or more genes on the X chromosome, particularly those that escape X-inactivation (that include the genes that have homologues on the Y) could be responsible for the abnormalities. Crow [2002a, 2002b] has further suggested that the gene *PCDH11X*, and its epigenetic regulation, is the leading candidate. This gene is part of the Xq21.3 block that was transposed to Yp11.2 six million years ago in the hominid lineage, and at that point *PCDH11X* acquired a homologue *PCDH11Y*. Thus humans have a gene pair *PCDH11X* and *PCDH11Y* whereas other mammals have only the gene *PcdhX*. Moreover this gene pair is expressed in the brain and the encoded protein as a member of the protocadherin family of cell surface adhesion molecules has the potential to play a role in cell-cell recognition. While males have the X and Y copies of the gene and females have copies on each of their two X chromosomes, it is not known if the copy on the inactive X escapes inactivation.

In this study we explore the possibility that there might be an association between an escape from inactivation by methylation of *PCDH11XY* (and thus an excess production of this gene's product) and the abnormal phenotypes reported by DeLisi et al. [2005]. Specifically we have examined the methylation status of the CpG islands of *PCDH11XY* in blood leukocytes of the KS individuals using bisulphite sequencing as well as the restriction enzyme McrBC which allowed quantification of unmethylated CpGs. In general, genes having unmethylated CpGs in their CpG islands are transcriptionally active whereas those containing methylated CpGs are silent. Exceptions have been reported by Anderson and Brown [2005] for cultured cell lines.

We hypothesize that if *PCDH11X* escapes inactivation, in the case of XXY individuals, all 3 alleles of *PCDH11XY* will be unmethylated whereas males and females of normal karyotype will have 2 unmethylated alleles.

We have included *SYBL1* in our study because associations with bipolar disease [Saito et al. 2000; Muller et al., 2002] have been reported. This gene like the *PCDH11X* and *Y* gene pair is present on both X and Y chromosomes in humans but only on the X chromosome in other primates. Moreover, the gene is unusual in that one allele (X or Y) is methylated and inactive in individuals with normal karyotype [Matarazzo et al. 2002]. It is not known how many *SYBL1* alleles are active in KS individuals. In addition to determining the methylation status of the gene in the probands we have also determined the parental origin of the X chromosomes and the level of skewed X-inactivation, in view of the earlier report of Iitsuka et al. [2001] who observed skewed X-inactivation in 5 of 16 KS cases and suggested that this phenomenon could explain the wide range of mental and phenotypic abnormalities observed among KS individuals.

The KS individuals, their parents and normal (XY) males have been described [DeLisi et al. 2005]. The results of structured psychiatric interviews summarised in Table I indicated that 4 of the 11 KS were psychotic (KS-PSY) and had hallucinations and delusions. A group of unrelated males and females from Oxford, UK whose Diagnostic interview for Genetic Studies failed to reveal any psychiatric diagnosis by DSM-IV category I or II criteria served as normal controls. DNA was extracted commercially (Tepnel Life Sciences plc, Manchester, UK) from blood leukocytes of these individuals using the Nucleon kit.

We carried out the human androgen receptor (HUMARA) methylation assay [Allen et al. 1992] to determine whether the parental origin of the extra X or skewed inactivation could explain the abnormalities reported in these KS individuals. One hundred ng of genomic DNA were digested overnight with 20 U of HpaII at 37 °C in 50 µl volumes. The isoschizomer MspI was routinely used as a control to check that digestion was complete. The digests were inactivated at 70 °C before amplification by PCR using CY5-labelled primers. Usually 10 µg of the digested DNA were amplified in a thermocycler (Eppendorf Mastercycler, Hamburg, Germany). Primers and conditions for amplification are summarised in Table II. Five µl of the amplified products were

denatured and separated by electrophoresis in polyacrylamide gels (Sequagel XR: National Diagnostics, East Riding, Yorkshire, UK) in an Alf Express sequencer (Amersham Biosciences, Amersham UK). Peak areas were computed using the Allelix software (Amersham Biosciences). Skewing of X-inactivation was calculated as described [Iitsuka et al. 2001]. Our results (Table I) are partly in agreement with Iitsuka et al. [2001] in that approximately 60% (5/8) of the KS had one maternal and one paternal X whereas in 40% (3/8) of cases both alleles were maternally derived. In contrast to Iitsuka et al. [2001], none of the 8 KS who had 2 distinguishable X alleles showed preferential inactivation of one allele (7 80% skewed) but two mothers had marked skewed inactivation. Their KS offspring inherited the active maternal allele and the paternal allele. In view of the small number of informative samples, we were unable to identify an association between the origin or selective inactivation of parental alleles and psychosis.

To determine how many *SYBL1* alleles are methylated in KS, we quantified the relative number of molecules containing methylated CmCGG at three HpaII sites in the *SYBL1* promoter. These correspond to CpGs 6, 19 and 29 [Matarazzo et al. 2002] and are known to be unmethylated on active X but methylated on inactive X and Y. Digests containing 10 ng of genomic DNA were amplified by PCR using fluorescently labelled primers (Table II). The digests were also amplified with primers for *MSX2*, an autosomal gene, which does not contain HpaII sites in the amplified region, used previously as an internal control for normalization [Ross et al. 2003]. All PCR assays were done separately in triplicate as described above and five μ l amounts of amplified products were quantified as described for the HUMARA assay. Results were expressed as ratios of *SYBL1/MSX2*. When the data for KS and KS with psychosis (KS-PSY) were pooled for analysis, the mean ratios of *SYBL1/MSX2* for the pooled KS group, and normal males and females were 0.36, 0.17 and 0.18 respectively. The 2 fold difference between the KS and normal controls was highly significant (One way ANOVA, $P < 0.001$) indicating that there are approximately twice as many methylated CmCGG in the KS group compared to normal males and females. *SYBL1/MSX2* ratios of mothers and fathers were lower than those of the KS group ($P < 0.001$) but did not differ significantly from those of normal males and females. We conclude from these results that KS have two methylated and one unmethylated *SYBL1* allele.

Preliminary experiments (not shown) using the methylation-sensitive restriction enzyme HpaII indicated that CpG islands 1 and 2 of *PCDH11XY* are not methylated. Bisulphite sequencing of genomic DNA from 4 KS and 2 KS-PSY individuals carried out essentially as described by Lopes et al. [2006] showed that the % of methylated CpGs in all clones examined ranged from 0-7% for fragment A (CpGs 1-16) and 0-3% for fragment B (CpGs 17-22). Data were obtained by sequencing a total of 20 clones (360 CpGs) from fragment A and 22 clones (132 CpGs) from fragment B. Similarly none of the CpGs examined from the 5' end of CpG island 2 was methylated. In view of the sparse, random distribution of methylated CpGs, we concluded that the CpG islands of *PCDH11XY* are unmethylated and that none of the alleles are silenced. Our results are consistent with those of Lopes et al. [2006] who concluded that both CpG islands are unmethylated in males and females of normal karyotype.

Since it was not possible to quantify methylated CpGs of *PCDH11XY*, we used a novel approach to determine whether KS had more unmethylated CpG alleles than controls of normal karyotype. The method consists essentially in digesting genomic DNA with the restriction enzyme McrBC which cuts methylated CpGs but fails to cut unmethylated CpGs [Yamada et al. 2004], followed by PCR amplification and quantification of the amplified products. We used *PGK1* for normalization since it is established that this gene is subject to X-inactivation [Pfeifer et al. 1990] and consequently both KS and normal karyotypes are expected to have one unmethylated allele (Fig 1). To test the validity of this assumption, we quantified methylated CmCGG in the CpG islands and promoter region of *PGK1* and *SYBL1* using primers and conditions summarized in Table II. The results showed that the ratios of methylated *PGK1* relative to *MSX2* for KS and normal female were 0.34 ± 0.05 (2xSE) and 0.38 ± 0.04 (2xSE) respectively, whereas the corresponding ratios for *SYBL1*/ *MSX2* were 0.34 ± 0.06 and 0.16 ± 0.02 respectively. Thus, KS contain approximately 2 fold more methylated CmCGG in *SYBL1* compared to normal females (t-test, $P=0.002$) whereas the relative number of molecules containing methylated sites in *PGK1* were the same. These findings are consistent with the following: (1) both KS and XX females have one active and one inactive *PGK1* allele, (2) KS have two inactive and one active *SYBL1* allele whereas normal females have one active and one inactive *SYBL1* allele. Accordingly we used *PGK1* to normalize quantitation of unmethylated *PCDH11XY* and *SYBL1* by the McrBc method.

One hundred ng amounts of genomic DNA were digested with 20 U of McrBC (New England Biolabs, Hitchin, Hertfordshire, UK) and 10 ng of the digested DNA were amplified using primers spanning 17 CpGs located at the 5' end of the CpG island 1 (Table II). For each experiment, aliquots of 10 ng of the digests were also amplified separately with the *PGK1* primers described above (Table II). All assays were done in triplicate and the products quantified as above. Results were expressed as ratios of unmethylated *PCDH11XY* / unmethylated *PGK1* (Fig 2A). The mean ratio of unmethylated *PCDH11XY* / *PGK1* of the pooled KS group was 1.4 times (0.57/0.4) greater than pooled normal controls (t-test; $P=0.001$) and is consistent with our hypothesis outlined in Fig 1 that *PCDH11XY* escapes inactivation and that all 3 alleles are unmethylated in KS. Similar results were obtained when unmethylated CpGs (CpGs 1-14) in CpG island 2 were quantified. The ratio of unmethylated *PCDH11XY/PGK1* for the pooled KS relative to controls of normal karyotype was 1.3 and the difference was significant (t-test, $P<0.05$). In contrast to *PCDH11XY*, we expected that ratios for unmethylated *SYBL1/PGK1* would be constant across genotypes. The results (Fig 2B) are consistent with this hypothesis. The mean ratios were 1.27 and 1.35 for the pooled KS and normal controls respectively and were not significantly different (t-test, $p>0.05$).

The main finding reported here is that KS individuals have 3 unmethylated alleles of *PCDH11XY* compared to the two unmethylated alleles that are in females and males of normal karyotype. Consistent results were obtained for both CpG islands using bisulphite sequencing on the one hand and the novel method for quantification of unmethylated CpGs in McrBC digests of genomic DNA on the other. The latter method has the advantage that it examines CpGs in both strands of DNA in a more representative sample of the genome (estimated to be approximately 1800 molecules based on the amount of DNA used for quantification). The implications of our results are that X-Y homologous genes such as *PCDH11XY* that escape X-inactivation have the potential to be expressed at higher levels in KS individuals compared to individuals with normal karyotype; these genes could therefore contribute to the abnormalities reported in KS. We realize that DNA methylation may not always be concordant with expression [Anderson and Brown 2005]. However it is of interest that microarray hybridization has shown elevated expression of *PCDH11XY* mRNA in lymphocytes from individuals with 4X or 5X chromosomes [Sudbrak et al. 2001]. Moreover Lopes et al. [2006] have shown by RT-PCR that expression of *PCDH11X* is twofold greater in

XX females than XY males and concluded that the gene escapes X-inactivation. Our analysis lacked power to demonstrate an association between psychosis among the KS and levels of *PCDH11XY* methylation. Clearly a larger number of cases should be examined in the future. In view of the evidence that sex chromosomal aneuploidies are associated with reciprocal (verbal with an extra X or Y vs non-verbal with absence of one X or Y) deficits in cognitive development, the concept that a determinant of the relative development of the two hemispheres is represented on the X and Y chromosomes [Crow, 1994;1998] is plausible. Further investigation of the identity and epigenetic control of dosage of such genes is indicated.

ACKNOWLEDGEMENTS

We are grateful to SANE and the T J Crow Psychosis Trust for support. Alexandra Lopes has a PhDgrant from the FCT Portugal.

REFERENCES

- Allen RC, Zoghbi HY, Moseley AB, Rosenblatt HM, Belmont JW. 1992. Methylation of HpaII and HhaI sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation. *Am J Hum Genet.* 51:1229-1239.
- Anderson CL, Brown CJ. 2005. Epigenetic predisposition to expression of TIMPI from the human inactive X chromosome. *BMC Genetics* 6: 48.
- Blanco-Arias P, Sargent CA, Affara NA. 2004. A comparative analysis of the pig, mouse, and human PCDHX genes. *Mammalian Genome* 15: 296-306.
- Crow TJ. 1994. The case for an X-Y homologous determinant of cerebral asymmetry. *Cytogenet. Cell Genet* 67: 393-394.
- Crow TJ. 1998. Why cerebral asymmetry is the key to the origin of *Homo sapiens*; how to find the gene or eliminate the theory. *Current Psychology of Cognition/Cahiers de Psychologie Cognitiv* 11:1237-1277.
- Crow TJ. 2002a. In Wray A (editor) *The Transition to language*: Oxford University Press pp 93-112.
- Crow TJ. 2002b. Handedness, language lateralisation and anatomical asymmetry: relevance of protocadherinXY to hominid speciation and the aetiology of psychosis. *Br J Psychiatry* 181:295-297.
- DeLisi LE, Freidrich U, Wahlstrom J, Boccio-Smith A, Forsman A, Eklund K, Crow TJ. 1994. Schizophrenia and sex chromosome anomalies. *Schizophrenia Bulletin* 20: 495-505.
- DeLisi LE, Maurizio AM, Svetina C, Ardekani B, Szulc K, Nierenberg J, Leonard J, Harvey PD. 2005. Klinefelter's syndrome (XXY) as a genetic model for psychotic disorders. *Am J Med Genet B Neuropsychiatr Genet* 135:15-23.

Geschwind DH, Gregg J, Boone K, Karrim J, Pawlikowska-Haddad A, Rao E, Ellison J, Ciccodicola A, D'Urso M, Woods R, Rappold GA, Swerdloff R, Nelson SF. 1998. Klinefelter's syndrome as a model of anomalous cerebral laterality: testing gene dosage in the X chromosome pseudoautosomal region using a DNA microarray. *Developmental Genetics* 23: 215-229.

Iitsuka Y, Bock A , Nguyen D , Samango-Sprouse CA, Simpson JL , Bischoff FZ. 2001. Evidence of skewed X-chromosome inactivation in 47, XXY and 48, XXYY Klinefelter patients. *Am J Hum Genet* 98: 25-31.

Lopes AM, Ross N, Close J, Dagnall A, Amorim A, Crow TJ. 2006. Inactivation status of PCDH11X: sexual dimorphisms in gene expression levels in brain. *Hum Genet Jan* 20:1-9 (Epub ahead of print).

Matarazzo MR, De Bonis ML, Gregory RI, Vacca M, Hansen RS, Mercadante G, D'Urso M, Feil R, D'Esposito M. 2002. Allelic inactivation of the pseudoautosomal gene *SYBL1* is controlled by epigenetic mechanisms common to the X and Y chromosomes. *Hum Mol Genet* 11: 3191-3198.

Muller DJ, Schulze TG, Jahnes E, Cichon S, Krauss H, Kesper K, Held T, Maier W, Propping P, Nothen MM, Rietschel M. 2002. Association between a polymorphism in the pseudoautosomal X linked gene *SYBL1* and bipolar affective disorder *Amer J Med Genet* 114:74-78

Pfeifer GP, Steigerwald SD, Hansen RS, Gartler SM, Riggs AD. 1990. Polymerase Chain Reaction-Aided Genomic Sequencing of an X Chromosome- linked CpG island: methylation patterns suggest clonal inheritance, CpG site autonomy, and an explanation of activity state stability. *Proc Natl Acad Sci USA* 87: 8252-8256.

Rezaie R, Roberts N, Cutter WJ, Murphy DGM, Robertson DMW, Daly EM, Maurizio A, DeLisi LE, Crow TJ. 2004. Anomalous asymmetry in Turner's and Klinefelter's syndromes-Further evidence for X-Y linkage of the cerebral dominance gene. *Am J Med Genet B Neuropsychiatr Genet*.130B:102-103.

- Rosen M, Bergman M, Telfer MA. 1970. MMPI profiles of males with abnormal sex chromosome complements. *J of Ment Deficiency Research* 15: 169-176
- Ross NL, Mavrogiannis LA, Sargent CA, Knight SJ, Wadekar R, DeLisi LE, Crow TJ. 2003. Quantification of X-Y homologous genes in patients with schizophrenia by multiplex polymerase chain reaction. *Psychiatr Genet.* 13:115-119.
- Saito T, Parsia S, Papolos DF, Lachman HM. 2000. Analysis of the pseudoautosomal X-linked gene *SYBL1* in bipolar affective disorder: description of a new candidate allele for psychiatric disorders. *Am J Med Genet* 96:317-23.
- Sudbrak R, Wieczorek G, Nuber UA, Mann W, Kirchner R, Erdogan F, Brown CJ, Wohrle D, Sterk P, Kalscheuer VM, Berger W, Lehrach H, Ropers HH. 2001. X-Chromosome-specific cDNA arrays: identification of genes that escape from X-inactivation and other applications. *Hum Mol Genet* 10:77-83.
- Van Rijn S, Swaab H, Aleman A, Kahn RS. 2005. Klinefelter Syndrome (47 XXY): Biological-Genetic vulnerability to schizophrenia and impairments in social-emotional processing. *Schizophrenia Bulletin* 31: 280.
- Warwick MM, Doody GA, Lawrie SM, Kestelman JN, Best JJ, Johnstone EC. 1999. Volumetric magnetic resonance imaging study of the brain in subjects with sex chromosome aneuploidies. *J Neurol Neurosurg Psychiatry* 66(5):628-632.
- Warwick MM, Lawrie SM, Beveridge A, Johnstone EC. 2003. Abnormal cerebral asymmetry and schizophrenia in a subject with Klinefelter's syndrome (XXY). *Biological Psychiatry* 53: 627-629.
- Yamada Y, Watanabe H, Miura F, Soejima H, Uchiyama M, Iwasaka T, Mukai T, Sakaki Y, Ito T. 2004. A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. *Genome Research* 14: 247-266.

Table I. Genotypes and X-inactivation status of the KS families.

Sample	Alleles	% Skewed	Remarks	Clinical diagnosis
01 XXY	282 ,285	59	No parental DNA	
02 M	248 ,269	58		
02 XXY	269 (M), 285 (F)	61		Delusions and hallucinations
03 M	267 , 273	90		
03 XXY	273 (M), 283 (F)	62		
04 XXY	264, 268	55	No parental DNA	Delusions
05 M	280 , 286	66		
05 XXY	286 (M) X 2	50	2 copies of same allele	
06 XXY	271	48	No parental DNA	Delusions and hallucinations
07 M	264, 277	80		
07 XXY	264 (M), 274 (F)	55		
08 XXY	264 (M), 290 (M)	69	Paternal DNA only	
12 M	267 , 283	68		
12 XXY	283 (M), 264 (F)	66		
14 XXY	266 (M) X 2	49*	Paternal DNA only	Delusions and hallucinations
15 XXY	266 (M) X 2	51*	Paternal DNA only	Delusions and hallucinations

Bold numbers indicate the size in bp of the most prominent allele (i.e methylated) in HpaII digests. (*) indicates that the maternal alleles were indistinguishable in size. It is therefore not known whether both alleles were digested equally by HpaII or whether X-inactivation was skewed in these cases. (M) and (F) indicate mother and father, respectively.

Table II. Summary of PCR primers and conditions for amplification.

	Forward primer (5'-3')	Reverse primer (5'-3')	Anneal Temp.
HUMARA	TAGGGCTGGGAAGGGTCTAC	GAACCATCCTCACCTGCT	60°C
<i>SYBL1</i>	TCGAACGAACGTGAAACACT	CACCCCACTCCATTACCTG	60°C
<i>PGK1</i>	GCATTTGCATTGATCCTGGT	AGGAACCTTCCCGACTTAGG	60°C
<i>PCDH11XY</i> CpG island1	GTGAGCAGAGGACCGAGGAGAAC	CAGCACCAACTGCTGAGTACC	60°C
<i>PCDH11XY</i> CpG island2	AAGCAAGTCGAAGTAATAGTGTATGTA	GGACTCAGCTTTTGGAAC	55°C

PCRs were carried out in 25 µl reactions using Red Taq Ready Mix PCR reaction kit (Sigma-Aldrich, Dorset, UK). After an initial denaturation step at 94°C for 5 min, samples were subjected to 20-25 cycles (predetermined to achieve linear amplification) for 40 sec at 94°C, 40 sec at the appropriate annealing temperature, extension for 1 min at 72°C, followed by a final step for 10 min at 72°C. Note that the primers for *PCDH11XY* do not differentiate between X and Y alleles.

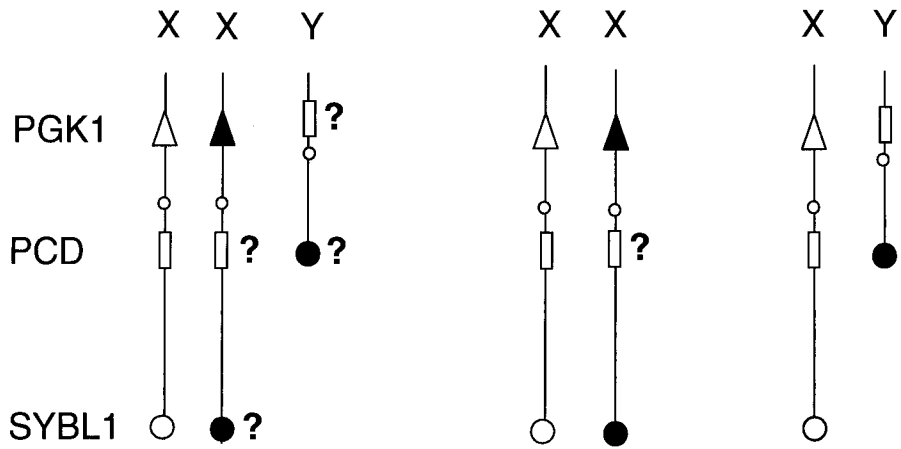
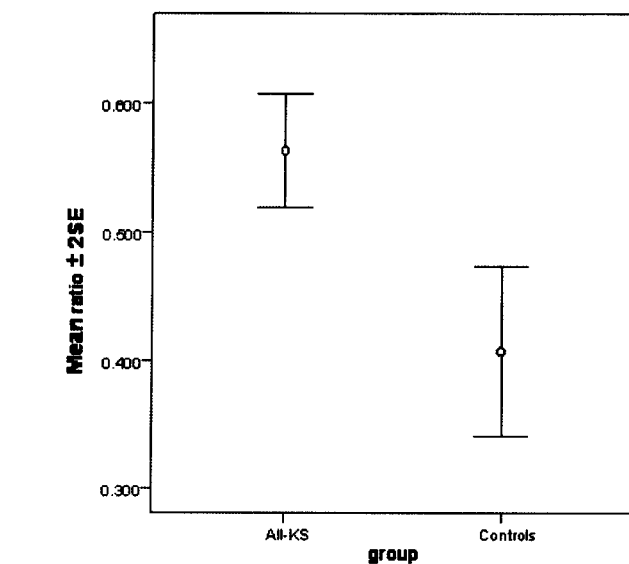
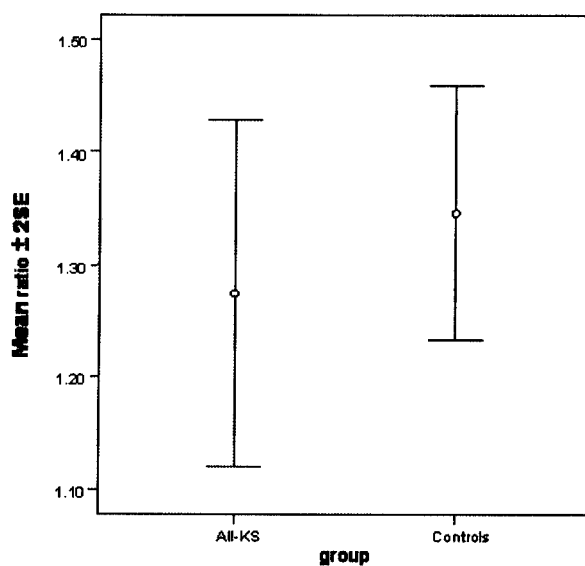


Fig 1. Schematic illustration (not to scale) of X and Y chromosomes showing gene dosage in KS (XXY) and normal males and females. Empty and filled symbols represent unmethylated and methylated alleles respectively. If *PCDH11X* escapes inactivation (not methylated), the ratio of unmethylated *PCDH11X* / *PGK1* is expected to be 2 in XX and XY individuals and 3 in XXY. Hence the ratio for XXY individuals relative to normal males and females should be 1.5. Ratios of unmethylated *SYBL1* / *PGK1* are expected to be 1 in each case. (?) indicates that the methylation status of the marked genes was unknown when this study was undertaken.



A



B

Fig 2. Quantification of unmethylated *PCDH11XY* (A) and unmethylated *SYBL1* (B). Graphs show ratios (means and 2 X SE) of peak areas relative to *PGKI*. All-KS (pooled KS and KS-PSY, n=9, excluding 2 outliers). Controls (n=9, pooled 4 males and 5 females).

DISCUSSION

1. Inactivation status and gene dosage of *PCDH11X*

Most X-linked genes which have lost their counterpart on the Y chromosome are subject to X-inactivation (Carrel and Willard, 2005). In rare instances in evolution an X-linked gene gained a homologue on the Y chromosome, by gene duplication events following structural rearrangements. The different modes of regulation found for these gene pairs highlight the existence of complex mechanisms of dosage adjustment to accommodate the gain of genes on the Y chromosome.

ProtocadherinX/Y (*PCDHX/Y* or *PCDH11X/Y*) is a human-specific gene pair, located in the Xq21.3/Yp11.2 homology block, since the X-linked gene was transposed to the Y chromosome after the human-chimpanzee lineages split (Page *et al.* 1984; Schwartz *et al.* 1998; Skaletsky *et al.* 2003). The recent acquisition of the Y homologue makes this gene pair an interesting model to explore the mechanisms by which dosage compensation copes with the gain of a Y-linked transcript but at the time this study was initiated there was only weak evidence in the literature supporting an escapee status for *PCDH11X* in humans (Sudbrack *et al.* 2001) and no such studies had been performed in any other species.

We aimed to determine the inactivation status of this gene in humans and in great apes (chimpanzee and gorilla), and thus contribute to a better understanding of the balance between sex chromosome rearrangements and adjustment of gene dosage by X-inactivation. Moreover, we intended to evaluate the impact of the Xq to Yp duplicative translocation on gene dosage differences between humans and chimpanzees and other great apes.

The results presented here indicate that this gene indeed escapes inactivation in humans and also that it might not be dosage compensated in chimpanzee and gorilla. As noted above, a number of X-linked genes that lack a homologue on the Y chromosome nevertheless escape inactivation. These genes are predicted to represent intermediates that have not yet acquired dosage compensation, due to loss of their Y homologues recently in evolution (Jegalian and Page, 1998). Concordantly, the great majority of this class of genes is located in the younger evolutionary *stratum* on Xp (Carrel and Willard, 2005), suggesting a relatively recent loss of their Y counterparts and supporting their status as intermediates. *PCDH11X* is located in Xq21.3 within evolutionary *stratum* 1, the conserved region of the X chromosome (XCR) (Graves, 1995; Lahn and Page, 1999; Kohn *et al.*, 2004) being also X-linked in the mouse. The elusion of *PCDH11X* from X-inactivation is thus unexpected in the view of its conserved X linkage and the absence of the ancestral Y homologue, which predates the radiation

of Eutherians. It would be interesting to determine its inactivation status in other mammals, namely the mouse, in order to better understand to what extent *PCDH11X* might represent an exception to the general model.

Analysis of methylation levels in the putative promoter region of *PCDH11X* revealed no differences between humans and great apes and suggests an escape from X-inactivation in all species considered. Discordantly, expression levels were found to be approximately 2-fold higher in human compared to chimpanzee brain, after correcting for between-species differences in expression levels of the endogenous control used in the present work (*GAPDH*). The nature of this difference is not obvious and apparently cannot be attributed to a generally higher rate of *PCDH11X* transcription in humans, since the expression levels of human males and chimpanzees (males and females) did not differ significantly. Discrepant results can be found in published datasets, concerning human-chimpanzee *PCDH11X* expression comparisons and the confounding factors might be diverse, as thoroughly discussed in Article 1.

The inter-species comparisons of levels of *PCDH11X* expression here presented more easily fit a model of *Protocadherin11X* escaping inactivation only in humans. This would imply that brain tissue and blood leukocytes are hypomethylated in non human primates or, alternatively, that *PCDH11X* expression is controlled by a mechanism independent of DNA methylation, in these species. Considering that CpG dinucleotides were found to be more methylated in brain tissue in humans compared to chimpanzees (Enard *et al.*, 2004) the last hypothesis seems less plausible. Further experiments focusing on other epigenetic modifications, such as increased histone methylation and deacetylation, should be conducted. The analysis presented here, using human and non human primate cell lines, was not conclusive possibly due to possible artefacts resultant from cell culture. It is therefore important to stress the necessity of performing this kind of analysis in native tissues, if biologically meaningful results are to be obtained. The challenge will be to improve existing methodologies in order to achieve reliable results from freeze-preserved primate brain tissues, which in our experience has revealed not to be a straightforward procedure (results not shown).

In summary, although we present some evidence for the existence of differences in *PCDH11X* expression levels between humans and chimpanzees, the possible contribution of the Xq21.3/Yp11.2 rearrangement to the observed differences is not clear. Moreover, due to the small number of individuals analysed, these findings must be interpreted carefully and further analysis are needed to resolve the matter.

Evidence of escape from inactivation in human brain is in agreement with the results of Sudbrak *et al.* (2001) in lymphoblastoid cell lines. By contrast Galfalvy *et al.* (2003) reported an up-regulation of *PCDH11X* expression in human prefrontal cortex in males, using the U133A array, which probably reflects cross hybridization of the probe with *PCDH11Y* transcripts. Assuming the expression levels reported by Galfalvy *et al.* (2003) represent the sum of X- and Y-linked products, an increased transcription rate is expected for male-specific (*PCDH11Y*) products, considering half dosage of *PCDH11X* transcripts are present in males. Thus the consensus of findings points to an imbalance of *PCDH11X/Y* gene products between the sexes.

Differences in *PCDH11X/Y* expression levels could relate to sexual dimorphism in brain. There are differences between males and females in aspects such as brain total size (Swaab and Hofman, 1984; Luders *et al.*, 2002), general cortical structure (Rabinowicz *et al.*, 2002) and hemispheric asymmetry (Barrick *et al.*, 2004) and it has been argued forcefully (Carruth *et al.* 2002; Arnold, 2004) that sex chromosomes act directly in the brain to determine a sexually dimorphic phenotype.

Therefore, a detailed evaluation of the expression levels of both genes in different brain areas would be of great interest. Moreover, the present results correspond to an average of the total *PCDH11X* mRNAs, from the high diversity of transcripts that are created in both the X- and Y-linked loci by alternative splicing (Blanco-Arias *et al.*, 2004a), since the primers used in the present study amplify a region that is common to virtually all *PCDH11X* transcripts. A detailed analysis of the expression levels of particular alternative transcripts in both sexes would be extremely valuable, since it could unravel additional sources of variation.

The final outcome of the referred gender specificities is largely dependent on the functional redundancy of X- and Y-linked isoforms in the central nervous system. In fact, equivalence of these proteins would result in higher overall expression levels in males, as indicated by the results of Galfalvy and co-workers, while an unbalance between sexes of X- and Y-linked isoforms is expected if these proteins have different roles. The impact of these alternative hypotheses on differences in brain physiology and/or structure in both normal individuals (between sexes) and in individuals with sex chromosomes aneuploidies (e.g. Klinefelter's syndrome) can only be a matter of speculation at the present time, urging for functional studies on *PCDHX/Y* protein products.

Finally, it should be pointed out that a considerable heterogeneity was found between individuals of the same gender and, since extensive variability in expression

levels has been reported by others in female lymphoblastoid cell lines (Carrel and Willard, 2005) and in brain tissue (Enard *et al.*, 2002b; Khaitovich *et al.*, 2004), it would be advisable to analyse a larger number of samples, in order to evaluate more accurately the extent of the observed variation.

2. Rates of molecular evolution of *PCDH11X* and its paralogue *PCDH11Y*

Apart from altering gene dosage and expression levels between species, gene duplications are also considered to be a major source of new protein functions. As originally postulated by Susumo Ohno more than three decades ago, gene duplication produces two functionally redundant paralogous genes, making one of them free from selective constraints (Ohno, 1970). Although the most likely outcome of such unconstrained evolution is for one of the paralogs to fix deleterious mutations and become a pseudogene, there is also the possibility of fixation of aminoacid changes that lead to a new function (e.g. Walsh, 1995; Wagner, 1998).

A violation of the molecular clock (Kimura, 1968) which predicts that protein's rate of evolution will be constant within different phylogenetic lineages is often detected in an early phase following gene duplication. Acceleration of the rate of protein evolution, as revealed by an elevation of the ratio of nonsynonymous substitutions per nonsynonymous site (K_a) to the synonymous substitutions per synonymous site (K_s), reflects more often a relaxation of negative selection or occasionally the action of positive selection.

A codon based maximum likelihood analysis of the relative evolutionary rate of *PCDH11X/Y* functional domains in hominoid lineages supports the existence of a period of accelerated evolution on the human and the chimpanzee-bonobo lineages on *PCDH11X* ectodomain and on *PCDH11Y* cytodomain, although not significantly above neutral expectations (N. Williams, J. Close and T.J. Crow, in preparation).

From the alignment of the predicted aminoacid sequence of great ape *PCDH11X* genes and of *PCDH11Y* to human *PCDH11X* translated transcripts, a higher number of amino acid changes is predicted to have occurred on the Y chromosome (Figure I.3), which is also evident in the correspondent phylogram (Figure I.4) generated by ClustalW (<http://www.ebi.ac.uk/clustalw/>).

Hsa	PCDH11X	MDLLSGTIVFAVLACVVFHSGAQEKNYTIREPENVLIGDLLKDLINSLIPNKSLLTAMQFKLVYKGDVPLIRIBEDTGEITFTTGARIDREKL	CAG	120
Hsa	PCDH11Y	MTVGFNSDISSVVRVNTTNCCHKC.....V.....I.....N.....T.....		99
Ptr	Pcdh11X			99
Ggo	Pcdh11X			99
Hsa	PCDH11X	IPRDEHCFYEVAIIPDEIRPLVKIRFIEDINDNAPLPATVINISIPENSAINSKYTIIPAAVDDVGINGVQNYELIKSQNIFGLDVIETPEGDKMPQLIVQKELDRCKDTYVMKV		219
Hsa	PCDH11Y			240
Ptr	Pcdh11X	T.....		219
Ggo	Pcdh11X	M.....K.....		219
Hsa	PCDH11X	KVEDGGFPQRSSTAIIQVSVTDNDNHVPFKETEIEVSIIPENAPVGSVTQLHATDADIGENAKIHFSNLSVNSIARELPHLNATTLGITIKEPLDRBETPNHKLVLASDGLMPARA		339
Hsa	PCDH11Y			360
Ptr	Pcdh11X			339
Ggo	Pcdh11X			339
Hsa	PCDH11X	MVLVNTDVNDNVPISIDIRYINPVNDTVLSENIPLNKIALITVTDKADNGRVTCTDHEIPFRLRPVFSNQFLETAAYLDYESTKEYAIIKLLAADAGKPLNQSAMLFIRVKDE		459
Hsa	PCDH11Y	N.....		480
Ptr	Pcdh11X			459
Ggo	Pcdh11X			459
Hsa	PCDH11X	NDNAPVFTQSFTVSIIPENNSPGIQLTKVSAMDADSGNAKINVLGPDAPEFSLDCRGTGMLTVVKKLDRKEDKYLFTILAKDNGVPPLTSNVTVFSIIDQNDNSPFVTHNEYNFVY		579
Hsa	PCDH11Y	M.....T.....E.....R.....		600
Ptr	Pcdh11X	T.....E.....R.....		579
Ggo	Pcdh11X	T.....E.....R.....		579
Hsa	PCDH11X	PENLPRHGTGVLITVTDPDYGDNSAVTLISILENDDFTIDSGTGVIRPNISIDREKQSGSYTFYKAEADGGVRSGRSSAKVTINVVDVNDNKPVIIPVPSNCSEYELVLPSTNPGTVVFEQVI		699
Hsa	PCDH11Y			720
Ggo	Pcdh11X	Y.....Y.....		699
Hsa	PCDH11X	AVDNDTGMAEVRYSIVGGNTRDLFAIDQETGNITLMEKCDVTDLGLHRLVLYKANDLGQDPSLFSWVIVNLFVNESVTNATLINELVRKSTEAPVTPTNTEIADVSSPTSDYVKILVAAVA		819
Hsa	PCDH11Y	I.....		840
Ptr	Pcdh11X			819
Ggo	Pcdh11X	I.....N.....		819
Hsa	PCDH11X	GTITVWVVFITAVVRCRQAPHLKAAQKNQKQSEWATPENPENRQIMMKKKKKKKHSPKNLLNFVTIIEETKADDVDSGNNRVTLDLPIDLEEQTGMGKNYVWTTPTTFKPDSPDLARHY		939
Hsa	PCDH11Y	M.....		960
Ptr	Pcdh11X	R.....		939
Ggo	Pcdh11X	R.....D.....		939

Hsa PCDH11X	KSASQPQAFQIQPETPLNSKHIIQELPLDNTFVACDSISKSCSSSDPSYSDCGYPVTTTFEYFYSVHTRP	1059
Hsa PCDH11Y	L.....	1043
Ptr Pcdh11X		1059
Ggo Pcdh11X		1059
Hsa PCDH11X	GSQESSDGLGDHDAGSLTSTSHGLPLGYQEEYFDRATPSNRTEGSDPESTFIPGLKAAEITVQPTVEASD	1179
Hsa PCDH11Y		1161
Ptr Pcdh11X		1179
Ggo Pcdh11X	L.....	1179
Hsa PCDH11X	ASTQHHSPRVTTQTIALCHSPPTVTTIALCHSPPIQVSALHHSPLVQATALHHSPPSAQAALCYSPPLAQ	1297
Hsa PCDH11Y	P.....V.....G.....S.....N.....	1279
Ptr Pcdh11X	P.....	1297
Ggo Pcdh11X	P.....R.....	1297
Hsa PCDH11X	ATSQFYTMSERLHPSDDSIKVIPLTTFTPRQARPSRSDSPIMEEHFL*	1347
Hsa PCDH11Y	A.....T.....	1329
Ptr Pcdh11X	V.....	1347
Ggo Pcdh11X	A.....	1347

Figure 1.3 - Alignment of predicted aminoacid sequences of great ape *Pcdh11X* genes and of *PCDH11Y* to human *PCDH11X*. Non-human sequences were determined from Genbank genomic sequences: *Ptr* (chimpanzee) – accession numbers AF461065-AF461072; *Ggo* (gorilla) - accession numbers AY249901-AY249908. Human genes are represented by the longest transcripts deposited in Genbank which include exons 4, 5, and 7-11 (*PCDH11X* – Accession number AF332218; *PCDH11Y* - Accession number AF33217). The signal peptide is underlined and is different in the X and Y isoforms due to a 13 bp deletion on the Y chromosome which brought in frame an upstream initiator, enlarging *PCDH11Y* signal peptide.

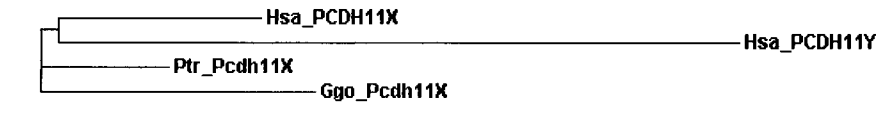


Figure I.4 – Phylogram generated in ClustalW from the alignment of predicted aminoacid sequences of great ape *Pcdh11X* genes and of *PCDH11Y* to human *PCDH11X*. (Figure I.3).

The K_a/K_s ratios (Table I.1) were calculated for the entire coding sequence of human *PCDH11X* and *PCDH11Y* long isoforms (Genbank accession numbers AF332218 and AF33217, respectively) and chimpanzee and gorilla *PCDH11X* (Genbank accession numbers AF461065-AF461072 and AY249901-AY249908, respectively), using the Nei and Gojobori method (1986) with the Jukes and Cantor (1969) correction, as implemented in DnaSP 4.0 (Rozas *et al.*, 2003). In these comparisons both the signal peptide region, which is not completely alignable between X and Y isoforms since a different start codon is used, and exons 7 and 8, which have been deleted on the Y chromosome, were not included. A higher rate of protein divergence is apparent between human *PCDH11X-PCDH11Y* ($K_a/K_s=0.546$), chimpanzee and human *PCDH11X* ($K_a/K_s=0.462$) or chimpanzee *PCDH11X* and *PCDH11Y* ($K_a/K_s=0.622$), compared to that found between gorilla and chimpanzee *Pcdh11X* ($K_a/K_s=0.228$).

Table I.1 - Rates of molecular evolution of *PCDH11X/Y*, and other X/Y genes, in hominoids.

Gene	Species	K_s	K_a/K_s	Divergence (%)		Sequence (bp)
				DNA	Protein	
<i>PCDH11X/Y</i>	<i>Hsa</i>	0.0141	0.546	1	1.8	3291
<i>PCDH11X/Y</i>	<i>Ptr/Hsa</i>	0.0065	0.462	0.4	1	3291
<i>PCDH11X/Y</i>	<i>Ptr/Hsa</i>	0.0119	0.622	0.8	1.6	3291
<i>Pcdh11X</i>	<i>Ggo/Ptr</i>	0.0162	0.228	0.7	1	3291
<i>TGIF2LX/Y</i>	<i>Hsa</i>	0.0180	0.500	1	1.7	723
<i>LOC392501X/Y*</i>	<i>Hsa</i>	0.0148	1.180	1.6	-	368
<i>FLJ37659</i>	<i>Ptr/Hsa</i>	0.0151	0.377	0.8	1	662

* Pseudogene

While these results are in agreement with accelerated evolution in *PCDH11X/Y*, the Ka/Ks value obtained for this gene pair is still more than 2-fold lower (0.546 versus 1.18) than that found for a X/Y pseudogene pair within the same duplicated region (*LOC392501X/Y*; Genbank accession number of the X-linked gene: XM_497132; Y-linked gene found by a Blat search available at <http://genome.ucsc.edu>), although it is similar to the rate of evolution ($Ka/Ks=0.500$) of the other duplicated protein coding gene pair in Xq21.3/Yp11.2 (*TGIF2LX/Y*; Genbank accession numbers: AF497480 and AF332223, respectively).

Kondrashov and colleagues (2002) have found that paralogous genes resulting from a relatively recent within-species duplication event evolve significantly faster than unduplicated orthologues with a similar level of divergence (Ks). Comparing the rate of *PCDH11X/Y* protein divergence to that of a pair of human-chimpanzee single copy orthologous genes within the same genomic region in Xq (*FLJ37659*; Genbank accession numbers: AK094978 and XM_521167 for human and chimpanzee mRNAs, respectively) and with a very similar level of synonymous divergence ($Ks=0.015$) there is indication that the former gene pair is diverging at a faster rate ($Ka/Ks=0.546$ versus $Ka/Ks=0.377$), which is in accordance with the observation of Kondrashov *et al.* (2002). Four other protein coding genes (*KLHL4*, *CPXCR1*, *NAP1L3* and *HSU24186*) are predicted in Xq21.3 but, due to the existence of frame-shift deletions in the chimpanzee sequence leading to very premature stop codons, *KLHL4* and *HSU24186* were not considered since they could represent non-functional genes in the chimpanzee. *CPXCR1* and *NAP1L3* have intact ORFs but present much lower levels of synonymous substitution rates than *PCDH11X/Y* (0.0049 and 0.0062, respectively, versus 0.014) and thus were not suitable for this type of analysis. Therefore, apparently there is a relative acceleration of the *PCDH11* rate of evolution after duplication, although this gene pair does not seem to be evolving under neutrality, which could indicate a process of pseudogeneization.

Interestingly, when performing a similar analysis using *Homo-Pan* ancestral *PCDH11X* sequence, inferred by parsimony from a human-chimpanzee-gorilla *PCDH11* alignment, a retention, to a higher extent, of the ancestral sequence is apparent in the branch leading to the human *Protocaderin11X* ($Ka/Ks=0.315$) compared to either chimpanzee *PCDH11X* ($Ka/Ks=0.619$) or *PCDH11Y* ($Ka/Ks=0.556$) branches (Table I.2). Therefore, the accelerated *PCDH11X/Y* divergence is common to the *Homo* and *Pan* lineages and in the human lineage is the outcome of an asymmetric evolution of the two paralogs, the Y-linked copy diverging faster.

There are contrasting reports on the literature concerning the relative rate of evolution of each gene copy after a gene duplication event, with some authors claiming

that both paralogs evolve at the same rate (Hughes and Hughes, 1993; Kondrashov *et al.*, 2002) while others describe the existence of a considerable percentage of gene pairs where one gene copy has a higher rate of protein divergence than the other (Robinson-Rechavi and Landet, 2001; Van de Peer *et al.*, 2001; Zhang *et al.*, 2003).

Table I.2 - Rates of molecular evolution of human *PCDH11X/Y* and chimpanzee *PCDH11X* since their divergence from a putative common ancestor.

		Species	Ks	Ka	Ka/Ks
Putative ancestral	<i>PCDH11X</i>	<i>Hsa</i>	0.0054	0.0017	0.315
<i>PCDH11X</i> sequence*	<i>PCDH11Y</i>	<i>Hsa</i>	0.0108	0.0060	0.556
compared to:	<i>Pcdh11X</i>	<i>Ptr</i>	0.0021	0.0013	0.619

* Inferred by parsimony from a human-chimpanzee-gorilla sequence alignment.

These comparisons rely on relative rates tests (Sarich and Wilson, 1973) where the relative rate of aminoacid substitutions on pairs of paralogs is compared to that of each gene copy to an outgroup gene. To ensure that such comparisons are robust, only relatively young duplicate gene pairs should be compared ($Ks < 0.5$), to avoid saturation in synonymous sites although those with $Ks < 0.05$ should not be considered since are subject to considerable statistical error (Lynch and Conery, 2000; Kondrashov *et al.*, 2002). Therefore, the statistical significance of the asymmetrical evolution on *PCDH11X/Y* cannot be accurately determined, since this very recent pair of duplicates is characterized by an extremely low level of divergence ($Ks \sim 0.014$).

A higher rate of protein divergence on Y-linked compared to X-linked homologous genes has been previously reported (Wyckoff *et al.*, 2002). The higher levels of divergence in the four gene pairs analysed allowed the authors to perform a relative rates test, using mouse orthologs as outgroups. The authors evoke the specific characteristics of the Y chromosome, such as a smaller population size and absence of recombination, as the driving force to the faster accumulation of amino acid substitutions on Y-linked genes due to a more drastic exposure to genetic drift. According to this view the observed higher divergence on the Y chromosome would reflect a process of degeneration, and in fact many ancestral homologues have been lost on this chromosome. However, the Y-linked genes under analysis in the referred do not seem to be in the process of becoming pseudogenes, since some apparently perform different functions from their X-linked counterparts, raising the possibility that accelerated evolution is coupled with a change in function. An alternative hypothesis

was proposed, where the elevated *Ka/Ks* values of genes on the Y chromosome might be driven by selection for male reproductive fitness, as has also been suggested to be the case on the chimpanzee Y (Hughes *et al.*, 2005).

The retention of *PCDH11Y* and *TGIF2LY* on the human Y chromosome, compared to the loss of *PABPC5* within the same X-transposed region, might also relate to a possible role in spermatogenesis for the proteins they encode since both of these genes are expressed in testis. In the case of *PCDH11Y*, expression in this tissue represents a novelty compared to the expression pattern of *PCDH11X* (Blanco *et al.*, 2000). This might be evidence of a neofunctionalization on the Y-linked copy by expression divergence, which could have been caused by regulatory mutations. The relevance of mutations in regulatory elements in the acquisition of new stage/tissue expression patterns by duplicate genes has been previously proposed under the DCC model (Duplication-Degeneration-Complementation; Force *et al.*, 1999). Although the proximal upstream region of *PCDH11X/Y*, including the CpG island and the first untranscribed exon, does not seem to harbour Y-specific differences of functional relevance, such as mutations disrupting transcription factor binding sites, the more distal region 5' of *PCDH11X*, including the marker DXS214, has been deleted on the Y chromosome and thus could have caused a alteration in *PCDH11Y* regulation. Overall these observations are in agreement with previous findings of acceleration of both regulatory and protein evolution in duplicate compared to orthologous genes (Castillo-Davis *et al.*, 2004) with an initial coupling of sequence divergence and divergence of spatial expression patterns between duplicates (Makova *et al.*, 2004). It is also possible that *PCDH11Y* went through a shift of function in the tissue where both duplicates are expressed, i.e. brain, since the genomic sequence of the Y-linked copy has accumulated not only sequence differences but also a few structural rearrangements.

Alternatively or yet complementarily to this hypothesis, a doubling of *PCDH11* and *TGIF2L* dosage in males might have been tolerated or even favoured, permitting their retention at an early phase. The long term fate of these two Y-linked genes is still dubious since it has been estimated a critical period defined as the "half-live" of a duplicate, which is on average approximately 4 Mya, and after which pseudogeneization becomes a highly probable scenario (Lynch and Conery, 2000).

CONCLUDING REMARKS

In spite of the general scenario of gene degeneration and loss that has characterized the evolution of the mammalian Y chromosome, there has been an episode in human evolution of gene recruitment into the NRY, interestingly, involving a brain-expressed gene (*PCDH11*). In the first few million years following this Xq to Yp transposition, there is some indication of neofunctionalization of the Y copy by gain of expression in a male specific tissue and by accumulation of structural changes in the genomic sequence which may affect the function of the protein in its initial site of expression. This regulatory innovation coupled with a faster rate of evolution of *PCDH11Y* protein compared to that of *PCDH11X*, yet not reaching the expectations under unconstrained evolution, might confer this gene the pace necessary to counterbalance a strong tendency to pseudogeneization in the next million years.

PCDH11X is apparently an old escapee that might defy dosage compensation in non-human primates and other mammals but has encountered a partner in recently in evolution. The relevance of these evolutionary novelties to human brain biology will depend on the point at which a new equilibrium, between X and Y expression and function, has been reached. Evidence for the existence of *PCDH11X* gene dosage differences between humans and chimpanzees must be considered preliminary and future research should clarify the extent of such differences as well as the possible mechanisms involved.

REFERENCES

- Adler DA, Bressler SL, Chapman VM, Page DC, Distèche CM (1991) Inactivation of the Zfx gene on the mouse X chromosome. *Proc Natl Acad Sci U S A* 88(11):4592-5.
- Adler DA, Rugarli EI, Lingenfelter PA, Tsuchiya K, Poslinski D, Liggitt HD, Chapman VM, Elliott RW, Ballabio A, Distèche CM (1997) Evidence of evolutionary up-regulation of the single active X chromosome in mammals based on Clc4 expression levels in *Mus spretus* and *Mus musculus*. *Proc Natl Acad Sci U S A* 94(17):9244-8.
- Allen RC, Zoghbi HY, Moseley AB, Rosenblatt HM, Belmont JW (1992) Methylation of HpaII and HhaI sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation. *Am J Hum Genet* 51:1229-1239.
- Anderson CL, Brown CJ (1999) Polymorphic X-chromosome inactivation of the human TIMP1 gene. *Am J Hum Genet* 65(3):699-708.
- Antequera F, Boyes J, Bird A (1990) High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines. *Cell* 62(3):503-14.
- Arnold AP (2004) Sex chromosomes and brain gender. *Nat Rev Neurosci* 5(9):701-8.
- Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. *Science* 297(5583):1003-7.
- Barrick TR, Mackay CE, Prima S, Vandermeulen D, Crow TJ et al. (2004) Automatic analysis of cerebral asymmetry: an exploratory study of the relationship between brain torque and planum temporale asymmetry. *Neuroimage* 24 (3): 678-691.
- Barr ML, Carr DH (1962) Correlations between sex chromatin and sex chromosomes. *Acta Cytol* 6:34-45.
- Bird AP (1986) CpG-rich islands and the function of DNA methylation. *Nature* 321(6067):209-13.
- Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev* 16(1):6-21.

References I

Birtle Z, Goodstadt L, Ponting CP (2005). Duplication and Positive Selection among Hominin-Specific PRAME genes. *BMC Genomics* 6(1):120.

Blanco P, Sargent CA, Boucher CA, Mitchell M, Affara NA (2000) Conservation of PCDHX in mammals; expression of human X/Y genes predominantly in brain. *Mamm Genome* 11(10):906-14.

Blanco P, Sargent CA, Boucher CA, Howell G, Ross M, Affara NA (2001) A novel poly(A)-binding protein gene (PABPC5) maps to an X-specific subinterval in the Xq21.3/Yp11.2 homology block of the human sex chromosomes. *Genomics* 74(1):1-11.

Blanco-Arias P, Sargent CA, Affara NA (2002) The human-specific Yp11.2/Xq21.3 homology block encodes a potentially functional testis-specific TGIF-like retroposon. *Mamm Genome* 13(8):463-8.

Blanco-Arias P, Sargent CA, Affara NA (2004a) Protocadherin X (PCDHX) and Y (PCDHY) genes; multiple mRNA isoforms encoding variant signal peptides and cytoplasmic domains. *Mamm Genome* 15(1):41-52.

Blanco-Arias P, Sargent CA, Affara NA (2004b) A comparative analysis of the pig, mouse, and human PCDHX genes. *Mamm Genome* 15(4):296-306.

Boggs BA, Chinault AC (1994) Analysis of replication timing properties of human X-chromosomal loci by fluorescence in situ hybridization. *Proc Natl Acad Sci U S A* 91(13):6083-7.

Boggs BA, Connors B, Sobel RE, Chinault AC, Allis CD (1996) Reduced levels of histone H3 acetylation on the inactive X chromosome in human females. *Chromosoma* 105(5):303-9.

Boggs BA, Cheung P, Heard E, Spector DL, Chinault AC, Allis CD (2002) Differentially methylated forms of histone H3 show unique association patterns with inactive human X chromosomes. *Nat Genet* 30(1):73-6.

Bonny C, Cooker LA, Goldberg E (1998) Deoxyribonucleic acid-protein interactions and expression of the human testis-specific lactate dehydrogenase promoter: transcription factor Sp1 plays a major role. *Biol Reprod* 58(3):754-9.

Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, Rastan S (1992) The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* 71(3):515-26.

Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, Lawrence J, Willard HF (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* 71(3):527-42.

Brown CJ, Miller AP, Carrel L, Rupert JL, Davies KE, Willard HF (1995) The DXS423E gene in Xp11.21 escapes X chromosome inactivation. *Hum Mol Genet* 4(2):251-5.

Brown CJ, Carrel L, Willard HF (1997) Expression of genes from the human active and inactive X chromosomes. *Am J Hum Genet* 60(6):1333-43.

Brown CJ, Robinson WP (2000) The causes and consequences of random and non-random X chromosome inactivation in humans. *Clin Genet* 58(5):353-63.

Bunick D, Johnson PA, Johnson TR, Hecht NB (1990) Transcription of the testis-specific mouse protamine 2 gene in a homologous in vitro transcription system. *Proc Natl Acad Sci U S A* 87(3):891-5.

Burgoyne PS (1982) Genetic homology and crossing over in the X and Y chromosomes of mammals. *Hum Genet* 61:85-90

Cáceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, Geschwind DH, Lockhart DJ, Preuss TM, Barlow C (2003) Elevated gene expression levels distinguish human from non-human primate brains. *Proc Natl Acad Sci U S A* 100(22):13030-5.

Carey M, Smale ST (2000) Promoters and the General Transcription Machinery. In: *Transcriptional Regulation in Eukaryotes: Concepts, Strategies and Techniques* (J Cudihy, ed) Cold Spring Harbour Laboratory Press, Cold spring Harbor, New York, pp 5-14.

Carrel L, Clemson CM, Dunn JM, Miller AP, Hunt PA, Lawrence JB, Willard HF (1996) X-inactivation analysis and DNA methylation studies of the ubiquitin activating enzyme E1 and PCTAIRE-1 genes in human and mouse. *Hum Mol Genet* 5(3):391-401.

References I

- Carrel L, Willard HF (1999) Heterogeneous gene expression from the inactive X chromosome: an X-linked gene that escapes X-inactivation in some human cell lines but is inactivated in others. *Proc Natl Acad Sci U S A* 96(13):7364-9.
- Carrel L, Cottle AA, Goglin KC, Willard HF (1999) A first-generation X-inactivation profile of the human X chromosome. *Proc Natl Acad Sci U S A* 96(25):14440-4.
- Carrel L, Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434(7031):400-4.
- Carruth LL, Reisert I, Arnold AP (2002) Sex chromosome genes directly affect brain sexual differentiation. *Nat Neurosci* 5(10):933-4.
- Charlesworth B (1991) The evolution of sex chromosomes. *Science* 251(4997):1030-3.
- Charlesworth B (1996) The evolution of chromosomal sex determination and dosage compensation. *Curr Biol* 6(2):149-62.
- Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. *Heredity* 95(2):118-28.
- Chao W, Huynh KD, Spencer RJ, Davidow LS, Lee JT (2002) CTCF, a candidate trans-acting factor for X-inactivation choice. *Science* 295(5553):345-7.
- Charchar FJ, Svartman M, El-Mogharbel N, Ventura M, Kirby P, Matarazzo MR, Ciccodicola A, Rocchi M, D'Esposito M, Graves JA (2003) Complex events in the evolution of the human pseudoautosomal region 2 (PAR2). *Genome Res* 13(2):281-6.
- Chen FC, Vallender EJ, Wang H, Tzeng CS, Li WH (2001) Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J Hered*. 92(6):481-9.
- Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Paabo S, Rocchi M, Eichler EE (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 37(7055):88-93.

Chen MW, Vacherot F, De La Taille A, Gil-Diez-De-Medina S, Shen R, Friedman RA, Burchardt M, Chopin DK, Buttyan R (2002) The emergence of protocadherin-PC expression during the acquisition of apoptosis-resistance by prostate cancer cells. *Oncogene*. 21(51):7861-71.

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437(7055):69-87.

Chou HH, Takematsu H, Diaz S, Iber J, Nickerson E, Wright KL, Muchmore EA, Nelson DL, Warren ST, Varki A (1998) A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. *Proc Natl Acad Sci U S A* 95(20):11751-6.

Ciccodicola A, D'Esposito M, Esposito T, Gianfrancesco F, Migliaccio C, Miano MG, Matarazzo MR, Vacca M, Franze A, Cuccurese M, Cocchia M, Curci A, Terracciano A, Torino A, Cocchia S, Mercadante G, Pannone E, Archidiacono N, Rocchi M, Schlessinger D, D'Urso M (2000) Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. *Hum Mol Genet* 9(3):395-401.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferreira S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302(5652):1960-3.

Clemson CM, McNeil JA, Willard HF, Lawrence JB (1996) XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J Cell Biol* 132(3):259-75.

Cooke HJ, Brown WR, Rappold GA (1985) Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. *Nature* 317: 687-692.

Cooper DW (1971) Directed genetic change model for X chromosome inactivation in eutherian mammals. *Nature* 230(5292):292-4.

Costanzi C, Pehrson JR (1998) Histone macroH2A1 is concentrated in the inactive X chromosome of female mammals. *Nature* 393(6685):599-601.

References I

Craig IW, Mill J, Craig GM, Loat C, Schalkwyk LC (2004) Application of microarrays to the analysis of the inactivation status of human X-linked genes expressed in lymphocytes. *Eur J Hum Genet* 12(8):639-46.

Crow TJ (2002) Sexual selection, timing and an X-Y homologous gene: did *Homo sapiens* speciate on the Y chromosome? In: *The Speciation of Modern Homo sapiens* (T Crow, ed) Proc Brit Acad, Oxford University Press, Oxford, pp 197-206.

Csankovszki G, Nagy A, Jaenisch R (2001) Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *J Cell Biol* 153(4):773-84.

D'Arcangelo G, Miao GG, Chen SC, Soares HD, Morgan JI, Curran T (1995) A protein related to extracellular matrix proteins deleted in the mouse mutant *reeler*. *Nature* 374(6524):719-23.

Davie JR, Candido EP (1978) Acetylated histone H4 is preferentially associated with template-active chromatin. *Proc Natl Acad Sci U S A* 75(8):3574-7.

Delbridge ML, Lingenfelter PA, Disteché CM, Graves JA (1999) The candidate spermatogenesis gene RBMY has a homologue on the human X chromosome. *Nat Genet* 22(3):223-4.

D'Esposito M, Ciccodicola A, Gianfrancesco F, Esposito T, Flagiello L, Mazzarella R, Schlessinger D, D'Urso M (1996) A synaptobrevin-like gene in the Xq28 pseudoautosomal region undergoes X-inactivation. *Nat Genet* 13(2):227-9.

D'Esposito M, Matarazzo MR, Ciccodicola A, Strazzullo M, Mazzarella R, Quaderi NA, Fujiwara H, Ko MS, Rowe LB, Ricco A, Archidiacono N, Rocchi M, Schlessinger D, D'Urso M (1997) Differential expression pattern of XqPAR-linked genes SYBL1 and IL9R correlates with the structure and evolution of the region. *Hum Mol Genet* 6(11):1917-23.

Emerson JJ, Kaessmann H, Betran E, Long M (2004) Extensive gene traffic on the mammalian X chromosome. *Science* 303(5657):537-40.

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Paabo S (2002a) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 418(6900):869-72.

Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Paabo S (2002b) Intra- and interspecific variation in primate gene expression patterns. *Science* 296(5566):340-3.

Enard W, Fassbender A, Model F, Adorjan P, Paabo S, Olek A (2004) Differences in DNA methylation patterns between humans and chimpanzees. *Curr Biol* 14(4):R148-9.

Esumi S, Kakazu N, Taguchi Y, Hirayama T, Sasaki A, Hirabayashi T, Koide T, Kitsukawa T, Hamada S, Yagi T (2005) Monoallelic yet combinatorial expression of variable exons of the protocadherin-alpha gene cluster in single neurons. *Nat Genet* 37(2):171-6. Epub 2005 Jan 9.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151(4):1531-45.

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, Karimpour-Fard A, Glueck D, McGavran L, Berry R, Pollack J, Sikela JM (2004) Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol* 2(7):E207.

Foster JW, Graves JA (1994) An SRY-related sequence on the marsupial X chromosome: implications for the evolution of the mammalian testis-determining gene. *Proc Natl Acad Sci U S A*. 91(5):1927-31.

Frank M, Kemler R (2002) Protocadherins. *Curr Opin Cell Biol* 14(5):557-62.

Fredga K (1983) Aberrant sex chromosome mechanisms in mammals. Evolutionary aspects. *Differentiation* 23 Suppl:S23-30.

Fredga K (1988) Aberrant chromosomal sex-determining mechanisms in mammals, with special reference to species with XY females. *Philos Trans R Soc Lond B Biol Sci* 322(1208):83-95.

References I

Freije D, Helms C, Watson MS, Donis-Keller H (1992) Identification of a second pseudoautosomal region near the Xq and Yq telomeres. *Science* 258: 1784-1787.

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A*. 89(5):1827-31.

Galfalvy HC, Erraji-Benchekroun L, Smyrniotopoulos P, Pavlidis P, Ellis SP, Mann JJ, Sibille E, Arango V (2003) Sex genes for genomic analysis in human brain: internal controls for comparison of probe level data extraction. *BMC Bioinformatics* 4:37.

Gidoni D, Kadonaga JT, Barrera-Saldana H, Takahashi K, Chambon P, Tjian R (1985) Bidirectional SV40 transcription mediated by tandem Sp1 binding interactions. *Science* 230(4725):511-7.

Giglion B, Comi P, Ronchi A, Mantovani R, Ottolenghi S (1989) The same nuclear proteins bind the proximal CACCC box of the human beta-globin promoter and a similar sequence in the enhancer. *Biochem Biophys Res Commun* 164(1):149-55.

Gilad Y, Man O, Paabo S, Lancet D (2003) Human specific loss of olfactory receptor genes. *Proc Natl Acad Sci U S A* 100(6):3324-7.

Giouze M, Williams NA, Lonie LJ, DeLisi LE, Crow TJ (2004) ProtocadherinX/Y, a candidate gene-pair for schizophrenia and schizoaffective disorder: a DHPLC investigation of genomic sequence. *Am J Med Genet B Neuropsychiatr Genet* 129(1):1-9.

Graves JA (1982) 5-azacytidine-induced re-expression of alleles on the inactive X chromosome in a hybrid mouse cell line. *Exp Cell Res* 141(1):99-105.

Graves JA (1995) The origin and function of the mammalian Y chromosome and Y-borne genes an evolving understanding. *Bioessays* 17(4):311-20.

Graves JA (1998) Evolution of the mammalian Y chromosome and sex-determining genes. *J Exp Zool* 281(5):472-81.

Graves JAM (2001) The rise and fall of SRY. *Trends Genet* 18(5):259-64.

Goodfellow PJ, Mondello C, Darling SM, Pym B, Little P, Goodfellow PN (1988) Absence of methylation of a CpG-rich region at the 5' end of the MIC2 gene on the active X, the inactive X, and the Y chromosome. *Proc Natl Acad Sci U S A* 85(15):5605-9.

Gu J, Gu X (2003) Induced gene expression in human brain after the split from chimpanzee. *Trends Genet.* 19(2):63-5.

Gumbiner BM (1996) Cell adhesion: the molecular basis of tissue architecture and morphogenesis. *Cell* 84(3):345-57.

Hansen RS, Canfield TK, Fjeld AD, Gartler SM (1996) Role of late replication timing in the silencing of X-linked genes. *Hum Mol Genet* 5(9):1345-53.

Hayakawa T, Angata T, Lewis AL, Mikkelsen TS, Varki NM, Varki A (2005) A human-specific gene in microglia. *Science* 309(5741):1693.

Heard E (2004) Recent advances in X-chromosome inactivation. *Curr Opin Cell Biol* 16(3):247-55.

Hedges SB, Kumar S (2004) Precision of molecular time estimates. *Trends Genet* 20(5):242-7.

Heissig F, Krause J, Bryk J, Khaitovich P, Enard W, Paabo S (2005) Functional analysis of human and chimpanzee promoters. *Genome Biol* 6(7):R57.

Hendriks RW, Chen Z-Y, Hinds H, Schuurman RKB, Craig IW (1992) An X chromosome inactivation assay based on differential methylation of a CpG island coupled to a VNTR polymorphism at the 5' end of the monoamine oxidase A gene. *Hum Mol Genet* 1: 187-194.

Hsieh WP, Chu TM, Wolfinger RD, Gibson G (2003) Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* 165(2):747-57.

References I

- Huber R, Hansen RS, Strazzullo M, Pengue G, Mazzarella R, D'Urso M, Schlessinger D, Pilia G, Gartler SM, D'Esposito M (1999) DNA methylation in transcriptional repression of two differentially expressed X-linked genes, GPC3 and SYBL1. *Proc Natl Acad Sci U S A* 96(2):616-21.
- Hughes JF, Skaletsky H, Pyntikova T, Minx PJ, Graves T, Rozen S, Wilson RK, Page DC (2005) Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* 437(7055):100-3.
- Hughes MK, Hughes AL (1993) Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol Biol Evol* 10(6):1360-9.
- Hurst LD (2001) Evolutionary genomics. Sex and the X. *Nature* 411(6834):149-50.
- Huynh KD, Lee JT (2003) Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos. *Nature* 426(6968):857-62.
- Huynh KD, Lee JT (2005) X-chromosome inactivation: a hypothesis linking ontogeny and phylogeny. *Nat Rev Genet* 6(5):410-8.
- Imataka H, Sogawa K, Yasumoto K, Kikuchi Y, Sasano K, Kobayashi A, Hayami M, Fujii-Kuriyama Y (1992) Two regulatory proteins that bind to the basic transcription element (BTE), a GC box sequence in the promoter region of the rat P-4501A1 gene. *EMBO J* 11(10):3663-71.
- Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals *Nature Genetics* 33:245 – 254.
- Jegalian K, Page DC (1998) A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* 394(6695):776-80.
- Jeppesen P, Turner BM (1993) The inactive X chromosome in female mammals is distinguished by a lack of histone H4 acetylation, a cytogenetic marker for gene expression. *Cell* 74(2):281-9.

Jones PA, Wolkowicz MJ, Rideout WM 3rd, Gonzales FA, Marziasz CM, Coetzee GA, Tapscott SJ (1990) De novo methylation of the MyoD1 CpG island during the establishment of immortal cell lines. *Proc Natl Acad Sci U S A* 87(16):6117-21.

Jukes TH, Cantor CR (1969). Evolution of protein molecules In: *Mammalian Protein Metabolism*. (Munro HN, ed), Academic Press, New York, pp 21-132.

Karaman MW, Houck ML, Chernick LG, Nagpal S, Chawannakul D, Sudano D, Pike BL, Ho VV, Ryder OA, Hacia JG (2003) Comparative analysis of gene-expression patterns in human and African great ape cultured fibroblasts. *Genome Res* 13(7):1619-30.

Khaitovich P, Muetzel B, She X, Lachmann M, Hellmann I, Dietzsch J, Steigle S, Do HH, Weiss G, Enard W, Heissig F, Arendt T, Nieselt-Struwe K, Eichler EE, Paabo S (2004) Regional patterns of gene expression in human and chimpanzee brains. *Genome Res* 14(8):1462-73.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S. (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309(5742):1850-4.

Khil PP, Smirnova NA, Romanienko PJ, Camerini-Otero RD (2004) The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nat Genet* 36(6):642-6.

Khil PP, Oliver B, Camerini-Otero RD (2005) X for intersection: retrotransposition both on and off the X chromosome is more frequent. *Trends Genet* 21(1):3-7.

Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217(129):624-6.

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science* 188(4184):107-16.

Kohmura N, Senzaki K, Hamada S, Kai N, Yasuda R, Watanabe M, Ishii H, Yasuda M, Mishina M, Yagi T (1998) Diversity revealed by a novel family of cadherins expressed in neurons at a synaptic complex. *Neuron* 20(6):1137-51.

References I

- Kohn M, Kehrer-Sawatzki H, Vogel W, Graves JA, Hameister H. (2004) Wide genome comparisons reveal the origins of the human X chromosome. *Trends Genet* 20(12):598-603.
- Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. *Genome Biol* 3(2).
- Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. *Nature* 392(6679):917-20.
- Kvaloy K, Galvagni F, Brown WR (1994) The sequence organization of the long arm pseudoautosomal region of the human sex chromosomes. *Hum Mol Genet* 3: 771-778.
- Lahn BT, Pearson NM, Jegalian K (2001) The human Y chromosome, in the light of evolution. *Nat Rev Genet* 2(3):207-16.
- Lahn BT, Page DC (1997) Functional coherence of the human Y chromosome. *Science* 278(5338):675-80.
- Lahn BT, Page DC (1999a) Four evolutionary strata on the human X chromosome. *Science* 286(5441): 964-7.
- Lahn BT, Page DC (1999b) Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. *Nat Genet* 21(4):429-33.
- Lee JT, Davidow LS, Warshawsky D (1999) Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat Genet* 21(4):400-4.
- Lee JT (2005) Regulation of X-chromosome counting by Tsix and Xite sequences. *Science* 309(5735):768-71.
- Li L, Hamer DH (1995) Recombination and allelic association in the Xq/Yq homology region. *Hum. Mol. Genet.* 4: 2013-2016.
- Lock LF, Takagi N, Martin GR (1987) Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell* 48(1):39-46.

- Luders E, Steinmetz H, Jancke L (2002) Brain size and grey matter volume in the healthy human brain. *Neuroreport* 13(17):2371-4.
- Luoh SW, Jegalian K, Lee A, Chen EY, Ridley A, Page DC (1995) CpG islands in human ZFX and ZFY and mouse Zfx genes: sequence similarities and methylation differences. *Genomics* 29(2):353-63.
- Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151-5.
- Lyon MF (1961) Gene action in the X-chromosome of the mouse (*Mus musculus* L.) *Nature* 190:372-3.
- Matarazzo MR, De Bonis ML, Gregory RI, Vacca M, Hansen RS, Mercadante G, D'Urso M, Feil R, D'Esposito (2002) M.Alelic inactivation of the pseudoautosomal gene SYBL1 is controlled by epigenetic mechanisms common to the X and Y chromosomes. *Hum Mol Genet* 11(25):3191-8.
- Mazeyrat S, Saut N, Mattei MG, Mitchell MJ (1999) RBMY evolved on the Y chromosome from a ubiquitously transcribed X-Y identical gene. *Nat Genet* 22(3):224-6.
- Mitchell MJ, Woods DR, Tucker PK, Opp JS, Bishop CE (1991) Homology of a candidate spermatogenic gene from the mouse Y chromosome to the ubiquitin-activating enzyme E1. *Nature*, 354, 483-486.
- Mitchell MJ, Wilcox SA, Watson JM, Lerner JL, Woods DR, Scheffler J, Hearn JP, Bishop CE, Graves JA (1998) The origin and loss of the ubiquitin activating enzyme gene on the mammalian Y chromosome. *Hum Mol Genet* 7(3):429-34.
- Mohandas T, Sparkes RS, Shapiro LJ (1981) Reactivation of an inactive human X chromosome: evidence for X-inactivation by DNA methylation. *Science* 211(4480):393-6.
- Mukhopadhyay D, Knebelmann B, Cohen HT, Ananth S, Sukhatme VP (1997) The von Hippel-Lindau tumor suppressor gene product interacts with Sp1 to repress vascular endothelial growth factor promoter activity. *Mol Cell Biol* 17(9):5629-39.

References I

- Muller HJ (1964) The relation of recombination to mutational advance. *Mutat Res* 106:2-9.
- Mumm S, Molini B, Terrell J, Srivastava A, Schlessinger D (1997) Evolutionary features of the 4-Mb Xq21.3 XY homology region revealed by a map at 60-kb resolution. *Genome Res* 7(4):307-14.
- Munton RP, Vizi S, Mansuy IM (2004) The role of protein phosphatase-1 in the modulation of synaptic and structural plasticity. *FEBS Lett* 567(1):121-8.
- Navarro A, Barton NH (2003) Chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes. *Science* 300(5617):321-4.
- Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 3: 418-426.
- Newman TL, Tuzun E, Morrison VA, Hayden KE, Ventura M, McGrath SD, Rocchi M, Eichler EE (2005) A genome-wide survey of structural variation between human and chimpanzee. *Genome Res* 15(10):1344-56.
- Nguyen DK, Disteché CM (2006) Dosage compensation of the active X chromosome in mammals. *Nat Genet* 38(1):47-53.
- Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, J Sninsky J, Adams MD, Cargill M (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* 3(6):e170.
- Nollet F, Kools P, van Roy F (2000) Phylogenetic analysis of the cadherin superfamily allows identification of six major subfamilies besides several solitary members. *J Mol Biol* 299(3):551-72.
- Nopoulos P, Flaum M, O'Leary D, Andreasen NC (2000) Sexual dimorphism in the human brain: evaluation of tissue volume, tissue composition and surface anatomy using magnetic resonance imaging. *Psychiatry Res* 98(1):1-13.

- Norris DP, Brockdorff N, Rastan S (1991) Methylation status of CpG-rich islands on active and inactive mouse X chromosomes. *Mamm Genome.*;1(2):78-83.
- Odorisio T, Mahadevaiah S K, McCarrey JR, Burgoyne PS (1996) Transcriptional analysis of the candidate spermatogenesis gene Ube1y and of the closely related Ube1x shows that they are coexpressed in spermatogonia and spermatids but are repressed in pachytene spermatocytes. *Dev Biol*, 180, 336-343.
- Ogawa Y, Lee JT (2003) Xite, X-inactivation intergenic transcription elements that regulate the probability of choice. *Mol Cell* 11(3):731-43.
- Ohno S (1967) Sex Chromosomes and Sex-linked Genes, Springer, Berlin.
- Okamoto I, Arnaud D, Le Baccon P, Otte AP, Disteché CM, Avner P, Heard E. (2005) Evidence for de novo imprinted X-chromosome inactivation independent of meiotic inactivation in mice. *Nature* 438(7066):369-73.
- Okamoto S, Sherman K, Bai G, Lipton SA(2002) Effect of the ubiquitous transcription factors, SP1 and MAZ, on NMDA receptor subunit type 1 (NR1) expression during neuronal differentiation. *Brain Res Mol Brain Res* 107(2):89-96.
- Olson MV (1999) When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet* 64(1):18-23.
- Orlando V, Strutt H, Paro R (1997) Analysis of chromatin structure by in vivo formaldehyde cross-linking. *Methods* 11(2):205-14.
- Page DC, Harper ME, Love J, Botstein D (1984) Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution. *Nature* 311(5982):119-23.
- Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N (1996) Requirement for Xist in X chromosome inactivation. *Nature* 379(6561):131-7
- Pfeifer GP, Tanguay RL, Steigerwald SD, Riggs AD (1990) In vivo footprint and methylation analysis by PCR-aided genomic sequencing: comparison of active and inactive X chromosomal DNA at the CpG island and promoter of human PGK-1. *Genes Dev* 4(8):1277-87.

References I

- Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA, Wang H, de la Cruz CC, Otte AP, Panning B, Zhang Y (2003) Role of histone H3 lysine 27 methylation in X-inactivation. *Science* 300(5616):131-5.
- Prestridge DS (1995) Predicting Pol II Promoter sequences using transcription factor binding sites. *J Mol Biol* 249: 923-32.
- Priddle TH, Cranfield M, Hewitson KS, Williams NA, Groome NP, Schofield CJ, Esiri MM, Crow TJ (2002) Demonstration of ProtocadherinX/Y immunoreactive neurones in human cerebral cortex. *Schizophr Res* 53(Supplement): 68-69.
- Priddle TH, Cranfield M, Hewitson KS, Groome NP, Schofield CJ, Esiri MM, Crow TJ (2004) Expression of protocadherinX/Y in the developing human brain." *Am J Med Genet B Neuropsychiatr Genet* 130B(1): 137-137.
- Priest JH, Heady JE, Priest RE (1967) Delayed onset of replication of human X chromosomes. *J Cell Biol* 35(2):483-7.
- Pugh BF, Tjian R (1991) Transcription from a TATA-less promoter requires a multisubunit TFIID complex. *Genes Dev* 5(11):1935-45.
- Qiu P, Benbow L, Liu S, Greene JR, Wang L (2002) Analysis of a human brain transcriptome map. *BMC Genomics* 3(1):10.
- Rabinowicz T, Petetot JM, Gartside PS, Sheyn D, Sheyn T, de CM (2002) Structure of the cerebral cortex in men and women. *J Neuropathol Exp Neurol* 61(1):46-57.
- Redies C, Vanhalst K, Roy F (2005) delta-Protocadherins: unique structures and functions. *Cell Mol Life Sci* 62(23):2840-52.
- Rice WR (1984) Sex chromosomes and the evolution of sex dimorphism. *Evolution* 38(735-742).
- RiceWR (1987) Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* 116(1):161-7.

- Rice WR (1992) Sexually antagonistic genes: experimental evidence. *Science* 256(5062):1436-9.
- Rice WR (1994) Degeneration of a nonrecombining chromosome. *Science* 263(5144):230-2.
- Riggs AD (1975) X-inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet* 14(1):9-25.
- Robinson-Rechavi M, Laudet V (2001) Evolutionary rates of duplicate genes in fish and mammals. *Mol Biol Evol* 18(4):681-3.
- Ross MT et al. (2005) The DNA sequence of the human X chromosome. *Nature* 434(7031):325-37.
- Rozas J, Sanchez-Delbarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496-2497.
- Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423 (6942):873-6.
- Saffer JD, Jackson SP, Annarella MB (1991) Developmental expression of Sp1 in the mouse. *Mol Cell Biol* 11(4):2189-99.
- Saifi GM, Chandra HS (1999) An apparent excess of sex- and reproduction-related genes on the human X chromosome. *Proc Biol Sci* 266(1415):203-9.
- Sano K, Tanihara H, Heimark RL, Obata S, Davidson M, St John T, Taketani S, Suzuki S (1993) Protocadherins: a large family of cadherin-related molecules in central nervous system. *EMBO J* 12(6):2249-56.
- Sargent CA, Briggs H, Chalmers IJ, Lambson B, Walker E, Affara NA (1996) The sequence organization of Yp/proximal Xq homologous regions of the human sex chromosomes is highly conserved. *Genomics* 32(2):200-9.

Sargent CA, Boucher CA, Blanco P, Chalmers IJ, Highet L, Hall N, Ross N, Crow T, Affara NA (2001) Characterization of the human Xq21.3/Yp11 homology block and conservation of organization in primates. *Genomics* 73(1):77-85.

Sarich VM, Wilson AC (1973) Generation time and genomic evolution in primates. *Science* 179(78):1144-7.

Saxena R, Brown LG, Hawkins T, Alagappan RK, Skaletsky H, Reeve MP, Reijo R, Rozen S, Dinulos MB, Disteche CM, Page DC (1996) The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. *Nat Genet* 14(3):292-9.

Schmidt M, Migeon BR (1990) Asynchronous replication of homologous loci on human active and inactive X chromosomes. *Proc Natl Acad Sci U S A* 87(10):3685-9.

Schneider-Gadicke A, Beer-Romero P, Brown LG, Nussbaum R, Page DC (1989) ZFX has a gene structure similar to ZFY, the putative human sex determinant, and escapes X-inactivation. *Cell* 57(7):1247-58.

Schwartz A, Chan DC, Brown LG, Alagappan R, Pettay D, Disteche C, McGillivray B, de la Chapelle A, Page DC (1998) Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum Mol Genet* 7(1):1-11.

Senzaki K, Ogawa M, Yagi T (1999) Proteins of the CNR family are multiple receptors for Reelin. *Cell* 99(6):635-47.

Sinclair AH, Berta P, Palmer MS, Hawkins JR, Griffiths BL, Smith MJ, Foster JW, Frischauf AM, Lovell-Badge R, Goodfellow PN (1990) A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* 346(6281):240-4.

Sharman GB (1971) Late DNA replication in the paternally derived X chromosome of female kangaroos. *Nature*. 230(5291):231-2.

Shmookler Reis RJ, Goldstein S (1982) Variability of DNA methylation patterns during serial passage of human diploid fibroblasts. *Proc Natl Acad Sci U S A* 79(13):3949-53.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang SP, Waterston RH, Wilson RK, Rozen S, Page DC (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423(6942):825-37.

Solari AJ (1974) The behavior of the XY pair in mammals. *Int Rev Cytol* 38(0):273-317.

Stevanovic M, Lovell-Badge R, Collignon J, Goodfellow PN (1993) SOX3 is an X-linked gene related to SRY. *Hum Mol Genet* 2(12):2013-8.

Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, Bridges CR, Shrager JB, Minugh-Purvis N, Mitchell MA (2004) Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* 428(6981):415-8.

Strahl BD, Allis CD (2000) The language of covalent histone modifications. *Nature* 403(6765):41-5.

Sudbrak R, Wieczorek G, Nuber UA, Mann W, Kirchner R, Erdogan F, Brown CJ, Wohrle D, Sterk P, Kalscheuer VM, Berger W, Lehrach H, Ropers HH (2001) X chromosome-specific cDNA arrays: identification of genes that escape from X-inactivation and other applications. *Hum Mol Genet* 10(1):77-83.

Suzuki S (1993) Protocadherins: a large family of cadherin-related molecules in central nervous system. *EMBO J* 12(6):2249-56.

Swaab DF, Hofman MA (1984) Sexual differentiation of the human brain. A historical perspective. *Prog Brain Res* 61:361-74.

Takagi N, Sasaki M (1975) Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. *Nature* 256(5519):640-2.

References I

- Takahata N, Satta Y (1997) Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proc Natl Acad Sci U S A* 94(9):4811-5.
- Tasic B, Nabholz CE, Baldwin KK, Kim Y, Rueckert EH, Ribich SA, Cramer P, Wu Q, Axel R, Maniatis T (2002) Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. *Mol Cell* 10(1):21-33.
- Tepass U, Truong K, Godt D, Ikura M, Peifer M (2000) Cadherins in embryonic and neural morphogenesis. *Nat Rev Mol Cell Biol* 1(2):91-100.
- Turner BM (2002) Cellular memory and the histone code. *Cell* 111(3):285-91.
- Uddin M, Wildman DE, Liu G, Xu W, Johnson RM, Hof PR, Kapatos G, Grossman LI, Goodman M (2004) Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. *Proc Natl Acad Sci U S A* 101(9):2957-62.
- Vacca M, Matarazzo MR, Jones J, Spalluto C, Archidiacono N, Ma P, Rocchi M, D'Urso M, Chen EY, D'Esposito M, Mumm S (1999) Evolution of the X-specific block embedded in the human Xq21.3/Yp11.1 homology region. *Genomics* 62(2):293-6.
- Vallender EJ, Lahn BT (2004) How mammalian sex chromosomes acquired their peculiar gene content. *Bioessays* 26(2):159-69.
- Van de Peer Y, Taylor JS, Braasch I, Meyer A (2001) The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J Mol Evol* 53(4-5):436-46.
- Vanhalst K, Kools P, Staes K, van Roy F, Redies C (2005) delta-Protocadherins: a gene family expressed differentially in the mouse brain. *Cell Mol Life Sci* 62(11):1247-59.
- Vallender EJ, Pearson NM, Lahn BT (2005) The X chromosome: not just her brother's keeper. *Nat Genet* 37(4):343-5.

- Vawter MP, Evans S, Choudary P, Tomita H, Meador-Woodruff J, Molnar M, Li J, Lopez JF, Myers R, Cox D, Watson SJ, Akil H, Jones EG, Bunney WE (2004) Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes. *Neuropsychopharmacology* 29(2):373-84.
- Vermeesch JR, Petit P, Kermouni A, Renauld JC, Van Den Berghe H, Marynen P (1997) The IL-9 receptor gene, located in the Xq/Yq pseudoautosomal region, has an autosomal origin, escapes X-inactivation and is expressed from the Y. *Hum Mol Genet* 6: 1-8.
- Vleminckx K, Kemler R (1999) Cadherins and tissue formation: integrating adhesion and signaling. *Bioessays* 21(3):211-20.
- Wagner A (1998) The fate of duplicated genes: loss or new function? *Bioessays* 20(10):785-8.
- Walsh JB (1995) How often do duplicated genes evolve new functions? *Genetics* 139(1):421-8.
- Wang PJ, McCarrey JR, Yang F, Page DC (2001) An abundance of X-linked genes expressed in spermatogonia. *Nat Genet* 27(4):422-6.
- Wang PJ, Page DC, McCarrey JR (2005) Differential expression of sex-linked and autosomal germ-cell-specific genes during spermatogenesis in the mouse. *Hum Mol Genet* 14(19):2911-8.
- Wang X, Su H, Bradley A (2002) Molecular mechanisms governing Pcdh-gamma gene expression: evidence for a multiple promoter and cis-alternative splicing model. *Genes Dev* 16(15):1890-905.
- Watson CM, Hughes RL, Cooper DW, Gemmell RT, Loebel DA, Johnston PG (2000) Sexual development in marsupials: genetic characterization of bandicoot siblings with scrotal and testicular maldevelopment. *Mol Reprod Dev* 57(2):127-34.

References I

- Winter H, Langbein L, Krawczak M, Cooper DN, Jave-Suarez LF, Rogers MA, Praetzel S, Heidt PJ, Schweizer J (2001) Human type I hair keratin pseudogene phihHaA has functional orthologs in the chimpanzee and gorilla: evidence for recent inactivation of the human gene after the Pan-Homo divergence. *Hum Genet* 108(1):37-42.
- Wu CI, Xu EY (2003) Sexual antagonism and X-inactivation--the SAXI hypothesis. *Trends Genet* 19(5):243-7.
- Wu Q, Maniatis T (1999) A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* 97(6):779-90.
- Wyckoff GJ, Li J, Wu CI (2002) Molecular evolution of functional genes on the mammalian Y chromosome. *Mol Biol Evol* 19(9):1633-6.
- Xue F, Tian XC, Du F, Kubota C, Taneja M, Dinnyes A, Dai Y, Levine H, Pereira LV, Yang X (2002) Aberrant patterns of X chromosome inactivation in bovine clones. *Nat Genet* 31(2):216-20.
- Yagi T, Takeichi M (2000) Cadherin superfamily genes: functions, genomic organization, and neurologic diversity. *Genes Dev* 14(10):1169-80.
- Yan GZ, Ziff EB (1997) Nerve growth factor induces transcription of the p21 WAF1/CIP1 and cyclin D1 genes in PC12 cells by activating the Sp1 transcription factor. *J Neurosci* 17(16):6122-32.
- Yang X, Chen MW, Terry S, Vacherot F, Chopin DK, Bemis DL, Kitajewski J, Benson MC, Guo Y, Buttyan R (2005) A human- and male-specific protocadherin that acts through the wnt signaling pathway to induce neuroendocrine transdifferentiation of prostate cancer cells. *Cancer Res* 65(12):5263-71.
- Yoo J, Jeong MJ, Kwon BM, Hur MW, Park YM, Han MY (2002) Activation of dynamin I gene expression by Sp1 and Sp3 is required for neuronal differentiation of N1E-115 cells. *J Biol Chem* 277(14):11904-9.
- Yoshida K, Sugano S (1999) Identification of a novel protocadherin gene (PCDH11) on the human XY homology region in Xq21.3. *Genomics* 62(3):540-3.

- Yunis JJ, Sawyer JR, Dunham K (1980) The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. *Science* 208(4448):1145-8.
- Zechner U, Wilda M, Kehrer-Sawatzki H, Vogel W, Fundele R, Hameister H (2001) A high density of X-linked genes for general cognitive ability: a run-away process shaping human evolution? *Trends Genet* 17(12):697-701.
- Zhang J (2003) Evolution of the human ASPM gene, a major determinant of brain size. *Genetics* 165(4):2063-70.
- Zhang LP, Stroud J, Eddy CA, Walter CA, McCarrey JR (1999) Multiple elements influence transcriptional regulation from the human testis-specific PGK2 promoter in transgenic mice. *Biol Reprod* 60(6):1329-37.
- Zhang P, Gu Z, Li WH (2003) Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol.* 2003;4(9):R56.
- Zhao H, Shiina H, Greene KL, Li LC, Tanaka Y, Kishi H, Igawa M, Kane CJ, Carroll P, Dahiya R (2005) CpG methylation at promoter site -140 inactivates TGFbeta2 receptor gene in prostate cancer. *Cancer* 104(1):44-52.

II

***PCDH11X/Y* EVOLUTIONARY PATTERNS IN HUMAN POPULATIONS**

INTRODUCTION

1. The significance of genetic variation

Differences between genomes may be informative over different resolutions and time-scales. Comparisons between two closely related species can reveal similarities reflecting a recently shared common ancestor, as well as differences due to an independent evolutionary trajectory of each species after their divergence. Variation between individuals within a species, i.e. polymorphism, is particularly relevant for elucidating species-specific evolutionary processes and demographic events in both ancient and historical times.

The study of human genetic variation relies on the analysis and interpretation of the patterns of allele frequency change at polymorphic sites, in relation to geography and time. In our species, although geographical structuring can be detected to a certain degree, most variation lies between individuals, within populations, rather than between populations (e.g. Barbujani *et al.*, 1997; Jorde *et al.*, 2000). Moreover, genetic diversity within and among continents seems to be better characterized by gradients of allele frequencies (Serre and Pääbo, 2004) than by sharp discontinuities. The concept of population is however important in studies of human genetic diversity since the dynamics of variation can only be assessed in groups of individuals, and different continents are often sampled in order to evaluate the relative weight of global and more geographically restricted trends.

2. Evolutionary and population genetics frameworks

Genetic diversity studies aim not only to describe the patterns of present diversity but also to infer about the processes that might have contributed to the observed patterns. Such inferences must rely on models of the evolutionary processes involved.

Human genetic diversity was defined to a great extent by the demographic scenario at the dawn of the species. This scenario is the framework upon which population genetics and evolutionary studies rely.

2.1. Genetic perspectives on human origins

The generally accepted "Out-of-Africa" or "Recent African Origin" (RAO) model postulates that all modern humans have originated recently in Africa, ~120-200 thousand years ago (Kya), from a small population of anatomically modern human ancestors that expanded in size and in geographical distribution, replacing the archaic

hominids already present in other continents. This model has been proposed following the first studies of human molecular diversity (Cann *et al.*, 1987) and based on the archaeological evidence for the presence of modern human morphology first in Africa (~130 Kya) and only substantially latter in other parts of the world (probably after 50 Kya). As corollary, low diversity is expected in human populations, due to a relatively recent common ancestor and a small initial effective population size (N_e) and greater diversity is expected within Africa, since all other continents are assumed to have been colonized by a small number of individuals with African origin. Accordingly, levels of human intraspecific diversity are generally low, compared to great apes (Gagneux *et al.*, 1999; Kaessmann *et al.*, 2001; Stone *et al.*, 2002) suggesting a small effective size for humans (~10 000; Harpending *et al.*, 1998). Comparisons of geographic distribution of genetic diversity in human populations have generally contributed evidence supporting this model, by showing the existence of higher levels of diversity in African populations for mtDNA sequences (e.g. Vigilant *et al.*, 1991; Ingman *et al.*, 2000), Y chromosome sequences (Shen *et al.*, 2000; Thomson *et al.*, 2000), X chromosome sequences (e.g. Kaessman *et al.*, 1999; Labuda *et al.*, 2000; Hammer *et al.*, 2004), autosomal microsatellites (e.g. Bowcock *et al.*, 1994; Jorde *et al.*, 2000) and autosomal sequences (e.g. Zhao *et al.*, 2000; Yu *et al.*, 2001) and by demonstrating that allelic diversity outside Africa often represents a subset of that found within Africa (e.g. Ingman *et al.*, 2000; Tishkoff *et al.*, 1996; Calafell *et al.*, 1998; Yu *et al.*, 2002). Phylogenies of non-recombining loci, mitochondrial DNA (mtDNA) and Y chromosome, have consistently revealed earliest branching lineages among African populations (Cann *et al.*, 1987; Ingman *et al.*, 2000; Underhill *et al.*, 2000; Semino *et al.*, 2002) and in general higher frequencies are often found in Africa for ancestral alleles.

Implicit within the model is the occurrence of major demographic changes underlying the origin of non African populations, namely an early population bottleneck related to the migration Out-of-Africa (~50 Kya) and one or more episodes of Pleistocene population expansion, 40-200 Kya. Such a demographic scenario would have affected the patterns of genetic diversity, being detectable by the existence of an excess of rare variants, unimodal mismatch distributions and star shaped genealogies at different unlinked loci (Harpending and Rogers, 2000). Signatures of global population expansions (Pritchard *et al.*, 1999) and more restricted events in Africa (Reich and Goldstein, 1998) or outside Africa (Rogers and Harpending, 1992; Kimmel *et al.*, 1998) have been detected. Often, however, the large confidence intervals implied in the data do not allow excluding the possibility of more recent expansions. The low diversity of human populations, as discussed previously, together with the finding of lower levels of linkage disequilibrium in African populations (Tishkoff *et al.*,

1996; Reich *et al.*, 2001; Gabriel *et al.*, 2002) have also been interpreted as evidence for a shorter history of diversity outside Africa as a consequence of a recent bottleneck.

Although the body of data are in support of the RAO model, some of its simplistic assumptions, i.e. an early population bottleneck associated to a specific migration event of an undifferentiated population out of Africa followed by a single large population expansion, have been challenged by recent results (reviewed in Excoffier, 2002 and in Goldstein and Chikhi, 2002). Evidence of very old global TMRCA ($>1\text{My}$) at many nuclear loci and relatively high values of worldwide population differentiation are not easily reconciled with a small and well mixed ancestral population having originated present human diversity (Excoffier, 2002 and references therein; Goldstein and Chikhi, 2002). The emerging view is incorporating in the RAO model the occurrence of early population substructuring and considering that the progressive colonization of the world by one of the subpopulations existing in Africa was achieved by a series of range expansions (Excoffier, 2002). A metapopulation model for human evolution has also been proposed where population extinction and rebirth by colonization are considered and where migration and admixture events are also taken into account (Harding and McVean, 2004).

2.2. The re-shaping of human diversity

Successive waves of dispersal from the African homeland are presumed from interpretation of archaeological evidence, a first one into south, reaching Australia 50-60 Kya and latter migrations into Asia and Europe ~40 Kya. The Americas were the last continent to be colonized ~14 Kya. Following this major movements, modern human populations have experienced local demographic events that have contributed to shape the patterns of diversity within each continent, such as population bottlenecks, expansions and range migrations.

Within Eurasia, following the original Paleolithic migration from Africa which reached Europe through the Middle East, two population expansions are noteworthy. The most ancient concern the re-expansions occurred after the Last Glacial Maximum, 23 Kya to 14 Kya, of small population isolates from southern refugia in the peninsulas of Iberia, Italy and in the Balkans. In more recent times, a population expansion related to the Neolithic advent of agriculture, during the last 10 Kya as taken place. East-West gradients of allele frequencies have been found within Europe and interpreted as signatures of the Neolithic movement of farmers from Near East, merging with local indigenous European Mesolithic populations (Cavalli-Sforza, Menozzi and Piazza, 1994; in Jobling, Hurles and Tyler-Smith, 2004; Semino *et al.*, 1996; Chikhi *et al.*, 1998;

Semino *et al.*, 2000). There is also some evidence for the contribution of Paleolithic post-glacial expansions and subsequent migrations to the existence of frequency clines in Central and Western Europe (Torroni *et al.*, 1998, 2001; Richards *et al.*, 2000; Semino *et al.*, 2000; Pereira *et al.*, 2005a). Although a substantial portion of the European gene pool is of Upper Paleolithic origin, the bottleneck associated with population contraction during the Last Glacial Maximum followed by localized re-expansions, and the latter contribution of Neolithic immigrants from Near East, are thought to have reshaped the genetic landscape of Europe (Richards *et al.*, 1996, 2000; Semino *et al.*, 2000; Pereira *et al.*, 2005a).

In Sub-Saharan Africa, a major population movement has shaped diversity in recent times - the Bantu agricultural expansion, in the past 3 000 years (Phillipson, 1993 in Jobling, Hurles and Tyler-Smith, 2004). This involves a complex pattern of dispersals of iron-working Bantu-speaking farmers from a centre in Nigeria, southwards, along the Eastern and Western coasts of Africa, which is traceable in present diversity patterns (Pereira *et al.*, 2001, 2002; Salas *et al.*, 2002; Cruciani *et al.*, 2002; Beleza *et al.*, 2005).

3. The sex chromosomes in population genetics

Genetic variation is created by mutation, reshuffled by recombination and modulated by gene flow within and between populations. Demographic events such as bottlenecks, expansions and migrations re-shape diversity patterns across the genome, as selective forces might do on a gene- or locus-specific basis. The particular characteristics of human sex chromosomes, acquired throughout mammalian evolution, define their dynamics within human populations and confer them specific properties concerning each of the referred population-genetics parameters.

3.1. Properties of the Y chromosome

The dramatic specialization of the Y chromosome in male sex determination and reproduction makes it the most notable outlier, being almost completely exempt from recombination in males. Notably, genetic isolation drastically diminishes the effective population size (N_e) of Y chromosomes to one-quarter of that of any autosome and one-third of that of the X chromosome. Within populations, Y chromosomes are thus more susceptible to stochastic sampling effects such as genetic drift, accelerating the differentiation of Y chromosome haplotypes between populations and contributing to geographic structuring.

Moreover, the absolute linkage of the NRY (or MSY) predicts that selective effects on any Y-linked locus will affect the whole chromosome, and purifying selection is expected to be relevant, due to the great number of genes on the Y chromosome related to male fertility (reviewed by Jobling and Tyler-Smith, 2003). Less clear is the importance of positive selection in shaping Y chromosome diversity since reports describing direct evidence of selection, or indicating possible mechanisms involved, are scarce in the literature. A characteristic which has been proposed to confer a moderate selective advantage, i.e., the inversion polymorphism in *PRKY* which is protective against X-Y ectopic recombination underlying an infrequent class of XX maleness leading to infertility (Jobling *et al.*, 1998), might not be sufficient to overweight the effect of genetic drift (Jobling and Tyler-Smith, 2003).

It has been argued that a recent common ancestry of human Y chromosomes, with an estimated TMRCA of 45,000-90,000 years, from microsatellite diversity (Pritchard *et al.*, 1999), and of 59,000 years, from sequence variation (Thomson *et al.*, 2000), as well as an excess of rare alleles on the Y chromosome (Underhill *et al.*, 1997; Pritchard *et al.*, 1999; Shen *et al.*, 2000; Thomson *et al.*, 2000) point to a role for selection in the shaping of Y chromosome diversity. However, these observations are also consistent with plausible demographic scenarios of recent population expansions. The existence of reproductive habits that confer higher reproductive variance to males, such as polygyny, can further reduce the effective population size of Y chromosomes, contributing to additional reduction of diversity in this chromosome (Charlesworth *et al.*, 2001; Dupanloup *et al.*, 2003; Wilder *et al.*, 2004).

Taking into account the small effective population size of Y chromosomes and increased power of both genetic drift and selection to eliminate or fixate new alleles, in the absence of recombination, lower levels of diversity are expected in this chromosome, compared to the rest of the genome. However, mutation must be also accounted for, considering that a male bias in mutation rates has been observed (Miyata *et al.*, 1987; Shimmin *et al.*, 1993; Bohossian *et al.*, 2000; Erlandsson *et al.*, 2000; Ebersberger *et al.*, 2002; Makova and Li, 2002) which would affect more intensely the male-specific Y chromosome. However, the higher mutation rate in the male germ line apparently has not been sufficient to prevail over the effects aforementioned, as lower levels of diversity on the Y chromosome have been described (Dorit *et al.*, 1995; Hammer, 1995; Thomson *et al.*, 2000).

Mutation rates at Y-specific tri- and tetranucleotide microsatellites have been determined directly, in father-son pairs (Kayser *et al.*, 2000; Gusmão *et al.* 2005) and in sperm pools (Holtkemper *et al.*, 2001) or estimated by indirect methods, using deep-rooting pedigrees (Heyer *et al.*, 1997), and within Y chromosome haplogroups in

chosen populations (Zhivotovsky *et al.*, 2004). An average rate of 3×10^{-3} (Kayser *et al.*, 2000) is often used as reference, although differences in relative diversity within different Y lineages has been observed (Bosh *et al.*, 1999; Carvalho-Silva *et al.*, 1999), likely due to marker- and allele-specific mutation rates (Ellegren, 2000). The fact that autosomal and Y-linked tri and tetranucleotide repeats have similar mutation rates and properties indicates that replication slippage (Ohta and Kimura, 1973; Di Rienzo *et al.*, 1994) might be the common mutation mechanism underlying diversity at these loci.

The recent finding of extensive gene conversion between paralogs on the Y chromosome (Rozen *et al.*, 200) has refined our view of the MSY as a non recombining genetic system. However, these events are restricted to a few gene families within this chromosome and their effects are distinct from X-Y crossing over. Recombination and gene conversion could occur between highly homologous X-Y sequences and would be detectable by a sequence homogenization between the sex chromosomes on the exchanged block, or in the converted sequence, and by higher divergence between Y haplotypes in those regions, since the minimum divergence between X and Y sequences is ~1%, while divergence between Y haplotypes is only ~0.05% (Jobling and Tyler-Smith, 2003). Gene conversion between X-Y gametologues has not been detected in highly homologous regions in human sex chromosomes in an extensive DHPLC survey (Giouzeli *et al.*, 2004) and was only detected twice in an interspecific phylogeny of the cat family (Pecon-Slattey *et al.*, 2000).

An additional outcome of the escape of the Y chromosome from recombination is that haplotypes are passed from father to son throughout generations without reshuffling of alleles and thus only changed by mutation. As a consequence, recurring to markers with low mutation rates, lineages are easily defined and the phylogeny of worldwide Y chromosomes can be reconstructed. A robust Y chromosome phylogenetic tree based on binary markers has been established, where groups of evolutionarily related lineages or haplogroups are organized into clades (YCC, 2002).

The Y chromosome phylogeny has been particularly useful in reconstructing, from a male perspective, the worldwide human past movements previously discussed (Underhill *et al.*, 2001; Jobling and Tyler-Smith, 2003), being often complemented with information obtained from mtDNA, which is specifically transmitted by females. However, the diversity within this locus has also been the more drastically reshaped by subsequent local expansions and migrations (Underhill *et al.*, 2001; Jobling and Tyler-Smith, 2003) and this is demonstrated by the presence of highly frequent derived Y chromosome lineages in populations where expansions have occurred, such as R(xR1a) haplogroups in Western Europe (Semino *et al.*, 2000), or in populations in the

route of migrations, such as E3a haplogroups in Sub-Saharan Africa (Underhill *et al.*, 2001; Cruciani *et al.*, 2002; Pereira *et al.*, 2002; Beleza *et al.*, 2005).

These examples clearly illustrate the difficulty in inferring ancient movements and demographic events from present Y chromosome diversity in human populations, since more recent and complex events might have distorted original diversity patterns.

3.2. Properties of the X chromosome

The X chromosome largely resembles an autosome, in the sense that it comprises an assembly of blocks of haplotypes reshuffled to different extents by recombination (Daly *et al.*, 2001; Phillips *et al.*, 2003). However, since it is present in single copy in males, X chromosomes have been less exposed to male germline mutagenic effects than autosomes, being characterized by lower mutation rates (Mcvean *et al.*, 1997) and an effective population size (N_e) that is three-quarters of that of the autosomes and thus threefold higher than the Y chromosome. Therefore, the lower diversity levels found on the X chromosome compared to the autosomes (Kaessman *et al.*, 1999) are not surprising, being in some cases close to half of that of the autosomes (Sachidanandam *et al.*, 2001), as theoretically predicted (Schaffner, 2004). Moreover, since X chromosomes recombine only in females, recombination rates are lower than the genome average and result in higher levels of linkage disequilibrium (Kong *et al.*, 2002). This effect is further enhanced by the reduced effective population size and coalescence time of extant X chromosomes, compared to autosomes. Nevertheless the TMRCA of X chromosomes is still substantially greater than that of Y chromosomes (1.8 My for *PDHA1*, Harris and Hey, 1999; ~500 Ky for Xq13.3 data, Kaessman *et al.*, 1999), which means that X-linked markers are informative at greater time-depths than Y-linked markers.

Hemizigosity of alleles in males allows selection to be more effective on X-linked genes, compared to autosomes. Although selective forces are not expected to have such a dramatic effect as predicted on the Y chromosome, there are a few reports in the literature relating low diversity levels on the X chromosome, particularly an excess of low frequency alleles, with the action of positive selection at linked loci (Harris and Hey, 1999 and 2001; Nachman and Crowell, 2000; Nachman *et al.*, 2004). In a recent survey including several X-linked genes Hammer and co-workers (2004) reported high heterogeneity in diversity patterns across loci, suggesting that a simple population bottleneck is not sufficient to explain the data and either complex demographic models or directional selection would have to be invoked to

accommodate the different degree of diversity reduction found in non-African populations.

3.3. Properties of the Xq21.3/Yp11.2 region

The contrasting properties of sex chromosomes imply that X- and Y-linked sequences will be exposed to different environments, regarding the occurrence of mutation and recombination, and to specific evolutionary dynamics, since genetic drift and demographic events will have a variable impact on genetic diversity. The study of patterns of variation within the non recombining homologous regions of sex chromosomes can thus provide new insights on the relative contribution of sex-specific processes to the generation and shaping of genome diversity. A major caveat of this type of comparisons is the considerable divergence that X-Y homologues have accumulated due to long periods of genetic isolation, in the absence of recombination (Ross *et al.*, 2005), rendering direct comparisons impractical and prone to biases. This limitation can however be circumvented, by analysing recently transposed highly homologous regions between human sex chromosomes (Skaletsky *et al.*, 2003). The homology block in Xq21.3/Yp11.2 is thus an optimal model, since it is characterized by high X-Y sequence identity (~99%; Skaletsky *et al.*, 2003) in the absence of recombination. Sequences within Xq21.3 undergo recombination in female meiosis and are more exposed to female than to male germline mutation rates, while the Yp11.2 region is genetically isolated and exclusively subject to male germline mutation rates. Moreover, X chromosome lineages should be older than Y chromosome lineages and X-linked markers in these regions are predicted to be more informative since they are less prone to stochastic or directional diversity loss. Finally, while in Yp11.2 only the male history is recorded, in Xq21.3 male and female histories are intermingled.

AIMS

1. To compare the levels of diversity in human populations of X-linked microsatellites and their counterparts on the Y chromosome, in a region of high X/Y sequence homology, the Xq21.3/Yp11.2 homology block.
2. To contrast the evolutionary patterns in human populations of *PCDH11X* and *PCDH11Y*, within this homology block, through the analysis of intragenic biallelic markers, in order to understand the evolutionary forces acting on each chromosome after the transposition event.
3. To allocate *PCDH11Y* SNPs within the Y chromosome phylogeny in order to evaluate the informativeness of these markers in further refining known Y chromosome clades.

RESULTS

1. Microsatellite diversity comparisons and evolutionary patterns of *PCDH11X/Y*

Article 3

Lopes AM, Calafell F, Amorim A (2004) Microsatellite variation and evolutionary history of *PCDHX/Y* gene pair within the Xq21.3/Yp11.2 hominid-specific homology block. *Mol Biol Evol* 21(11):2092-2101.

Microsatellite Variation and Evolutionary History of *PCDHX/Y* Gene Pair Within the Xq21.3/Yp11.2 Hominid-Specific Homology Block

Alexandra M. Lopes,*[‡] Francesc Calafell,[†] and António Amorim*[‡]

*IPATIMUP, Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Portugal; [†]Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona; and [‡]Faculdade de Ciências, Universidade do Porto, Portugal

To better understand the evolutionary dynamics of repetitive sequences in human sex chromosomes, we have analyzed seven new X/Y homologous microsatellites located within *PCDHX/Y*, one of the two recently described gene pairs in the Xq21.3/Yp11.2 hominid-specific homology block, in samples from Portugal and Mozambique. Sharp differences were observed on X/Y allele distributions, concerning both the presence of private alleles and a different modal repeat length for X-linked and Y-linked markers, and this difference was statistically significant. Higher diversity was found in X-linked microsatellites than in their Y chromosome counterparts; when comparing populations, Mozambicans showed more allele diversity for the X chromosome, but the contrary was true for the Y chromosome microsatellites. Evolutionary patterns, relying on intragenic *PCDHX/Y* SNPs, also revealed distinct scenarios for X and Y chromosomes. Greater microsatellite diversity was displayed by African X chromosomes within the most common haplotypes shared by both populations, whereas higher microsatellite diversity was found in Portugal for the ancestral Y chromosome haplotype. The most frequent *PCDHX/Y* haplotype in Portuguese was the derived one, and it was not found in Mozambicans. TMRCA estimated by the ρ parameter resulted in 13,700 years (7,500–20,000 years), which is consistent with a recent, post-Out-of-Africa origin for this haplotype. In conclusion, the newly described microsatellite loci generally displayed greater X-linked to Y-linked diversity and this pattern was also detected with slower evolving markers, with a remarkable differentiation between populations observed for Y chromosome haplotypes and, thus, greater divergence among Y chromosomes in human populations.

Introduction

The study of patterns of variation in nonrecombining homologous regions of the sex chromosomes provides the unique opportunity of approaching the mechanisms underlying the generation and shaping of diversity in two DNA sequences evolving in different genomic environments. The existence of particular genomic environments in these regions is due to both the differentiation experienced by these chromosomes during evolution, outside the pairing regions (Graves 1995), and to the fact that males and females bear different sex chromosome complements. This implies that, within these regions, recombination events are restricted to the X chromosome on females and that sequences located on the X and Y chromosomes are subject to sex-specific dynamics. Moreover, because of a 3:1 ratio of X/Y chromosome population effective sizes, sex chromosomes are exposed to genetic drift to different extents (Jobling and Tyler-Smith 2003).

Until now, few X/Y homologous microsatellite markers have been described (Scozzari et al. 1997; Karafet et al. 1998; Carvalho-Silva and Pena 2000; Dupuy, Gedde-Dahl, and Olaisen 2000; Kersting et al. 2001). Moreover, the markers analyzed were not always the most convenient to perform direct X-Y comparisons because of limitations concerning either the different complexity of repeat structure between sex chromosomes (Carvalho-Silva et al. 1999; Dupuy, Gedde-Dahl, and Olaisen. 2000) or the assignment of alleles to the respective sex chromosome by male/female frequency-based inferences (Karafet et al.

1998; Kersting et al. 2001), factors that may have blurred the patterns of diversity.

In the present study, we analyzed, in a chromosome-specific manner, seven new X/Y homologous microsatellites located within *PCDHX/Y*, one of the two recently described genes in the Xq21.3/Yp11.2 hominid-specific homology block (Blanco et al. 2000; Blanco-Arias, Sargent, and Affara 2002). Divergence between X and Y sequences in these regions began 3 to 4 MYA, after the translocation that isolated one of the blocks on the nonrecombining region of the Y chromosome (NRY) (Schwartz et al. 1998). Maintenance of high X-Y sequence identity (99%) (Skaletsky et al. 2003) makes this region an optimal model for the comparisons we aimed to perform, because we can in this way minimize the confounding effects of extensive divergence between X-Y homologs. Both diversity comparisons and chromosome-specific patterns of evolution in these regions in human populations are particularly interesting, because they belong to the restricted group of rearranged chromosomal segments that could have been involved in human speciation (Navarro and Barton 2003; Rieseberg and Livingstone 2003).

To perform these comparisons we have analyzed samples from two available and well-characterized populations, Mozambique (African) and Portugal (European). In each of these populations, a different and highly frequent Y lineage can be found as a result of specific demographic events. These events include, in Europe, the expansion from isolated population nuclei in refuges after the Last Glacial Maximum, which occurred 20,000 to 13,000 years ago and, in Africa, the more recent Bantu agricultural expansion, in the past 3,000 years. These populations would nevertheless guarantee the inclusion of ancient as well as younger Y chromosome lineages in this study (Pereira et al. 2000, 2002).

Key words: Microsatellite variation, sex chromosomes, *ProtocadherinX/Y*, *Homo sapiens*, evolution.

E-mail: alopes@ipatimup.pt.

Mol. Biol. Evol. 21(11):2092–2101. 2004

doi:10.1093/molbev/msh218

Advance Access publication August 5, 2004

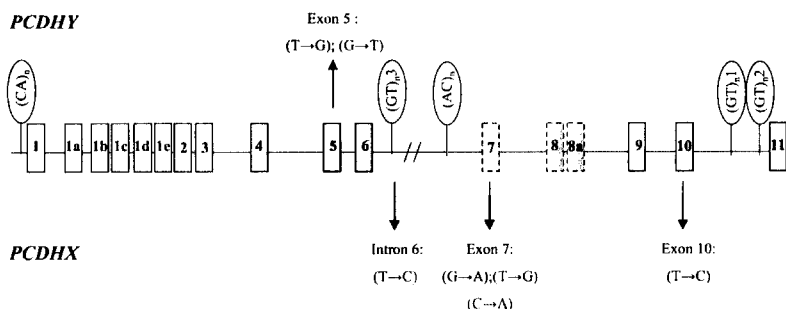


Fig. 1.—Relative positions of the polymorphisms analyzed, within *PCDHX/Y* genes. Bars represent previously described exons: dark gray bars, coding exons; bars enclosed with a broken line, exons absent in *PCDHX*. X/Y microsatellites are represented inside circles. Binary markers are above (*PCDHX*) and below (*PCDHX*) the scheme. Maximum distance between *PCDHX* markers is approximately 700 kb.

To further clarify the evolutionary forces acting on these homologous regions in the framework of the relationships between the two human populations under study and to gain in this way some perspective on global and population-specific X-Y chromosome comparisons, we proceeded to the analysis of haplotypes defined by the microsatellite markers presently described and anchored the microsatellite information on previously reported binary markers within *PCDHX/Y* genes (Giouze et al. 2004).

Materials and Methods

Sampling

Blood samples were collected from 112 healthy unrelated males from one European population (Portugal) and one African population (Mozambique). Mozambique sampling was performed in Maputo (capital of Mozambique) and included only the prevailing dialects or languages of the Bantu group. Genomic DNA was extracted using a Chelex resin method (Lareu et al. 1994) from humans. A *Pan troglodytes* male sample was provided by Dr. Angel Carracedo from the Institute of Legal Medicine, University of Santiago de Compostela (Spain) and one *Gorilla gorilla* male sample was provided by Patricia Blanco from the Department of Pathology, University of Cambridge (UK).

Microsatellite Detection and Selection

Search for new microsatellites within the *PCDHX* genes was performed in silico in *Homo sapiens* chromosome X and Y genomic contigs (GeneBank accession numbers NT_011651 and NT_011896, respectively) using Tandem Repeat Finder (Benson 1999) available on (<http://tandem.biomath.mssm.edu/trf/trf.html>) in a region comprising *PCDHX/Y* genes and approximately 10 kb of flanking sequences. Potential polymorphic repeats were selected throughout the gene by X/Y homology in flanking regions and maximum (always greater than eight) number of repeated motifs on both sex chromosomes (fig. 1).

PCR Amplification

The PCR amplification was performed with chromosome-specific primers, taking advantage of single base

sequence differences between the X and Y chromosomes in the flanking regions of each marker of interest. Typically 15 ng of genomic DNA were amplified in a 12.5 μ l reaction volume comprising 1.5 mM $MgCl_2$, 20 mM ammonium sulfate, 1U *Taq* DNA polymerase (Promega), 200 μ M of each dNTP, and 0.3 μ M of each primer (see Supplementary Material online). PCR conditions were the following: 5 min preincubation step at 94°C, 32 to 35 cycles of denaturation at 94°C for 30 sec, annealing for 30 sec at the respective AT for each primer pair, and extension at 72°C for 30 sec, followed by a final extension step at 72°C for 10 min.

Microsatellite Typing

For microsatellite typing, forward or reverse primers were fluorescently labeled, allowing the distinction of overlapping fragments from different chromosomes. Separation and fragment size analysis of PCR products were performed in an ABI 310 sequencer using the Genescan version 2.1 analysis software.

Single Nucleotide Polymorphisms (SNP) Typing

Amplification products were submitted to restriction endonuclease digestion with the enzymes listed in table 1, according to the manufacturer's instructions (MBI Fermentas). The digested products were then electrophoresed on horizontal polyacrylamide gels (T9C5) and visualized with silver staining (Budowle et al. 1991). Presence of either of the alleles in each DNA molecule produced different number and/or sizes of bands in the gel. The intronic T $\rightarrow$ C polymorphism in the X chromosome found while typing (GT)_n 3 was assessed directly by allele-specific PCR performed to differentiate X/Y products, because the X chromosome presented either the same nucleotide in that position as the one present on the Y chromosome (C) or a T. Different primers were then designed to accomplish X-Y specificity for (GT)_n 3 microsatellite typing. Three X chromosome's binary markers a few base pairs apart, in *PCDHX* exon 7 (G $\rightarrow$ A) and intron 7 (T $\rightarrow$ G) and (C $\rightarrow$ A), were typed by sequencing.

Sequencing

Extensive sequencing was performed on a total of 67 male individuals (56 Portuguese and 11 Mozambican), one *Pan troglodytes* and one *Gorilla gorilla* male sample, in 528 bp of a noncoding region of *PCDHXY*. Further sequencing was conducted on selected human samples of virtually all allelic states to confirm RFLP results and to verify repeat structure and assign the respective number of repeats to each fragment size detected in the ABI 310 sequencer. Sequencing of one male *Pan troglodytes* sample for X chromosome microsatellites was also performed. As a standard procedure only one DNA strand was sequenced. PCR fragments were purified with Microspin S-300 HR columns (Pharmacia). A dideoxy cycle sequencing reaction was carried out using the Big Dye Terminator Cycle Kit (PE Applied Biosystems). The products were purified using SigmaSpin Post-Reaction Clean-Up columns (Sigma) and run in an ABI 3100 sequencer (PE Applied Biosystems). The results were analyzed using the Data Collection software.

Data Analysis

Allelic and haplotype frequencies were estimated by direct allele/haplotype counting because only males were sampled. Gene diversity across loci for all populations, population comparisons (differentiation values), and average pairwise differences within populations were calculated with the Arlequin version 2.000 software (Schneider, Roessli, and Excoffier 2000).

Nonparametric tests were used to assay X/Y equality of (1) allelic microsatellite distributions (Kolmogorov-Smirnov) and (2) gene diversity values across microsatellite loci (Mann-Whitney).

Phylogenetic analysis of both binary and microsatellite variation were performed in Network version 4.0, applying the Reduced-Median (Bandelt et al. 1995) and Median-Joining (Bandelt, Forster, and Rohl 1999) network methods sequentially to resolve extensive reticulation at microsatellite loci (Qamar et al. 2002). In all calculations, ϵ was set to zero and differential microsatellite weighting was applied to achieve most parsimonious networks. Weights for each microsatellite were always inversely proportional to the ratio of the variance displayed by each marker in the respective population and the average variance value across loci in that population, within a fourfold range. A null weight was attributed to (AC)_n repeat, considering that this was the only Y microsatellite displaying repeat length frequency distributions that deviate from the Gaussian-like shape expected under the stepwise mutation model. Binary markers' weights were always set to the maximum value (99).

Relative age for the Y chromosome haplotype carrying the derived allelic state at both SNPs analyzed was estimated by a phylogenetic approach, using the ρ statistic, which represents the average number of mutational changes between a set of selected haplotypes and the ancestral one (Forster et al. 1996). The ancestral node was defined as the most frequent one that, although presenting the derived SNP haplotype, displayed a STR haplotype

Table 1
Allele Frequency Distributions of *PCDHXY* Binary Markers in Portugal and Mozambique

Gene	Exon/Intron ^a	Restriction Endonuclease	Allele	Frequency ^b	
				Portugal (n = 112)	Mozambique (n = 112)
<i>PCDHX</i>	I6	---	T	0.79	0.78
	(ss23133234)		C	0.21	0.22
	E7 ^c	---	G	0.98	1
	(ss5608039)		A	0.02	0.00
	I7	---	T	0.96	0.79
	(ss5608040)		G	0.04	0.21
	I7	---	C	0.89	0.97
	(ss5608041)		A	0.11	0.03
	E10 ^c	(Cfr13I)	T	0.98	1
	(ss5608042)		C	0.02	0.00
<i>PCDIY</i>	E5 ^c		T	0.57	1
	(ss5608044)	(TspEI)	G	0.43	0.00
	E5 ^c		G	0.57	1
	(ss5608045)		T	0.43	0.00

^a NCBI assay IDs are indicated inside brackets.

^b n = number of chromosomes.

^c Nonsynonymous substitutions.

shared between ancestral and derived chromosomes. Because STR mutation rates were essential to obtain absolute time estimates, and considering the scarcity of this kind of data on Y chromosome dinucleotide repeats, autosomal STR variation at corresponding loci and the respective effective mutation rate estimate (w) of 1.52×10^{-3} (Zhivotovsky et al. 2000) was used to perform variance comparisons that would allow obtaining mutation rate estimates specific for the presently described Y markers, as suggested by Zhivotovsky et al. (2004). Values for the effective mutation rate were then obtained for each Y locus, using the respective repeat length variance within the population used to perform relative age estimates. One absolute time estimate was then generated, multiplying the average effective mutation rate across Y loci (6.0×10^{-4}) by p and a generation time of 25 years.

Results

Molecular Characterization of the X/Y Markers

All the markers analyzed revealed a simple (CA)_n, (GT)_n, or (AC)_n repeat structure in both sex chromosomes in humans and in the X chromosome of chimpanzee, differing only in the number of repetitive units and, in some cases, in point mutations and/or small insertions/deletions in the flanking regions of the repeats (see Supplementary Material online). In the specific case of the (GT)_n 2 repeat we detected some degree of variation in a Poly (T) stretch and a (GTT)_n trinucleotide in the immediate 5' sequence on the X chromosome but not on the Y chromosome (fig. 2). Despite the physical proximity between these variable units, we decided to take them into account individually for further analysis because, attending to their heterogeneity, it seemed more consistent to consider the existence of three different repeat motifs, rather than a complex repeat on the X chromosome. Nevertheless more caution is needed in interpreting X-Y comparisons concerning these microsatellites, because there could be some mutual influence on the evolutionary

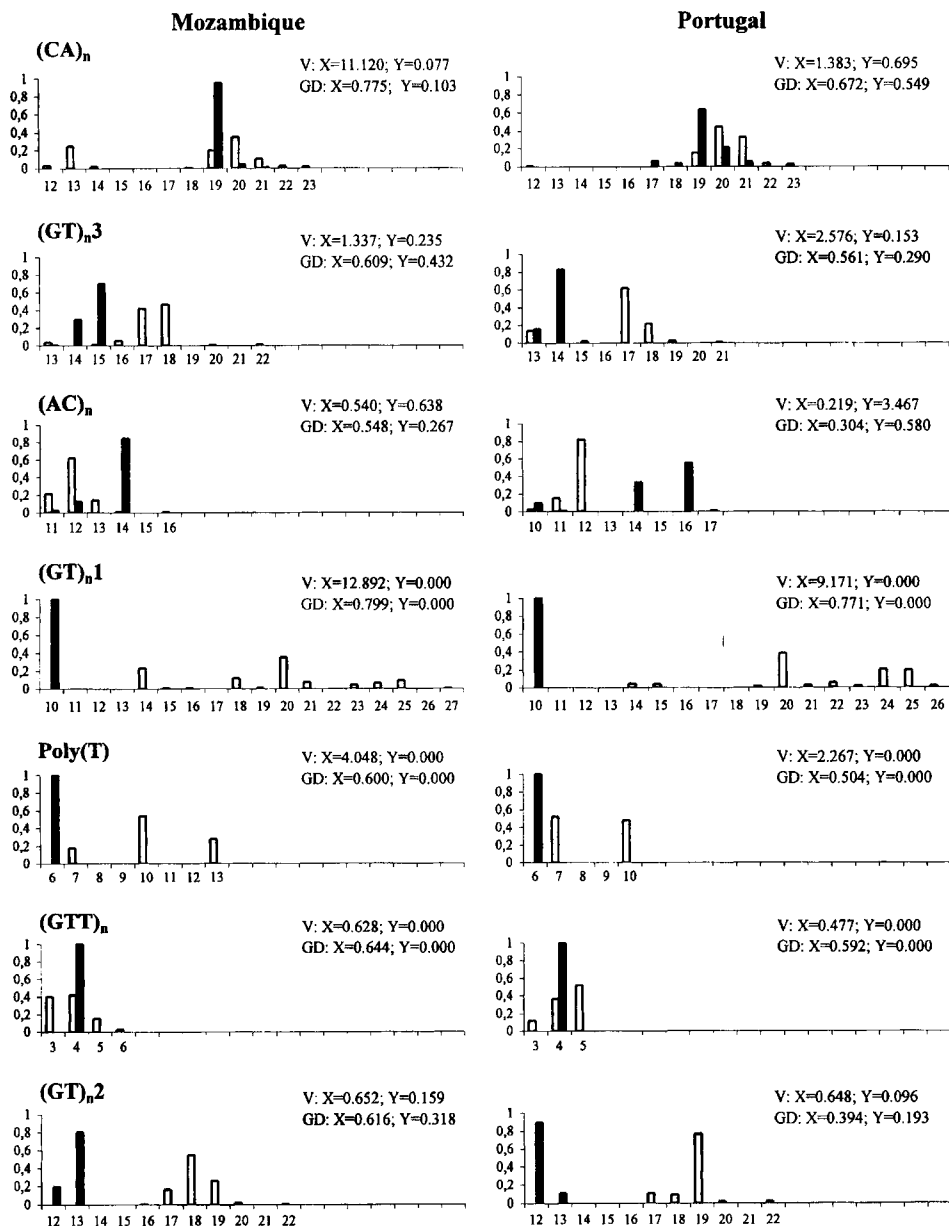


FIG. 2.—Allele frequency distributions of X/Y microsatellites in male individuals from Portugal ($n = 112$) and Mozambique ($n = 112$). V = variance in repeat number, GD = gene diversity; white bars X chromosome, black bars Y chromosome.

dynamics of these three repetitive sequences that would differ on the X and Y chromosomes. While analyzing variation in (GT)_n 3 dinucleotide on the X chromosome, one polymorphic T → C substitution, 190 bp downstream of the repeat, was found (ss23133234).

Microsatellite Markers

Comparison of Homologous Allelic Distributions

The most striking feature of all X-Y homologous markers in both of the populations sampled was the different allelic distributions displayed, according to the

chromosome of origin of the microsatellite, concerning both the presence of private alleles and a different modal repeat length for X-linked and Y-linked markers (fig. 2), as reported for other markers (Karafet et al. 1998). These differences were reflected in a statistically significant value in a two-sample Kolmogorov-Smirnov test ($P < 0.001$ in all X-Y pairs).

Comparisons of gene diversity between sex chromosomes, although with some variation across loci, has shown an overall trend to reduced levels of intrapopulation gene diversity for Y chromosome versus X chromosome microsatellites in both European and African populations, consistent with the results obtained for another Y-linked marker (Karafet et al. 1998) but revealing the opposite X-Y diversity patterns observed for other markers (Scozzari et al. 1997). The most extreme difference between X and Y dinucleotide microsatellite distributions was found at (GT)_n 1, Poly (T) stretch, and (GTT)_n trinucleotide repeats, where the Y repeat number was constant, but the respective X counterpart was moderate to highly polymorphic in both populations analyzed (fig. 2). However, the opposite trend was observed at (AC)_n in the Portuguese population. This marker is also exceptional in the fact that the alleles found in the chimpanzee sample for (AC)_n and also for the Poly (T) stretch are within the range detected in the human Y chromosome sample, whereas for the other microsatellites analyzed, the primate alleles fall in X-specific regions of the human distribution (see Supplementary Material online).

Overall, the X-linked markers displayed in both populations higher diversity values when compared with their Y counterparts, and this difference was statistically significant (Mann-Whitney: $P = 0.025$ in Portugal; $P = 0.002$ in Mozambique). This greater variability is also evident in the higher average repeat length variance values obtained for the X-linked markers (2.39 versus 0.63 in Portugal; 4.45 versus 0.16 in Mozambique).

Average among-population diversity values revealed contrasting patterns for X-linked and Y-linked markers, being higher in the African population when compared with the European one if X chromosome microsatellites are considered (1.21 ratio) but pointing to a greater European diversity for Y-linked markers (1.44 ratio).

Plotting within-population variance values for X versus Y homologous microsatellites, we observed that the samples clustered in two groups distinguished mainly by very high or moderately low values of variance for the X chromosome repeats and extremely low values for Y chromosome repeats, with only (AC)_n standing out (fig. 3). It is also interesting to notice that we find the samples distributed throughout all the extremes of the quadrants in the graph, except the one representing high values of variance for both X and Y chromosome microsatellites.

Haplotype Analysis

For direct comparisons of X-Y microsatellite haplotype diversity, all markers were considered for both chromosomes. Concordantly with the results obtained in the locus-by-locus analysis, the X chromosome presented higher haplotype diversity than the Y chromosome in both populations, this effect being more pronounced in the

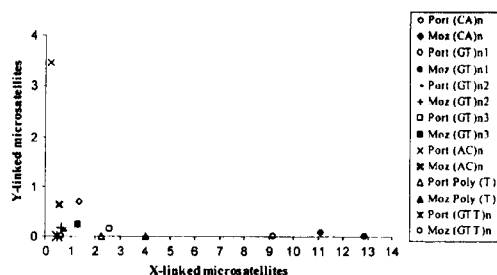


FIG. 3.—Plots of within-population variances for X/Y homologous microsatellites in Portugal and Mozambique.

African population (0.995 versus 0.560 in Mozambique; 0.987 versus 0.746 in Portugal). This reduced variation observed for the Y chromosome is reflected in lower diversity values for the African population when compared with the European one for haplotypes of Y-linked markers, whereas the opposite is observed for haplotypes defined by X-linked markers, as it has been described for other markers in the literature (Scozzari et al. 1997).

No X-Y shared haplotypes were observed, consistent with our previous findings of mostly chromosome-private alleles for individual markers. Microsatellite haplotypes were grouped according to population and chromosome and compared with each other by mean of exact tests of population differentiation. All comparisons were statistically significant ($P < 0.001$).

Binary Markers

Allele Frequencies

The allele frequencies of two binary markers for the Y chromosome and five for its X counterpart were determined (table 1). All of these markers were previously described (Giouzezi et al. 2004), except (T → C) substitution in intron 6, which was found during the course of this study and are included in NCBI database. Determination of the nucleotide present in the orthologous position in chimpanzee in the last marker allowed us to define the ancestral allele.

The X-linked SNP found by us was moderately polymorphic, with the derived allele (C) present at a frequency of approximately 20% in both populations surveyed. As for the other markers, the frequencies found in the Portuguese population were quite similar to the ones reported in the previously mentioned study, as expected because the vast majority of samples in both cases belonged to populations with European ancestry, while some differences in allele frequencies were observed in Mozambique (table 1).

For the Y chromosome, the two previously described allelic combinations of exon 5 base substitutions, T and G or G and T, were found only in the Portuguese population, in a proportion of 43% to 57%, respectively. In Mozambique, only T-G bearing Y chromosomes, corresponding both to X-like and, thus, to ancestral allelic states, were observed.

Table 2
Observed Haplotype Distribution at *PCDHX/Y* Genes in Portugal and Mozambique

Gene	Marker ^a						Haplotype Frequency ^b (n = 112)	
	I6 (ss23133234)	E7 (ss5608039)	I7 (ss5608040)	I7 (ss5608041)	E10 (ss5608042)		Portugal	Mozambique
<i>PCDHX</i>								
H1	T	G	T	C	T		0.69	0.54
H2	T	G	T	A	T		0.05	0.03
H3	T	G	G	C	T		0.02	0.21
H4	C	G	T	C	T		0.15	0.22
H5	T	G	T	C	C		0.02	0.00
H6	T	A	G	C	T		0.02	0.00
H7	C	G	T	A	T		0.05	0.00
Gene	Marker				Haplotype Frequency (n = 112)			
	E5 (ss5608044)		E5 (ss5608045)		Portugal		Mozambique	
<i>PCDHY</i>								
H1	T		G		0.43		1	
H2	G		T		0.57		0	

^a NCBI assay IDs are indicated inside brackets.

^b n = number of chromosomes.

Evolutionary Patterns

The X-linked binary markers segregated as four haplotypes in Mozambique, whereas in Portugal, three additional haplotypes were found (table 2). Because the ancestral alleles at all binary positions were known, evolutionary relationships between all molecular variants found were predicted (fig. 4A). Two mutation events in the ancestral background need to be considered to embrace all haplotypes found, although recombination events could also be involved.

Adding the microsatellite variation data gathered on the X chromosome to each haplotypic background, Median-joining networks generated gave us an insight on the relative diversity of each molecular variant, revealing a greater diversity for the ancestral haplotype (H1) in both populations, when compared with the derived ones (fig. 4B and C). Considering the geographic distribution and frequencies displayed in each population by the X-linked haplotypes, it is plausible to assume the existence of three older haplotypes (H1, H3, and H4), present both in Africa and Europe. Hints for African ancestry of these shared haplotypes are a higher microsatellite haplotype diversity within Mozambique compared with Portugal of H1 (0.993 versus 0.982) and H3 (0.972 in Mozambique; in Portugal H3 diversity is 1.00 but only two chromosomes can be found). H4 displays similar diversity values in both populations (0.984 in Portugal and 0.980 in Mozambique), but a higher number of average pairwise differences within Mozambican chromosomes (12.027 versus 7.676) reveals the presence of more diverged and, thus, older haplotypes in Africa. The fourth haplotype found to be shared by both populations (H2) is present in very low frequencies (3% in Mozambique and 5% in Portugal) and exhibits equal microsatellite diversity in Portugal and in Mozambique, being more difficult to infer its antiquity. The remaining haplotypes (H5, H6, and H7) are European-specific and have, most probably, a relatively recent origin, taking into account both their peripheral distribution in the network and the low frequencies at which they are found.

The most frequent *PCDHY* haplotype in Portuguese, namely H2, was not found in Mozambicans. This both

hints at a recent age and allows estimating it by a simple method, which need not take into account population substructure or the action of recombination (as in the X chromosome). The amount of diversity accumulated in the STRs by mutation from a putative ancestral haplotype (which was found in over 80% of the H2 chromosomes [see fig. 4D]) was measured by means of the ρ parameter, which translates linearly into time. The time to most recent common ancestor (TMRCA) of H2 was, thus, estimated as 13,700 years (7,500 to 20,000 years), which is consistent with a recent post-Out-of-Africa origin for this haplotype.

Sequence Comparisons with Primates

Given that some of the polymorphisms analyzed are nonsynonymous substitutions (table 1), we decided to include in this study sequence comparisons in 528 bp of noncoding region, in humans and nonhuman primates.

Five mutations occurred in this region in the human lineage, three of which became fixed as X-Y differences (see Supplementary Material online). From the latter, two are undoubtedly attributable to the Y chromosome and one to the X chromosome, whereas in the remaining two X-Y differences, we found variability within human X chromosomes, presenting in those positions either the Y characteristic allele (T) or an alternative nucleotide (C, A). If we consider that in the orthologous positions of both chimpanzee and gorilla X chromosomes we found the same alleles as in the human Y chromosome, then we must account for two more mutations occurring in the human X chromosome, which have not reached fixation yet. The evenness in changes accumulated in either human sex chromosome is in conformity with previous estimates in the gene free Xq-Yp homology region (Bohossian, Skaletsky, and Page 2000).

Ancient polymorphism in the X chromosome should also be considered as a possible source of variation that could have become fixed on the Y chromosome after the translocation (Makova and Li 2002), reducing the number of mutations that should be ascertained to the human Y chromosome, although none of the differences observed between human Y chromosome and chimpanzee X was found to be polymorphic in present human X chromosomes.

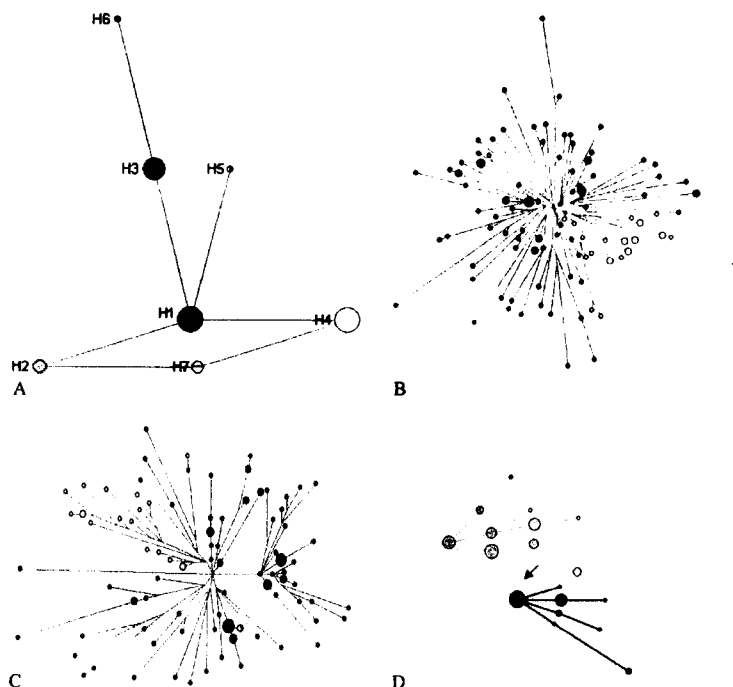


FIG. 4.—Median-joining networks of X chromosome (A, B, and C) and Y chromosome (D) haplotypes. Circle area is proportional to frequency and branch length is proportional to the number of STR mutations. Different circle fillings symbolize different SNP haplotypes. (A) X-linked haplotypes defined by binary markers in both populations analyzed, where H1 is the ancestral. (B and C) X chromosome haplotypes inferred from variation at seven microsatellites in Mozambique (B) and Portugal (C). (D) Y chromosome haplotypes defined by five microsatellites in Portugal. The arrow indicates the ancestral node used for ρ estimate and derived haplotypes, considered for this estimation, are in black and connected by solid lines.

Nucleotide diversity (π) for human X sequences (7.3×10^{-4}) was twice that previously described for a region of low recombination (Kaessmann et al. 1999) but in agreement with other values described for the X chromosome (Jaruzelska, Zietkiewicz, and Labuda 1999) and for autosomes (Zhao et al. 2000). As for the Y chromosome, no variation was found in human sequences, except a 10-bp deletion in one individual, which was not accounted for in these calculations, in accordance with previously described reduced diversity levels in this chromosome (Dorit, Akashi, and Gilbert 1995; Hammer 1995; Underhill et al. 1996).

Discussion

We have analyzed seven new X/Y homologous microsatellites in samples from Portugal and Mozambique and found higher diversity in the X chromosome microsatellites than in their Y chromosome counterparts; when comparing populations, Mozambicans showed more allele diversity for the X chromosome, but the contrary was true for the Y chromosome microsatellites.

In a previous study, Scozzari et al. (1997) found the same patterns for X/Y microsatellites when comparing African to European internal diversity but, overall, found more diversity in the two Y than in the X microsatellites.

This may be the product of random deviation from an expected model of greater diversity on the X than on the Y chromosome, as discussed below. The X/Y homologous loci studied by Karafet et al. (1998) matched our overall pattern of greater X than Y diversity but showed more diversity for Africans than for Europeans.

The inconsistency of X and Y microsatellites in attributing more diversity to African versus European populations, as would be expected under the Out-of-Africa model and as observed for other microsatellite loci and mtDNA (Vigilant et al. 1991; Bowcock et al. 1994; Jorde et al. 1995) deserves further investigation as it cannot be resolved by the present study, given the restricted distribution of the African samples analyzed. The reduced diversity observed by us for Y-linked markers in Africa is most probably related to the exclusively Bantu-speaking origin of the individuals sampled, which have been shown to exhibit lower levels of diversity than Europeans (Lucotte et al. 1994; Pereira et al. 2002).

A number of different factors may account for the larger diversity observed by us on the X chromosome compared with the Y chromosome. First, founding Y chromosome alleles translocated from the X chromosome may have been on the short end of the X microsatellite distribution, leading to the stabilization or even loss of polymorphism on the Y locus (Carvalho-Silva et al. 1999).

However, for some loci, the Y microsatellite displayed reduced or absent variability, although presenting allele lengths only one step apart or even in the same allele length range as the X alleles.

Second, the X and Y homologous regions differ in their genome dynamics: recombination operates on the X chromosome but not on Yp11.2. Recombination hot spots have been reported to associate with long (GT)_n repeats (Majewski and Ott 2000). Unequal crossing-over might explain the bimodal, non-Gaussian allele length distribution at the (GT)_n 1 locus (which contrasts starkly with its monomorphic Y counterpart). However, it is unlikely that recombination contributes much allele diversity as compared with replication slippage, given that the empirically determined mutation rates for Y chromosome and trinucleotide and tetranucleotide autosomal microsatellites are very similar (Kayser et al. 2000; Zhivotovsky et al. 2004), even though these pending observations are to be confirmed with dinucleotide repeats. Sequence diversity is generated in the amplicon regions of the Y chromosome by gene conversion (Skaletsky et al. 2003; Bosch et al. 2004). *PCDHY* is, however, a single-copy gene outside extensive repetitive regions, and no such phenomenon has been reported in this gene, despite the extensive sequence survey by DHPLC undertaken by Giouzeli et al. (2004).

Third, different male/female demographic and reproductive patterns may be the main cause of the difference in genetic diversity among the two sex chromosomes. The effective population size of the Y chromosome would be one-third that of the X chromosome if both sexes contributed equally to the next generation. Reproductive habits conferring higher reproductive variance for males, such as polygyny, can further reduce Y chromosome *N_e* (Underhill et al. 1996; Charlesworth 2001; Dupanloup et al. 2003), leading to a reduction of polymorphism caused by genetic drift. In fact, the pronounced reduction of Y-linked microsatellite diversity when compared with the X chromosome in Mozambique is most probably explained by a model of asymmetric gene flow of paternal and maternal lineages between food-producer populations (Bantu-speakers) and hunter-gatherers within Africa, jointly with different levels of polygyny and patrilocality in each population, as proposed by Destro-Bisol et al. (2004) and supported by other data (Pereira et al. 2001, 2002).

A recent selective sweep may have also contributed to a general reduction of variation in the Y chromosome, although there is little direct evidence for such an event (Pérez-Lezaun et al. 1997; Thomson et al. 2000). Our sequence data, even though not very powerful, show that although substitutions have accumulated evenly between the X and Y chromosome, nucleotide diversities found are extremely lower for the Y chromosome, reinforcing the view of a reduction of diversity on the Y chromosome.

Evolutionary patterns relying on intragenic *PCDHX/Y* SNPs also revealed distinct scenarios for X and Y chromosomes. Although a higher number of X and Y haplotypes were found in Europe compared with Africa, reflecting the ascertainment in European populations of all polymorphisms analyzed, greater microsatellite diversity was displayed by African X chromosomes within the three most common haplotypes shared by both populations,

whereas for the widespread ancestral Y chromosome haplotype, higher microsatellite diversity was found in Portugal—although both populations presented the same number of haplotypes, 65% of Mozambican Y chromosomes shared the same STR haplotype. As the T-G ancestral haplotype (H1) must be older in Africa, an elimination of STR lineages within this Y chromosome group in Mozambique must have taken place, most probably by the Bantu migration and expansion along the Eastern coast of Africa (Salas et al. 2002).

The biased ascertainment of the binary markers under consideration must also be evoked to justify the lack of putative recombinant X chromosome haplotypes in the older Mozambican population, where lower levels of LD are expected (Reich et al. 2001). SNPs with low minor allele frequencies tend to present higher LD (Ke et al. 2004); ascertainment bias implies that SNPs tend to be less polymorphic in populations other than those where the SNPs were ascertained, thus increasing LD. However, from microsatellite data, it is possible to infer an African ancestry for X chromosome SNP-defined haplotypes shared by both populations, because of higher diversity within Mozambique for these haplotypes, results which are in accordance with an early population bottleneck scenario shaping European diversity, as expected under the Out-of-Africa model. The frequency distribution of X chromosome haplotypes based on binary markers alone is suggestive of a population expansion within Europe with two frequent haplotypes in Portugal followed by a tail of rare ones.

The asymmetric distribution of *PCDHY* haplotypes between Portugal and Mozambique seems justifiable by a recent origin of the derived one (H2), approximately 7,500 to 20,000 years ago as suggested by the Y STR haplotype phylogeny. This estimate partially overlaps with the Last Glacial Maximum, which occurred in Europe 20,000 to 13,000 years ago and points to a more recent origin for this haplotype, compared with the M173 mutation (~30,000) (Semino et al. 2000) that defines most of the Y lineages within clade R, highly frequent in Western Europe. The ancestral *PCDHY* haplotype (H1) must be found in several Y chromosome lineages because the Mozambican Y chromosome pool, which is characterized by its exclusive presence, comprises lineages belonging to some of the older African Y chromosome clades as well as more recent and widespread ones (Pereira et al. 2002). Thus, the considerable difference in frequency of the two observed *PCDHY* haplotypes is not surprising in view of their different antiquity and bearing in mind that genetic drift is particularly powerful in shaping diversity on the NRY, accelerating differentiation between populations. These two haplotypes, which are different at both SNPs, are a further example of “yin yang haplotypes” in the human genome (Zhang et al. 2003). The intervening haplotypes were not detected in the present survey, although we cannot rule out their existence in other populations, namely in Asia.

Conclusion

The newly described microsatellite loci generally displayed greater X-linked than Y-linked diversity, and this

pattern was also detected with slower evolving markers, with a remarkable differentiation between populations observed for Y chromosome haplotypes. Overall, information drawn from Yp11.2 polymorphisms is in accordance with the diversity patterns observed for other Y-linked loci, which is expected because of the absolute linkage that characterizes the NRY. The analyzed binary markers may also be enlightening in refining the Y chromosome world phylogeny, namely distinguishing lineages within clade R.

Its parental sequence in Xq21.3 records the major signatures of African and European population movements from both maternal and paternal point of view, and diversity patterns in this locus seem to be in agreement with the long-standing view of European diversity representing a subset of the ancestral gene pool within Africa, although more informative markers are needed for a detailed picture.

Differences observed for the presently described X/Y microsatellites are, thus, expected to reflect both chromosome-specific mutation rates and historical contexts of the populations surveyed, within the complex framework of male/female specificities, which cannot be disentangled in an obvious way. Empirical determination of mutation rates for these markers on both sex chromosomes would be of great interest to quantify the relative contribution of each chromosome in creating new alleles.

Supplementary Material

Details on primers used in this study, as well as all sequence alignments are supplied in the MBE Web site as Supplementary Material online.

Acknowledgments

The authors would like to thank Maria Giouzeli, from the Department of Psychiatry of the University of Oxford, for providing all information available on the SNPs within *PCDHX1/Y* and for her help with primer design. We thank Dr. Albertino Damasceno and Dr. Benilde Soares of the Eduardo Mondlane University (Maputo) for kindly providing the Mozambican samples. We also would like to acknowledge Patricia Blanco for her scientific advice. This work was partially supported by Fundação para a Ciência e Tecnologia (through grant SFRH/BD/7006/2001 and POCTI, Programa Operacional Ciência, Tecnologia e Inovação).

Literature Cited

- Bandelt, H. J., P. Forster, and A. Rohl. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**:37–48.
- Bandelt, H. J., P. Forster, B. C. Sykes, and M. B. Richards. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* **141**:743–753.
- Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**:573–580.
- Blanco, P., C. A. Sargent, C. A. Boucher, M. Mitchell, and N. A. Affara. 2000. Conservation of *PCDHX* in mammals: expression of human X/Y genes predominantly in brain. *Mamm. Genome* **11**:906–914.
- Blanco-Arias, P., C. A. Sargent, and N. A. Affara. 2002. The human-specific Yp11.2/Xq21.3 homology block encodes a potentially functional testis-specific TGIF-like retroposon. *Mamm. Genome* **13**:463–468.
- Bohossian H. B., H. Skaletsky, and D. C. Page. 2000. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature* **406**:622–625.
- Bosch, E., M. E. Hurles, A. Navarro, and M. A. Jobling. 2004. Dynamics of a human inter-paralog gene conversion hotspot. *Genome Res.* **14**:835–844.
- Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd, and L. L. Cavalli-Sforza. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**:455–457.
- Budowle, B., R. Chakraborty, A. M. Giusti, A. J. Eisenberg, and R. C. Allen. 1991. Analysis of the VNTR locus D1S80 by the PCR followed by high-resolution PAGE. *Am. J. Hum. Genet.* **48**:137–144.
- Carvalho-Silva, D. R., and S. D. Pena. 2000. Molecular characterization and population study of an X chromosome homolog of the Y-linked microsatellite DYS391. *Gene* **247**:233–240.
- Carvalho-Silva, D. R., F. R. Santos, M. H. Hutz, F. M. Salzano, and S. D. Pena. 1999. Divergent human Y-chromosome microsatellite evolution rates. *J. Mol. Evol.* **49**:204–214.
- Charlesworth, B. 2001. The effect of life-history and mode of inheritance on neutral genetic variability. *Genet. Res.* **77**:153–166.
- Destro-Bisol, G., F. Donati, V. Coia, I. Boschi, F. Verginelli, A. Caglia, S. Tofanelli, G. Spedini, and C. Capelli. 2004. Variation of female and male lineages in sub-Saharan populations: the importance of socio-cultural factors. *Mol. Biol. Evol.* (in press).
- Dorit R. L., H. Akashi, and W. Gilbert. 1995. Absence of polymorphism at the ZFY locus on the human Y chromosome. *Science* **268**:1183–1185.
- Dupanloup I., L. Pereira, G. Bertorelle, F. Calafell, M. J. Prata, A. Amorim, and G. Barbujani. 2003. A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. *J. Mol. Evol.* **57**:85–97.
- Dupuy, B. M., T. Gedde-Dahl, and B. Olaisen. 2000. DXYS267: DYS393 and its X chromosome counterpart. *Forensic Sci. Int.* **112**:111–121.
- Forster, P., R. Harding, A. Torroni, and H. J. Bandelt. 1996. Origin and evolution of Native American mtDNA variation: a reappraisal. *Am. J. Hum. Genet.* **59**:935–945.
- Giouzeli, M., N. A. Williams, L. J. Lonie, L. E. DeLisi, and T. J. Crow. 2004. *ProtocadherinX/Y*, a candidate gene-pair for schizophrenia and schizoaffective disorder: a DHPLC investigation of genomic sequence. *Am. J. Med. Genet.* **129B**:1–9.
- Graves, J. A. 1995. The origin and function of the mammalian Y chromosome and Y-borne genes—an evolving understanding. *Bioessays* **17**:311–320.
- Hammer, M. F. 1995. A recent common ancestry for human Y chromosomes. *Nature* **378**:376–378.
- Jaruzelska, J., E. Zietkiewicz, and D. Labuda. 1999. Is selection responsible for the low level of variation in the last intron of the ZFY locus? *Mol. Biol. Evol.* **16**:1633–1640.
- Jobling, M. A., and C. Tyler-Smith. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.* **4**:598–612.
- Jorde, L. B., M. J. Bamshad, W. S. Watkins, R. Zenger, A. E. Fraley, P. A. Krakowiak, K. D. Carpenter, H. Soodyall, T. Jenkins, and A. R. Rogers. 1995. Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. *Am. J. Hum. Genet.* **57**:523–538.

- Kaessmann, H., F. Heissig, A. von Haeseler, and S. Paabo. 1999. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat. Genet.* **22**:78–81.
- Karafet, T., P. de Knijff, E. Wood, J. Ragland, A. Clark, and M. F. Hammer. 1998. Different patterns of variation at the X- and Y-chromosome-linked microsatellite loci DXYS156X and DXYS156Y in human populations. *Hum. Biol.* **70**: 979–992.
- Kayser, M., L. Roewer, M. Hedman et al. (11 co-authors). 2000. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am. J. Hum. Genet.* **66**:1580–1588.
- Ke, X., S. Hunt, W. Tapper et al. (12 co-authors). 2004. The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.* **13**:577–588.
- Kersting, C., C. Hohoff, B. Rolf, and B. Brinkmann. 2001. Pentanucleotide short tandem repeat locus DXYS156 displays different patterns of variations in human populations. *Croat. Med. J.* **42**:310–314.
- Lareu, M. V., C. P. Phillips, A. Carracedo, P. J. Lincoln, D. Syndercombe Court, and J. A. Thomson. 1994. Investigation of the STR locus HUMTH01 using PCR and two electrophoresis formats: UK and Galician Caucasian population surveys and usefulness in paternity investigations. *Forensic Sci. Int.* **66**:41–52.
- Lucotte, G., N. Gerard, R. Krishnamoorthy, F. David, A. Aouizerate, and P. Galzot. 1994. Reduced variability in Y-chromosome-specific haplotypes for some Central African populations. *Hum. Biol.* **66**:519–526.
- Majewski, J., and J. Ott. 2000. GT repeats are associated with recombination on human chromosome 22. *Genome Res.* **10**: 1108–1114.
- Makova, K. D., and W. H. Li. 2002. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**: 624–626.
- Navarro, A., and N. H. Barton. 2003. Chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes. *Science* **300**:321–324.
- Pereira, L., L. Gusmão, C. Alves, A. Amorim, and M. J. Prata. 2002. Bantu and European Y-lineages in sub-Saharan Africa. *Ann. Hum. Genet.* **66**:369–378.
- Pereira, L., V. Macaulay, A. Torroni, R. Scozzari, M. J. Prata, and A. Amorim. 2001. Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann. Hum. Genet.* **65**:439–458.
- Pereira, L., M. J. Prata, M. A. Jobling, and A. Amorim. 2000. Analysis of the Y-chromosome and mitochondrial DNA pools in Portugal. Pp. 191–195 in C. Renfrew and K. Boyle, eds. *Archaeogenetics: DNA and the population prehistory of Europe*. McDonald Institute for Archeological Research, University of Cambridge, Cambridge, UK.
- Pérez-Lezaun, A., F. Calafell, M. Seielstad, E. Mateu, D. Comas, E. Bosch, and J. Bertranpetit. 1997. Population genetics of Y-chromosome short tandem repeats in humans. *J. Mol. Evol.* **45**:265–270.
- Qamar, R., Q. Ayub, A. Mohyuddin, A. Helgason, K. Mazhar, A. Mansoor, T. Zerjal, C. Tyler-Smith, and S. Q. Mehdi. 2002. Y-chromosomal DNA variation in Pakistan. *Am. J. Hum. Genet.* **70**:1107–1124.
- Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. 2001. Linkage disequilibrium in the human genome. *Nature* **411**:199–204.
- Rieseberg, L. H., and K. Livingstone. 2003. Evolution. Chromosomal speciation in primates. *Science* **300**:267–268.
- Salas, A., M. Richards, T. De la Fe, M. V. Lareu, B. Sobrino, P. Sanchez-Diz, V. Macaulay, and A. Carracedo. 2002. The making of the African mtDNA landscape. *Am. J. Hum. Genet.* **71**:1082–1111.
- Schneider, S., D. Roessli, and L. Excoffier. 2000. *Arlequin ver. 2.000*: a software for population genetics data analysis. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Switzerland.
- Schwartz, A., D. C. Chan, L. G. Brown, R. Alagappan, D. Pettay, C. Disteche, B. McGillivray, A. de la Chapelle, and D. C. Page. 1998. Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum. Mol. Genet.* **7**:1–11.
- Scozzari, R., F. Cruciani, P. Malaspina et al. (18 co-authors). 1997. Differential structuring of human populations for homologous X and Y microsatellite loci. *Am. J. Hum. Genet.* **61**:719–733.
- Semino, O., G. Passarino, P. J. Oefner et al. (14 co-authors). 2000. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science* **290**:1155–1159.
- Skaletsky, H., T. Kuroda-Kawaguchi, P. J. Minx et al. (37 co-authors). 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**:825–837.
- Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner, and M. W. Feldman. 2000. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Natl. Acad. Sci.* **97**:7360–7365.
- Underhill, P. A., L. Jin, R. Zemans, P. J. Oefner, and L. L. Cavalli-Sforza. 1996. A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. *Proc. Natl. Acad. Sci.* **93**:196–200.
- Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson. 1991. African populations and the evolution of human mitochondrial DNA. *Science* **253**:1503–1507.
- Zhang, J., W. L. Rowe, A. G. Clark, and K. H. Buetow. 2003. Genomewide distribution of high-frequency, completely mismatching snp haplotype pairs observed to be common across human populations. *Am. J. Hum. Genet.* **73**:1073–1081.
- Zhao, Z., L. Jin, Y. X. Fu et al. (13 co-authors). 2000. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proc. Natl. Acad. Sci.* **97**: 11354–11358.
- Zhivotovsky, L. A., L. Bennett, A. M. Bowcock, and M. W. Feldman. 2000. Human population expansion and microsatellite variation. *Mol. Biol. Evol.* **17**:757–767.
- Zhivotovsky, L. A., P. A. Underhill, C. Cinnioglu et al. (14 co-authors). 2004. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am. J. Hum. Genet.* **74**:50–61.

Lisa Matisoo-Smith, Associate Editor

Accepted July 19, 2004

2. *PCDH11Y* biallelic markers in a large sample of well characterized Portuguese Y chromosome lineages

Article 4

Beleza S, Gusmão L, **Lopes A**, Alves C, Gosmes I, Giouzei M, Calafell F, Carracedo A, Amorim A. Micro-phylogeographic and demographic history of Portuguese male lineages. *Ann Hum Genet* (in press)

Micro-Phylogeographic and Demographic History of Portuguese Male Lineages

Sandra Beleza^{1,2*}, Leonor Gusmão¹, Alexandra Lopes^{1,3}, Cíntia Alves¹, Iva Gomes¹, Maria Giouzeli⁵, Francesc Calafell⁴, Angel Carracedo² and António Amorim^{1,3}

¹IPATIMUP, Instituto de Patologia e Imunologia da Universidade do Porto, R. Dr. Roberto Frias s/n, 4200 Porto, Portugal

²Unidad de Genética Forense, Instituto de Medicina Legal, Universidad de Santiago de Compostela, E-15705 Santiago de Compostela, Galicia, Spain

³Faculdade de Ciências da Universidade do Porto, 4050 Porto, Portugal

⁴Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona 08003, Catalonia, Spain

⁵Department of Psychiatry, POWIC SANE Research Centre, University of Oxford, Warneford Hospital, Oxford OX3 7JX, United Kingdom

Summary

The clinal pattern observed for the distribution of Y-chromosome lineages in Europe is not always reflected at a geographically smaller scale. Six hundred and sixty-three male samples from the 18 administrative districts of Portugal were typed for 25 Y-chromosome biallelic and 15 microsatellite markers, in order to assess the degree of substructuring of male lineage distribution. Haplogroup frequency distributions, Analysis of Molecular Variance (AMOVA) and genetic distance analyses at both Y-SNP and Y-STR levels revealed a general genetic homogeneity of Portuguese sub-populations. The traditional division of the country in north, central and south, which is usually considered in studies addressing questions of the genetic variation distribution in Portugal, was not reflected in the Y-haplotype distribution. Instead, just one sub-region (Alentejo) stood out due to the presence of high diversity levels and a higher number of different lineages, at higher frequencies than in other regions. These results are reconciled with the historical evidence available, assuming that from prehistorical times down to the end of the medieval period this region harboured the most diverse groups of people and, because of economic depression, remained relatively isolated from recent homogenisation movements. The finding of a broadly homogeneous background for the Portuguese population has vast repercussions in forensic, epidemiological and association studies.

Keywords: Y-SNPs, Y-STRs, population structure, phylogeny, Portugal.

Introduction

Portugal, a country situated in the Iberian Peninsula, constitutes the south-western European edge, and faces the Atlantic Ocean to the west and south, just ~300 km from North Africa. The present Portuguese genetic landscape is the outcome of an old and slow process of gene flow, admixture with many different populations, and local differentiation. These include the expansion

from isolated population nuclei in refuges following the Last Glacial Maximum (LGM), the movement of peoples related to the introduction of agriculture, and subsequent Roman and Germanic invaders, which may have influenced the distribution of genetic diversity in the territory. In addition, some admixture events took place in historical epochs that were reported to have left imprints in the Portuguese genetic background, namely of North and sub-Saharan African origin, resulting from the Moslem invasion during the 8th century, and the slave trade, particularly important from the 15th to 18th centuries (Pereira *et al.* 2000a,b; Spinola *et al.* 2002).

*Corresponding author: Sandra Beleza, IPATIMUP, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal. Tel: +351 22 5570700; Fax: +351 22 5570799. E-mail: sbeleza@ipatimup.pt

Previous analyses of the Portuguese Y-SNP diversity (Pereira *et al.* 2000b) found statistically significant differences between the north and south of Portugal. In that study one haplogroup, E* (xE3a), showed an increasing north-to-south frequency gradient, which was associated with the influx of Berbers during the Islamic period in Iberia in the 8th to 14th centuries. The analysis of mtDNA variability failed to reveal any statistically significant differences (Pereira *et al.* 2000a), although a different study showed opposite frequency gradients of the Paleolithic mtDNA lineage H* (without CRS) and the Neolithic J2 (González *et al.* 2003). These were explained as the result of a higher impact of the Neolithic in the south.

In this work we have conducted a study of the distribution of the Y-chromosome in Portugal, by extending the sampling to all Portuguese districts and by increasing the number of analysed markers. The combined study of slow and fast evolving Y-specific markers provides a suitable approach to analyse the demographic histories of populations on both micro-geographic level and in historical times, at least from a male perspective. In this study we use, for the first time, joint information from both types of markers in order to better understand how the different sequence of historical events interfered with the genetic background of Portugal, and to determine if these events led to a relevant geographic structuring of the male lineages in the territory.

Material and Methods

DNA Samples

A total of 663 blood samples were collected in the 18 districts of Portugal from healthy unrelated individuals born in each area, after informed consent (Figure 1). The samples analysed for Y-STRs (657) and Y-SNPs (658) did not always concur, either because the condition of the sample did not allow further typing or due to the lack of available sample. During the statistical analyses, three categories of samples were considered: either (1) local samples were considered independently, or (2) they were pooled into North, Central and South, taking the rivers Douro and Tagus as barriers, as previously considered by others (Pereira *et al.* 2000a,b; Spinola *et al.* 2002; González *et al.* 2003), or (3) they were as-

signed to provinces, traditional regions defined according to geography and climate criteria (Amorim Girão, 1941) (Figure 1). Genomic DNA was extracted with the chelex extraction method (Lareu *et al.* 1994).

Marker Typing

Twenty-five Y-chromosome SNP markers were typed in order to define the most frequent male haplogroups in Portugal (Figure 2, plus two SNPs within exon 5 of the *PCDHY* gene). The typing strategy involved several standard methods. The biallelic markers SRY 10831 1/2, YAP, SRY4064, M2, 12f2, M9, 92R7, SRY2627, LLY22g and Tat were typed as previously (Rosser *et al.* 2000). For the remaining 14 Y-SNP markers, the following hierarchical scheme was used (see Table 1 for methods):

- M35, M78, M81 and M123 were tested in individuals that fell within the E* (xE3a);
- M170, M26, M172, M62 and M201 were tested in individuals that fell within haplogroup F* (xK*);

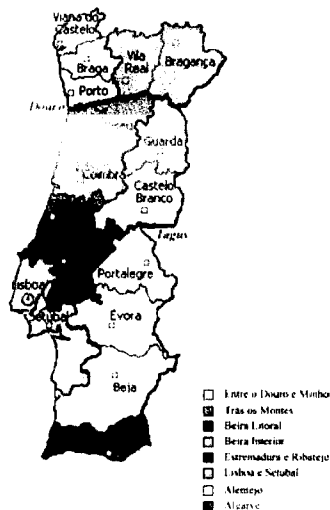


Figure 1 Map of Portugal. Names in the map correspond to the district names. Provinces are individualized by different grey colours. North, centre and south comprise the regions separated by the main Portuguese rivers displayed in the figure, Douro and Tagus.

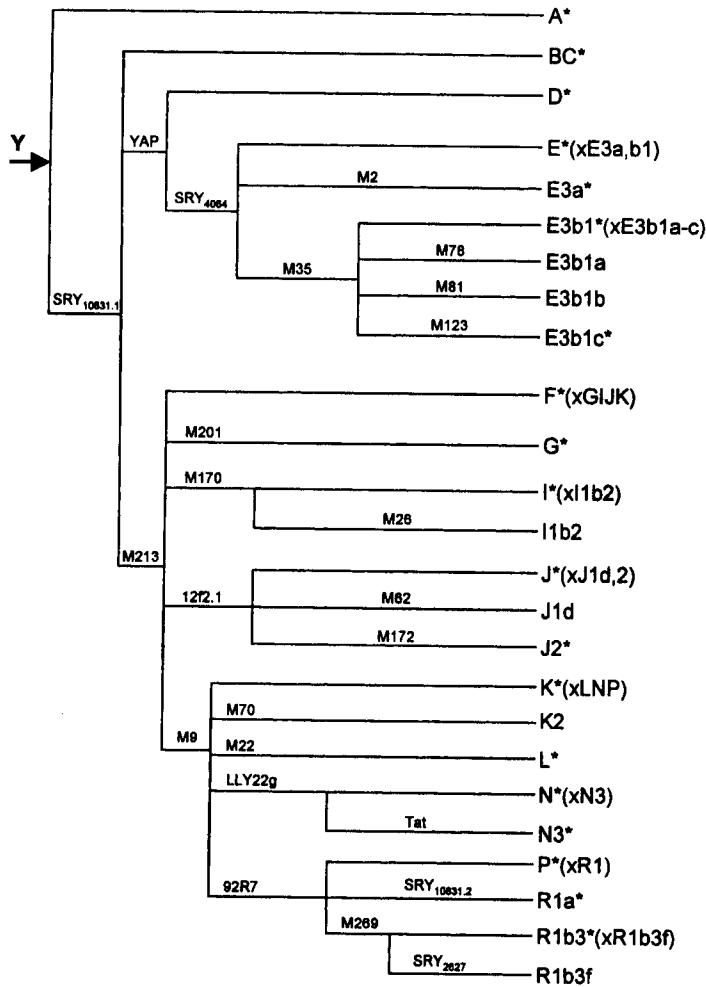


Figure 2 Phylogenetic tree of Y-SNP haplogroups in Portugal. Biallelic markers are displayed in each branch. Note: following the work of Cinnioglu *et al.* (2004) and Semino *et al.* (2004), "J1d" was attributed to the J*-M62 lineage.

- one BCF* (xGIJK) individual was tested for M213;
- K* (xP) individuals were typed for M70;
- M269 was typed in individuals recognised as R1b* (xR1b3f);
- two individuals that could not be assigned to any of these lineages were later allocated to haplogroup L* due to the presence of a derived state at M22 (see Table 1);
- finally, two SNPs located within exon 5 of the *PCDH1Y* gene (NCBI assay IDs: ss5608044 and

ss5608045), recently described by Giouzezi *et al.* (2004), were typed according to Lopes *et al.* (2004).

The nomenclature given is according to the YCC (2002) and Jobling & Tyler-Smith (2003). The nomenclature of haplogroups E3b and J was updated according to Cinnioglu *et al.* (2004), Cruciani *et al.* (2004) and Semino *et al.* (2004).

Fifteen Y-STRs (DYS19, DYS389 I and II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438,

Table 1 Typing strategy and sequence of newly constructed primers used in Y-SNP typing.

Marker	Typing	Ref	Forward and Reverse primers (5'-3')	Ref	Tm	Enzyme
M35	RFLP	-	GCA TGG TCC CTT TCT ATG GAT GAG AAT GAA TAG GCA TGG GTT C	-	60°C	BmrI
M78	RFLP	-	CTT CAG GCA TTA TTT TTT TTG GT ATA GTG TTC CTT CAC CTT TCC TT	[1] [1]	-	-
M81	RFLP	-	GCA CTA TCA TAC TCA GCT ACA CAT CTC AAC CAT TGT GTT ACA TGG CCT A	-	60°C	MaeII
M123	RFLP	-	ATT CAT GCT CTC AGG GGA AA AGC AAA GTT GAG GTT GCA CA	-	59°C	Sfa NI
M170	SNaPshot	[2]	Multiplex 2	[2]	-	-
M26	SNaPshot	[2]	Multiplex 2	[2]	-	-
M172	SNaPshot	[2]	Multiplex 2	[2]	-	-
M62	SNaPshot	[2]	Multiplex 2	[2]	-	-
M201	SNaPshot	[2]	Multiplex 2	[2]	-	-
M213	Sequencing	-	GGCCATATAAAAACGCAGCA TGAATGGCAAATTGATTCCA	[2] [2]	59°C	-
M70	Sequencing	-	TCATAGCCCACTATACTTTGGAC CTGAGGGCTGGACTATAGGG	[2] [2]	59°C	-
M22	SNaPshot	[2]	Multiplex 1	[2]	-	-
M269	Allele-specific PCR	-	AAT GAT CAG GGT TTG GTT AAA T/C ^{a,b} ACT ATA CTT CTT TTG TGT GCC TTC	- -	59°C	-

[1] Underhill *et al.* 2000; [2] Brion *et al.* 2005.
^aThe two forward primers were labeled with different fluorescent dyes.
^bUnderlined base is a mismatched base, intentionally inserted to prevent the annealing of the non-homologous primer during PCR amplification.

DYS439, DYS460, DYS461, DYS635 (GATA C4), GATA A10 and GATA H4) were typed in two multiplex reactions, using the same primer sequences and amplification and detection conditions as described by Beleza *et al.* (2003). Allele nomenclature was as proposed in the Y Chromosome Haplotype Reference Database (YHRD; www.yhrd.org) and by Gusmão *et al.* (2002).

Statistical Analysis

Haplogroup/haplotype frequencies were estimated by haplogroup/haplotype counting. The heterogeneity in the haplogroup frequencies within the Portuguese territory was assessed by an exact test using a Markov chain procedure, as implemented in the STRUC program included in GENEPOP 3.1c computer package (Raymond & Rousset, 1995). Spatial autocorrelation analysis was done by AIDA (Bertorelle & Barbujani, 1995) for the entire data set. Autocorrelation coefficients (II) were calculated by subdividing population pairs into five up to ten geographic distance classes; in each analysis, classes were defined so that each would contain roughly the same number of population pairs. The same protocol was applied to individual haplogroups,

which were subjected to classical autocorrelation analysis (Sokal & Odden, 1978) by means of the PASSAGE software (Rosenberg, 2001).
For both Y-SNP and Y-STR haplotypes, genetic distances, genetic diversities, and the analysis of molecular variance (AMOVA) were calculated with the ARLEQUIN ver 2.000 software (Schneider *et al.* 2000). The distance method used for Y-STRs assumed the stepwise mutational model (R_{ST}; Slatkin, 1995). Another measure for Y-STR diversity, a variance ratio, was also estimated by dividing the variance of each microsatellite in each population by the average variance of the microsatellite across the 18 sub-populations (Jorde *et al.* 2000). A multidimensional scaling (MDS) analysis with pairwise R_{ST} distances was performed using the STATISTICA software package.
Mantel tests were performed in the Mantel Nonparametric Test Calculator, version 2.0 for Windows, developed by Adam Liedloff (<http://www.sci.qut.edu.au/NRS/mantel.htm>), in order to check for correlation between geographical coordinates of each locality and genetic distances.
Relative age of the main Y-chromosome lineages (with a sample size >7) present in Portugal were

estimated using a Bayesian approach incorporated in the program *Batwing* (Wilson *et al.* 2003). We considered a model of exponential growth from initial constant population size, using the effective population size and the population growth rate priors specified in Weale *et al.* (2001). We considered two different sets of mutation rates as mutation priors: those estimated from pedigree data (Kayser *et al.* 2000) and the effective mutation rate estimated in phylogenetic analyses (Zhivotovsky *et al.* 2004). All other parameters were given flat, uninformative priors to minimise their impact on the results. The median and the 5% and 95% quantiles for the posterior distribution of the time since the most recent common ancestor of the sample (TMRCA or T) were estimated. Calculations were based on 50,000 runs of MCMC estimator after a 10,000 run "burn in" time. Generation time was set at 25 years.

Results

Y-SNP Haplogroup Variation

Of the 25 haplogroups discriminated by the 25 Y-SNPs that were shown to be polymorphic in European populations (Rosser *et al.* 2000; Semino *et al.* 2000; Underhill *et al.* 2000; 2001; Jobling & Tyler-Smith, 2003) only 17 were present in our sample (Figure 2). Haplogroup frequencies observed in each population and their respective diversities are shown in Table 2. Portugal has a typical Western European haplogroup composition, characterized by high levels of haplogroup R1b3*-M269 and smaller levels of haplogroups I*-M170, G*-M201 and E3b1*-M35 (Semino *et al.* 2000; Bosch *et al.* 2001; Cruciani *et al.* 2002; Flores *et al.* 2004; Brion *et al.* 2005). The only exception is the higher frequency of haplogroup E3b1b, which was already reported to be a characteristic of Southwestern European populations, namely Iberia and Sicily (Rosser *et al.* 2000; Cruciani *et al.* 2004; Semino *et al.* 2004). The study of two recently described SNPs within exon 5 of the *PCDH1Y* gene, 288 bp apart, in a sample of 69 male individuals of European descent (Giouzezi *et al.* 2004) and in a sample of 112 Portuguese male individuals (Lopes *et al.* 2004) showed them to be polymorphic, but presenting only two allelic combinations: one ancestral (TG) and one doubly derived (GT). A rough estimation of these mutation ages suggested a more recent origin for

the derived haplotype compared to M173 (Lopes *et al.* 2004). In our sample, we found that only R1a* and R1b3* chromosomes harboured the derivate state at both *PCDH1Y* SNPs, meaning that these SNPs could be phylogenetically equivalent to either M173 or M207 mutations. However, more lineages within clade R and within P(xR) need to be typed for these *PCDH1Y* SNPs to be precisely allocated in the Y phylogenetic tree. The most frequent haplogroup was R1b3* (xR1b3f), presenting frequencies higher than 50% in almost all sub-populations except Alentejo. The other haplogroup frequencies remained lower than 10% in the total sample and within the provinces, but exhibited the opposite trend in some districts.

No significant differences in haplogroup frequencies were found between districts ($P = 0.09$, exact test), as well as between north, central and south ($P = 0.64$, exact test), or between provinces ($P = 0.66$, exact test). Furthermore, and contrary to previous findings (Pereira *et al.* 2000b), no correlations were found between the frequency of any haplogroup and latitude or longitude. Spatial autocorrelograms were also computed by means of AIDA (Bertorelle & Barbujani, 1995). Different ways of dividing pairs of populations into distance classes yielded consistently non-significant correlograms (data not shown). The same was true for the frequencies of all haplogroups present in the Portuguese sample.

The haplogroup diversity values ranged between 0.402 and 0.893. No differences in these values were found between the North, Centre and South (Mann-Whitney test, $P > 0.05$). Curiously, the districts of southern Alentejo show higher diversity values than the other provinces, with marginally significant differences from the North and Central coastal provinces Entre Douro e Minho and Beira Litoral (Mann-Whitney test, $P = 0.049$). To exclude a spurious effect due to different sample sizes, we tested to see if the haplogroup diversities and their standard errors were correlated across provinces, which was shown not to be the case (Spearman's $\rho = 0.19$, $P = 0.445$).

Y-STR Haplotype Variation

The typing of a 15 Y-STR set allowed the detection of 577 different haplotypes, 36 of them being shared by two sub-populations, nine by three, two by four and one by five sub-populations (STR haplotype

Table 2 Y-SNP frequency distribution and haplogroup diversity per population in Portugal.

Population, code	n	H	E*	E3a*	E3b1*	E3b1a	E3b1b	E3b1c*	F*	G*	I*	J*	J2*	K2	L*	R1a*	R1b3*	R1
			(x/E3a, b1)		(x/E3b1-c)				(x/GJ)		(x/Ib)	(x/J1d, 2)					(x/R1b 3f)	b3f
Entre Douro e Minho																		
Viana do Castelo, VC	59	0.695	1.7	0	3.4	0	6.8	0	0	5.1	8.5	0	3.4	15.3	1.7	0	52.5	1.7
Braga, B	51	0.705	0	2.0 ^a	0	2.0	5.9	2.0	0	7.8	15.7	2.0	3.9	0	3.9	0	51.0	0
Porto, P	118	0.652	0	0	0	4.2	6.8	1.7	0	5.9	3.4	0.9	3.4	8.5	2.5	0.9	57.6	2.5
Total	228	0.676	0.4	0.4	0.9	2.6	6.6	1.3	0	6.1	7.5	0.9	3.5	8.3	2.6	0.4	55.3	1.8
Trás os Montes																		
Bragança, BN	25	0.590	0	0	0	4.0	4.0	4.0	0	4.0	12.0	4.0	0	4.0	0	0	64.0	0
Vila Real, VR	39	0.673	0	0	0	2.6	2.6	0	0	2.6	7.7	5.1	0	7.7	7.7	0	56.4	2.6
Total	64	0.636	0	0	0	3.1	3.1	1.6	0	3.1	9.4	4.7	0	6.3	4.7	0	59.4	1.6
Beira Litoral																		
Aveiro, AV	66	0.626	0	0	0	4.6	3.0	3.0	0	3.0	3.0	1.5	6.1	4.6	6.1	0	60.7	3.0
Viseu, VI	30	0.547	0	0	0	0	13.3	0	0	0	3.3	0	0	6.7	3.3	0	66.7	3.3
Coimbra, Co	20	0.742	0	0	5.0	0	0	5.0	0	5.0	5.0	0	15.0	10.0	0	5.0	50.0	0
Total	116	0.625	0	0.9	0.9	2.6	5.2	2.6	0	2.6	3.4	0.9	6.0	6.0	4.3	0.9	60.3	2.6
Beira Interior																		
Guarda, Gu	30	0.708	0	0	0	10.0	3.3	0	3.3	10.0	0	0	3.3	6.7	3.3	0	53.3	3.3
Castelo Branco, CB	28	0.489	0	0	0	3.6	7.1	0	0	7.1	7.1	0	0	0	0	0	71.4	3.6
Total	58	0.605	0	0	0	7.0	5.3	0	1.8	8.8	3.5	0	1.8	3.5	1.8	0	62.3	3.5
Estremadura																		
Leiria, Le	35	0.629	0	0	2.9	5.7	5.7	0	0	8.6	11.4	2.9	0	0	0	0	60	2.9
Santarém, Sa	8	0.821	0	0	0	25.0	0	0	0	0	0	12.5	0	25.0	0	0	37.5	0
Total	43	0.674	0	0	2.3	9.3	4.6	0	0	7.0	9.3	4.7	0	4.7	0	0	55.8	2.3
Lisboa e Setúbal																		
Lisboa, Li	35	0.692	0	0	2.9	2.9	11.4	0	0	5.7	2.9	0	0	8.6	0	0	54.3	5.7
Setúbal, Se	27	0.402	3.7	0	3.7	3.7	0	0	0	3.7	0	0	3.7	3.7	0	0	77.8	0
Total	62	0.577	1.6	0	3.2	3.2	6.5	0	0	4.8	1.6	0	1.6	6.5	0	0	64.5	3.2
Alentejo																		
Portalegre, Po	28	0.759	0	0	0	14.3	10.7	0	0	3.6	7.1	0	3.6	14.3	0	0	44.8	0
Évora, Ev	29	0.805	0	0	0	6.9	3.5	0	0	13.8	10.3	3.5	10.4	3.5	3.5	0	41.4	3.5
Beja, Be	8	0.893	12.5	0	0	0	12.5	12.5	0	0	0	12.5	0	12.5	0	0	37.5	0
Total	65	0.790	1.5	0	0	9.2	7.7	1.5	0	7.7	7.7	3.1	6.2	9.2	1.5	0	43.1	1.5
Algarve																		
Faro, Fa	21	0.424	0	0	0	0	0	0	0	4.8	4.8	0	4.8	9.5	0	0	76.2	0
Total Portugal	657	0.652	0.5	0.2	0.9	4.1	5.6	1.2	0.2	5.5	6.1	1.5	3.4	7.0	1.6	0.3	57.7	2.2

^aThe only E3a* chromosome found holds a STR-haplotype typical of E3b1b (see Table S1).

Table 3 Y-STR haplotype diversity and average variance ratio per population in Portugal.

Population	Viana do Castelo					Vila			Castelo					Faro			
	Braga	Porto	Bragança	Real	Aveiro	Viseu	Coimbra	Guarda	Branco	Leiria	Santarém	Lisboa	Setúbal		Portalegre	Évora	Beja
N	60	53	119	45	39	48	30	21	30	28	31	7	35	26	28	8	23
Haplotype diversity	1.0000	0.9993	0.9997	0.9990	1.0000	0.9991	0.9954	1.0000	0.9977	1.0000	0.9957	1.0000	0.9966	1.0000	0.9938	0.9974	1.0000
Average variance ratio	1.0827	1.0601	0.9851	0.9170	0.9057	0.9023	1.0109	1.0205	1.0234	0.6616	1.1215	1.2802	0.9138	0.8144	1.1186	1.0879	1.1053

data within each haplogroup is given in supplementary material, Table S1). Haplotype diversity was over 99% in all districts (Table 3), and shown to be uncorrelated with haplogroup diversity ($r = 0.02$). However, estimating a Y-STR variance ratio for each district (Jorde *et al.* 2000) in order to remove the possible bias arising from the contribution of highly mutating Y-STRs, the overall correlation increased to $r = 0.613$. No differences in the Y-STR diversity values were found between the North, Centre and South (Mann-Whitney test, $P > 0.05$). No differences were also observed between the different provinces, although the samples from Alentejo showed consistently higher diversity values.

Genetic Distances and MDS Analysis

Haplogroup distributions in Portuguese sub-populations did not show a clear differentiation (Figure 3). Nevertheless, since the occurrence of unique mutation events at the biallelic loci is much older than the establishment of the Portuguese population, and Y-STR loci possess a higher evolutionary rate, Y-STR haplotypes are more informative in testing whether some heterogeneity exists among these sub-populations as a consequence of either genetic drift or low gene flow. Obviously, given the non-recombinant nature of the NRY, differentiation between haplogroups and haplotypes is correlated (Mantel test between haplogroup Fst and haplotype Rst: $r = 0.554$, $P < 0.01$). However, this value implies that only $\sim 1/4$ of the STR haplotype variation can be explained by haplogroup differentiation; the remaining variation may be a consequence of drift (affecting STR haplotypes more intensely, each of which tends to be relatively infrequent) and/or mutation locally generating STR variation. R_{ST} pairwise comparison values were low (data not shown), as could be predicted from such a microgeographic study of neighbouring populations. The MDS plot of Figure 3 also shows that there is not clear geographic heterogeneity in the Portuguese Y-chromosomal variation. However, with respect to the Y-STR MDS plot, the samples from Alentejo tend to separate from a common cluster composed of the other Portuguese sub-populations (Figure 3).

Moreover, the Mantel test revealed no correlation between the geographical distance and Y-SNP based

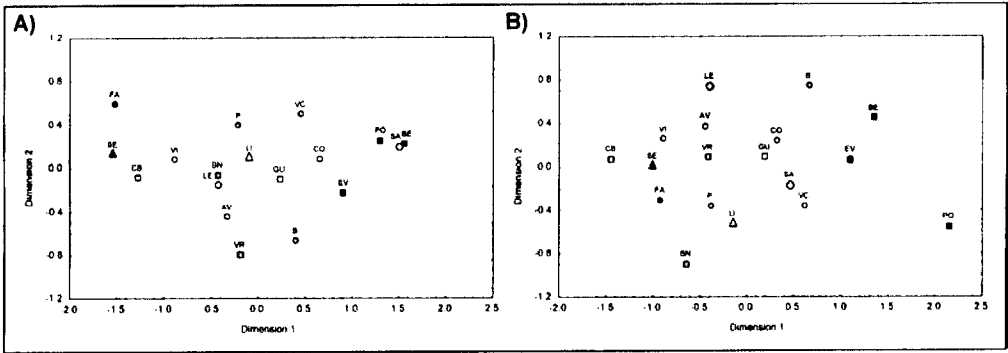


Figure 3 Multidimensional scaling plot of A) F_{ST} and B) R_{ST} pairwise values found when analysing, respectively, Y-SNP and Y-STR variation in Portugal. Stress value = 0.104 in A) and 0.155 in B). Grey, white and black symbols represent northern, central and southern Portuguese districts, respectively. Circles, squares, triangles and crosses represent provinces. District codes and corresponding provinces are indicated in Table 2.

Table 4 AMOVA results of Y-chromosome lineages in Portugal.

Grouping and Source of Variation	Y-SNPs (%)	Y-STRs (%)
No Grouping		
Among populations	0.64	0.67
Within populations	99.36	99.33
North/Centre/South		
Among groups	−0.25	0.05
Among populationsWithin groups	0.81	0.64
Within populations	99.44	99.31
Provinces		
Among groups	0.52	0.52
Among populationsWithin groups	0.18	0.21
Within populations	99.29	99.26
Alentejo vs Portugal		
Among groups	2.99*	3.86*
Among populationsWithin groups	0.04	−0.07
Within populations	96.97	96.21

*Statistically significant values after Bonferroni correction for multiple tests.

genetic distances ($r = 0.0751$, $P = 0.23$) or with Y-STR based genetic distances ($r = -0.0053$, $P = 0.46$).

AMOVA Analysis

Analysis of molecular variance (AMOVA) was computed from haplogroup and haplotype frequencies (Table 4). The overall value obtained between the 18 districts with both data sources was low and non-significant, consistent with the values found for other regions in Iberia (Brion *et al.* 2004; Flores *et al.* 2004).

Furthermore, when grouping the districts according to latitude (North, Central and South Portugal), the percentage of variation explained by differences among groups was even smaller. A higher between-group component was found when dividing sub-populations according to the administrative region to which they belong (Table 4).

According to all the evidence obtained until now, namely the higher genetic diversities and genetic distances of populations from Alentejo, we sought to find out the proportion of variation present between Alentejo and the remainder of the Portuguese regions. This criterion led to the highest and most significant value found for the “among groups” component, and to the smallest for the “among populations within groups” component (Table 4).

Phylogeographic Analysis

In order to determine the major components that have left their imprint on the Portuguese male gene pool during its demographic history, we estimated the relative age of the major lineages present in the territory; by considering two different mutation rate estimates (Kayser *et al.* 2000; Zhivotovsky *et al.* 2004) we have obtained an upper and a lower boundary for the age of each lineage in Portugal (Table 5).

Previous phylogeographic analyses allowed us to conclude that haplogroup R1*(xR1a)-M173 has been spread throughout Europe since Paleolithic times

Table 5 Y-chromosome haplogroup variance and age estimates in Portugal based on 15 Y-STR loci.

Haplogroup	n	Variance	Posterior Probabilities (in Ky) ^a	
			T _{lower} ¹	T _{upper} ²
R1b3*(xR1b3f)-M269	335	0.286	14.1 (7.7, 34.3)	38.3 (22.9, 77.4)
R1b3f-SRY ₂₆₂₇	14	0.172	3.1 (1.7, 6.9)	9.9 (5.9, 19.2)
R1a*-SRY _{10831.2}	13	0.517	13.6 (7.6, 29.1)	40.8 (26.2, 77.2)
K2-M70	14	0.299	5.6 (3.1, 13.1)	17.5 (10.7, 36.5)
J*(xJ1d,2)-12f2	21	0.329	8.7 (4.5, 21.0)	27.7 (15.7, 59.7)
J2*-M172	44	0.607	11.9 (7.0, 25.2)	37.5 (25.1, 67.0)
I*(xI1b2)-M170	38	0.521	10.3 (6.1, 20.8)	33.5 (22.2, 61.0)
I1b2-M26	10	0.376	6.6 (3.7, 13.9)	20.2 (12.8, 37.0)
G*-M201	36	0.452	8.5 (4.9, 20.1)	27.8 (18.0, 57.8)
E3b1a-M78	26	0.383	5.5 (3.3, 10.5)	17.2 (11.9, 28.8)
E3b1b-M81	37	0.129	2.8 (1.4, 7.1)	8.0 (4.8, 18.1)
E3b1c*-M123	8	0.426	7.8 (4.4, 16.6)	23.4 (14.8, 43.4)

^aMedian (95% equal-tailed intervals).

¹Estimated with the mutation rate obtained from pedigree data (Kayser *et al.* 2000).

²Estimated with the effective mutation rate obtained in phylogenetic analyses (Zhivotovsky *et al.* 2004).

(Semino *et al.* 2000). Later, M269 was found to be present in all European R1*(xR1a)-M173 individuals by Cruciani *et al.* (2002), and STR variance analysis suggested that the distribution of R1b3*-M269 in this continent is more the result of its diffusion from the Iberian refugia after the LGM, during the Late Upper Paleolithic (Cinnioğlu *et al.* 2004). The same analysis of variance made with the Portuguese R1b3*(xR1b3f) individuals corroborates these findings (STR variance of the Portuguese R1b3*(xR1b3f) equals 0.29, a smaller value when compared with that obtained by Cinnioğlu *et al.* [2004] in other Iberian samples [0.32]). Age estimates obtained from the STR variation associated with R1b3*(xR1b3f) chromosomes in Portugal (Table 5) point to a pre-Neolithic age for this haplogroup, and set an upper limit for its introduction to Portugal (Barbujani *et al.* 1998). Another lineage also found to be characteristic of LGM expansions from refuge areas, I*-M170 (Rootsi *et al.* 2004; see also Table 5), was present in our sample at lower levels.

Due to their decreasing frequencies from the Middle East to Europe, haplogroups J, G and E have been associated with the demic diffusion of Neolithic farmers into Europe (Semino *et al.* 2000; Underhill *et al.* 2001). J lineages were found to have a greater impact in the eastern and central part of the Mediterranean basin (Scozzari *et al.* 2001; Di Giacomo *et al.* 2003) than in Iberia and northeastern Europe (Scozzari *et al.* 2001; Flores *et al.* 2004). Later, the refinement of the phylogeny of haplogroups J and E allowed the discrim-

ination of a collection of subclades of Neolithic and post-Neolithic origin, reflecting very different evolutionary histories and processes of migration (Cruciani *et al.* 2004; Semino *et al.* 2004). Indeed, the age estimates of J*(xJ1d,2)-12f2, G*-M201 and E3b1c*-M123 (Table 5) are consistent with an introduction to Portugal no earlier than the Neolithic. J2*-M172 presents higher age estimates (Table 5) but, judging from its wide variance, this must be a consequence of the coexistence of more than one subclade within this haplogroup (mainly J2-M172 and J2*-M67, both of Neolithic origin according to Semino *et al.* 2004) that was not discriminated with the set of Y-SNP markers typed. This set of post-Paleolithic Y chromosomes comprises 17% of the male Portuguese background, a frequency that is smaller than that observed in Italy or Greece for the same component (Semino *et al.* 2000; Di Giacomo *et al.* 2003) but higher than in northeastern Europe (Semino *et al.* 2000; Wells *et al.* 2001). Within the Portuguese territory, this component is present at 17.5% in the north, 15% in the centre and 20.4% in the south.

The distribution of E3b1a-M78 (E3b1a) in Western Europe was described as resulting from late demographic expansions from the Balkans along the Mediterranean Sea (Cruciani *et al.* 2004; Semino *et al.* 2004), and E3b1a was considered to be a signature of Greek colonists in south Italy (Semino *et al.* 2004). Its presence in the Portuguese territory might have the same origin (see Table 5). In fact, within Portugal there is a higher frequency of E3b1a in the south, which decreases

toward the north (6.2%, 4.8% and 2.7% in south, centre and north, respectively). Nevertheless, and given the current sample sizes, these values are not significantly different from each other ($P = 0.64$, exact test), and the autocorrelation analysis was not significant (data not shown).

E3b1b-M81 was found at 1.5–11.5% in a number of Spanish populations (Bosch *et al.* 2001; Cruciani *et al.* 2004; Flores *et al.* 2004; Brion *et al.* 2005), being as high as 20%–40% in Pasiegos from Cantabria, Spain (Brion *et al.* 2004; Cruciani *et al.* 2004). The distribution of these E3b1b chromosomes and respective age estimates in Europe (e.g. Table 5) led to the suggestion that its presence in the continent was mediated by the Islamic occupation of the Iberian Peninsula (Bosch *et al.* 2001; Cruciani *et al.* 2004). The degree of North African contribution was found to be highly variable across different Iberian populations (Cruciani *et al.* 2004; Flores *et al.* 2004). The same holds true in Portugal, where no clear pattern can be observed (Table 2). Considering the division of the territory by latitude, we find 5.8% of E3b1b lineages in the north, 6% in the centre and 4.4% in south. Pereira *et al.* (2000b) speculated that the observation of a north to south increasing gradient of E* (xE3a) might be associated with the patterns of Islamic occupation in Portugal reported in the historical records (stronger and longer-lasting in the south). Our results do not support this hypothesis, and the higher frequencies of E* (xE3a) in the south observed by the authors may have two non-exclusive explanations. Firstly, they might be a consequence of the higher frequency of E3b1a rather than E3b1b in this region. Secondly, these observations were based on a relatively small sample size for the south compared with the northern one, and the southern samples were mainly from Alentejo, which possesses higher frequencies of both haplogroups. Estimating the frequency of the same lineage in our extended sample we also obtained an increasing gradient of 11.0% in the north, 13.1% in the centre and 14.4% in the south, although this was not significant. One fact worth mentioning is the complete absence of this lineage in the most southern Portuguese region, Algarve, the area that is described as one of those most influenced by the Islamic rule (Torres, 1993), and its presence at 6.6% in the northern region, Entre Douro e Minho.

Interestingly, the North African mtDNA counterpart, U6, is also found at 6% in the same northern region, the highest frequency found in Portugal (Pereira *et al.* 2000a; González *et al.* 2003), being rarer in southern areas.

When analysing globally the European Y lineages, we observe that all of them, R1b3* (xR1b3f) excepted, are observed at higher frequencies in Alentejo (Table 2).

Finally, we have detected two L chromosomes, a haplogroup that is characteristic of Southern Indian populations (Kivisild *et al.* 2003) but extremely rare in Europe (Semino *et al.* 2000; Scozzari *et al.* 2001; Flores *et al.* 2004; Brion *et al.* 2005).

Discussion

Y-chromosome population studies on a large geographic scale (Rosser *et al.* 2000; Semino *et al.* 2000) have revealed a significant substructuring of Y-chromosome variation and important clinal patterns for some male lineages in Europe. However, research in more restricted geographical contexts has shown that this pattern is not always reflected on a smaller scale (Di Giacomo *et al.* 2003; Flores *et al.* 2004), revealing important genetic drift and founder effects (Di Giacomo *et al.* 2003). In this work we have analysed for the first time joint information given by Y-SNPs, and Y-STRs in order to assess the degree of micro-geographic substructuring of the Portuguese male lineages. It is known from archaeological data (Ribeiro, 1966) that, throughout the demographic history of Portugal, there was a clear differentiation between cultures that existed to the north of the river Douro and those that influenced the south, beyond the Tagus River. However, our results failed to detect any significant male gene pool heterogeneity between the north, central and south regions. Conversely, our analysis revealed a general genetic homogeneity of Portuguese sub-populations, with the exception of those belonging to Alentejo. The relative differentiation of Alentejo comes from the presence and abundance of a higher number of different lineages of recent ancestry than are found in other sub-populations (Table 2). On the other hand, the lineage showing an opposite clinal distribution to the arrival of new influences in

Western Europe, R1b3*-M269, presents an unusually low frequency in Alentejo: 43.1% vs values higher than 50% in other Western European populations (Rosser *et al.* 2000; Semino *et al.* 2000). This seems rather curious since populations from Alentejo nowadays are the smallest in size and are still decreasing. However, this was not the case up to the end of the 15th century, when they harboured important rural and urban populations.

As in other European populations, almost all Portuguese Y chromosomes belong to haplogroups R1b3*-M269, I*-M170, J*-12f2, G*-M201 and E3b1*-M35.

One haplogroup that deserves special consideration is E3b1b-M81. Historical records show that in Portugal (and in all Iberia) the Muslim influence was more limited to the south, but occupying a vast area that had the Iberian central mountain range as its northern limit (Torres, 1993). However, what we observe today at the genetic level, both in the paternal (Table 2; Brion *et al.* 2004; Cruciani *et al.* 2004; Flores *et al.* 2004) and the maternal gene pool (Pereira *et al.* 2000a; González *et al.* 2003; Maca-Meyer *et al.* 2003), is that this North African influence is well spread all over Iberia. However, the later Christian Reconquest is described (Bartlett, 1993) as a period of extensive migrations all over the Iberian Peninsula, involving the resettlement of Iberia on a vast scale. These might have erased any differential distribution of the Y-chromosome that could have existed in the past, especially for E3b1b; and today we witness a rather homogeneous genetic landscape in Portugal (and probably also in most of Iberia; Flores *et al.* 2004), at least from the male perspective.

Finally, our findings for the Portuguese paternal background and the ones of Pereira *et al.* (2000a; 2000b) for the paternal and maternal components of the Portuguese population do not reconcile with the clines detected by González *et al.* (2003). A more detailed survey of the distribution pattern of female lineages in Portugal is still needed, with a larger sample size and a more in-depth phylogenetic analysis.

In conclusion, we show the usefulness of a detailed Y-chromosome analysis to the understanding of the demographic history of human populations in the recent past, and at a micro-geographic level, a population attribute that has had a great impact on forensic and epidemiological

applications. The finding of a rather homogeneous background for the Portuguese Y-chromosome gene pool, and all the strong indications that the same characteristic is found for its female counterpart, are of great relevance for association studies. Although genetic differentiation follows a random distribution, and certainly some genes of genetic epidemiological interest are bound to show more stratification than Y lineages do, it is nonetheless true that the Y chromosome is the genomic region most sensitive to drift, and thus, with the largest opportunity for stratification. The low levels of differentiation for the Y chromosome in Portugal add additional confidence to the use of general Portuguese samples in case-control designs.

Acknowledgements

This work was supported by POCTI, Programa Operacional Ciência, Tecnologia e Inovação. Sandra Belez and Alexandra Lopes have a PhD grant (SFRH/BD/860/2000 and SFRH/BD/7006/2001, respectively) from FCT. Two anonymous reviewers made significant contributions to the improvement of this manuscript. We wish to thank local hospitals from the whole country for providing the samples that made this study possible.

References

- Amorim Girão, A. (1941) *Geografia de Portugal*, p. 330. Portucalense Editora, Porto.
- Barbujani, G., Bertorelle, G. & Chikhi, L. (1998) Evidence for Paleolithic and Neolithic gene flow in Europe. *Am J Hum Genet* **62**, 488–492.
- Bartlett, R. (1993) *The Making of Europe: Conquest, Colonization, and Cultural Change, 950–1350*. New Jersey: Princeton University Press.
- Beleza, S., Alves, C., González-Neira, A., Lareu, M., Amorim, A., Carracedo, A. & Gusmão, L. (2003) Extending STR markers in Y chromosome haplotypes. *Int J Legal Med* **117**, 27–33.
- Bertorelle, G. & Barbujani, G. (1995) Analysis of DNA diversity by spatial autocorrelation. *Genetics* **140**, 811–819.
- Bosch, E., Calafell, F., Comas, D., Oefner, P. J., Underhill, P. A. & Bertranpetit, J. (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am J Hum Genet* **68**, 1019–1029.
- Brion, M., Quintans, B., Zarrabeitia, M., Gonzalez-Neira, A., Salas, A., Lareu, V., Tyler-Smith, C. & Carracedo,

- A. (2004) Micro-geographical differentiation in Northern Iberia revealed by Y-chromosomal DNA analysis. *Gene* **329**, 17–25.
- Brion, M., Sobrino, B., Blanco-Verea, A., Lareu, M. V. & Carracedo, A. (2005) Hierarchical analysis of 30 Y-chromosome SNPs in European populations. *Int J Legal Med* **119**, 10–15.
- Cinnioglu, C., King, R., Kivisild, T., Kalfoglu, E., Atasoy, S., Cavalleri, G. L., Lillie, A. S., Roseman, C. C., Lin, A. A., Prince, K., Oefner, P. J., Shen, P., Semino, O., Cavalli-Sforza, L. L. & Underhill, P. A. (2004) Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet* **114**, 127–148.
- Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., Modiano, D., Holmes, S., Destro-Bisol, G., Coia, V., Wallace, D. C., Oefner, P. J., Torrioni, A., Cavalli-Sforza, L. L., Scozzari, R. & Underhill, P. A. (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet* **70**, 1197–1214.
- Cruciani, F., La Fratta, R., Santolamazza, P., Sellitto, D., Pascone, R., Moral, P., Watson, E., Guida, V., Colomb, E. B., Zaharova, B., Lavinha, J., Vona, G., Aman, R., Cali, F., Akar, N., Richards, M., Torrioni, A., Novelletto, A. & Scozzari, R. (2004) Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. *Am J Hum Genet* **74**, 1014–1022.
- Di Giacomo, F., Luca, F., Anagnou, N., Ciavarella, G., Corbo, R. M., Cresta, M., Cucci, F., Di Stasi, L., Agostiano, V., Giparaki, M., Loutradis, A., Mammi, C., Michalodimitrakis, E. N., Papola, F., Pedicini, G., Plata, E., Terrenato, L., Tofanelli, S., Malaspina, P. & Novelletto, A. (2003) Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. *Mol Phylogenet Evol* **28**, 387–395.
- Flores, C., Maca-Meyer, N., Gonzalez, A. M., Oefner, P. J., Shen, P., Perez, J. A., Rojas, A., Larruga, J. M. & Underhill, P. A. (2004) Reduced genetic structure of the Iberian peninsula revealed by Y-chromosome analysis: implications for population demography. *Eur J Hum Genet* **12**, 855–863.
- Giouzeli, M., Williams, N. A., Lonie, L. J., DeLisi, L. E. & Crow, T. J. (2004) ProtocadherinX/Y, a candidate gene-pair for schizophrenia and schizoaffective disorder: a DH-PLC investigation of genomic sequence. *Am J Med Genet* **129B**, 1–9.
- González, A. M., Brehm, A., Perez, J. A., Maca-Meyer, N., Flores, C. & Cabrera, V. M. (2003) Mitochondrial DNA affinities at the Atlantic fringe of Europe. *Am J Phys Anthropol* **120**, 391–404.
- Gusmão, L., González-Neira, A., Lareu, M., Costa, S., Amorim, A. & Carracedo, A. (2002) Chimpanzee homologous of human Y specific STRs: a comparative study and a proposal for nomenclature. *Forensic Sci Int* **126**, 129–136.
- Jobling, M. & Tyler-Smith, C. (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet* **4**, 598–612.
- Jorde, L. B., Watkins, W. S., Bamshad, M. J., Dixon, M. E., Ricker, C. E., Seielstad, M. T. & Batzer, M. A. (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am J Hum Genet* **66**, 979–988.
- Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Kruger, C., Krawczak, M., Nagy, M., Dobosz, T., Szibor, R., de Knijff, P., Stoneking, M. & Sajantila, A. (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am J Hum Genet* **66**, 1580–1588.
- Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., Metspalu, E., Adojaan, M., Tolk, H. V., Stepanov, V., Golge, M., Usanga, E., Papiha, S. S., Cinnioglu, C., King, R., Cavalli-Sforza, L., Underhill, P. A. & Villems, R. (2003) The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet* **72**, 313–332.
- Lareu, M. V., Phillips, C. P., Carracedo, A., Lincoln, P. J., Syndercombe Court, D. & Thomson, J. A. (1994) Investigation of the STR locus HUMTH01 using PCR and two electrophoresis formats: UK and Galician Caucasian population surveys and usefulness in paternity investigations. *Forensic Sci Int* **66**, 41–52.
- Lopes, A. M., Calafell, F. & Amorim, A. (2004) Microsatellite Variation and Evolutionary History of PCDHX/Y Gene Pair Within the Xq21.3/Yp11.2 Hominid-Specific Homology Block. *Mol Biol Evol* **21**, 2092–2101.
- Maca-Meyer, N., Sanchez-Velasco, P., Flores, C., Larruga, J. M., Gonzalez, A. M., Oterino, A. & Leyva-Cobian, F. (2003) Y chromosome and mitochondrial DNA characterization of Pasiegos, a human isolate from Cantabria (Spain). *Ann Hum Genet* **67**, 329–339.
- Pereira, L., Prata, M. J. & Amorim, A. (2000a) Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation. *Ann Hum Genet* **64**, 491–506.
- Pereira, L., Prata, M. J., Brion, M., Jobling, M. A., Carracedo, A. & Amorim, A. (2000b) Clinal variation of YAP+ Y-chromosome frequencies in Western Iberia. *Hum Biol* **72**, 937–944.
- Raymond, M. & Rousset, F. (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J Heredity* **86**, 248–249.

- Ribeiro, O. (1966) Portugal, A Formação de. In: *Dicionário da História de Portugal* (eds Iniciativas Editoriais), Vol. V, pp. 130–149. Lisboa, Portugal.
- Roots, S., Magri, C., Kivisild, T., Benuzzi, G., Help, H., Bermisheva, M., Kutuev, I., Barac, L., Pericic, M., Balanovsky, O., Pshenichnov, A., Dion, D., Grobei, M., Zhivotovsky, L. A., Battaglia, V., Achilli, A., Al-Zahery, N., Parik, J., King, R., Cinnioglu, C., Khusnutdinova, E., Rudan, P., Balanovska, E., Scheffrahn, W., Simonescu, M., Brehm, A., Goncalves, R., Rosa, A., Moisan, J. P., Chaventre, A., Ferak, V., Furedi, S., Oefner, P. J., Shen, P., Beckman, L., Mikerezi, I., Terzic, R., Primorac, D., Cambon-Thomsen, A., Krumina, A., Torroni, A., Underhill, P. A., Santachiara-Benerecetti, A. S., Vilems, R. & Semino, O. (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. *Am J Hum Genet* **75**, 128–137.
- Rosenberg, M. S. (2001) *PASSAGE. Pattern Analysis, Spatial Statistics, and Geographic Exegesis. Version 1.1*. Tempe, AZ: Department of Biology, Arizona State University.
- Rosser, Z. H., Zerjal, T., Hurles, M. E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., Barbujani, G., Beckman, G., Beckman, L., Bertranpetit, J., Bosch, E., Bradley, D. G., Brede, G., Cooper, G., Corte-Real, H. B., de Knijff, P., Decorte, R., Dubrova, Y. E., Evgrafov, O., Gilissen, A., Glisic, S., Golge, M., Hill, E. W., Jeziorowska, A., Kalaydjieva, L., Kayser, M., Kivisild, T., Kravchenko, S. A., Krumina, A., Kucinskas, V., Lavinha, J., Livshits, L. A., Malaspina, P., Maria, S., McElreavey, K., Meitinger, T. A., Mikesaar, A. V., Mitchell, R. J., Nafa, K., Nicholson, J., Norby, S., Pandya, A., Parik, J., Patsalis, P. C., Pereira, L., Peterlin, B., Pielberg, G., Prata, M. J., Previdere, C., Roewer, L., Roots, S., Rubinsztajn, D. C., Saillard, J., Santos, F. R., Stefanescu, G., Sykes, B. C., Tolun, A., Vilems, R., Tyler-Smith, C. & Jobling, M. A. (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet* **67**, 1526–1543.
- Schneider, S., Roessli, D. & Excoffier, L. (2000) *Arlequin: A software for population genetics data analysis. Ver 2.000*. University of Geneva.
- Scozzari, R., Cruciani, F., Pangrazio, A., Santolamazza, P., Vona, G., Moral, P., Latini, V., Varesi, L., Memmi, M. M., Romano, V., De Leo, G., Gennarelli, M., Jaruzelska, J., Vilems, R., Parik, J., Macaulay, V. & Torroni, A. (2001) Human Y-chromosome variation in the western Mediterranean area: implications for the peopling of the region. *Hum Immunol* **62**, 871–884.
- Semino, O., Magri, C., Benuzzi, G., Lin, A. A., Al-Zahery, N., Battaglia, V., Maccioni, L., Triantaphyllidis, C., Shen, P., Oefner, P. J., Zhivotovsky, L. A., King, R., Torroni, A., Cavalli-Sforza, L. L., Underhill, P. A. & Santachiara-Benerecetti, A. S. (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am J Hum Genet* **74**, 1023–1034.
- Semino, O., Passarino, G., Oefner, P. J., Lin, A. A., Arbuzova, S., Beckman, L. E., De Benedictis, G., Francalacci, P., Kouvatsi, A., Limborska, S., Marcikiae, M., Mika, A., Mika, B., Primorac, D., Santachiara-Benerecetti, A. S., Cavalli-Sforza, L. L. & Underhill, P. A. (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science* **290**, 1155–1159.
- Slakutin, M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462.
- Spinola, H., Brehm, A., Williams, F., Jesus, J. & Middleton, D. (2002) Distribution of HLA alleles in Portugal and Cabo Verde. Relationships with the slave trade route. *Ann Hum Genet* **66**, 285–296.
- Sokal, R. R. & Oden, N. L. (1978) Spatial autocorrelation in biology I. *Methodology* **10**, 199–228.
- Torres, C. (1993) O Garb-Al-Andaluz. In: *História de Portugal* (ed J. Mattoso), Vol. I, pp. 361–439. Lisboa, Portugal: Circulo dos Leitores.
- Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., Kauffman, E., Bonne-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J. R., Mehdi, S. Q., Seielstad, M. T., Wells, R. S., Piazza, A., Davis, R. W., Feldman, M. W., Cavalli-Sforza, L. L. & Oefner, P. J. (2000) Y chromosome sequence variation and the history of human populations. *Nat Genet* **26**, 358–361.
- Underhill, P. A., Passarino, G., Lin, A. A., Shen, P., Mirazon Lahr, M., Foley, R. A., Oefner, P. J. & Cavalli-Sforza, L. L. (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* **65**, 43–62.
- Weale, M. E., Yepiskoposyan, L., Jager, R. F., Hovhannissyan, N., Khudoyan, A., Burbage-Hall, O., Bradman, N. & Thomas, M. G. (2001) Armenian Y chromosome haplotypes reveal strong regional structure within a single ethnolinguistic group. *Hum Genet* **109**, 659–674.
- Wells, R. S., Yuldasheva, N., Ruzibakiev, R., Underhill, P. A., Evseeva, I., Blue-Smith, J., Jin, L., Su, B., Pitchappan, R., Shannugalakshmi, S., Balakrishnan, K., Read, M., Pearson, N. M., Zerjal, T., Webster, M. T., Zhoshchvili, I., Jamarjashvili, E., Gambarov, S., Nikbin, B., Dostiev, A., Aknazarov, O., Zalloua, P., Tsoy, I., Kitaev, M., Mirrakhimov, M., Chariev, A. & Bodmer, W. F. (2001) The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci U S A* **98**, 10244–10249.
- Wilson, I., Balding, D. & Weale, M. (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J R Stat Soc Ser A* **166**, 155–188.

- Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res* **12**, 339–348.
- Zhivotovsky, L. A., Underhill, P. A., Cinnioglu, C., Kayser, M., Morar, B., Kivisild, T., Scozzari, R., Cruciani, E., Destro-Bisol, G., Spedini, G., Chambers, G. K., Herrera, R. J., Yong, K. K., Gresham, D., Tournev, I., Feldman, M. W. & Kalaydjieva, L. (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet* **74**, 50–61.

Received: 25 January 2005

Accepted: 14 June 2005

Supplementary Material

The following material is available for this article online:

Table S1. Y-chromosome haplotypes in 657 individuals from Portugal.

DISCUSSION

1. Xq21.3/Yp11.2 microsatellite diversity in two human populations

1.1. The genetic background of the populations under study

Levels of polymorphism are a balance between the rate of generation of new variants and the rate of loss or fixation of these variants in populations. On the Y chromosome, genetic drift increases the rate of the latter processes, contributing to lower levels of polymorphism. Further reduction of diversity can be caused by demographic events such as population expansions, migrations and bottlenecks. All these processes have operated in the shaping of present Y chromosome diversity and are reflected in high frequencies of particular Y lineages in most human populations (Jobling and Tyler-Smith, 2003). In fact, genetic diversity on the Y chromosome seems to be more deeply structured by lineage than by population, due to the fact that mutations defining Y haplogroups are probably older than the populations they are found in and no recombination occurs between haplogroups (Bosch *et al.*, 1999). As the authors point out, a population may be better understood as an association of old genealogical units (lineages) rather than as a complete evolutionary unit.

The X chromosome is less prone to diversity reduction by genetic drift and presents less geographical structuring than the Y chromosome, although the scarcity of well resolved gene phylogenies hinders the direct comparison of these systems. Considering that the X chromosome is more sensitive to the female history since it is inherited twice as often from females as it is from males, the comparison of Y chromosome with a female specific marker system such as the mtDNA, characterized by equivalent properties as an evolutionary marker to the male specific system, is useful to contrast the male and female genetic backgrounds in the populations under study. The populations sampled in the present study are characterized by a high frequency of specific male and female lineages of different relative antiquity.

1.1.1. Male and female lineages in Mozambique

In Mozambique the E3a sub-clade reaches 73.5% (Pereira *et al.*, 2002) which is characterized by recent strong founder effects since a high percentage of E3a Y chromosomes carry the M191 mutation are thought to have been dispersed within Africa with the expansion of Bantu farmers in the last 3, 000 years (Underhill *et al.*, 2001; Cruciani *et al.*, 2002). Deeply divergent Y chromosomes representing older African lineages, later replaced by the Bantu arrival, are also represented in

Mozambique, although in much lower frequencies (8.8% of AB* lineages; Pereira *et al.*, 2001), as well as other widespread clade E African lineages (11.8% of E*(xE3a)). In sum, two highly frequent haplotypes, as defined by binary markers, comprise almost three quarters of Mozambique Y chromosome gene pool and this is also reflected in low levels of microsatellite haplotype diversity, with the putative Bantu founder haplotype and its one step neighbours representing ~28% of the haplotypes defined with commonly used Y STRs (DYS19/390/391/392/393; Pereira *et al.*, 2002). Parallel characterization of female lineages in Mozambique (Pereira *et al.*, 2001; Salas *et al.*, 2002) revealed the existence of a high percentage of mtDNA haplogroups implicated in the Bantu expansion since the sum of L1a, L2a, L3b and L3e1a lineages comprises ~59% of the sample of Pereira *et al.* (2001) and ~53% of the sample of Salas *et al.* (2002). Remarkably, although there is evidence for strong founder effects in some lineages, substantial diversity of female lineages has been maintained in this Southeastern Bantu population, with most of the Sub-Saharan mtDNA haplogroups being represented (Salas *et al.*, 2002). A reduced frequency (5-7%) of Khoisan mtDNAs is also present in Mozambique (Pereira *et al.*, 2001; Salas *et al.*, 2002) and is interpreted as being the result of maternal gene flow, subsequent to the Bantu arrival (Salas *et al.*, 2002). Several factors might have contributed to the more dramatic diversity reduction of male lineages compared to the female ones in populations in the path of the Bantu expansion, such as male-biased gene flow, contributing to a more severe replacement of ancestral Y chromosome lineages (Hammer *et al.*, 2001), together with the retention of a higher number of ancestral female lineages by higher levels of assimilation en route on the female side or larger female effective population size in the dispersing groups (Salas *et al.*, 2002). Recently, a complex model has been proposed according to which the existence of asymmetric male and female gene flow between hunter-gatherers and Bantu food producers (pastoralists and agriculturalists), as well as high levels of polygyny and patrilocality (female migration) practiced among Bantu farmers has contributed to more effective genetic drift on paternal lineages of these populations (Destro-Bisol *et al.*, 2004). This model has been supported by further analysis of African genetic and linguistic diversity (Wood *et al.*, 2005).

1.1.2. Male and female lineages in Portugal

In Portugal R1b3*(xR1b3f), an old European lineage which probably spread from the Iberian refugia after the LGM, during the Late Upper Paleolithic (Semino *et al.*, 2000; Cinnioğlu *et al.*, 2004), displays an average frequency of ~58% (Pereira *et al.*, 2000; Article 4). A less frequent haplogroup I*(xI1b2) (Rootsi *et al.*, 2004) which is

present in approximately 6% of the Portuguese Y chromosomes (Article 4) has a similar origin. Age estimates obtained from Y STR variation associated with R1b3*(xR1b3f) and I*(xI1b2) chromosomes in Portugal point to a pre-Neolithic age for these lineages (Article 4). Lineages of post-Paleolithic origin (haplogroups J, G and E) comprise ~17% of the Portuguese Y gene pool, and other minor components brought into the territory in historical times sum ~10% of the chromosomes (Article 4). These observations have to some extent a parallel at the level of female lineages found in Portugal (Pereira *et al.*, 2000, 2005a,b), with the presence of the highly frequent haplogroup H (41%), which comprises two relatively frequent European lineages, H1 (~19%) and H3 (~8%), and the less frequent haplogroups K (~6%), T* (~7%) and V (~5%), all expanded during the Late Upper Paleolithic, following the LGM (Torroni *et al.*, 1998, 2001; Richards *et al.*, 1998, 2000; Pereira *et al.*, 2005a). The Neolithic component (haplogroups J, T1 and U3; Richards *et al.*, 1998, 2000) as well as other more recent introgressions (North African and Sub-Saharan), reaches approximately 10% of the Portuguese mtDNA pool (Pereira *et al.*, 2000, 2005b). Middle Upper Paleolithic haplogroups (H*, I, U2 and U4; Richards *et al.*, 2000; Pereira *et al.*, 2005b) reach 12.8% in Portugal and female lineages dating to the Early Upper Paleolithic are represented by haplogroup U and U5 (Richards *et al.*, 1998, 2000), which comprises ~10% of the mtDNAs (Pereira *et al.*, 2000, 2005b). In sum, although the major components of both male and female lineages in the Portuguese population are of post-glacial Upper Paleolithic origin, the female gene pool has kept a small proportion of older pre-LGM lineages, while the signal of such ancestral lineages has been erased from present Portuguese Y chromosomes by the recent expansion of more derived chromosomes.

1.2. Comparison of X- and Y-linked microsatellite variation

The evolutionary dynamics of microsatellites is quite complex and although some mutational models have been advanced (Ohta & Kimura, 1973; Di Rienzo *et al.*, 1994) this issue is still hotly debated (Gonser *et al.*, 2000). As for the role of recombination in this process, some authors claim there is no correlation between mutation rate and recombination rate in autosomal microsatellites, although not discarding the possibility of unequal crossing over contributing to large jumps in repeat length (Huang *et al.*, 2002). Moreover, there is some evidence for an association between recombination hotspots and long (GT)_n repeats (Majewski & Ott, 2000) as well as with minisatellites (Jeffreys *et al.*, 1998).

As for a possible male mutation bias in microsatellites, contradictory data has been advanced. While some studies report higher mutation rates in the male germline (Brinkmann *et al.*, 1998; Ellegren, 2000), others found no difference between sexes (Huang *et al.*, 2002). Moreover, similar mutation rates for autosomal and Y chromosome microsatellites (Zhivotovsky *et al.*, 2004) have been described.

Homologous microsatellites outside the pairing regions of sex chromosomes are differently exposed to recombination and the analysis of these markers in populations can give us some insights on the contribution of this phenomenon to generate diversity. In fact, the shape of allele frequency distributions of microsatellites reflects the mode by which new variants are created. Moreover, comparison of diversity indices in these regions represents an indirect approach to compare male and female mutation rates.

Although a few X/Y markers have been described (Scozzari *et al.*, 1997; Karafet *et al.*, 1998; Carvalho-Silva *et al.*, 2000; Dupuy *et al.*, 2000) most display different repeat structure between homologues. Only two pairs of the homologous *loci* described present an uninterrupted simple repeat structure on both sex chromosomes: the Yq11-Xp22 dinucleotide repeats, DY413-DXS8174/5, which have been diverging for the past ~20 Mya (Malaspina *et al.* 1997; Scozzari *et al.* 1997) and the pentanucleotide DXYS156, located in the same regions analyzed in the present study, Xq21.3/Yp11.2 (Karafet *et al.* 1998).

In the study of Scozzari and co-workers (1997) haplotype diversity was calculated for DY413, since this system comprises two polymorphic *loci*, and compared to heterozygosity values for each X-linked locus (DXS8174 and DXS8175). Thus, although overall higher diversity was found for Y-linked repeats than for the X-linked ones, this comparison is not informative since the Y-linked diversity is likely overestimated. The results of Karafet *et al.* (1998) are in agreement with those obtained by us in the same region, showing higher diversity for DXYS156X compared to DXYS156Y, both in Africa and in Europe. These results point to higher diversity for X chromosome microsatellites compared to their Y chromosome counterparts.

The generally low Y chromosome diversity in Africa, compared to other genetic systems, is well documented and can explain to a great extent the results obtained by Karafet *et al.* (1998) and in the present study (Article 3), for comparisons of Xq21.3/Yp11.2 microsatellite allele distributions and haplotypes in Mozambique.

The lower levels of Y-linked diversity obtained in the Portuguese population should be interpreted taking into account the composition of maternal and paternal

lineages in this population, concerning their relative antiquity, a factor that determines the opportunity for microsatellite mutations to occur and diversity to accumulate. Therefore, the fact that the paternal and the maternal gene pools in the Portuguese population are characterized by similar components in comparable frequencies is expected not to contribute to inequalities of X and Y-linked microsatellite diversity. The existence of a small percentage of older female lineages could be reflected in higher X-linked diversity, although this effect should be weak since the X chromosome has a miscellaneous history. However, although the mtDNA analysis contributes information on the composition of the population on the female perspective, to which the history of X chromosomes is more related, this marker system shares many of the properties of the Y chromosome such as a reduced effective population size and absence of recombination. The parallel between mtDNA and X chromosome markers is thus limited since the latter are expected to be more diverse than either of the monoparental systems and to remain informative at greater time-depths.

The contribution of older female lineages to higher levels of diversity on the X chromosome would be detected by a higher number of alleles in the microsatellite allele distribution and by the existence of relatively frequent step neighbours of a putative founder allele, since there would have been more time for mutation to occur. This effect is clear in microsatellite frequency distributions within each of the *PCDHY* lineages in Portugal, with higher variance of repeat length associated with the ancestral haplotype (TG), but is harder to be demonstrated within the different *PCDHX* lineages, likely due to the occurrence of recombination between markers (data not shown).

The greater diversity of X-linked repeats in the Portuguese population, although less pronounced than in Mozambique, is still approximately twice of that found for the Y-linked microsatellites. However this disparity is mainly due to the fact that three of the repeats analysed are monomorphic on the Y, while presenting moderate diversity on the X chromosome. In fact, the ratio of average X/Y diversity, or of average variance, across the repeats that show variation on both chromosomes is approximately 1. These observations do not seem to reflect greater time for the accumulation of diversity on the X chromosome, specially having in mind that these markers are monomorphic in both populations, Portugal and Mozambique.

Alternatively, the threshold of repeat length from which mutation rate is higher than genetic drift might not have been reached on the Y, although this hypothesis seems only plausible in the case of (GT)_n1 repeat, where Y and X-linked alleles are four steps apart. This marker is also the best candidate for having a contribution of recombination to the generation of diversity on the X chromosome, taking into account

its ragged distribution in both populations (Article 3) and within different SNP-defined backgrounds (data not shown), and the stark contrast with its Y counterpart.

In conclusion, the Portuguese population represents a reasonably good model to compare the diversity of X- and Y-linked markers and points to higher diversity for *PCDH11X* microsatellites, a result that does not seem to relate to a major heterogeneity between the male and the female backgrounds. More likely, the stronger effect of genetic drift on the Y chromosome might be disproportionately erasing Y-linked microsatellites diversity, compared to their X-linked counterparts.

As a future perspective, considering that population diversity surveys are difficult to interpret since male and female gene pools may not share the same history, the analysis of diversity within X and Y chromosome lineages defined by biallelic markers should be performed. Applying this strategy, the effects of the structuring of short tandem repeat variation on the Y chromosome by lineage (Bosch *et al.*, 1999) and allele-specific mutation rates (Brinkmann *et al.*, 1998; Dupuy *et al.*, 2004) are taken into account. Moreover, the confounding effects of demographic events such as recent admixture of different genetic backgrounds within populations (Gusmão *et al.*, 2003; Kittler *et al.*, 2003) would be removed from the analysis since in this way we would be able to focus on the evolution of each founder allele within a lineage and to directly compare lineages from both chromosomes. The challenge will be to establish X chromosome phylogenies, using markers within blocks of high linkage disequilibrium within the model region on Xq21.3 that may parallel the resolution of the Y chromosome phylogeny. The determination of the mutation rate of the X-linked markers within this region in males and females and of their homologues on Yp11.2 in males would also provide direct evidence for the contribution of each sex chromosome in the generation of diversity.

1.3. Comparison of African *versus* European X/Y microsatellite diversity

Although most genetic systems attribute greater diversity to African populations compared to European ones, Y chromosome polymorphisms have not always agreed with this pattern. For instance Hammer *et al.* (1997) have found greater haplotype diversity in Africa, using a marker that is highly frequent in African populations, the YAP insertion. Discordant results were obtained with the analysis of less biased sequence polymorphisms (Hammer *et al.*, 2001) since, although African populations exhibited a higher level of mean pairwise differences between haplotypes, higher levels of heterozygosity were found in Central Asia. Ascertainment bias is indeed a major concern when performing this kind of comparisons. Microsatellites, due to their higher

mutation rates, are less subject to this bias (Rogers and Jorde, 1996) and are more adequate when comparing African and European relative diversity. Even so, different results have been reported using Y chromosome microsatellites. While Seielstad *et al.* (1999) reported higher average gene diversity in Africa compared to the rest of the world, data from Pritchard *et al.* (1999) failed to show a clear excess of African variability since, although variance in repeat number was higher in this continent, the observed heterozygosity was lower. In the study by Jorde *et al.* (2000) the highest average repeat variance, as well as the highest gene diversity, was present in African populations. As pointed out by Kayser *et al.* (2001), who found lower haplotype diversity but higher average gene diversity in Africa, the number and origin of the African populations sampled may be crucial in attributing higher diversity to this continent.

A discordant picture also emerges when comparing African and European relative diversity recurring to microsatellites on the sex chromosomes, with X-linked markers attributing higher diversity to African populations, while Y-linked markers are more diverse in European populations (Scozzari *et al.* 1997 and the present study). By contrast, DXYS156 repeats were always more diverse in Africa than in Europe (Karafet *et al.* 1998). This discrepancy is not easily explained since several African populations were sampled in two of these studies (Scozzari *et al.* 1997; Karafet *et al.* 1998), including not only populations with expected low Y chromosome variation (Bantu-speaking), but also populations with considerable Y chromosome diversity (Khoisan). The fact that in the present study only X chromosome markers attributed higher diversity to the African population is widely explained by the little age and reduced diversity of the Bantu Y chromosomes sampled (Gusmão *et al.*, 2001; Pereira *et al.*, 2002) compared to the more diverse Portuguese Y chromosomes which comprise one most frequent relatively old lineage as well as several other minor components.

2. *PCDH11Y* haplotypes within the Y chromosome phylogeny

The age estimate of *PCDH11Y* haplotypes carrying the derived allelic state at both SNPs (GT haplotype) using microsatellite variation within the gene, pointed to a recent post-Out-of-Africa origin for this haplotype (7,500 to 20,000 years; Article 3). This led us to propose that these mutations were probably more recent than the M173 mutation (~30,000; Semino *et al.*, 2000).

In a larger Portuguese sample (Article 4), only R1a* and R1b3* chromosomes harboured the derived state at both *PCDH11Y* SNPs (GT haplotype) and thus, according

to our results, these mutations could be phylogenetically equivalent to either M173 or M207 (Figure II.1).

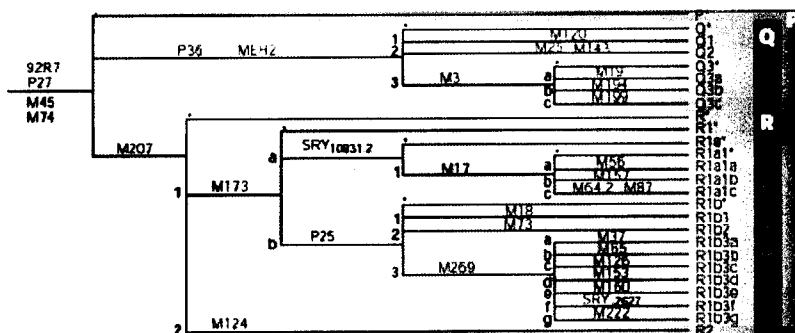


Figure II.1 – Y chromosome lineages carrying the derived allele at M9 position which defines clades P, Q and R of the Y chromosome phylogeny (YCC 2003 tree; adapted from Jobling and Tyler-Smith, 2003).

A DHPLC analysis of these 2 SNPs in representative samples of all haplogroups in the Y chromosome phylogeny showed the presence of both mutations in all samples of haplogroup R subclades and their absence in all samples of haplogroups A-P and Q (P. Underhill, personal communication). Therefore, the two *PCDHY*-exon 5 SNPs lie at the root of haplogroup R and are phylogenetically equivalent to M207. Recently, similar results have been published by Durand *et al.* (2006).

No Y chromosomes were found bearing an intermediate state at these SNPs (corresponding to TT and GG haplotypes) in either Portuguese or Mozambican samples (Articles 3 and 4) or in the multi-population representative samples analysed by DHPLC (P. Underhill, personal communication). Hence, it is possible that both mutations have occurred on the same ancestral Y chromosome or, alternatively, low frequency intermediate chromosomes may exist in populations with representation of haplogroups A-P and Q.

3. Diversity patterns within *PCDH11X*

Giouzeli *et al.* (2004) have performed an extensive survey within the approximately 4 Kb of *PCDH11X* coding sequence, in individuals of mainly European descent and reported the existence of three low frequency polymorphisms (rs4252202, rs4252205 and rs4252208). The haplotype analysis of these polymorphisms and of one

additional intronic SNP, in one African and one European population, revealed the presence of a higher number of chromosomal variants outside Africa, including one putative recombinant (Article 3). These findings most certainly reflect an ascertainment bias in the analysed polymorphisms and this effect was evident when comparing microsatellite diversity within the most frequent common haplotypes between the two populations, since it was possible to infer African ancestry of all haplotypes. Therefore, in contrast to the sharp differentiation of *PCDH11Y* haplotypes between African and European populations, the most frequent *PCDH11X* haplotypes are shared between the two populations and have originated in Africa. The reduction of microsatellite diversity in Europe, of which the shift from a bimodal to a unimodal distribution at the Ca_n repeat is particularly illustrative, could relate to the occurrence of a bottleneck on the origin of non-African populations, as recently detected on genome-wide SNP data (Marth *et al.*, 2003; Gabriel *et al.*, 2002).

In view of the scarcity of described polymorphisms within *PCDH11X*, most population-specific, and considering that this gene spans approximately 800 Kb, it is crucial to analyse further variation in this region in order to obtain a more detailed picture. This was attempted by retrieving and analysing the information available from whole-genome SNP discovery and genotyping projects (unpublished results).

The HapMap Project (International HapMap Consortium, 2003) is a public resource of genome-wide genetic variation (<http://www.hapmap.org>) which, with the completion of Phase I (International HapMap Consortium, 2005), has achieved a resolution of one SNP every 5 Kb across 270 individuals from four geographical populations: 30 mother-father-offspring trios from the Yoruba people from Ibadan Peninsula in Nigeria (referred to as YRI), 30 trios from the CEPH (Centre d'Etude du Polymorphisme Humaine) project in Utah (CEU), 45 unrelated individuals from the Han Chinese population of Beijing (HCB) and 44 unrelated individuals of Japanese ancestry from the Tokyo area (JPT). The inclusion of parent-offspring trios and the use of PHASE algorithm (Stephens and Donnelly, 2003) allowed a remarkably accurate statistical reconstruction of long-range haplotypes for each individual, which are available for download at the Project website. Since Phase I was targeted to common variation in the human genome, only SNPs polymorphic across the four populations and displaying a minor allele frequency of 0.05 or greater were selected for genotyping, both from the public SNP database (dbSNP) and from newly discovered variants by HapMap and Perlegen Sciences' projects (Hinds *et al.*, 2005). A choice of markers without complete resequencing of a number of chromosomes from different human populations, as aimed in the Phase II of the project, implies a certain degree of ascertainment bias.

The project includes data on *PCDH11X*, although the SNP resolution is lower than the autosome average. As a first approach to characterize the diversity patterns within the gene, it is important to evaluate if this region stands out in the genome-wide distribution in terms of specific summary statistics, such as the allele frequency spectrum, since an excess of low frequency variants could indicate the action of positive selection. The genomic region containing *PCDH11X* does not seem to be an outlier in the context of global X chromosome variation and thus it does not present evidence of a deviation from neutral expectations (Figure II.2; S. Schaffner, personal communication).

Although applying a Tajima's D test to the variation in the coding region of *PCDH11X* Giouzeli *et al.* (2004) obtained a negative value ($D = -1.218$), this does not exclude the possibility of neutral evolution. On one hand the result was non-significant at a confidence limit of 90%; moreover, the referred statistic is sensitive to both selective pressures and changes in population size such that negative values, caused by an excess of rare sequences, might reflect the action of positive selection or population growth (Tajima, 1989a and b). All the coding polymorphisms found are present at very low frequencies in European populations (~1%, 1-2% and ~2%, respectively; Article 3 and Giouzeli *et al.*, 2004) and two (rs4252205 and rs4252208) were not found in a sample of 112 African X chromosomes (Article 3), indicating their recent origin. Therefore it might be that the negative Tajima's D based on the frequency spectrum of these variants is a signal of demographic changes within Europe, such as population expansion, and not of diversity-reducing effects related to a selective sweep within *PCDH11X*.

Variation in this locus thus seems to be suitable for general inferences on human diversity. Genotype data and phased *PCDH11X* haplotypes for 4 populations from three continents were retrieved from the HapMap website. Because the HCB and JPT allele frequencies are generally very similar the two datasets were combined in several analyses. To avoid the confounding effects of contamination of the X-linked polymorphism data with fixed differences between X and Y chromosomes sequence, which apparently was not totally eliminated in the HapMap dataset, only polymorphisms encompassing exons 7 and 8 of *PCDH11X*, a region that has been deleted from the Y chromosome and is thus X-specific, were considered.

Genotype data were used to obtain linkage disequilibrium plots and to define haplotype blocks with the algorithm of Gabriel *et al.* (2002), using Haploview 3.2 software (Barret *et al.*, 2005). Only those SNPs for which allele frequencies had been determined in all populations and included in the phased haplotypes were considered

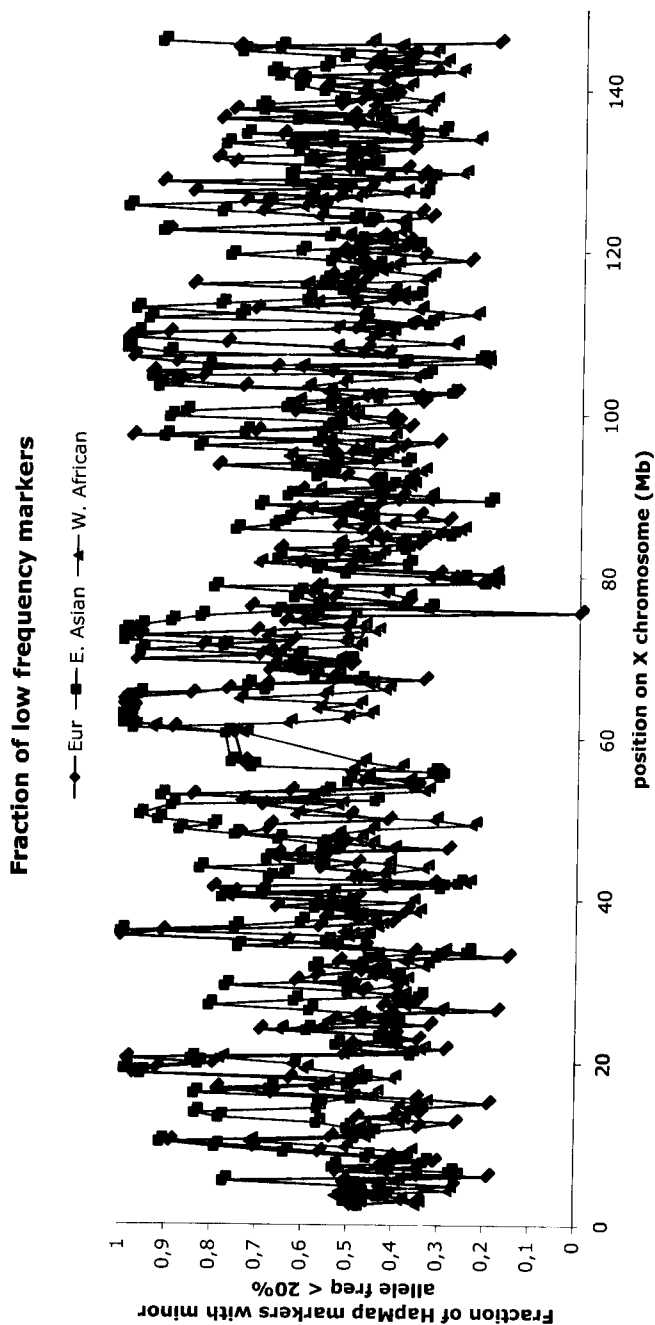


Figure II.2 - Fraction of markers on the X chromosome, from the HapMap project, that have a minor allele frequency that is smaller than 20%, in each population indicated. A region with a high value for this statistic (ex: around 20 Mb and 60-65 Mb) would be a candidate for evaluate for the action of positive selection. *PCDH11X* is situated at approximately 90 Mb. (Courtesy of S. Schaffner)

for further analysis, making a total of 50 SNPs. The ancestral allele at each SPN was determined by comparison with the corresponding position in chimpanzee by a blast search against the draft chimpanzee genome sequence (available at <http://genome.ucsc.edu/>). When this information was not available, the most frequent allele shared by African and non-African populations and/or associated with a larger number of haplotypes was considered to be the older and thus the ancestral. Phased haplotypes, directly downloaded from the HapMap website were used to determine several diversity indices in Arlequin 2.000 (Schneider *et al.*, 2000).

Within this region, and in contrast with the results previously obtained with a reduced number of markers throughout the gene (Article 3), the African population (YRI) displayed the highest number of haplotypes (41 versus 33 in HCB, 26 in JPT and 13 in CEU) and of pairwise differences between haplotypes and consequently higher haplotype diversity (0.968 versus 0.927 in HCB, 0.935 in JPT and 0.718 in CEU). The higher number of segregating sites found in HCB compared to the YRI population (50 versus 45) might relate to a yet not ideal ascertainment of markers in the present dataset. Reflecting a common ancestry and/or gene flow, the two populations with higher level of haplotype-sharing are Europeans (CEU) and Asians (HCB+JPT) with 6 haplotypes in common, followed by Africans (YRI) and Asians sharing 4 haplotypes and finally by Europeans and Africans who share only 3 haplotypes. A high number of haplotypes found within this region (92) is suggestive of reshuffling of alleles by recombination, which is not surprising since the markers analysed span over 100 Kb. The fact that only three haplotypes are shared between continents and an abundance of population-specific low frequency haplotypes might thus reflect the existence of many geographically restricted recombinants. Such extensive recombination obscures the relationship between haplotypes and hinders a clear definition of gene genealogy.

Recent reports have highlighted the heterogeneity of the recombination rate across the genome and have described the existence of sizeable regions where there is little evidence for historical recombination and where a reduced haplotype diversity is observed (Daly *et al.*, 2001; Gabriel *et al.*, 2002; Hinds *et al.*, 2005). These regions are referred to as haplotype blocks and, while recombination events are expected to have occurred in the intervals between blocks, little or no exchange was detected within blocks.

In order to restrict the number of recombination events and to allow a better reconstruction of the haplotypes' phylogeny, a ~29 Kb region starting at 33,093 bp downstream *PCDH11X* of exon 7 was analysed, which is contained within a single haplotype block common to all the populations (Figure II.3).

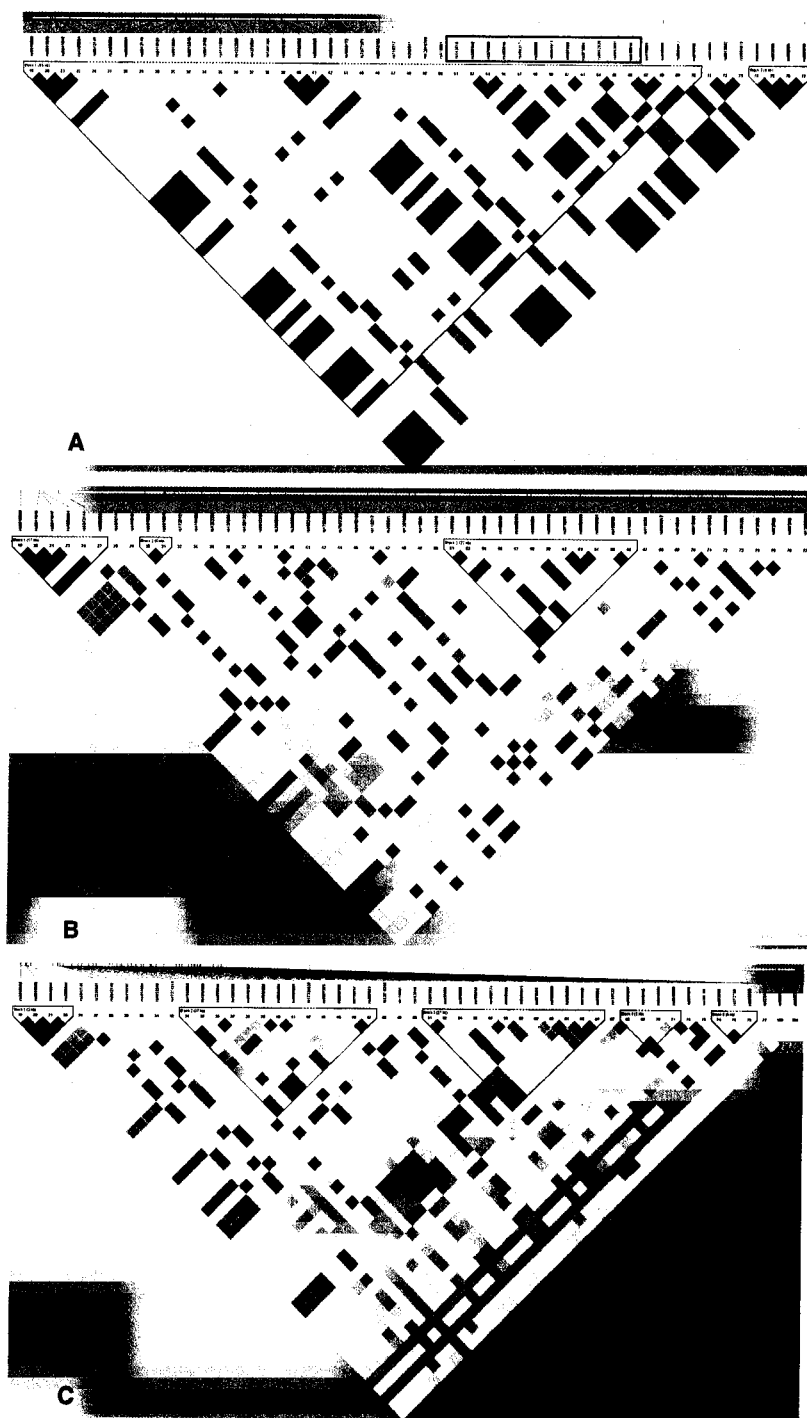


Figure II.3 (Legend in the next page)

Legend of **Figure II.3** - LD plot of r^2 values within the X-specific region comprising exons 7 and 8 of *PCDH11X*, in CEU (A), HCB (B) and YRI (C), generated in Haploview 3.2 with genotype data directly downloaded from the HapMap website. The colour scheme represents weak and strong association between pairs of markers (white, $r^2=0$; black, $r^2=1$; shades of grey, $0 < r^2 < 1$); unfilled areas represent markers that are not polymorphic in each population. Haplotype blocks, represented by black lined triangles, were defined with the algorithm of Gabriel *et al.* (2002). The 29 Kb region analysed is contained a single block in all populations and the markers analysed are inside a black lined square in the CEU population.

In Africans and Asians all the analysed positions within this region are polymorphic, while in Europe there are only ten segregating sites (Table II.1). Ten different haplotypes were found in Africans (YRI), seven in Asians (HCB+JPT) and three in Europeans (CEU).

Table II.1 - Haplotype frequency in a 29 Kb region within *PCDH11X*, across three analysis panels, determined from HapMap phased haplotypes.

	Markers ^a												Haplotype frequency ^b		
	rs443810	rs37902	rs370928	rs453810	rs4893174	rs382096	rs5984915	rs397580	rs371681	rs6652628	rs6652631	rs419619	YRI (n=90)	HCB + JPT (n=134)	CEU (n=90)
►	G	C	T	C	G	T	A	G	C	T	A	G			
H1	G	T	T	T	G	T	A	G	T	A	A	G	0.09	0.19	-
H2	G	C	G	C	C	T	A	G	T	A	A	G	0.01	-	-
H3	G	T	T	C	G	C	A	G	T	A	A	G	0.08	0.25	-
H4	G	C	T	T	G	T	A	G	T	A	A	G	0.01	-	-
H5	G	T	T	C	G	T	A	G	T	A	A	G	0.04	0.21	0.30
H6	G	C	T	C	C	T	A	C	C	T	A	G	0.43	0.22	0.03
H7	G	C	T	C	C	T	A	G	T	A	A	G	0.06	-	-
H8	A	C	G	C	C	T	G	C	T	A	C	A	0.13	0.13	0.67
H9	A	C	T	C	C	T	A	C	T	A	C	A	0.13	-	-
H10	A	C	G	C	C	T	A	C	T	A	C	A	0.01	-	-
H11	G	C	G	C	C	T	G	C	T	A	C	A	-	0.01	-
H12	G	C	G	C	C	T	A	C	C	T	A	G	-	0.01	-

^a NCBI IDs are indicated

^b n= number of chromosomes

► bases shown after this symbol are the putative ancestral allele

To better analyse the evolutionary relationships between all haplotypes in the sample, a Median-Joining (Bandelt *et al.*, 1999) network (Figure II.4) was generated in Network 4.112 (www.fluxus-engineering.com). The haplotype more closely related to the putative ancestral (H6) is the most frequent in Africa, comprising 43% of the

chromosomes, and is present in all populations. The three most frequent haplotypes in the global sample (H5, H6 and H8), comprising 70% of all chromosomes analysed, are common to all populations and are generally the most frequent in each population. While H5 and H6 derive by a reduced number of mutations from the ancestral (3 and 2 mutations, respectively), H8 is highly diverged, with nine different positions from the ancestral.

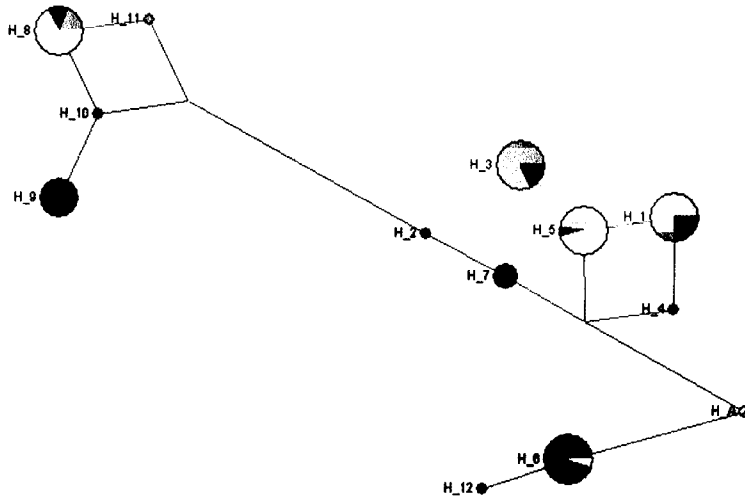


Figure II.4 - Median-Joining network of *PCDH11X* minimal haplotypes in three analysis panels (YRI - African; HCB+JPT - Asian; CEU - European). The percentage of circle area filled with a specific colour is proportional to the frequency of that haplotype in the corresponding population (black, YRI; grey, HCB+JPT; white, CEU) and branch length is proportional to the number of mutations. The putative ancestral haplotype (H_A) is not present in the sample.

Two relatively frequent haplotypes (H1 and H3) are shared between the African and Asian populations and are absent from Europe. The YRI population is the one presenting a higher number of private haplotypes (5 haplotypes), followed by the Asian population which harbours two low frequency haplotypes restricted to this continent. Asian- and African-specific haplotypes are closely related to the most frequent and globally dispersed ones. Surprisingly, having in consideration that the SNPs analysed are restricted to a single haplotype block in all populations, the Median-Joining network displays a certain degree of reticulation, which can be due to either recurrent mutation or recombination events. Since recurrent mutation is apparently rare in the HapMap dataset (International HapMap Consortium, 2005) it seemed reasonable

to integrate the possibility of historical recombination in the analysis of this group of haplotypes.

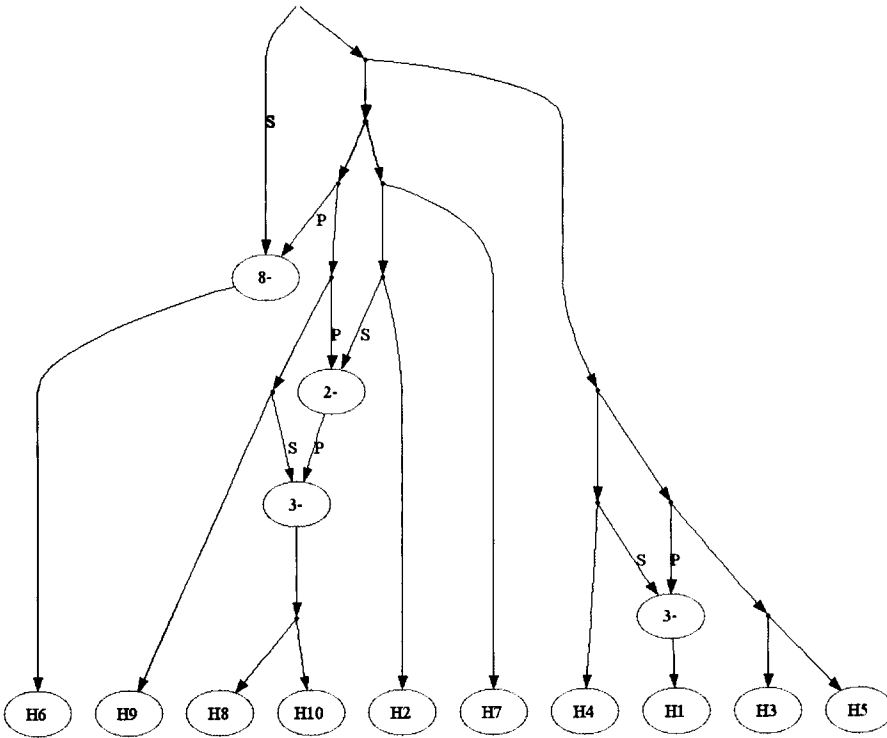


Figure II.5 - Recombination map for the twelve haplotypes found in the YRI sample. Each blue circle represents a recombination event at a specific site (corresponding to marker order in Table II.1) between sequence P (prefix) and S (suffix) from which H6, H8 and H10, and H1 have originated. The minimum number of recombinations in this dataset is four, since H8 and H10 are double recombinants. Green dots represent coalescence of two different sequences in one ancestral chromosome and the yellow dot represents the ancestral sequence that, although not present in the sample, was assumed as known. Although mutations have occurred in some branches, branch length is not proportional to the number of mutational steps.

For this purpose the branch and bound method implemented in Beagle (Lyngsø *et al.*, 2005) was used to determine the exact minimum number of recombinations required for each continental dataset. The programme generates a sequence of evolutionary events with a minimum number of recombinations and outputs an ancestral recombination map, as depicted in Figure II.5 for the YRI haplotypes assuming the ancestral as known, where a minimum of four historical recombination events must be considered to explain currently observed haplotype diversity.

In agreement with the homoplasy patterns observed in the Median-Joining network haplotypes H1, H8 and H10 have been originated by single (H1) or double (H8 and H10) recombination events. Some of these events might be relatively old, since not all variants involved in sequence exchange are represented in the extant diversity. Although differing from the ancestral by only two mutations, H6 has apparently been derived from a quite ancient recombination. In the CEU and HCB+JPT datasets a single recombination event must be evoked (data not shown), which in both cases has also been detected in the YRI population and thus most probably reflects the presence of an ancient African recombinant in the non-African set. The different groups of African chromosomes are moderately divergent and some coalesce very close to the root of the phylogeny.

From the analyses of extended and "minimal" haplotypes within *PCDH11X* greater diversity was found in the African population and is clear that non-African chromosomes, including recombinants, are a subset of the African variants, in this dataset. These observations are consistent with the results obtained at the genome level with a large number of SNPs (Gabriel *et al.*, 2002; Hinds *et al.*, 2005) and are in support of an African origin of the bulk of present day human diversity in the three continents sampled.

However, in view of the recent concerns on the influence of marker ascertainment, spacing and selection on haplotype block definition and observed haplotypic patterns (Phillips *et al.*, 2003; Nothnagel and Rohde, 2005) the present analysis can only be regarded as preliminary until higher marker densities are achieved and ascertainment bias is negligible.

CONCLUDING REMARKS

Analysis of single locus or non-recombining units might be misleading especially in the case of such a particular genetic system as the Y chromosome. The presented data pictures two very different scenarios for male-specific and X chromosome diversity patterns, within a region of high sequence homology on the sex chromosomes. The merging of male and female histories on extant X chromosomes confers greater informativeness to X-linked loci in diversity studies due to the maintenance of a broader spectrum of variants on the population, compared to the Y-linked counterparts. As allele frequencies of polymorphisms on the X chromosome are less influenced by stochastic fluctuations they are thus more reliable on deriving global inferences on human diversity levels in different populations. This effect is apparent in different types of markers, such as single nucleotide polymorphisms and microsatellites, which result from different mutational processes.

The results obtained from Xq21.3 diversity are compatible with the general assumptions of the "Out of Africa" model of human evolution. The contrasting patterns of evolution within the homologous region on Yp do not necessarily defy the referred model, if the specificities inherent to Y-linked markers are properly accounted for.

REFERENCES

- Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P; International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437(7063):1299-320.
- Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16:37-48.
- Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. *Proc Natl Acad Sci U S A* 94(9):4516-9.
- Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21(2):263-5.
- Beleza S, Gusmao L, Amorim A, Carracedo A, Salas A (2005) The genetic legacy of western Bantu migrations. *Hum Genet* 117(4):366-75.
- Bohossian HB, Skaletsky H, Page DC (2000) Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature* 406(6796):622-5.
- Bosch E, Calafell F, Santos FR, Perez-Lezaun A, Comas D, Benchemsi N, Tyler-Smith C, Bertranpetit J (1999) Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. *Am J Hum Genet* 65:1623-38.
- Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368:455-7.
- Brinkmann B, Klitsch M, Neuhuber F, Huhne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet* 62:1408-15.
- Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK (1998) Short tandem repeat polymorphism evolution in humans. *Eur J Hum Genet* 6(1):38-49.
- Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature* 325(6099):31-6.

Carvalho Silva DR, Pena SD (2000) Molecular characterization and population study of an X chromosome homolog of the Y-linked microsatellite DYS391. *Gene* 247:233-40

Charlesworth B (2001) The effect of life-history and mode of inheritance on neutral genetic variability. *Genet Res* 77:153-166.

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G (1998) Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. *Proc Natl Acad Sci U S A* 95(15):9053-8.

Cinnioglu C, King R, Kivisild T, Kalfoglu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K, Oefner PJ, Shen P, Semino O, Cavalli-Sforza LL, Underhill PA (2004) Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet.* Jan 114(2):127-48.

Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet* 70(5):1197-214.

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

Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Caglia A, Tofanelli S, Spedini G, Capelli C (2004) Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. *Mol Biol Evol* 21(9):1673-82.

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB. (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proc Natl Acad Sci* 91:3166-70.

Dorit RL, Akashi H, Gilbert W (1995) Absence of polymorphism at the ZFY locus on the human Y chromosome. *Science* 268:1183-5.

- Dupanloup I, Pereira L, Bertorelle G, Calafell F, Prata MJ, Amorim A, Barbujani G (2003) A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosomediversity. *J Mol Evol* 57:85-97.
- Dupuy BM, Gedde-Dahl T, Olaisen B (2000) DXYS267: DYS393 and its X chromosome counterpart. *Forensic Sci Int* 112:111-21.
- Dupuy BM, Stenersen M, Egeland T, Olaisen B (2004) Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Hum Mutat* 23:117-24.
- Durand CM, Kappeler C, Betancur C, Delorme R, Quach H, Goubran-Botros H, Melke J, Nygren G, Chabane N, Bellivier F, Szoke A, Schurhoff F, Rastam M, Anckarsater H, Gillberg C, Leboyer M, Bourgeron T (2006) Expression and genetic variability of PCDH11Y, a gene specific to Homo sapiens and candidate for susceptibility to psychiatric disorders. *Am J Med Genet B Neuropsychiatr Genet* 141(1):67-70.
- Ebersberger I, Metzler D, Schwarz C, Paabo S (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet* 70(6):1490-7.
- Ellegren H (2000) Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat Genet* 24:400-2.
- Erlandsson R, Wilson JF, Paabo S (2000) Sex chromosomal transposable element accumulation and male-driven substitutional evolution in humans. *Mol Biol Evol* 17(5):804-12.
- Excoffier L (2002) Human demographic history: refining the recent African origin model. *Curr Opin Genet Dev* 12(6):675-82.
- 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 (2002) The structure of haplotype blocks in the human genome. *Science* 296(5576):2225-9.

Gagneux P, Wills C, Gerloff U, Tautz D, Morin PA, Boesch C, Fruth B, Hohmann G, Ryder OA, Woodruff DS (1999) Mitochondrial sequences show diverse evolutionary histories of African hominoids. *Proc Natl Acad Sci U S A* 96(9):5077-82.

Giouze M, Williams NA, Lonie LJ, DeLisi LE, Crow TJ (2004) ProtocadherinX/Y, a candidate gene-pair for schizophrenia and schizoaffective disorder: a DHPLC investigation of genomic sequence. *Am J Med Genet B Neuropsychiatr Genet* 129(1):1-9.

Goldstein DB, Chikhi L (2002) Human migrations and population structure: what we know and why it matters. *Annu Rev Genomics Hum Genet*;3:129-52.

Gonser R, Donnelly P, Nicholson G, Di Rienzo A (2000) Microsatellite mutations and inferences about human demography. *Genetics* 154:1793-807.

Gusmão L, Alves C, Amorim A (2001) Molecular characteristics of four human Y-specific microsatellites (DYS434, DYD437, DYS438, DYS439) for population and forensic studies. *Ann Hum Genet* 65(Pt 3):285-91.

Gusmão L, Krawczak M, Sanchez-Diz P, Alves C, Lopes A, Beleza S, Carracedo A, Amorim A (2003) Bimodal allele frequency distribution at Y-STR loci DYS392 and DYS438: no evidence for a deviation from the stepwise mutation model. *Int J Legal Med* 117(5):287-90.

Gusmão L, Sanchez-Diz P, Calafell F, Martín P, Alonso CA, Alvarez-Fernandez F, Alves C, Borjas-Fajardo L, Bozzo WR, Bravo ML, Builes JJ, Capilla J, Carvalho M, Castillo C, Catanesi CI, Corach D, Di Leonardo AM, Espinheira R, Fagundes de Carvalho E, Farfan MJ, Figueiredo HP, Gomes I, Lojo MM, Marino M, Pinheiro MF, Pontes ML, Prieto V, Ramos-Luis E, Riancho JA, Souza Goes AC, Santapa OA, Sumita DR, Vallejo G, Vidal Rioja L, Vide MC, Vieira da Silva CI, Whittle MR, Zabala W, Zarrabeitia MT, Alonso A, Carracedo A, Amorim A (2005) Mutation rates at Y chromosome specific microsatellites. *Hum Mutat* 26(6):520-8.

Hammer MF (1995) A recent common ancestry for human Y chromosomes. *Nature* 378(6555):376-8.

- Hammer MF, Spurdle AB, Karafet T, Bonner MR, Wood ET, Novelletto A, Malaspina P, Mitchell RJ, Horai S, Jenkins T, Zegura SL (1997) The geographic distribution of human Y chromosome variation. *Genetics* 145(3):787-805.
- Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, Zegura SL (2001) Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol* 18(7):1189-203.
- Hammer MF, Garrigan D, Wood E, Wilder JA, Mobasher Z, Bigham A, Krenz JG, Nachman MW (2004) Heterogeneous patterns of variation among multiple human X-linked Loci: the possible role of diversity-reducing selection in non-africans. *Genetics* 167(4):1841-53.
- Harding RM, McVean G (2004) A structured ancestral population for the evolution of modern humans. *Curr Opin Genet Dev* 14(6):667-74.
- Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST (1998) Genetic traces of ancient demography. *Proc Natl Acad Sci U S A* 95(4):1961-7.
- Harpending H, Rogers A (2000) Genetic perspectives on human origins and differentiation. *Annu Rev Genomics Hum Genet* 1:361-85.
- Harris EE, Hey J (1999) X chromosome evidence for ancient human histories. *Proc Natl Acad Sci U S A* 96(6):3320-4.
- Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum Mol Genet* 6(5):799-803.
- Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* 307(5712):1072-9.
- Holtkemper U, Rolf B, Hohoff C, Forster P, Brinkmann B (2001) Mutation rates at two human Y-chromosomal microsatellite loci using small pool PCR techniques. *Hum Mol Genet* 10(6):629-33.

Huang QY, Xu FH, Shen H et al. (2002) Mutation patterns at dinucleotide microsatellite loci in humans. *Am J Hum Genet* 70:625-34.

Ingman M, Kaessmann H, Paabo S, Gyllenstein U (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* 408(6813):708-13.

Jeffreys AJ, Murray J, Neumann R (1998) High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol Cell* 2(2):267-73.

Jobling MA, Williams GA, Schiebel GA, Pandya GA, McElreavey GA, Salas GA, Rappold GA, Affara NA, Tyler-Smith C (1998) A selective difference between human Y-chromosomal DNA haplotypes. *Curr Biol* 8(25):1391-4.

Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet* 4(8):598-612.

Jobling MA, Hurles ME, Tyler-Smith C (2004) *Human Evolutionary Genetics: Origins, Peoples and Disease*. (Garland Science, eds) New York, Abingdon.

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am J Hum Genet* 66(3):979-88.

Kaessmann H, Heissig F, von Haeseler A, Paabo S (1999) DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat Genet* 22(1):78-81.

Kaessmann H, Wiebe V, Weiss G, Paabo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat Genet* 27(2):155-6.

Karafet T, de Knijff P, Wood E, Ragland J, Clark A and Hammer MF (1998) Different patterns of variation at the X- and Y-chromosome-linked microsatellite loci DXYS156X and DXYS156Y in human populations. *Hum Biol* 70:979-92.

- Karger B, Francalacci P, Rudan P, Janicijevic B, Rickards O, Savontaus ML, Huoponen K, Laitinen V, Koivumaki S, Sykes B, Hickey E, Novelletto A, Moral P, Sellitto D, Coppa A, Al-Zaheri N, Santachiara-Benerecetti AS, Semino O, Scozzari R (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. *Am J Hum Genet* 69(4):844-52.
- Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, Kruger C, Krawczak M, Nagy M, Dobosz T, Szibor R, de Knijff P, Stoneking M, Sajantila A (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am J Hum Genet* 66(5):1580-8.
- Kayser M, Krawczak M, Excoffier L, Dieltjes P, Corach D, Pascali V, Gehrig C, Bernini LF, Jespersen J, Bakker E, Roewer L, de Knijff P (2001) An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. *Am J Hum Genet* 68(4):990-1018.
- Kimmel M, Chakraborty R, King JP, Bamshad M, Watkins WS, Jorde LB (1998) Signatures of population expansion in microsatellite repeat data. *Genetics* 148(4):1921-30.
- Kittler R, Erler A, Brauer S, Stoneking M, Kayser M (2003) Apparent intrachromosomal exchange on the human Y chromosome explained by population history. *Eur J Hum Genet* 11(4):304-14.
- Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. *Nat Genet* 31(3):241-7.
- Labuda D, Zietkiewicz E, Yotova V (2000) Archaic lineages in the history of modern humans. *Genetics* 156(2):799-808.
- Lyngsø RB, Song YS, Hein J (2005) Minimum Recombination Histories by Branch and Bound In: *Proceedings of Workshop on Algorithms in Bioinformatics, Lecture Notes in Computer Science* 3692, pp 239-250.

Majewski J and Ott J (2000) GT repeats are associated with recombination on human chromosome 22. *Genome Res* 10:1108-14.

Makova KD, Li WH (2002) Strong male-driven evolution of DNA sequences in humans and apes. *Nature* 416(6881):624-6.

Malaspina P, Ciminelli BM, Viggiano L, Jodice C, Cruciani F, Santolamazza P, Sellitto D, Scozzari R, Terrenato L, Rocchi M, Novelletto A (1997) Characterization of a small family (CAIII) of microsatellite-containing sequences with X-Y homology. *J Mol Evol* 44(6):652-9.

Marth G, Schuler G, Yeh R, Davenport R, Agarwala R, Church D, Wheelan S, Baker J, Ward M, Kholodov M, Phan L, Czabarka E, Murvai J, Cutler D, Wooding S, Rogers A, Chakravarti A, Harpending HC, Kwok PY, Sherry ST (2003) Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc Natl Acad Sci U S A* 100(1):376-81.

McVean GT, Hurst LD (1997) Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature* 386(6623):388-92.

Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T (1987) Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb Symp Quant Biol* 52:863-7.

Nothnagel M, Rohde K (2005) The effect of single-nucleotide polymorphism marker selection on patterns of haplotype blocks and haplotype frequency estimates. *Am J Hum Genet* 77(6):988-98.

Ohta T, Kimura K (1973) The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic sample. *Genet Res* 22:201-204.

Pecon Slattery J, Sanner-Wachter L, O'Brien SJ (2000) Novel gene conversion between X-Y homologues located in the nonrecombining region of the Y chromosome in Felidae (Mammalia). *Proc Natl Acad Sci U S A* 97(10):5307-12.

- Pereira L, Prata MJ, Jobling MA, Amorim A (2000) Analysis of the Y-chromosome and Mitochondrial DNA Pools in Portugal. in *Archaeogenetics: DNA and the population prehistory of Europe*. (Renfrew C, Boyle K, eds). McDonald Institute for Archeological Research, University of Cambridge, pp 191-195.
- Pereira L, Macaulay V, Torroni A, Scozzari R, Prata MJ, Amorim A (2001) Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann Hum Genet* 65:439-58.
- Pereira L, Gusmão L, Alves C, Amorim A, Prata MJ (2002) Bantu and European Y-lineages in Sub-Saharan Africa. *Ann Hum Genet* 66:369-78.
- Pereira L, Richards M, Goios A, Alonso A, Albarran C, Garcia O, Behar DM, Golge M, Hatina J, Al-Gazali L, Bradley DG, Macaulay V, Amorim A (2005a) High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. *Genome Res* 15(1):19-24.
- Pereira L, Cunha C, Alves C, Amorim A (2005b) African female heritage in Iberia: a reassessment of mtDNA lineage distribution in present times. *Hum Biol* 77(2):213-29.
- Perez-Lezaun A, Calafell F, Seielstad M, Mateu E, Comas D, Bosch E, Bertranpetit J (1997) Population genetics of Y-chromosome short tandem repeats in humans. *J Mol Evol* 45(3):265-270.
- Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, Ankener WM, Alfisi SV, Kuo FS, Camisa AL, Pazorov V, Scott KE, Carey BJ, Faith J, Katari G, Bhatti HA, Cyr JM, Derohannessian V, Elosua C, Forman AM, Grecco NM, Hock CR, Kuebler JM, Lathrop JA, Mockler MA, Nachtman EP, Restine SL, Varde SA, Hozza MJ, Gelfand CA, Broxholme J, Abecasis GR, Boyce-Jacino MT, Cardon LR (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet* 33(3):382-7.
- Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* 16(12):1791-8.
- Reich DE, Goldstein DB (1998) Genetic evidence for a Paleolithic human population expansion in Africa. *Proc Natl Acad Sci U S A* 95(14):8119-23.

References II

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

Richards M, Corte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, Hedges R, Bandelt HJ, Sykes B (1996) Paleolithic and neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 59(1):185-203.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Norby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt HJ (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* 67(5):1251-76.

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* 9(3):552-69.

Rogers AR, Jorde LB (1996) Ascertainment bias in estimates of average heterozygosity. *Am J Hum Genet* 58(5):1033-41.

Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, Barac L, Pericic M, Balanovsky O, Pshenichnov A, Dion D, Grobei M, Zhivotovsky LA, Battaglia V, Achilli A, Al-Zahery N, Parik J, King R, Cinnioglu C, Khusnutdinova E, Rudan P, Balanovska E, Scheffrahn W, Simonescu M, Brehm A, Goncalves R, Rosa A, Moisan JP, Chaventre A, Ferak V, Furedi S, Oefner PJ, Shen P, Beckman L, Mikerezi I, Terzic R, Primorac D, Cambon-Thomsen A, Krumina A, Torroni A, Underhill PA, Santachiara-Benerecetti AS, Villems R, Semino O (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. *Am J Hum Genet* 75(1):128-37.

Ross MT et al. (2005) The DNA sequence of the human X chromosome. *Nature* 434(7031):325-37.

- Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423(6942):873-6.
- Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D; International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409(6822):928-33.
- Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo A (2002) The making of the African mtDNA landscape. *Am J Hum Genet* 71(5):1082-111.
- Schaffner SF (2004) The X chromosome in population genetics. *Nat Rev Genet* 5(1):43-51.
- Schneider S, Roessli D, Excoffier L (2000) Arlequin ver. 2.000: a software for population genetics data analysis. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Switzerland
- Scozzari R, Cruciani F, Malaspina P, Santolamazza P, Ciminelli BM, Torroni A, Modiano D, Wallace DC, Kidd KK, Olckers A, Moral P, Terrenato L, Akar N, Qamar R, Mansoor A, Mehdi SQ, Meloni G, Vona G, Cole DE, Cai W, Novelletto A (1997) Differential structuring of human populations for homologous X and Y microsatellite loci. *Am J Hum Genet* 61(3):719-33.
- Seielstad M, Bekele E, Ibrahim M, Toure A, Traore M (1999) A view of modern human origins from Y chromosome microsatellite variation. *Genome Res* 9(6):558-67.
- Semino O, Passarino G, Brega A, Fellous M, Santachiara-Benerecetti AS (1996) A view of the neolithic demic diffusion in Europe through two Y chromosome-specific markers. *Am J Hum Genet* 59(4):964-8.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA (2000) The genetic legacy of Paleolithic Homo sapiens in extant Europeans: a Y chromosome perspective. *Science* 290(5494):1155-9.

Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL, Underhill PA (2002) Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am J Hum Genet* 70(1):265-8.

Serre D, Paabo S (2004) Evidence for gradients of human genetic diversity within and among continents. *Genome Res* 14(9):1679-85.

Shen P, Wang F, Underhill PA, Franco C, Yang WH, Roxas A, Sung R, Lin AA, Hyman RW, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ (2000) Population genetic implications from sequence variation in four Y chromosome genes. *Proc Natl Acad Sci U S A* 97(13):7354-9.

Shimmin LC, Chang BH, Li WH (1993) Male-driven evolution of DNA sequences. *Nature* 362(6422):745-7.

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 73(5):1162-9.

Stone AC, Griffiths RC, Zegura SL, Hammer MF (2002) High levels of Y-chromosome nucleotide diversity in the genus Pan. *Proc Natl Acad Sci U S A* 99(1):43-8.

The International HapMap Consortium (2003) The International HapMap Project. *Nature* 426(6968):789-96.

Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc Natl Acad Sci* 97(13):7360-5.

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 271(5254):1380-7.

Torroni A, Bandelt HJ, D'Urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, Forster P, Savontaus ML, Bonne-Tamir B, Scozzari R (1998) mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *Am J Hum Genet* 62(5):1137-52.

Torroni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, Villems R, Kivisild T, Metspalu E, Parik J, Tolk HV, Tambets K, Forster P, Karger B, Francalacci P, Rudan P, Janicijevic B, Rickards O, Savontaus ML, Huoponen K, Laitinen V, Koivumaki S, Sykes B, Hickey E, Novelletto A, Moral P, Sellitto D, Coppa A, Al-Zaheri N, Santachiara-Benerecetti AS, Semino O, Scozzari R (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. *Am J Hum Genet* 69(4):844-52.

Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res* 7(10):996-1005.

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ (2000) Y chromosome sequence variation and the history of human populations. *Nat Genet* 26(3):358-61.

Underhill PA, Passarino G, Lin AA, Shen P, Mirazon Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* 65(Pt 1):43-62.

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. *Science* 253:1503-7.

Wilder JA, Mobasher Z, Hammer MF (2004) Genetic evidence for unequal effective population sizes of human females and males. *Mol Biol Evol* 21(11):2047-57.

Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, Louie L, Bamshad M, Strassmann BI, Soodyall H, Hammer MF (2005) Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur J Hum Genet* 13(7):867-76.

References II

Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res* 12(2):339-48.

Yu N, Chen FC, Ota S, Jorde LB, Pamilo P, Patthy L, Ramsay M, Jenkins T, Shyue SK, Li WH (2002) Larger genetic differences within africans than between Africans and Eurasians. *Genetics* 161(1):269-74.

Zhao Z, Jin L, Fu YX, Ramsay M, Jenkins T, Leskinen E, Pamilo P, Trexler M, Patthy L, Jorde LB, Ramos-Onsins S, Yu N, Li WH (2000) Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proc Natl Acad Sci* 97:11354-8.

Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G, Chambers GK, Herrera RJ, Yong KK, Gresham D, Tournev I, Feldman MW, Kalaydjieva L (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet* 74(1):50-61.