



Department of Experimental and Health Science

Universitat Pompeu Fabra

# Human Genetic Diversity In Genes Related To Host-Pathogen Interactions

Tesi Doctoral

Anna Ferrer i Admetlla

Evolutionary Biology Unit

Experimental and Health Science Department

Pompeu Fabra University





Department of Experimental and Health Science

Universitat Pompeu Fabra

# Human Genetic Diversity In Genes Related To Host-Pathogen Interactions

Memòria presentada per Anna Ferrer i Admetlla per optar al grau de doctora per la Universitat Pompeu Fabra. La present tesi doctoral ha estat realitzada sota la codirecció del Dr Ferran Casals i López (actualment a la Université de Montréal) i del Dr Jaume Bertranpetit i Busquets (Universitat Pompeu Fabra) a la Unitat de Biologia Evolutiva del Departament de Ciències Experimentals i de la Salut de la Universitat Pompeu Fabra, dins del programa de doctorat: *PhD Program in Health and Life Science* (2004-2006).

**Ferran Casals i Lòpez**

**Director**

**Jaume Bertranpetit i Busquets**

**Director**

**Anna Ferrer i Admetlla**

**Doctoranda**









A Tots vosaltres, des de Pares fins a Amic

pel bon rotllo que m'heu transmès durant aquests quatre anys

i, molt especialment, a aquells qui em van obrir les portes del món de la ciència:

l'Arcadi Navarro i l'Anna Barceló



## Agraïments

Tot comença el primer dilluns de juliol del 2003. Aquell dia fa un sol rabiós i l'aprofito baixant a la platja, fa 2 dies que he fet l'últim exàmen de la carrera i en fa només un que he enviat el meu primer currículum, quan torno de la platja em trobo amb un correu electrònic citant-me a una entrevista de feina. L'entrevista en questio és per cobrir la baixa maternal de l'Anna Pérez al Servei de Genòmica de la UPF (visca l'Anna Pérez i, visca la Laia!!; si no fos per ella jo ara potser no estaria aquí). Així, treballant al servei de genòmica, és com vaig coneixent tota la gent que d'una manera tan pròxima i alegre m'acompanyarà al llarg d'aquests quatre anys. Aquesta gent sou vosaltres: els del grup o bioevos (això com vulgeu), els "cegeneros", els del zulo, el servei...

Tornant al relat i anant per ordre conològic he de dir que al servei hi passo uns mesos gloriosos. Amb el Roger m'hi estic la mar de bé (rodolí). Em fa sentir molt còmode, tenim molt bon rotllo. Però val a dir que això és bastant fàcil perquè és algú molt agradable, obert i "xispós" (menys quan acaba de córrer perquè se li estanca la sang a les cames ☺). De fet, entre ell, la María Jiménez, el fricandó matiner i unes actuacions de màgia molt particulars a la porta del servei el temps passa volant. En realitat el temps va passar tan de pressa que el febrer de l'any següent jo diria que va arribar més aviat del normal (!). Va ser aquell febrer quan vaig incorporar-me (en esperit però no en cos, perquè encara no tenia cadira on seure) a Bioevo. A la famosa Bioevo, un aiguabarreig d'interessats en zoologia, biòlegs nedant al mar de l'evolució i enamorats de la bioestadística (si em permeteu dir-ho així). Ah! I la Mònica que constitueix una categoria independent. La categoria de "la cunya", de les històries de la gata i del veí, però també la categoria d'algú que es preocupa per les coses, que escolta, que s'entrega i fa favors sempre que pot. Gràcies Mònica per ser-hi i sobretot per preguntar, perquè de vegades un:- Com estàs?- val tot l'or del món (gràcies pel "com estàs?" d'aquests últims dies!).

(Tornant al relat) Aquesta és l'última època del barri xino, de la feina "més de poyata", del Marmi, l'Olga, el Flopez, l'Ainhoa... (per cert; visca la Patum!), i del català als seminaris de grup. Durant aquest període es constitueix el zulo, arranca el CeGen, els projectes pilot i arriben el C.Morcillo, el Pep, el Ricard i... el Ferran. Pels que no el conegueu; algú encantador! (però això ho reservo pel final). Amb el temps els antics becaris del grup van marxant; primer l'Aida i després l'Óscar, la Marta, el Tomàs... tots ens deixen una mica de la seva essència, si més no en el meu record...: els riures de l'Aida, la paciència i l'entrega de l'Oscar, que humilment sempre ajudava i ajuda a tothom sense queixar-se ni un pel. També penso molt en la Marta, que em va ensenyar a fer servir l'EndNote (que

tan bé m'ha anat durant el període d'escriptura d'aquesta tesi) ah! I també me'n recordo molt de la seva faceta de guardiana del patrinomi "esses sonores" (que sempre em va semblar una tasca útil, i no ho dic amb conya!) i del Tomàs que va ser el meu company i mestre de pràctiques zoològia i que sempre m'acaba contagiats amb la seva empenta i motivació per la ciència.

Durant el període doctor-aiguader-80 visc al fons del laboratori entre la Lourdes i la Kristin, primer, i després en el seu lloc el Martin. Durant aquest temps em vaig fer molt amb la Lourdes, que anyoro molts cops. És una persona fantàstica, molt noble i tranquil·la que no té mai un no per resposta. Realment treballar al seu costat va ser un gust. Una altra persona fantàstica que també va passar per bioevo en aquella època és el Joan Pons, algú extremadament agradable i simpàtic!. D'ell n'aguado, a banda d'alguns consells científics, la recepta del cocarrois que per cert (ehm, ehm...) encara tinc pendents de fer.

Aquella també va ser l'època de les esquíades amb el Jesús, la Glòria, la Monix, el Roger...L'època de l'oro el moro el mico i el senyor de "Puerto Rico", de mesurar les alçades al marc de la porta del despatx David-CeGen i de segrestar "trofeus" de vollei. Arribats a aquest punt esbombaré que el David, que és super-simpàtic (com diria ell), va ser un dels segrestadors.

En aquells moments comencen a arracar els projectes pilot. Ara que hi som vull agrair a l'Elena la seva contribució en la posada a punt d'aquests projectes, la seva actitud franca i la seva voluntat d'ajudar. No em voldria descuidar l'Arcadi, ell va ser la persona que em va persuadir per fer el doctorat a la casa i a ell li vull agrair, a banda d'aquesta oportunitat, la d'ampliar els meus horitzons musicals i l'ajut en algun moment puntual (allò que si necessites alguna cosa exemple una carta... l'Arcadi t'ajuda). El Francesc, també m'ha ajudat durant aquest temps, bé de fet, diria que el Francesc m'ha ensenyat bastantes coses a base d'explicar-me una petita part de tot allò que sap, i jo no (ups!).

Tornant al relat, el temps passa i com aquell qui no vol arribar al moment de traslladar-nos al PRBB. Entorn d'aquest moment (mesos amunt, mesos avall segons el cas) la plantilla del laboratori s'amplia amb becaries, però també algun becari, informàtics i post-docs tots la mar d'oberts i fantàstics. L'Ixa, l'Urko, la Belén i la Marta (que, de fet, són anteriors a la migració al PRBB), el Rui, el Txema, l'Àngel, el Hafid, l'Araceli, l'Òscar R, el Fernando, el Doménech, el Marc, la Begoña, la Laura, el Johannes, la Mireia, la Fleur i la Judit (no m'agradaria deixar-me a ningú). Aquesta és una altra gran època del

laboratori, molta gent, molts riures, els crits aguts i de puntualitat britànica de l'1xa cada dia a 2/4 de 7 de la tarda (que per mi són un misteri... potser són un exorcisme o potser un costum basc (¿), un dia d'aquests li pregunto ☺). Les històries de la Belén i la seva actitud "carinyosa"; Belén et fas estimar.

No voldria oblidar-me de l'Urko que és algú amb extramada bona fe i carregat d'una actitud integradora i idees genials, com el club de "lectura i tertúlia entorn de l'evolució" o el futbol. Tampoc em voldria deixar el Hafid que, també és un company magnífic, sempre disposat a donar un cop de mà. A mi m'ha recolzat molt durant els últims dos mesos i d'ell en puc dir ben clar que és un company amb qui s'hi pot comptar de veïtat (i de qui a més, se n'aprén molt). Ni l'Elodie i el C.Lalueza, gràcies als dos per petites ajudes. No em voldria oblidar de l'Anna Pérez (ja sé que l'he ennomenat abans, però ho torno a fer), l'Araceli, la Judit i la Valeria ja que són les persones que de llarg i, mooolt llarg, s'han preocupat de mi durant la penitència d'escriure la tesi. És recomfortant i necessari que algú et preguntis com et van les coses. Moltíssimes gràcies noies. Us asseguro que m'heu ajudat a sobreviure aquest període, que per cert ha estat molt dur. També vull mostrar el meu agraïment als companys del "despacho perdido" sobretot pel "Yes, you can!!" "Vielen Dank, Mitstreiter!!!!" En particular, gràcies Martin per les mil col·laboracions, pels cracks d'alguns programes, per les teves explicacions pacients, pel sopar a soles a París (sempre em "partiré de riure" recordant com el cambre et deixava el tiquet a tu!!!) i gràcies Johannes pel teu interès en la meva feina, per aportar-me idees i pels "bailoteos" del sopar de Nadal. Ah! I a tu Giovanni gràcies pel teu esperit de col·laboració i per preocupar-te del com m'anava tot.

Finalment, voldria dedicar unes paraules als meus directors de tesi: el Jaume i el Ferran, Persones amb qui m'hi he entés amb facilitat, però sobretot amb qui m'hi he sentit molt a gust. En concret, del Jaume n'admiro l'energia i l'impetu, que sovint ha aconseguit transmetre'm. També n'admiro la fortalesa per avançar malgrat els entrebancs, els horaris apretats, les agendes impossibles i altres temes diversos. D'ell especialment, m'agrada el que comparteix amb el Ferran: el to proper. Ja que parlem del Ferran... d'ell n'admiro el bon humor (sempre tranquil i content) i el fet que saps de sobres que amb ell sempre hi pots comptar. Durant aquests quatre anys m'ha ajudat enormement i ho ha fet, com apuntava abans, d'una manera molt propera i reposada que he de dir que m'ha encantat. Aquest ha estat un bon tàndem de directors, molt compensat. De fet, un per l'altre han creat un efecte Ying-Yang. Gràcies, moltíssimes, a tots dos!!!

Encara em quedem moltes més coses a dir i a recordar d'aquest temps. No voldria deixar de banda per res del món als companys que vàreu començar el doctorat amb mi i encara menys als membres de l'Home de caramel: l'Ivan, la Vane, la Selma, l'Alberto... tots heu jugat un paper improtantíssim en la meva felicitat, amb les festes i sopars, els viatges accidentats amb bici, amb l'escriptura del "Capuxet Vermell" (eh, Selma!) i la representació teatral i estal·lar d'aquell Nadal...

Voldria sortir del cercle de la ciència per dedicar unes paraules als meus fantàstics amics, que m'han fet costat com ningú. Sense cap mena de dubte enpaçalen la llista l'Esther i la Txell, però també l'Agnès, la Gore, l'Anna Vila, la Isabel, la Carol, el Lluís, el Roger, l'Òscar, el Joan, l'Isma, la Núria, la Rebe... I tots els que segueixen: el Damià, l'Isaac, el Rúben... i el Raül que em va ajudar amb la portada de la tesi!! En definitiva tots el Vilanovins i Vilanovines Il·lunàtics que he tingut al voltant.

També vull recordar els no-vilanovis, que no per ser d'una altra banda es mereixen menys. Així doncs, visca la Joana, la Laura, la Míriam, la Julie, l'Arnau, la Laia, la Marta, el Quim, l'Imma... a qui he "fotut el sol al cap" més d'un cop amb històries del laboratori.

Ja per acabar voldria dedicar unes paraules als de casa: al meu pare (l'Enric) i a la meua mare (la Lourdes) que es mereixen un tros immens d'aquesta tesi perquè realment l'han viscut com en pròpia carn i m'han ajudat amb tot el que han pogut; al Xavier (el meu germà) que no ha parat de fer soroll i tocar la pera mentre escrivia exercint a la perfecció el rol de germà tal com Déu mana i la iaia (l'Anna) que m'ha entès més que ningú i ha respectat meua la clausura i meu el vot de silenci mentre escrivia (com els meus amics entendran, "lo" del vot de silenci és conya)

En un capítol especial vull fer referència a l'Oriol. Ell m'ha ajudat moltíssim tot superant la distància que invariablement sempre ho complica. M'ha fet costat i m'ha comprés, sobretot durant aquesta última època que ha estat especialment difícil. A banda d'això, és ell qui m'ha acompanyat en mil aventures (de vegades arriscades) que han fet que aquests anys fossin ben divertits. Oriol ets un company ben intrepit tu!!! Però fantàstic!

I finalment, només permeteu-me afegir dues coses:

Visca el vinagre Vilanoví! I visca el Transport de Panses Armènies!!!



## **Dissertation Summary**

The present thesis includes four studies with a common objective: determining whether pathogens (virus, bacteria, parasites...) have exerted selective pressures on the genome of their hosts (for example, humans).

Detecting signatures of positive selection is a useful tool to identify functionally relevant genomic regions since selection locally shapes the functional variation. Based on this premise, we have studied the possible signatures of selection in genes related to host-pathogen interactions. Specifically, we have analyzed those genes encoding: a) components of the innate immunity response; and ii) glycosylation enzymes most of them involved in four major glycan biosynthesis pathways, in different human populations.

The main conclusion obtained from these studies is that both studied gene categories show clear signatures of selection. Moreover, we have determined that according to their biological context certain genes are more prone to the action of selection.

## **Resum de la Tesi Doctoral**

La tesi que teniu a les mans recull quatre treballs amb un objectiu comú; determinar si els patògens (virus, bacteris, paràsits...) han exercit pressions selectives sobre els genomes dels seus hostes (com per exemple els humans).

Sabent que la detecció de l'empremta de la selecció permet identificar aquelles regions del genoma que han estat rellevants al llarg de l'evolució d'una espècie, ja que a nivell local és la variació funcional qui acaba essent objecte de la selecció, ens hem disposat a estudiar els possibles senyals de selecció en gens relacionats amb la interacció hoste-patògen. En concret, hem analitzat gens que codifiquen per: a) components del sistema immunitari innat i, b) enzims de glicosilació, la majoria dels quals s'inclouen en quatre de les principals rutes biosintètiques de glicans, en diferents poblacions humanes.

Com a conclusió principal; ambdós conjunts de gens mostren clars senyals de selecció. A més hem vist que segons el context biològic on és troben certs gens és veuen més afectats per l'acció de la selecció natural.



## **Preface**

In the past few years the study of genetic variation has been used not only to understand the history of human populations but also to identify those regions in the genome that have played an important role in the evolution of our species. So far this strategy has been useful to identify loci involved in adaptation to pathogens, climate or even diet.

The present thesis focuses on the study of human genetic variation in genes presumably susceptible to be under pathogen-mediated selection. Specifically, our study aims to identify the signature of selection in two groups of genes encoding: i) components of the innate immunity response; and ii) glycosylation enzymes most of them involved in four major glycan biosynthesis pathways.

During the last four years we have been mainly focused in the study of human lineage, using a set of globally distributed human populations. Notice, though, that some of the works constituting this manuscript do also include different non-human primate species of Old World and New World monkeys.

It is important to mention that, although we have taken advantage of the large repository of information available on line, we have produced most of the data used in this thesis by means of resequencing and genotyping techniques. Moreover, we have moved from single gene analysis to the study of whole gene categories and biosynthesis pathways. This has caused a dramatic increase in the number of candidate genes making necessary a computational approach to handle data. In consequence this thesis represents an example of well integration of wet lab (data production) and subsequent dry lab (data analysis). To fully understand the evolutionary pattern of candidate loci we have tested for positive, and also for balancing and purifying selection, when possible.

The most important contribution of this thesis to the scientific community is the proposal of studying evolutionary genetics from an holistic point of view. This concept specifically implies the need of knowing the pathways or biological processes our candidate genes are involved in order to perform contextualized studies. It is worth pointing that proceeding in this way requires gathering information from several fields, such as molecular biology and biochemistry which provide fundamental information to establish the study's framework.



---

<b>1</b>	<b>INTRODUCTION.....</b>	<b>1</b>
1.1	HUMAN POPULATION HISTORY .....	3
1.1.1	The origins of modern humans.....	3
1.1.1.1	The multiregional model .....	4
1.1.1.2	The out of Africa model .....	5
1.2	HUMAN GENETIC DIVERSITY.....	7
1.2.1	Mutation .....	7
1.2.1.1	Genotyping.....	8
1.2.1.2	High-throughput resequencing .....	9
1.2.2	Natural selection .....	10
1.2.2.1	Positive selection.....	10
1.2.2.2	Balancing selection .....	11
1.2.2.3	Purifying selection .....	12
1.2.3	Genetic drift.....	12
1.2.4	Migration .....	13
1.3	CURRENT STUDIES ON HUMAN GENETIC DIVERSITY AND THEIR IMPLICATIONS.....	14
1.3.1	High-throughput studies .....	14
1.3.2	Public available data .....	15
1.3.3	Analytical methods to detect selection .....	16
1.4	HUMAN EVOLUTION AND HOST-PATHOGEN INTERACTIONS .....	19

1.4.1	Signatures of selection mediated by pathogens.....	19
1.4.2	Infectious diseases.....	19
1.4.3	Genes related to host-pathogen interactions .....	20
1.4.4	Immune system.....	20
1.4.4.1	Innate immunity .....	22
1.4.4.2	Acquired immunity.....	26
1.4.5	Glycosylation.....	27
1.4.5.1	Glycoconjugate classes and metabolic pathways .....	27
1.4.5.2	Glycan biosynthesis .....	29
1.4.5.3	Fucosyltransferases .....	32
1.4.5.4	Sialyltransferases .....	33
1.4.5.5	Galactosyltransferases.....	33
1.4.6	Blood groups.....	34
1.4.6.1	ABO blood group.....	35
1.4.6.2	Pseudogenes .....	36
1.4.6.3	The ABO precursor gene FUT2.....	36
<b>2</b>	<b>MATERIAL AND METHODS.....</b>	<b>39</b>
2.1	Samples and data.....	42
2.1.1	The Human Genome Diversity Project (HGDP-CEPH) .....	42
2.1.2	International HapMap Project.....	43
2.1.3	650K SNPs in HGDP-CEPH .....	44

---

2.1.4	PGA Innate Immunity.....	46
2.1.5	SNPlex genotyping.....	47
2.2	Analysis tools .....	48
2.2.1	SNPator .....	48
2.2.2	Arlequin.....	49
2.2.3	DnaSP.....	50
2.2.4	Phase.....	51
<b>3</b>	<b>OBJECTIVES .....</b>	<b>55</b>
<b>4</b>	<b>RESULTS .....</b>	<b>59</b>
4.1	Chapter 1.....	61
	"Balancing selection is the main force shaping the evolution of innate immunity genes"	
4.2	Chapter 2.....	71
	"A natural history of <i>FUT2</i> polymorphism in humans"	
4.3	Chapter 3.....	93
	"Evolutionary analysis of human pseudogenes of the ABO family show a complex picture in their dynamics and function loss"	
4.4	Chapter 4.....	115
	"Worldwide human genetic diversity in four major pathways of glycan biosynthesis"	
<b>5</b>	<b>DISCUSSION.....</b>	<b>147</b>

<b>6</b>	<b>REFERENCES.....</b>	<b>163</b>
<b>7</b>	<b>APPENDIX.....</b>	<b>189</b>
7.1	Chapter 5.....	191
	"Is there selection for the pace of successive inactivation of the arpAT gene in primates?"	
7.2	Chapter 6.....	199
	"Neuropathologic findings in an aged albino gorilla"	



# <sup>1</sup> INTRODUCTION

---



## 1.1 HUMAN POPULATION HISTORY

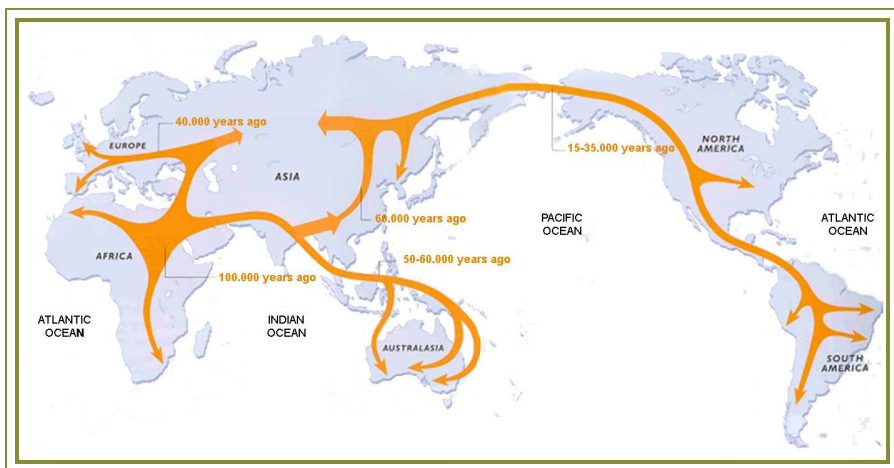
### 1.1.1 The origins of modern humans

After human and chimpanzee lineages split about 5-7 MYA (millions years ago) the earliest hominid species were originated. But very antique hominid fossils are scant and difficult to date, so when the first hominid species appeared remains uncertain. So far the existent fossil record have been sufficient to reveal the existence of: four hominid species originated between 7 and 4,2 MYA; five species from *Australopithecus* genus, three from *Paranthropus* genus and one from *Kenyanthropus* genus originated between 4,2 and 2 MYA; and four species belonging to the genus *Homo* (humans) originated between 2 MYA and 141 KYA, when our species (*Homo sapiens*) arose. Anthropologists have made the distinction among above cited hominid species based on differences of morphological traits since modern humans or AMH (anatomically modern humans) differ from archaic humans (earlier hominid species) because of two main traits; 1) extent of the globular shape of the skull and 2) retraction of the face. All modern human fossils have been found either in Africa and elsewhere or outside Africa, in contrast to archaic human fossils only found in Africa. This fact points to an African origin of our species.

From the *Homo* species described until date, our species (*Homo sapiens*) was originated from an archaic *Homo sapiens* species coming from the second out of Africa wave and, not from *Homo erectus* as it was accepted until recently. Therefore, the name *Homo erectus* would strictly refer to (and, then should be used for) the Asian focus of the first hominids. In order to explain the transition to modern *Homo sapiens*, two different models based on paleontological, archeological and/or genetic data have been proposed: the Multiregional model and the Out of Africa model. Although both models have been under vigorous debate due to the lack of conclusive evidences for any of them, the Out of Africa model is more supported by genetic data (mtDNA, chromosome Y, lack of Neanderthal mtDNA genes in modern humans and gradients of nuclear genetic diversity from Africa to America) (Jobling et al. 2004). Consequently, the current

best explanation for the beginning of modern humans seems to be the Out of Africa Model since it postulates a single and recent African origin for *Homo sapiens* (Figure 1).

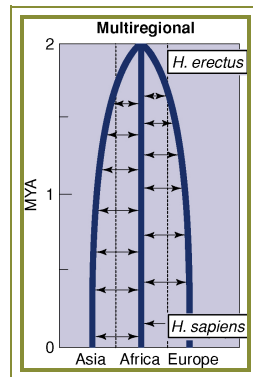
Getting a demographic model able to explain neutral variability is of major importance because past demographic events might have affected the current pattern of genetic diversity (Fagundes et al. 2007).



**Figure 1.** The migration of modern *Homo sapiens*.

### 1.1.1.1 The multiregional model

The multiregional model proposes that earlier *Homo erectus* (also called *Homo ergaster*) left Africa 2 MYA and dispersed into other parts of the Old World. Then *Homo erectus* would have slowly evolved into modern humans in the different regional populations in a synchronized manner. This model assumes some level of gene flow between geographically separated populations that would have prevented speciation after the dispersal (Figure 2).



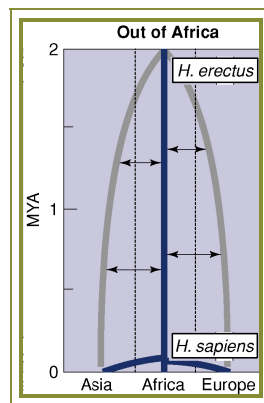
**Figure 2.** Multiregional evolution model for the origin of modern humans. Arrows indicate gene flow between populations on continents; Black lines represent the ancestors of modern humans (Jobling et al 2004)

Thus, genetic exchanges between populations, migration and gene flow across a worldwide network of relationships would let most adaptively advantageous genes, or gene combinations, spread to all groups. It would indicate that selection, and not recent origin, accounts for human similarities across human species. This scenario implies long term interconnection between populations by gene flow and therefore high population sizes which is unlikely.

#### 1.1.1.2 The out of Africa model

The out of Africa model claims that the transition to modern *Homo sapiens* took place ~150 KYA and proposes that ~60 KYA modern humans migrate out of Africa (Figure 1). In accordance to this model anatomically modern humans originated from a small African population that afterwards colonized the globe replacing the hominid species already present in other continents (Figure 3). Although this model assumes a reduced or inexistent inbreeding between modern humans and archaic populations, recent studies have revealed some polymorphism patterns difficult to explain when taking into account a pure African replacement scenario. Despite of this, the out of Africa model is the most plausible. In fact, it has been proved to be a robust model thanks to the study of more genetic loci, the improvement of the techniques to analyze genetic data and better mathematical

models which have been substantiated that: humans lack genetic diversity (reflecting an origin in a small population), the global diversity is a subset of African diversity (supporting the possibility of the original population being from Africa) and, Neanderthals lie outside of the range of human diversity and show they evolutionary history. There exist two demographic variants of the out of Africa model; the “strong garden of Eden hypothesis” and the “weak garden of Eden hypothesis”. The former stands for a sub-population of earlier *Homo sapiens*, perhaps a newly formed species, which colonized the Old World. The latter also refers to a sub-population that colonized the Old World, but in a slower manner taking tens of thousands of years and producing separate daughter populations that latter re-expanded (Harpending 1993).



**Figure 3.** Out of African replacement model; Arrows indicate gene flow between populations on continents; Black lines represent the ancestors of modern humans and lighter gray represents lineages that are not ancestors of humans. (Jobling et al 2004)

## 1.2 HUMAN GENETIC DIVERSITY

Humans are different from each other both at phenotypic and genetic level. These differences are due to environmental and genetic factors. In terms of genetic differences, any two non-related individuals present among each other an average of one substitution every 1000 base pairs. In fact, population genetic studies have reported that the major proportion of human variation (80%) is explained by differences between individuals of the same population, whereas a minor part of the variation is explained in one part by differences between populations of the same continent (5-10%) and in the other by differences among continents (10-15%) (Barbujani et al. 1997).

Human genetic variation is the result of mutation, natural selection, genetic drift and migration. It is relevant to mention that demographic factors such as population expansions, bottlenecks and subdivision of populations might have also left a signature across the genome that can interfere with the detection of the footprint of selection at particular genes (Balaesque et al. 2007; Harris and Meyer 2006).

### 1.2.1 Mutation

Mutation is a permanent structural alteration in DNA (Wright 1931). DNA alterations can be generated chemically as is the case of cytosine deamination; physically by the break down of the DNA double helix for generating an insertion; or enzymatic by the slippage of the DNA polymerase. DNA mutations either have no effect or cause harm, but occasionally a mutation can improve an organism's chance of survival. DNA changes can occur in any cell line, but just those mutations appearing in the germ line will have evolutionary consequences, since they can pass to the following generations.

Once mutation occurs it gives place to a variable position as for example single nucleotide polymorphisms (SNPs). These are the specific type of genetic variants

studied in the present thesis. There are several methods to detect single nucleotide polymorphisms on DNA sequences. Some of them require previous knowledge about SNPs characteristics. The first method (Restriction Fragment Length Polymorphism or RFLP) used for both detecting and characterizing variation in the DNA sequence was based on the use of restriction enzymes. Due to restriction enzymes sensitivity this method is able to detect specific changes in fragments of DNA constituted usually of 3 to 6 nucleotides. A later method to detect and characterize mutations is direct sequencing (Sanger and Coulson 1975; Sanger et al. 1977). This method is based on a first elongation of the DNA template using PCR (polymerase chain reaction) and a second PCR step in which the incorporation of fluorescent ddNTPs that stop the DNA chain elongation after each new nucleotide addition is performed. Nucleotides constituting the DNA sequence are determined depending on their fluorescence emissions and, their positions are determined according to the fragment length.

Once mutations are characterized other techniques, less laborious for routine screening, have been traditionally used to determine which allele or alleles are present in a certain polymorphic position, like ASO and ARMS. Allele Specific Oligonucleotide (ASO) (Wallace et al. 1979) consists on designing and using a short piece of synthetic DNA (oligonucleotide of 15-20 nt) complementary to the sequence of a variable target DNA, which is specific for only one allele. These oligonucleotides act as a probe for the presence of the target in a Southern blot. In the case of ARMS (Amplification Regulatory Mutation System) mutations are detected by hybridizing a probe either complementary to the mutation or to the reference sequence. Being amplified just that DNA template matching perfectly to the probe. Currently, a number of genotyping techniques have been inspired in the detection of mutations by using specific probes.

#### **1.2.1.1 Genotyping**

In the last few years different genotyping techniques have been under fervent development. Some of them allow users to select the SNPs to type (TaqMan, Kaspas, SNIPlex, MassArray, BeadArray) whereas others consist on pre-designed



SNP panels either covering the whole genome (Illumina Sentrix HumanHap550 Duo or Affymetrix Genome-Wide Human SNP Array 6.0) or covering certain genes or gene categories (Illumina MHC panels for 1290 and 2360 loci or Affymetrix GeneChip® Human Immune and Inflammation 9K SNP Kit). All these genotyping techniques differ among each other in the chemistry used and on the number of individuals and polymorphic positions they can handle simultaneously. Nevertheless, all of them contribute in making everytime much feasible the obtainance of large amounts of genotypes. All works constituting the present thesis include genotype data we have obtained taking advantage of one of these genotyping techniques: SNPlex.

### **1.2.1.2 High-throughput resequencing**

Over the last years and similarly to the situation concerning genotyping techniques, a considerable number of faster and more cost-effective sequencing technologies have been developed due to an increasing demand (Hall 2007; Shendure et al. 2005). Many of the new high-throughput methods parallelize the sequencing process in order to produce thousands or millions of sequences at once.

To improve sequencing, some of those techniques use a method to isolate and increase the number of DNA molecules (Emulsion PCR). This is the case of sequencing technologies commercialized by 454 Life Sciences (Roche) (Margulies et al. 2005), "polony sequencing" (Shendure et al. 2005). and SOLiD sequencing (Applied Biosystems) (Biosystems). Another method for increasing the number of DNA molecules is "bridge PCR", where fragments are amplified upon primers attached to a solid surface, method used by Solexa sequencing. Besides these two amplification methods there is one more method that skips the amplification step by directly fixing DNA molecules to a surface (Braslavsky et al. 2003). To determine the DNA sequence each individual DNA molecule not only has to be amplified and fixed on a surface but also has to be sequenced in parallel though one of the following approaches; sequencing by synthesis (as done with automated sequencing machines based on Sanger sequencing); reversible terminator methods (used by Illumina and Helicos) (<http://www.illumina.com>);

pyrosequencing (as done by 454 Sequencing technology), and, sequencing by ligation which is used in the polony method.

Futhermore, there are new proposals for DNA sequencing, like labeling the DNA polymerase, reading the sequence while a DNA strand transits through nanopores, and microscopy-based techniques, such as AFM or electron microscopy that are used to identify the positions of individual nucleotides within long DNA fragments (>5,000 bp) by nucleotide labeling with heavier elements (e.g., halogens), all them still in development. Two of the works included in this manuscript include resequencing data which has been obtained by means of Sanger resequencing techniques.

### **1.2.2 Natural selection**

Natural selection is the phenomena changing the frequency of a mutation in the descendents by increasing or decreasing the fitness of the organism carrying it. In turn, the concept fitness is described as the ability of an individual genotype to survive and reproduce, which is in part dependent on the environment. Three types of selection have been defined to shape genetic variation within and between species: positive selection, balancing selection and purifying selection.

#### **1.2.2.1 Positive selection**

The term positive selection refers to the case in which a novel DNA variant increases the fitness of the carriers in comparison to the rest of individuals by conferring some advantage in the adaptation to the environment. There is a renewed interest in finding targets of positive selection in the human genome because this kind of selection delimit regions of the genome that are, or have been, functionally important and, therefore such regions can help in identify the genetic variation that contributes to phenotypic diversity (Biswas and Akey 2006; Sabeti et al. 2006). An example of this would be the study made by Wang et al. on Perlegene data using LD decay (LDD) test, a method specifically testing for positive selection. This study revealed that 1799 human genes showing signatures

of positive selection clustered in 5 categories: host-pathogen interactions, reproduction, DNA metabolism and/or cell cycle, protein metabolism and neuronal function (Wang et al. 2006). As cited, genes involved in the interaction between pathogens and hosts constitute a candidate gene category where looking for the signature of positive selection. Another example of this is the gene encoding the Duffy antigen (FY), a red cell surface protein that *Plasmodium vivax* malaria parasite uses to enter cells. A mutation in the promoter region of *FY* that disrupts Duffy antigen expression conferring protection against *Plasmodium vivax* has been shown to be under positive selection in African regions where this parasite is endemic (Hamblin and Di Rienzo 2000).

### 1.2.2.2 Balancing selection

Balancing selection is the regime maintaining multiple alleles at a locus in a population. So, this type of selection increases genetic variation within a species. Several different biological processes can generate balanced polymorphisms. There are different types of balancing selection; 1) overdominant selection which refers to the situation in which heterozygotes have higher fitness than homozygote individuals; 2) frequency dependent selection that is the term given to an evolutionary process where the fitness of a phenotype is dependent on its frequency relative to other phenotypes in a given population. This type of selection is subdivided into positive frequency dependent selection when the fitness of a phenotype increases as it becomes more common or negative frequency dependent selection if the fitness of the phenotype increases as it becomes less common, as it is the case for *MHC*; and 3) environmental heterogeneity which favors one allele or the other according to conditions or time. It favors different alleles at different time intervals, as selective regimes changes.

Some well known examples of balancing selection are: *MHC* (Solberg et al. 2008) and *HbS* (Kwiatkowski 2005). The former is an acquired immunity gene involved in the recognition and presentation of bacterial and viral antigens to T-cells. Individuals with heterogeneous *MHC* haplotypes can recognize a broader spectrum of antigens, thus heterogeneous individuals are favored with respect to

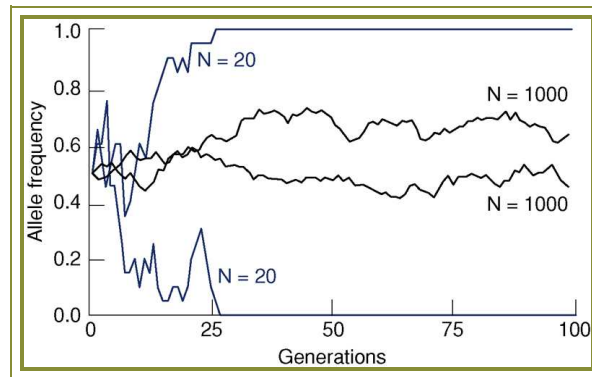
homozygotes. The later, *HbS*, is the classic paradigm of balanced polymorphisms in human populations (Feng et al. 2004; Hedrick 2004). It is a variant of the *HBB* gene, encoding  $\beta$ -globin, that confers 10-fold reduced risk of severe malaria when in heterozygosis, whereas in homozygosis it cause sickle-cell disease (Ackerman et al. 2003; Hill et al. 1991). This variant has arisen independently in different locations and is maintained at 10% frequency in many malaria-endemic regions (Flint et al. 1998).

### **1.2.2.3 Purifying selection**

The term purifying selection refers to selective regimes that reduce the fitness of the individuals carrying a disadvantageous mutation with respect to other individuals. So the carriers have less probability of having descendents than non-carriers. Purifying selection is expected to be the most frequent type of selection keeping disadvantageous mutations at low frequencies or even removing them. This is, for example, the kind of selection that has been proposed to be the force removing mutations that result in loss-of-function alleles at *MCR1*, a gene necessary for eumelanin production and responsible of dark hair or skin color in Africans (Harding et al. 2000; Rana et al. 1999).

### **1.2.3 Genetic drift**

Genetic drift is the variation in allele frequencies between generations due to a random process in a finite population (Jobling et al. 2004). The magnitude of genetic drift relates to the effective population size ( $N_e$ ), which is the number of chromosomes that pass to the next generation. The greater the  $N_e$  is, lower the magnitude of the genetic drift will be. This phenomenon is described in a theoretical way by the Wright-Fisher model assuming that 1) populations are of constant size, 2) there are nonoverlapping generations and 3) individuals have the same probability of contributing to the next generation (Figure 4):



**Figure 4.** Genetic drift of a binary marker in populations of different sizes (Jobling et al 2004)

Actually, effective population size ( $N_e$ ) is not constant. It fluctuates as a consequence of variation in population size ( $N$ ). In fact, many human populations experienced in the past a reduction of population size which shaped their current genetic diversity. Such a size reduction can be caused by two important demographic processes: bottlenecks and founder effects. The first refers to the reduction of genetic diversity of the whole ancestral population by a drastic reduction of the number of individuals. The second refers to the separation of a subset of the genetic diversity from the source population, for example when colonizing new regions. Bottlenecks produces a reduction of the initial population size implies the loss of the prior diversity.

#### 1.2.4 Migration

Migration is the movement of individuals (migrants) from one occupied place to another that can be occupied or not. This process causes gene flow whenever a migrant contributes to the next generation in the new location, if occupied. Gene flow tends to homogenize the amount of genetic diversity between both populations in relation to the time before the migration process began. So, detection of migration require from observing allele frequencies before the homogenization processes finishes.

### **1.3 CURRENT STUDIES ON HUMAN GENETIC DIVERSITY AND THEIR IMPLICATIONS**

The study of human population genetics has become a field of emergent interest in the recent years specially because of its potential in detecting adaptation to new environments and diseases. Recently the detection of the signatures that the process of natural selection left in our genome is used as a clue to identify regions that have been functionally relevant during human evolution and underlie variation in disease resistance or drug metabolism (Bamshad and Wooding 2003).

#### **1.3.1 High-throughput studies**

Until very recently studies based on human genetic variation were examining candidate genes individually. This was motivated by *a priori* information about the function of the gene and its association with a variable trait. But studying a single candidate gene gave rise to difficulties with assessing statistical significance. Therefore, in order to solve this issue statistical significance were evaluated by comparing empirical data to theoretical models assuming neutral evolution. With the improvement of sequencing technologies (as cited in section 2.1.2) and genotyping methodologies (as cited in section 2.1.1) plus the irruption of microarrays as DNA genotyping tools the amount of empirical data has notably increased. All this new data helps to evaluate and refine the theoretical models used to assess significance and can be used to obtain background distributions to compare results to (Sabeti et al. 2006).

Lately, the detection of natural selection from genetic data has been carried out in a genome-scan manner. It has been possible mainly thanks to emergence of a broad range of sequencing and genotyping methodologies and, in some cases, to the availability of DNA collections. These studies scope go from hundreds to hundred of thousands of genotypes. The existing DNA collections cover human population diversity to different extends. They either go from a set of populations covering most of the extant worldwide diversity (HGDP-CEPH panel) (Cann et al. 2002) used in many studies from our Unit at UPF (Ferrer-Admetlla et al. 2008;

Gardner et al. 2008; Gardner et al. 2006; Sikora et al. 2008) to some populations with different origin as the ones used for HapMap project: Yoruba in Ibadan, Nigeria; Japanese in Tokyo, Japan; Han Chinese in Beijing, China; and the CEPH (U.S. Utah residents with ancestry from northern and western Europe) (de Bakker et al. 2005; Hinds et al. 2005; Sabeti et al. 2007; Sikora et al. 2008; Walsh et al. 2006), or the ones used in SeattleSNPs database: African-Americans and European-Americans from Coriell Institute collection (National Institutes of et al. 1982).

### **1.3.2 Public available data**

Large publicly available sources of data are increasing in number and quality promoting and facilitating genome-wide studies on natural selection (Sabeti et al. 2006; Sabeti et al. 2007). At intraespecific level the major contributor to this fact is the increasing number of genome sequences from different species. Until date, 39 species have been sequenced being the order of primates and rodents and the superorder of *Laurasiatheria* the ones with more species sequenced: 6, 7 and 7 respectively. Furthermore, the genome sequence of 5 more species is in process (<http://www.ensembl.org/index.html>).

At interspecific level, the near-complete sequence of the human genome, the genotyping data sets as the ones produced by the International HapMap Project (Consortium 2005; Frazer et al. 2007) or Perlegen Sciences (Hinds et al. 2005) or the recently released 650K (Affimetrix) genotype data for HGDP-CEPH panel samples plus publicly available resequencing data as the one produced by the innate immunity Program for Genome Applications (<http://innateimmunity.net/>) are sources of data stimulating genome-wide studies on natural selection. No doubt, the future data from “the 1000 genome project” will also contribute to this fact.

### 1.3.3 Analytical methods to detect selection

Since 1950, when Wright developed a test based on allele frequencies to detect selection (Wright 1950), a variety of new statistical tests and approaches to detect natural selection at DNA level have arisen (Akey et al. 2002; Fay and Wu 2000; Fu and Li 1993; Hanchard et al. 2006; Hastbacka et al. 1994; Hudson et al. 1987; Hughes and Nei 1988; Kim and Nielsen 2004; Lewontin and Krakauer 1973; Li et al. 1985; McDonald and Kreitman 1991; Sabeti et al. 2002; Tajima 1989; Toomajian and Kreitman 2002). These tests can be divided in two categories. In one hand, the tests based on polymorphisms within species and on the other, those based on polymorphisms within species and divergence between species.

The former category includes Tajima's D statistic that measures the difference between two estimators of the population mutation rate,  $\theta$  and  $\pi$ . Under neutrality these two estimators are expected to be equal. Therefore, positives and negatives values of Tajima's D indicated departures from neutrality. Specifically, positive values of Tajima's D arise from an excess of intermediate frequency alleles as a result of bottlenecks, structure and/or balancing selection, whereas negative values of Tajima's D are due to an excess of low frequency alleles as the result of population expansions or positive selection (Tajima 1989); Fu and Li's D and F tests also detect a skew in the allele frequency spectrum, but unlike Tajima's D they take into account whether mutations are old or recent. In the same way as Tajima's, both Fu and Li D and F positive and/or negative values are informative about distinct demographic and/or selective events (Fu and Li 1993); Fay and Wu's H statistic detects the presence of an excess of high-frequency-derived mutations, which are expected to be in excess in positive selection scenarios. This statistic requires the use of an outgroup species (Fay and Wu 2000).

Long range haplotype test (LRH) detects alleles of high frequency with long-range linkage disequilibrium (LD). This approach looks for recent positive selection since this selective regime is expected to accelerate the frequency of an advantageous allele faster than recombination can break down the LD at the selected haplotype.

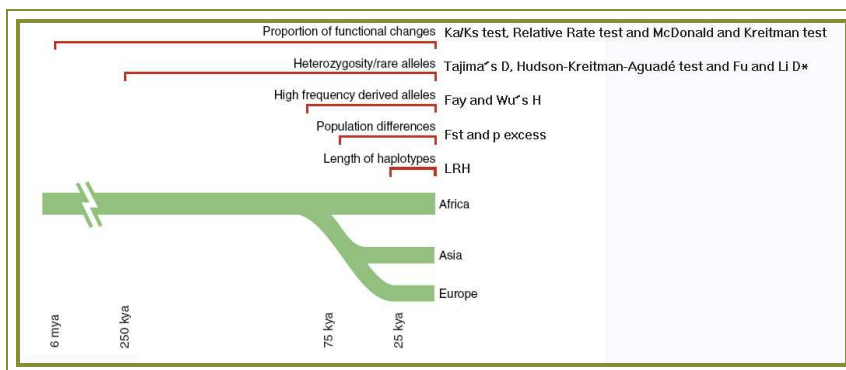


To capture the signature of positive selection this method selects a “core” haplotype or a SNP and assesses the LD decay for flanking markers by calculating EHH (Extended Haplotype Homozygosity). Positive selection is then determined when a core haplotype presents an elevated EHH relative to other EHH core haplotypes (Sabeti et al. 2002). iHS is also based on the LD extension; this method calculates the integrated EHH from a specific core allele until EHH reaches a frequency of 0.05. Large negative values of iHS indicate long haplotypes carrying the ancestral allele, whereas large positive values indicate long haplotypes that carry the derived allele (Voight et al. 2006). LD decay (LDD) is another approximation based on the LD measure. In this case an homozygote SNP is selected, then a statistic based on the fraction of chromosomes for all adjacent SNPs and the distance from them to the target SNP is calculated (ALnLH). High ALnLH values for one specific allele of the target SNP compared with the other allele indicates selection (Wang et al. 2006).

$F_{ST}$  statistic quantifies the level of population differentiation between subpopulations. There are several methods for estimating  $F_{ST}$  from samples of a group of populations, but probably the simplest is the difference between an estimate of the total heterozygosity and the average heterozygosity across subpopulations corrected by the estimation of the total heterozygosity. Since  $F_{ST}$  statistic looks for the reduction in heterozygosity among subpopulations relative to what is expected under random mating, under neutrality large  $F_{ST}$  values are produced by genetic drift and migration and, whereas really accentuated values could indicate positive selection. One more issue concerning  $F_{ST}$  is the diversity in structure between regions of the genome in different populations, since the precise genealogy is not the same for each chromosome or part thereof, and thus  $F_{ST}$  values become more similar as more linked are the regions of the genome (Weir et al. 2005). Nonetheless,  $F_{ST}$ -based methods are good in detecting ancient selective events, regardless their strength, since they can reveal the effects of natural selection in global population differentiation over the last 75.000 years (Barreiro et al. 2008).

The tests based on polymorphism and divergence includes: Hudson-Kreitman-Aguade (HKA) that tests whether levels of nucleotide variation within and between species at two or more loci positively correlates, which is what is expected under the neutral theory. It indicates adaptive evolution when showing an excess or reduction of polymorphism levels in one locus, or excess or limited divergence at the other (Hudson et al. 1987); McDonald and Kreitman (MK) test also looks for polymorphism and divergence data at a specific locus, but it compares different type of mutations (synonymous versus non-synonymous) (McDonald and Kreitman 1991), and finally there is a method to tests for signatures of selection specifically between species. It is the  $K_A/K_S$  test. It computes the ratio between synonymous and non-synonymous mutations in a protein. This test can distinguish among positive selection ( $K_A/K_S > 1$ ), neutrality ( $K_A/K_S = 1$ ) and purifying selection ( $K_A/K_S < 1$ ) (Nei 1986; Nielsen 1998)

These tests can be divided in five categories according to their strength and the time scale in which they can be used to detect selection (Figure 5).



**Figure 5.** The five categories into which positive selection tests can be classified; their time depths and some statistical tests examples within each category. Notice, LRH stands for Long Range Haplotype test (modified from Sabeti et al. 2006)

## 1.4 HUMAN EVOLUTION AND HOST-PATHOGEN INTERACTIONS

### 1.4.1 Signatures of selection mediated by pathogens

The genetic diversity we observe nowadays is the result of demographic history and selection events that have acted to adapt different populations to their environment (Balaresque et al. 2007). Different studies have exploited the availability of large data sets (HapMap, PerleGene... ) to look for signatures of adaptation. The classes of genes lying in the selected regions from these studies turned out to be broadly concordant (Consortium 2005; Tang et al. 2007; Voight et al. 2006), thus, revealing that most of the genes were involved in: chemosensory perception, olfaction, immunity, reproduction, fertility and carbohydrate metabolism. In summary, the most important selective pressures human encountered during the colonization process of new territories have been exerted by pathogens, climate, dietary and cognitive genetic adaptation (Balaresque et al. 2007; Hill 2006). Some of them have given place to infectious diseases that might have exerted selective pressures on their hosts and even caused their mortality. So, pathogens are considered one of the major selective agents of human history.

### 1.4.2 Infectious diseases

Infectious diseases as malaria, HIV/AIDS, tuberculosis, invasive pneumococcal disease, leprosy and chronic hepatitis B virus have been reported as clear examples of association between gene variants and susceptibility or resistance to infection (Galvani and Slatkin 2003; Khor et al. 2007; Meccas et al. 2004; Ohashi et al. 2004; Tishkoff et al. 2001; Wood et al. 2005). The most striking from the above mention cases is malaria, which is causing from 350 to 500 million clinical episodes per year, and over a one million of deaths. Malaria is one of the strongest known forces for evolutionary selection in the recent history of humans. This fact has motivated research on selection signatures on the human genome exerted by *Plasmodium falciparum* infection. Results have disclosed many resistance variants (Duffy O allele at the *FY*, HbC and HbE at *HBB*, *HBA*, *SLC4A1* and *G6PD*), but probably the most important is HbS, which constitutes a classic

paradigm of balancing selection (Feng et al. 2004; Hedrick 2004). This variant (as explained in section 1.4.2) reduces risk of severe malaria when in heterozygosis (Kwiatkowski 2005). Malaria constitutes a proof of how pathogens can affect host genomes.

### **1.4.3 Genes related to host-pathogen interactions**

As mentioned above (see section 3.1) some gene categories are more prone to be related to host pathogen interactions than others. Among the gene categories showing signatures of selection in several of its members there is immunity. Most of the previous studies on natural selection have considered acquired immunity genes. However, innate immunity genes constitute a promising category of genes when looking for selective signatures, since some of its members as Toll-like receptors (TLRs) and other members of their signaling pathway are directly involved in the interaction with and the response to a wide variety of pathogens (Ferrer-Admetlla et al. 2008; Khor et al. 2007). Besides innate immunity many more gene categories could be considered candidates for pathogen mediated selection. This is the case for genes involved in glycosylation processes, since some glycosylated structures interact with pathogens either because of protecting the cell against invaders or because acting as pathogen receptors. Thus, these genes are also exposed to action of selection exerted by pathogens.

### **1.4.4 Immune system**

The immune system is the mechanism that defends our organism against disease-causing agents and even against our own proteins in autoimmunity and eliminates our own aberrant cells in tumor immunity. This mechanism is constituted by a serial of cells, tissues and proteins which detect and kill invader organisms or, as mentioned above, tumor cells. The immune system is capable to recognize a wide spectrum of agents; viruses, bacteria, protozoa, fungus, parasites... through a variety of receptors (Table 1). Traditionally, the immune system has been divided into two major subdivisions: innate immunity/non-specific immunity and

adaptive/acquired/specific immunity, each with different function and role (Medzhitov and Janeway 2000a).

Receptor	Ligand	Origin of ligand
TLR1	Triacyl lipopeptides	Bacteria and mycobacteria
	Soluble factors	Neisseria meningitidis
TLR2	Lipoprotein/lipopeptides	Various pathogens
	Peptidoglycan	Gram-positive bacteria
	Lipoteichoic acid	Gram-positive bacteria
	Lipoarabinomannan	Mycobacteria
	Phenol-soluble modulín	Staphylococcus epidermidis
	Glycoinositolphospholipids	Trypanosoma cruzi
	Glycolipids	Treponema maltophilum
	Porins	Neisseria
	Atypical lipopolysaccharide	Leptospira interrogans
	Atypical lipopolysaccharide	Porphyromonas gingivalis
	Zymosan	Fungi
	Heat-shock protein 70*	Host
TLR3	Double-stranded RNA	Viruses
TLR4	Lipopolysaccharide	Gram-negative bacteria
	Taxol	Plants
	Envelope protein	Respiratory syncytial virus
	Heat-shock protein 60*	Chlamydia pneumoniae
	Heat-shock protein 70*	Host
	Type III repeat extra domain A of fibronectin*	Host
	Oligosaccharides of hyaluronic acid*	Host
	Polysaccharide fragments of heparan sulfate*	Host
Fibrinogen*	Host	
TLR5	Flagellin	Bacteria
TLR6	Diacyl lipopeptides	Mycoplasma
	Lipoteichoic acid	Gram-positive bacteria
	Zymosan	Fungi
TLR7	Imidazoquinoline	Synthetic compounds
	Loxoribine	Synthetic compounds
	Bropiramine	Synthetic compounds
	Single-stranded RNA	Viruses
TLR8	Imidazoquinoline	Synthetic compounds
	Single-stranded RNA	Viruses
TLR9	CpG-containing DNA	Bacteria and viruses
TLR10	N.D.	N.D
TLR11	N.D.	Uropathogenic bacteria

**Table 1.** Toll-like receptor and their ligands. \*Need of more precise analysis to conclude that TLRs recognize these endogenous ligands. ND, not determined and TLR, Toll-like receptor.

#### **1.4.4.1 Innate immunity**

Innate immunity constitutes the first line of defense against infectious agents. Mainly, three features differentiate innate immunity from acquired: 1) Innate immunity is continually ready to respond to invasion because it includes defenses that are constitutively present and ready to be mobilized upon infection. 2) It is not antigen specific and reacts equally well to a variety of organisms and 3) It does not demonstrate immunological memory. Moreover, it is present in all eukaryotes. Innate immunity is able to confer protection against a variety of organisms because it recognizes Pathogen Associated Molecular Patterns (PAMPs) which are molecular structures common to and essential for the survival of a wide variety of pathogens (Janeway and Medzhitov 2002; Medzhitov and Janeway 2000b). The concept PAMPs includes: LPS from the gram-negative cell wall; peptidoglycan and lipoteichoic acids from the gram-positive cell wall; the sugar mannose (a terminal sugar common in microbial glycolipids and glycoproteins but rare in those of humans); bacterial and viral unmethylated CpG DNA; bacterial flagellin; the amino acid N-formylmethionine found in bacterial proteins; double-stranded and single-stranded RNA from viruses and glucans from fungal cell walls. In addition, unique molecules displayed on stressed, injured, infected, or transformed human cells also act as PAMPs.

Innate immunity capability to discriminate between self and nonself molecules is fine tuned because it has been selected over evolutionary time. Its activation is immediate as it depends on receptors predetermined by the genome. Once activated, innate immunity responds to the infection releasing cytokines, chemokines and co-stimulatory molecules (Janeway and Medzhitov 2002), which activate other immune genes and trigger the inflammation response. Innate immunity is constituted by three kinds of barriers to infection: anatomical, humoral and cellular barriers.

### a) Anatomical barriers to infection

Our body is protected against infectious agents through mechanical, chemical and biological factors which act as anatomical barriers. Mechanical factors include 1) epithelial surface that acts as a first line barrier to infection as in turn, desquamation helps in removing bacteria and other agents adhered to the epithelial surfaces, 2) the movement of cilia or peristalsis which helps to keep air passages and the gastrointestinal tract free of microorganisms, 3) tears and saliva that prevent infection of eyes and mouth and, 4) mucus that covers respiratory and gastrointestinal tract to protect lungs and digestive system from infection. The chemical factors helping us in protection to infectious agents are 1) Fatty acids in sweat that inhibit the growth of bacteria, 2) Lysozyme and phospholipase found in tears, saliva and nasal secretions that break down bacterial cell wall, 3) Low pH in sweat and gastric secretions which prevent bacterial growth, 4) Defensins present in the lungs and in gastrointestinal tract which have antimicrobial activity, and, 5) Sufactants, a kind of opsonin, present in the lung that promote phagocytosis. Finally, the most relevant biological factor contributing to pathogen defense is the normal flora of the skin, which secretes toxic substances and competes with bacteria for nutrients and for attachment to cell surfaces.

### b) Humoral barriers to infection

The anatomical barriers are very effective in preventing the colonization of tissues by microorganisms. If, however, these barrier layers are penetrated, the infection occurs and then another innate defense mechanism comes into play: the acute inflammation. Humoral factors play a crucial role in inflammation that is edema and recruitment of phagocytic cells. These humoral factors can be either present in serum or formed at the site of the infection. Here we summarize the different humoral non-specific defense mechanisms:

The major humoral non-specific defense mechanism is the complement system. It triggers the increase of vascular permeability, the recruitment of phagocytic cells and lysis and opsonization of bacteria. Besides this, there are several other

mechanisms also involved in the humoral response: The coagulation system releases products some of which increase vascular permeability and act as chemotactic agents (components that help movement toward an increasing chemical gradient) for phagocytic cells while some others are antimicrobical as  $\beta$ -lysin, able to lyse Gram+ bacteria. Cytokines are also members of the humoral non-specific defense mechanism. They include lactoferrin and transferrin that limit bacterial growth by binding iron and thus depriving bacteria of this essential nutrient; interferons which are proteins that can limit virus replication; lysozyme that breaks down bacterial wall and interleukin-1 which induces fever and the production of antimicrobial proteins.

c) Cellular barriers to infections

Part of the inflammatory response is the recruitment of certain cells; neutrophils, macrophages, natural killer (NK) and eosinophils, to sites of infection. These cells are the main line of defense in the innate immune system because of their capacity of engulfing foreign organisms and killing them without the need for antibodies. Neutrophils phagocytose invading organisms and kill them intracellularly. Macrophages also phagocytose invading organisms and kill them intracellularly, although they can also kill extracellularly infected or altered self target cells. Furthermore, macrophages contribute to tissue repair and act as antigen presenting cells, which are required for the induction of specific immune responses. Natural killer (NK) and lymphokine activated killer (LAK) can nonspecifically kill tumor and virus infected cells. These cells are not part of the inflammatory response but they are important in nonspecific immunity to viral infections and tumor surveillance. Eosinophils release proteins in granules that are effective in killing parasites.

All phagocytic cells contain granules which are involved in their antimicrobial properties. Neutrophils present two kinds of such granules. The first kind of them, are abundant in young neutrophils and contain cationic proteins and defensins that can kill bacteria, proteolytic enzymes like elastase and cathepsin G that break down proteins, lysozyme that break down bacterial cell walls, and

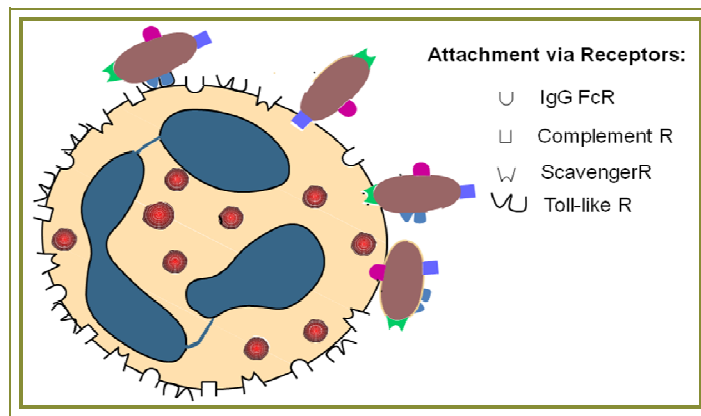


myeloperoxidase that is involved in generation of bacteriocidal compounds. The second type of granules is found in more mature cells. These granules contain NADPH oxidase components involved in the generation of toxic oxygen products, lactoferrin, B12-binding protein and, they also present lysozyme. In turn macrophages have numerous lysozymes with content similar to the neutrophils granules.

#### d) Response of phagocytes to infection

Circulation neutrophils, eosinophils and macrophages respond to danger signals generated at the site of an infection. These signals include N-formyl-methionine containing peptides released by bacteria, clotting system peptides, complement products and cytokines released from tissue macrophages that have encountered bacteria in tissues. Some danger signals stimulate endothelial cells to express adhesion molecules to make phagocytes adhere to endothelium, and then vasodilators produced at the side of the infection create channels between endothelial cells that help phagocytes to cross endothelial barrier. Once in the tissue spaces phagocytes move to infection attracted by chemotaxis, produced by danger signals. These danger signals can also increase phagocytosis.

Phagocytosis is the process of engulfing and killing infectious agents. Phagocytic cells bind pathogens thanks to a variety of receptor lying on their cells membrane. These receptors include; Fc receptor that can bind to a Fc region from the IgG antibody that some bacteria carry on their surface; complement receptor that can bind C3b-coated bacteria (C3b is the 3th component of complement); scavenger receptor which bind a variety of polyanions on bacterial surface and, toll-like receptors which recognize broad molecular patterns called PAMPs.



**Figure 6.** Phagocytic receptors

Binding of infectious agents via toll-like receptors results in phagocytosis and in the release of inflammatory cytokines, like IL6. After the attachment of a pathogen, the phagocyte extends pseudopods surrounding it. Then the phagocyte engulfs the infectious agent and encloses it in a phagosome. During the phagocytosis granules or lysosomes of the phagocyte fuse with phagosome and empty their content which results in killing the infectious agent.

#### **1.4.4.2 Acquired immunity**

In contrast to innate immunity, acquired immunity is a second-line barrier against pathogens, it is exclusive of vertebrates and it demonstrates immunological memory. Acquired immune system is antigen specific and thus reacts only with the organism that induced the response. This happens because it is composed of highly specialized systemic cells able to recognize concrete molecular structures from infectious agents (peptides, carbohydrates and proteins). Acquired immunity is selected in individual somatic cells, which makes discrimination between self and nonself components in a less refined manner than in the case of innate immunity, which, as said above, has been selected over evolutionary time. Besides, acquired immunity requires some time to react to an invader organism since its activation depends on innate immunity signals, on the rearrangement of

the different gene segments coding for receptors and on antigen presentation which delays the action of acquired immunity in respect to the infection time (Janeway and Medzhitov 2002).

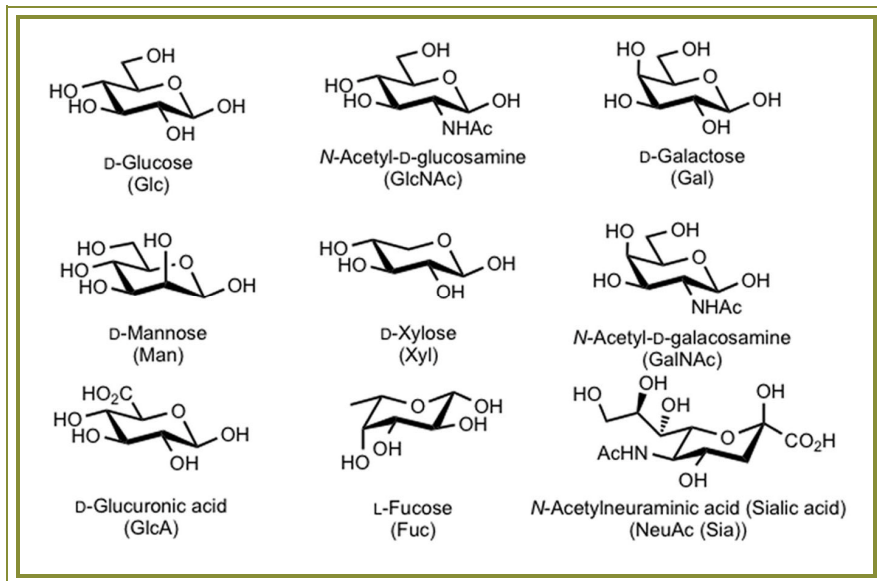
#### **1.4.5 Glycosylation**

The term glycobiology refers to the study of the structure, biosynthesis, biology and evolution of saccharides (sugar chains or glycans) that are widely distributed in nature, and in the proteins that recognize them (Varki et al. 1999). The process of formation of these glycan structures is named glycosylation.

Glycans have an important biological role in multicellular organs and organisms, since they are mediating the interaction between cells and the surrounding matrix. Specifically, secreted or outer cell surface glycans can modulate or mediate cell-cell, cell-matrix or cell-molecule interactions, and they can also mediate the interaction between different organisms (for example between host and parasite or symbiont).

##### **1.4.5.1 Glycoconjugate classes and metabolic pathways**

Monosaccharides are the basic structural units of glycans. In other words, any carbohydrate which can not be hydrolyzed into a simpler form is considered a monosaccharide. Monosaccharides are constituted by a potential carbonyl group: aldehyde (when the carbonyl group is at the end of the chain) or ketone (when the carbonyl group is attached to an inner carbon). In nature there exist thousands of different monosaccharides. Curiously, common animal glycans are constituted by a small fraction of them: Pentoses (D-xylose), hexoses (D-glucose, D-galactose, D-mannose), hexosamines (N-acetyl-D-glucosamine, N-acetyl-D-galactosamine), deoxyhexoses (L-fucose), uronic acids (D-glucuronic acid, L-iduronic acid) and sialic acids (N-acetylneuraminic acid) (Figure 7).



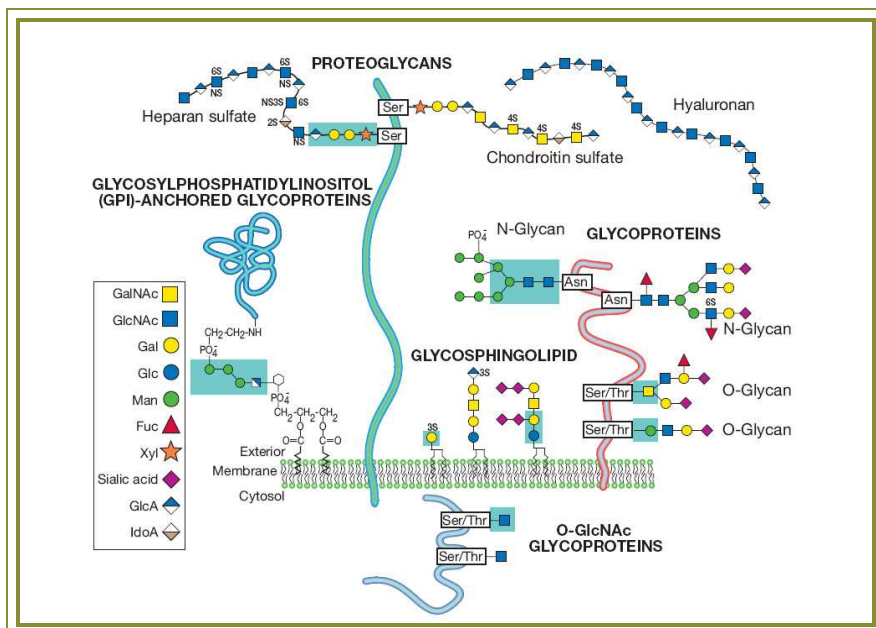
**Figure 7.** Ring forms of common animal monosaccharides

Glycoconjugates are the combination of glycans with other macromolecules like proteins (giving place to glycoproteins) or lipids (giving place to glycolipids). Many different classes of glycoconjugates can be found in eukaryotic cells. Traditionally, these classes have been defined accordingly to the anchor between the glycan and the non-glycan (protein or lipid) units. Thus, glycoproteins are polypeptides carrying glycans covalently linked to their backbone usually with N- or O- linkage. N-glycans are characterized by N-linkage, a covalent union between an N-acetylglucosamine of the glycan chain and an asparagine residue of the polypeptide. O-glycans are characterized by the linkage of an N-acetylgalactosamine of the glycan with a serine or threonine residue. There are different types of cores constituted by O-glycans, like mucins that are long and dense O-glycan chains. There are other types of glycoproteins different that N- and O-glycans, like C-glycans and P-glycans. In the former the glycan is attached via mannose to a triptophan residue and in the latter the 1-phosphate-N-acetylglucosamine is attached to a serine. Proteoglycans are glycoproteins

constituted by chains of glycosaminoglycans (GAGs) anchored through a xilose to the serine residue of the polypeptide.

Moreover, glycosphingolipids or glycolipids (synonymous terms) are glican chains united via glucose or galactose to the primary hidroxil group of a ceramide, a lipid composed by sphingosine and fatty acid (Figure 8).

The surface of all cell types is completely covered by a layer of different kinds of glycoconjugates, named glycocalyx.



**Figure 8.** Common classes of animal glycans.

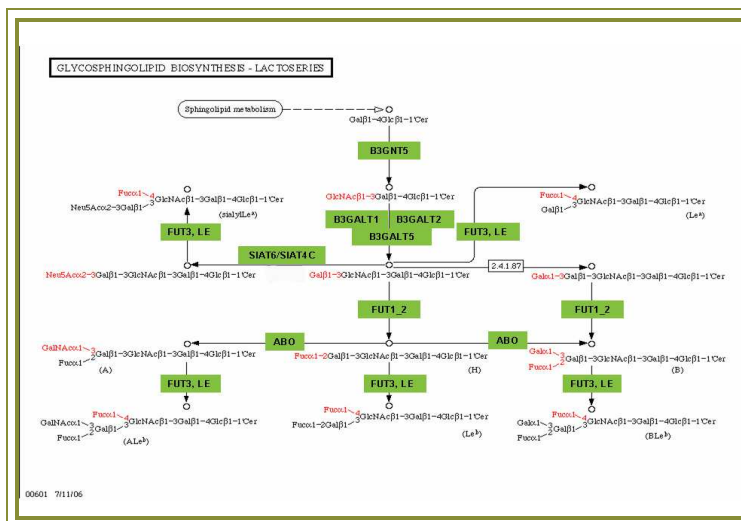
#### 1.4.5.2 Gycan biosynthesis

The biosynthesis of glycans is carried out by glycosyltransferases. They assemble monosaccharide units into linear or branched glycan chains. Glycosyltransferases constitute a very large family of enzymes, representing 1-2% of the genome. They require from a donor (ex. Dilochol-phosphate, GDP, UDP and CMP) carrying the

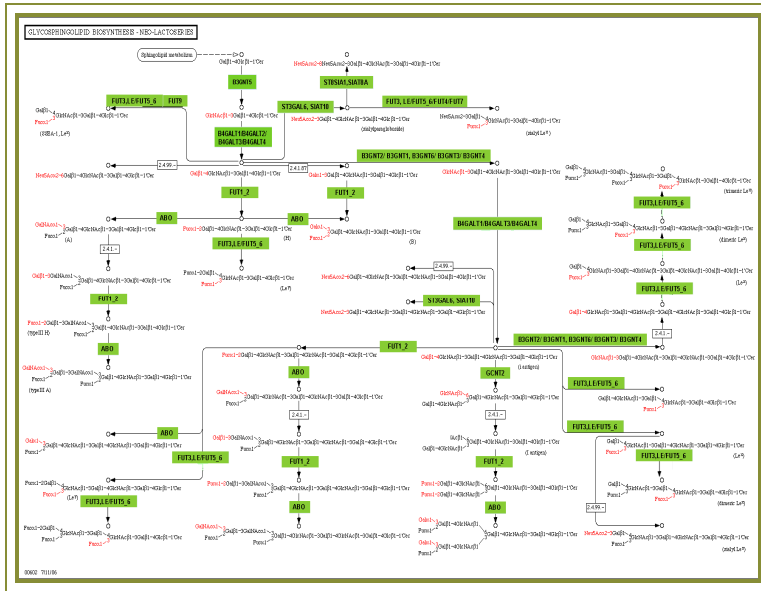
sugar and an acceptor substrate (oligosaccharides, monosaccharides, lipids, proteins, and even DNA) to transfer the sugar to. Their specificity depends on such sugar donors and acceptors. In general each glycosyltransferase recognize one donor and one acceptor and perform a specific linkage, though different enzymes can catalyze the same reaction, as is the case of fucosyltransferases III-VII, all of which attach fucose in  $\alpha$ 1-3 linkage to an N-acetylglucosamine. Seldom can a single enzyme catalyze more than one reaction.

The combined and ordered sequential action of glycosidases (enzymes that remove monosaccharides) and glycosyltransferases in mammalian cell compartments leads to the addition of glycans to proteins and lipids (Gerber-Lemaire and Juillerat-Jeanneret 2006). Glycosylation pathways are then the sequential addition of monosaccharides to form the different glycan molecules: O-glycan biosynthesis; N-glycan biosynthesis; Chondroitin; Heparan and Keratan sulfate biosynthesis; Lipopolysaccharide biosynthesis; Peptidoglycan biosynthesis; and Glycosphingolipid biosynthesis. In the present work we analyse four glycan biosynthesis pathways (Figure 9).

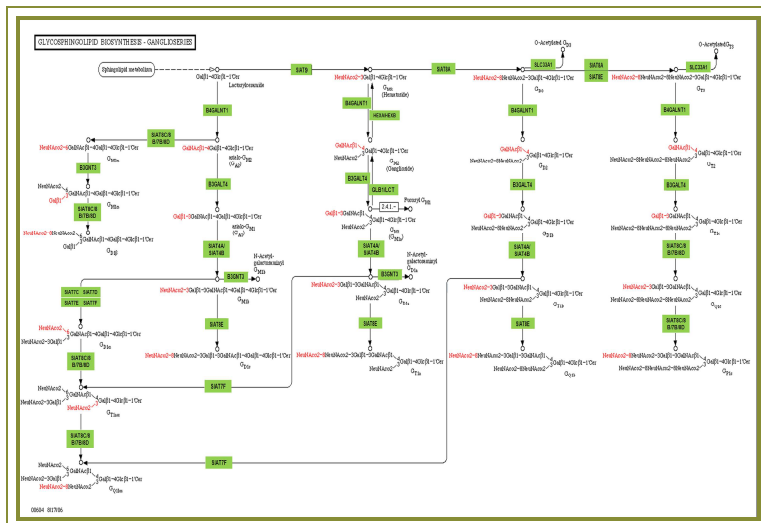
a)



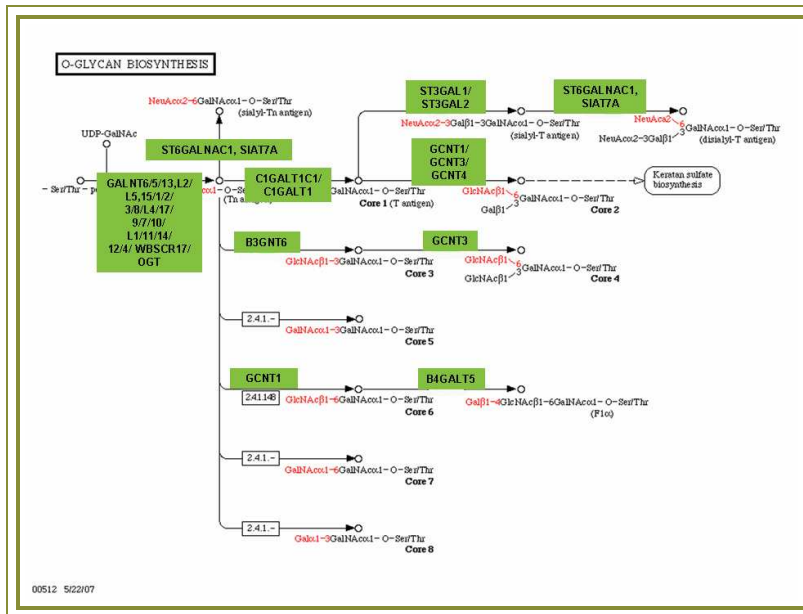
b)



c)



d)



**Figure 9.** From top to the bottom the four pathways analyzed; a, Lactoseries biosynthesis; b, neo-lactoseries biosynthesis; c, ganglioseries biosynthesis and d, O.-glycans biosynthesis. Pathways have been retrieved from KEGG (Kyoto Encyclopedia for Genes and Genomes) bioinformatic resource (<http://www.genome.jp/kegg/>). In green boxes the name of the enzymes; (/) separating enzymes when more than one operate in the same step, and (,) separating same enzyme aliases.

### 1.4.5.3 Fucosyltransferases

Fucosyltransferases are typically transferring fucose from a GDP-donor to N-linked type complex glycopeptides. Fucose is frequently found as a terminal modification of these branched chain glycoconjugates. Fucosyltransferases are also capable of adding fucose directly to serine or threonine residues via O-glycosidic linkage, as it happens on epidermal growth factor or thrombospondin repeats. In humans, the fucosyltransferase family is constituted by 11-13 members. Some of them (*FUT1*, *FUT2*, *FUT6* and *FUT7*) are Golgi stack membrane proteins, whereas others as O-fucosyltransferase (*FUT12* and *FUT13*)



are soluble enzymes in the endoplasmatic reticulum (ER). Fucosyltransferases are not very specific since some of them use the same acceptor to do the same linkage, as it is the case of fucosyltransferases III-VII which all attach a fucose in  $\alpha$ -1,3-linkage to N-acetyllactosamine glycan moieties. They have a broad range of functions: catalyzing the last step of Lewis antigen biosynthesis (*FUT3*), having a role during the first 5 to 10 weeks of development (*FUT4*) and in ligand-induced receptor signaling (O-fucose glycans).

#### **1.4.5.4 Sialyltransferases**

Sialyltransferases are the enzymes responsible of the addition of sialic acids on the terminating branches of N-glycans, O-glycans and glycosphingolipids. This family of enzymes is constituted by around 50 members, whose expression is tissue-specific and developmentally regulated. Thus, different cell types of an organism express different sialic acids. Even the same glycoconjugate can present different sialic acid modifications according to the position of that particular sialic acid within the glycan. Sialic acids are abundant on outer cell membranes, on the interior of lysosomal membranes and on secreted glycoproteins suggesting they have an important role in the stabilization of molecules and membranes and in modulating the interaction with environment. So far, some of the most important functions attributed to them are: enhancement of the viscosity of mucins, protection of molecules and cells to the attack of glycosidases or proteases, regulation of the affinity of receptors and masking antigenic sites to protect cells to macrophage degradation. It has been postulated that the species-specific variation of sialic acid linkages and modifications might be the signature of evolutionary history of species in relation to sialic-binding preferences of its pathogens.

#### **1.4.5.5 Galactosyltransferases**

Galactosyltransferases are the enzymes adding an N-acetylgalactosamine (GalNAc) moiety to the -OH of serine or threonine residue of polypeptides, via O-glycan bond. At least, there exist 21 different galactosyltransferases, All them recognizing the same donor (UDP-GalNAc), but transferring the GalNAc moiety to

different acceptor substrates. Galactosyltransferases (ppGalNAcT) seem to work hierarchically, so one ppGalNAcT attaches a GalNAc to the serine or threonine residue only if the adjacent serine or threonine has been previously glycosylated by a different ppGalNAcT. Immunological studies have demonstrated some ppGalNAcTs localize to the cis-Golgi in submaxillary glands, but ppGalNAcTs can also be present in later compartments of Golgi. The subcellular localization of ppGalNAcTs, plus the presence of other glycosyltransferases, critically determines the range of O-glycans synthesized by a cell. ppGalNAcTs expression levels varies considerably between cell types and mammalian tissues. There are different types of O-glycans according to both the linkage type ( $\alpha$ -linkage when the configuration between the anomeric carbon and the further stereogenic center is the same or  $\beta$ -linkage when it is different) and the type of sugar that is transferred (glucose, xylose, galactose...). Thus, the different types of O-glycans include:  $\alpha$ -linked O-fucose,  $\beta$ -linked O-xylose,  $\alpha$ -linked O-mannose,  $\beta$ -linked O-GlcNAc,  $\alpha$ - or  $\beta$ -linked O-galactose and  $\alpha$ - or  $\beta$ -linked O-glucose and  $\alpha$ -linked O-GalNAc. An example of proteins heavily O-glycosylated via the last mention linkage are mucins. These proteins are located in mucous secretions, on epithelial cell surfaces and in body fluids, making them a shield against physical and chemical damage and protecting cells against pathogen infections.

#### **1.4.6 Blood groups**

Many blood groups are formed by the action of glycosylation enzymes. This is the case of ABO, which was characterized at the beginning of this century (Landsteiner 1900). From then on, up to 30 major blood group systems have been described, including ABO. Apart from these major blood groups there are more than two hundred minor blood groups types, which are also known as rare blood groups. Blood groups are defined as single gene locus or two or more very closely linked homologous genes with little or no observable recombination between them that control the expression of one or more antigens. Thus, the classification of blood group depends on the absence or presence of inherited antigens on the surface of red blood cells. These antigens may be proteins, carbohydrates,

glycoproteins, or glycolipids that can be present on cell surfaces other than red blood cells of various tissues.

#### **1.4.6.1 ABO blood group**

ABO was the first blood group ever described (Landsteiner 1900). Moreover, it was also the first human genetic system where human genetic studies were applied. That revealed the variation in A, B and O allele frequencies among populations (Mourant 1954). From these three alleles, two (A and B) are codominant, whereas O is a recessive silent allele. The combination of these alleles gives place to four different phenotypes: A, B, AB and O, which determine the presence or absence of A and B antigens on the red blood cells surface and in turn the absence of antibodies against the expressed antigen/s in serum. Both, A and B alleles, code for glycosyltransferases that add N-acetylgalactosamine and galactose respectively to the H substrate (see below). In contrast, the O allele presents a deletion of a G at position 261 in exon 6. This deletion ( $\Delta 261$ ) induces a frameshift and gives place to a premature stop codon that truncates the protein making it unable to develop its glycosylation activity.

ABO is a high polymorphic locus, possibly one of the higher polymorphic loci in the human genome. Currently, more than 70 different alleles have been described among the three major antigenic classes (A, B and O). Polymorphisms at this locus seem to be advantageous in relation to infectious diseases. Specifically, various infectious agents express A and B antigens, whereas some other infectious agents use A and B human substances as receptors. That is, individuals not expressing A or B are producing anti-A or anti-B natural antibodies which potentially will protect the individual from pathogens expressing A and B motifs. Thus, individuals homozygous for the O allele are protected against pathogens using A and B antigens as receptors and at the same time are producing anti-A and anti-B antibodies which protect them against infectious agents expressing A and B antigens. An example of that is the protection against malaria conferred by  $\Delta 261$  allele (Fry et al. 2008). Although O-homozygotes are resistant to some

infectious diseases, they have been shown to be susceptible to others like *Helicobacter pylori* and some severe forms of cholera.

#### 1.4.6.2 Pseudogenes

ABO belongs to the glycosyltransferase 6 (GT6) family, its members transfer a galactose or an N-acetylgalactosamine to different substrates. In placental mammals, the GT6 family is constituted by seven functional members: *ABO*, Forssman (FS), *iGB3*, *GGTA1*, *GT6m5*, *GT6m6*, and *GT6m4*. These genes show a complex phylogenetic pattern of activity / inactivity with several episodes of gain and loss along the different species (Turcot-Dubois et al. 2007). In humans, *ABO* is the only functional member of this family. The lack of function has been sometimes related to an immunological advantage against pathogens because, as commented before, some pathogens use carbohydrates at the cell surface as receptors (Gagneux and Varki 1999; Sharon 1996). However, after the recent findings describing transcriptional activity across all the genome (Cheng et al. 2005; Hubbard et al. 2007; Zheng et al. 2007) and proving that some transcribed pseudogenes have biological function (Hirotsune et al. 2003; Korneev et al. 1999; Lee 2003) as well as some reported evidences indicating a putative function in some of these pseudogenes as the sequence conservation and widespread tissue expression of *FS* (Xu et al. 1999), we undertook a study concerning human GT6 family pseudogenes to evaluate the intraspecific and interspecific variability pattern of these sequences to elucidate whether they are conserved, since conservation may be indicative of some functional role and a source of genetic diversity by using them in recombination events with functional genes (Balakirev and Ayala 2003; Vargas-Madrado et al. 1995).

#### 1.4.6.3 The ABO precursor gene FUT2

The A, B and O antigens are formed by the sequential action of glycosyltransferases encoded in three loci (*ABO*, *H* and *FUT2/secretor* loci). The pathway of *ABO* blood group antigen synthesis begins with the modification of type-1 N-acetyllactosamine (galactose added in  $\beta$ 1-3 linkage to GlcNAc) and type-

2 N-acetyllactosamine (galactose added in  $\beta$ 1-4 linkage to GlcNAc). This modification consists on transferring a fucose to type-1 or type-2 N-acetyllactosamine forms by  $\alpha$ -1.2-fucosyltransferases. Such transfer gives place to H determinant. The human genome contains two different  $\alpha$ -1.2-fucosyltransferases; one encoded by the *H* and the other by the *FUT2/Se* blood group loci. The former is expressed in red cells and transfers fucose in type-2 and type-4 glycan units from H antigen; the latter is expressed in epithelial cells and transfers a fucose in type-1 and type-3 N-acetyllactosamines from the H determinant. The second and last step of *ABO* antigen synthesis is getting the A and B determinants. These antigens are formed by the subsequently action of glycosyltransferases, encoded in *ABO*, on type-1, type-2, type-3 and type-4 H determinants. Specifically, the antigen A is formed by the action of  $\alpha$ -1,3GalNAcT encoded in the A allele of the *ABO* blood group locus, whereas the B antigen is formed by  $\alpha$ -1,3GalT encoded in the B allele.

*ABO* antigens not only locate at the surface of red blood cells but also on other cells in many tissues, as in the vascular endothelium and in epithelium. And, epithelial cells from the urinary, digestive, respiratory and reproductive tract and epithelia cells of some salivary and other exocrine glands can secrete *ABO* antigens. This is a genetically determined trait depending on *Se* locus, since in these tissues *H* gene is not expressed.

Some individuals can not express *ABO* antigens in secretions and body fluids because of a null-allele disrupting the transcription of *FUT2*. These individuals, named non-secretors, are homozygotes for such null-allele. Until now two null-alleles have been described as the most common cause of the non-secretor status; the non-functional alleles *se*<sup>428</sup> and *se*<sup>385</sup>. The former gives place to a stop codon at position 143 (Trp-ter) whereas the later reduces  $\alpha$ (1,2)fucosyltransferase activity due to a missense mutation at codon 129 (Ile-Phe) (Henry et al. 1996; Koda et al. 1996; Pang et al. 2001; Yu et al. 1995). These mutations have been reported to have different frequencies in different populations in the world. Moreover they have been related to susceptibility to some infections as Norwalk-

like viruses (Marionneau et al. 2005) and HIV-1 (Kindberg et al. 2006). In this thesis we describe the worldwide variation at the sequence level to depict the presence of null-alleles in a global scenario, determine possible presence of other null-alleles and, analyze which selective pressures, probably related with the different pathogenic environments existing in the different geographic areas around the world, have acted on this locus during modern human history.

## <sup>2</sup> MATERIAL AND METHODS

---





The four studies included in the present dissertation have been carried out on data from two natures: genotyping and resequencing. These data have been mainly produced by our selves on HGDP-CEPH panel. Additionally, we have also used data from three public sources; HapMap database, 650K SNPs in HGDP-CEPH data and innate immunity PGA database. HGDP-CEPH panel is a resource of 1064 DNA samples from individuals in different world populations, covering most of the extant human diversity. We have genotyped this sample set for the genes included in three of the works presented in this thesis (chapters 1, 3 and 4), whereas we resequenced it in the case of *FUT2* study (chapter 2). The public data sources have been used for different aims. Thus, HapMap database was used for SNPs selection prior to the genotyping probes design. The genotypes from 650K SNPs in HGDP-CEPH panel have been used to complete the original data: gaps covering, flanking regions enlargement and generating background distributions for several parameters. Finally, the innate immunity PGA database, which contain resequencing data for 133 genes, was used to download the sequences of eight out of the 15 genes related to the innate immune function studied here (chapter 1).

Genotyping data has been uploaded, handled and stored in SNPator, by means of which preliminary tests were performed and the data was phased. The main further calculations have been done by using 1) Arlequin v3.11 software that allowed us to calculate the interpopulation differentiation statistic ( $F_{ST}$ ), 2) DNAsp v4.50.3 by which genetic diversity and neutrality test statistics calculation were computed and, 3) Sweep v1.1 that helped us in the detection of high frequency alleles with long-range linkage disequilibrium (LD) for the primary purpose of detecting evidence of natural selection.

## 2.1 Samples and data

### 2.1.1 The Human Genome Diversity Project (HGDP-CEPH)

Science 12 April 2002:  
Vol. 296. no. 5566, pp. 261 - 262  
DOI: 10.1126/science.296.5566.261b

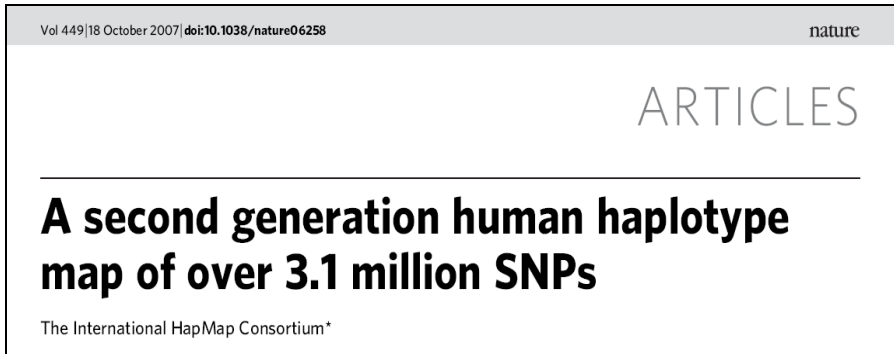
#### A Human Genome Diversity Cell Line Panel

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL.

#### Description:

A resource of 1064 cultured lymphoblastoid cell lines (LCLs) (1) from individuals in different world populations and corresponding milligram quantities of DNA is deposited at the Foundation Jean Dausset (CEPH) (2) in Paris. LCLs were collected from various laboratories by the Human Genome Diversity Project (HGDP) (3) and CEPH to provide unlimited supplies of DNA for studies of sequence diversity and history of modern human populations. Information for each LCL is limited to sex, population, and geographic origin of the individual. Each LCL is registered in a project-specific database and provided with a CEPH-generated, numeric identifier that labels the LCL, its corresponding DNA, and subject information. Every product derived from an LCL is labeled with a unique alpha-numeric bar-code. LCLs were tested for *mycoplasma* infestation by a polymerase chain reaction-based test (4). All lines showing infestation were discarded. LCLs were expanded for storage at -196°C and for production of milligram quantities of DNA. The panel contains LCLs from populations living on all continents. Details of the LCL collections in the resource and a world map showing the 60 positions of the 51 different population samples that contributed blood specimens for production of the cell lines are given in the supplementary material (5).

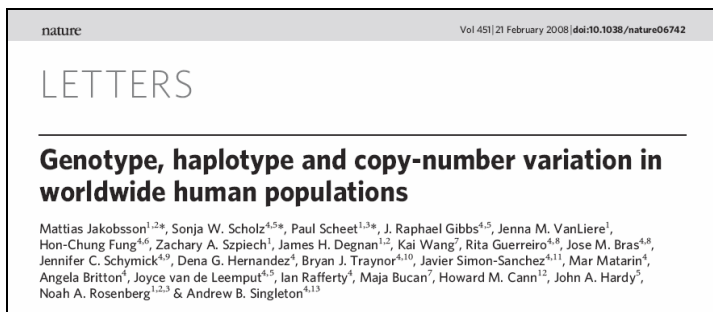
### 2.1.2 International HapMap Project



#### Abstract:

We describe the Phase II HapMap, which characterizes over 3.1 million human single nucleotide polymorphisms (SNPs) genotyped in 270 individuals from four geographically diverse populations and includes 25-35% of common SNP variation in the populations surveyed. The map is estimated to capture untyped common variation with an average maximum  $r^2$  of between 0.9 and 0.96 depending on population. We demonstrate that the current generation of commercial genome-wide genotyping products captures common Phase II SNPs with an average maximum  $r^2$  of up to 0.8 in African and up to 0.95 in non-African populations, and that potential gains in power in association studies can be obtained through imputation. These data also reveal novel aspects of the structure of linkage disequilibrium. We show that 10-30% of pairs of individuals within a population share at least one region of extended genetic identity arising from recent ancestry and that up to 1% of all common variants are untaggable, primarily because they lie within recombination hotspots. We show that recombination rates vary systematically around genes and between genes of different function. Finally, we demonstrate increased differentiation at non-synonymous, compared to synonymous, SNPs, resulting from systematic differences in the strength or efficacy of natural selection between populations.

### 2.1.3 650K SNPs in HGDP-CEPH



#### Abstract:

Genome-wide patterns of variation across individuals provide a powerful source of data for uncovering the history of migration, range expansion, and adaptation of the human species. However, high-resolution surveys of variation in genotype, haplotype and copy number have generally focused on a small number of population groups. Here we report the analysis of high-quality genotypes at 525,910 single-nucleotide polymorphisms (SNPs) and 396 copy-number-variable loci in a worldwide sample of 29 populations. Analysis of SNP genotypes yields strongly supported fine-scale inferences about population structure. Increasing linkage disequilibrium is observed with increasing geographic distance from Africa, as expected under a serial founder effect for the out-of-Africa spread of human populations. New approaches for haplotype analysis produce inferences about population structure that complement results based on unphased SNPs. Despite a difference from SNPs in the frequency spectrum of the copy-number variants (CNVs) detected—including a comparatively large number of CNVs in previously unexamined populations from Oceania and the Americas—the global distribution of CNVs largely accords with population structure analyses for SNP data sets of similar size. Our results produce new inferences about inter-population variation, support the utility of CNVs in human population-genetic research, and serve as a genomic resource for human-genetic studies in diverse worldwide populations.

---

SCIENCE VOL 319 22 FEBRUARY 2008

## **Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation**

Jun Z. Li,<sup>1,2\*†</sup> Devin M. Absher,<sup>1,2\*</sup> Hua Tang,<sup>1</sup> Audrey M. Southwick,<sup>1,2</sup> Amanda M. Casto,<sup>1</sup> Sohini Ramachandran,<sup>4</sup> Howard M. Cann,<sup>5</sup> Gregory S. Barsh,<sup>1,3</sup> Marcus Feldman,<sup>4†</sup> Luigi L. Cavalli-Sforza,<sup>1†</sup> Richard M. Myers<sup>1,2†</sup>

Abstract:

Human genetic diversity is shaped by both demographic and biological factors and has fundamental implications for understanding the genetic basis of diseases. We studied 938 unrelated individuals from 51 populations of the Human Genome Diversity Panel at 650,000 common single-nucleotide polymorphism loci. Individual ancestry and population substructure were detectable with very high resolution. The relationship between haplotype heterozygosity and geography was consistent with the hypothesis of a serial founder effect with a single origin in sub-Saharan Africa. In addition, we observed a pattern of ancestral allele frequency distributions that reflects variation in population dynamics among geographic regions. This data set allows the most comprehensive characterization to date of human genetic variation.

#### 2.1.4 PGA Innate Immunity

<http://innateimmunity.net/>

##### **Innate Immunity in Heart, Lung and Blood Disease**

##### **Programs for Genomic Applications**

#### Description:

The innate immunity PGA (IIPGA) is a collaboration between the Respiratory Sciences Center at the University of Arizona, The Respiratory and Genetics Research Group at the Channing Laboratory at Brigham and Women's hospital and The Boston Children's Hospital Informatics Program (CHIP). The IIPGA is funded as part of the National Heart, Lung, and Blood Institutes (NHLBI) Programs for Genomic Applications (PGA).

The goal of the IIPGA is to discover and model the associations between nucleotide sequence variations, primarily Single Nucleotide Polymorphisms (SNPs) and Insertion/Deletion polymorphisms (Indels), in the genes of the innate immunity pathway in humans. The Program is two-fold: Part 1 identifies the common variable sites through sequencing and establishes their relative allele frequencies and haplotypes in four human populations having different evolutionary histories. Part two utilizes this data in case-control association studies of airways disease (i.e., asthma and chronic obstructive pulmonary disease). We have available to us a full array of bioinformatics tools for performing association analysis in human genetics. These tools and resources are important for any association study relating innate immune responses to disease in a whole host of complex common human disorders such as: airways disease, myocardial infarction, hypertension, stroke, Crohn's Disease, allergy, and diabetes to name a few.

### 2.1.5 SNPLex genotyping

SNPLex™ Genotyping System

#### Description

The SNPLex™ Genotyping System enables the simultaneous genotyping of up to 48 SNPs (single nucleotide polymorphisms) against a single biological sample. This system is ideal for fine mapping and candidate gene analysis, population stratification and microarray replication studies.

The system enables to 1) Streamline and improve multiplex design by using genomic screening capabilities for additional organisms such as bovine, canine, influenza, chicken, chimpanzee, corn, rice, and more; 2) Run assays with confidence -SNPLex Design Pipeline v3.0 improves assay conversion, design and the experimental success rate; 3) Analyze SNPs or insertions and deletions, up to 6 bases in length; 4) Reduce costs by selecting and designing SNPLex assays for only the SNPs you need. 5) Get faster results with our improved master mix and an abbreviated assay workflow. 6) Customize sample throughput - choose manual pipetting or liquid-handling robotics to perform hundreds to 450,000 genotypes per day. 6) Perform linkage mapping using informative SNP clusters spaced by genetic distance with the SNPLex™ Linkage Mapping Set.

The SNPLex™ Genotyping System uses Oligonucleotide Ligation Assay/PCR technology for allelic discrimination and ligation product amplification. Genotype information is then encoded into a universal set of dye-labeled, mobility modified fragments, called Zipchute™ Mobility Modifiers, for rapid detection by capillary electrophoresis. The same set of Zipchute™ Mobility Modifiers are used for every SNPLex™ pool, regardless of which SNPs are chosen, you can achieve extremely reproducible detection and identification of assay products.

## 2.2 Analysis tools

### 2.2.1 SNPator

**BIOINFORMATICS APPLICATIONS NOTE** Vol. 24 no. 14 2008, pages 1643–1644  
doi:10.1093/bioinformatics/btn241

*Genetics and population analysis*

#### **SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data**

Carlos Morcillo-Suarez<sup>1,2,3</sup>, Josep Alegre<sup>1,2,3</sup>, Ricardo Sangros<sup>1,2,3</sup>, Elodie Gazave<sup>1</sup>, Rafael de Cid<sup>2,4</sup>, Roger Milne<sup>2,5</sup>, Jorge Amigo<sup>2,6</sup>, Anna Ferrer-Admetlla<sup>1</sup>, Andrés Moreno-Estrada<sup>1</sup>, Michelle Gardner<sup>1</sup>, Ferran Casals<sup>1</sup>, Anna Pérez-Lezaun<sup>1,2</sup>, David Comas<sup>1,7</sup>, Elena Bosch<sup>1,7</sup>, Francesc Calafell<sup>1,7</sup>, Jaume Bertranpetit<sup>1,2,7</sup> and Arcadi Navarro<sup>1,2,3,7,8,\*</sup>

#### Abstract:

Single nucleotide polymorphisms (SNPs) are the most widely used marker in studies to assess associations between genetic variants and complex traits or diseases. They are also becoming increasingly important in the study of the evolution and history of humans and other species. The analysis and processing of SNPs obtained thanks to high-throughput technologies imply the time consuming and costly use of different, complex and usually format-incompatible software. SNPator is a user-friendly web-based SNP data analysis suite that integrates, among many other algorithms, the most common steps of a SNP association study. It frees the user from the need to have large computer facilities and an in depth knowledge of genetic software installation and management. Genotype data is directly read from the output files of the usual genotyping platforms. Phenotypic data on the samples can also be easily uploaded. Many different quality control and analysis procedures can be performed either by using built-in SNPator algorithms or by calling standard genetic software. AVAILABILITY: Access is granted from the SNPator webpage <http://www.snpator.org>.



### 2.2.2 Arlequin

Evolutionary Bioinformatics Online 2005:1 47-50

APPLICATION NOTE

#### **Arlequin (version 3.0): An integrated software package for population genetics data analysis**

Laurent Excoffier, Guillaume Laval, Stefan Schneider


#### Abstract:

Arlequin ver 3.0 is a software package integrating several basic and advanced methods for population genetics data analysis, like the computation of standard genetic diversity indices, the estimation of allele and haplotype frequencies, tests of departure from linkage equilibrium, departure from selective neutrality and demographic equilibrium, estimation or parameters from past population expansions, and thorough analyses of population subdivision under the AMOVA framework. Arlequin 3 introduces a completely new graphical interface written in C++, a more robust semantic analysis of input files, and two new methods: a Bayesian estimation of gametic phase from multi-locus genotypes, and an estimation of the parameters of an instantaneous spatial expansion from DNA sequence polymorphism. Arlequin can handle several data types like DNA sequences, microsatellite data, or standard multilocus genotypes. A Windows version of the software is freely available on <http://cmpg.unibe.ch/software/arlequin3>.

### 2.2.3 DnaSP

**BIOINFORMATICS APPLICATIONS NOTE** Vol. 19 no. 18 2003, pages 2496–2497  
DOI: 10.1093/bioinformatics/btg359

---



***DnaSP, DNA polymorphism analyses by the coalescent and other methods***

*Julio Rozas<sup>1,\*</sup>, Juan C. Sánchez-DelBarrio<sup>2,†</sup>, Xavier Messeguer<sup>2</sup>  
and Ricardo Rozas<sup>1</sup>*

**Abstract:**

**SUMMARY:** DnaSP is a software package for the analysis of DNA polymorphism data. Present version introduces several new modules and features which, among other options allow: (1) handling big data sets (approximately 5 Mb per sequence); (2) conducting a large number of coalescent-based tests by Monte Carlo computer simulations; (3) extensive analyses of the genetic differentiation and gene flow among populations; (4) analysing the evolutionary pattern of preferred and unpreferred codons; (5) generating graphical outputs for an easy visualization of results. **AVAILABILITY:** The software package, including complete documentation and examples, is freely available to academic users from: <http://www.ub.es/dnasp>

### 2.2.4 Phase

*Am. J. Hum. Genet.* 68:978–989, 2001

#### **A New Statistical Method for Haplotype Reconstruction from Population Data**

Matthew Stephens,<sup>1,3</sup> Nicholas J. Smith,<sup>2</sup> and Peter Donnelly<sup>1</sup>

Departments of <sup>1</sup>Statistics and <sup>2</sup>Biochemistry, University of Oxford, Oxford; and <sup>3</sup>Department of Statistics, University of Washington, Seattle

#### Abstract:

Current routine genotyping methods typically do not provide haplotype information, which is essential for many analyses of fine-scale molecular-genetics data. Haplotypes can be obtained, at considerable cost, experimentally or (partially) through genotyping of additional family members. Alternatively, a statistical method can be used to infer phase and to reconstruct haplotypes. We present a new statistical method, applicable to genotype data at linked loci from a population sample, that improves substantially on current algorithms; often, error rates are reduced by > 50%, relative to its nearest competitor. Furthermore, our algorithm performs well in absolute terms, suggesting that reconstructing haplotypes experimentally or by genotyping additional family members may be an inefficient use of resources

**A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase**

Paul Scheet and Matthew Stephens

**Abstract:**

We present a statistical model for patterns of genetic variation in samples of unrelated individuals from natural populations. This model is based on the idea that, over short regions, haplotypes in a population tend to cluster into groups of similar haplotypes. To capture the fact that, because of recombination, this clustering tends to be local in nature, our model allows cluster memberships to change continuously along the chromosome according to a hidden Markov model. This approach is flexible, allowing for both "block-like" patterns of linkage disequilibrium (LD) and gradual decline in LD with distance. The resulting model is also fast and, as a result, is practicable for large data sets (e.g., thousands of individuals typed at hundreds of thousands of markers). We illustrate the utility of the model by applying it to dense single-nucleotide-polymorphism genotype data for the tasks of imputing missing genotypes and estimating haplotypic phase. For imputing missing genotypes, methods based on this model are as accurate or more accurate than existing methods. For haplotype estimation, the point estimates are slightly less accurate than those from the best existing methods (e.g., for unrelated Centre d'Etude du Polymorphisme Humain individuals from the HapMap project, switch error was 0.055 for our method vs. 0.051 for PHASE) but require a small fraction of the computational cost. In addition, we demonstrate that the model accurately reflects uncertainty in its estimates, in that probabilities computed using the model are approximately well calibrated. The methods described in this article are implemented in a software package, fastPHASE, which is available from the Stephens Lab Web site.

Sweep

*Nature* 419, 832-837 (24 October 2002)

## **Detecting recent positive selection in the human genome from haplotype structure**

**Pardis C. Sabeti<sup>\*,†,‡</sup>, David E. Reich<sup>\*</sup>, John M. Higgins<sup>\*</sup>  
Haninah Z. P. Levine<sup>\*</sup>, Daniel J. Richter<sup>\*</sup>, Stephen F. Schaffner<sup>\*</sup>,  
Stacey B. Gabriel<sup>\*</sup>, Jill V. Platko<sup>\*</sup>, Nick J. Patterson<sup>\*</sup>, Gavin J. McDonald<sup>\*</sup>,  
Hans C. Ackerman<sup>‡</sup>, Sarah J. Campbell<sup>‡</sup>, David Altshuler<sup>\*,§</sup>,  
Richard Cooper<sup>||</sup>, Dominic Kwiatkowski<sup>‡</sup>, Ryk Ward<sup>†</sup> & Eric S. Lander<sup>\*,¶</sup>**

Abstract:

The ability to detect recent natural selection in the human population would have profound implications for the study of human history and for medicine. Here, we introduce a framework for detecting the genetic imprint of recent positive selection by analysing long-range haplotypes in human populations. We first identify haplotypes at a locus of interest (core haplotypes). We then assess the age of each core haplotype by the decay of its association to alleles at various distances from the locus, as measured by extended haplotype homozygosity (EHH). Core haplotypes that have unusually high EHH and a high population frequency indicate the presence of a mutation that rose to prominence in the human gene pool faster than expected under neutral evolution. We applied this approach to investigate selection at two genes carrying common variants implicated in resistance to malaria: G6PD and CD40 ligand. At both loci, the core haplotypes carrying the proposed protective mutation stand out and show significant evidence of selection. More generally, the method could be used to scan the entire genome for evidence of recent positive selection.



## **3 OBJECTIVES**

---





**OBJECTIVES**

The general goal of the present thesis is the understanding of the evolution pattern of two groups of genes, both of them encoding molecules used in the interaction of pathogens and hosts, through the study of their genetic diversity in different widely distributed human populations. In both cases the grouping of genes has been based on either function or functional pathway approach.

Below we enumerate the specific objectives for each chapter of the manuscript:

**Innate immunity genes**

Firstly, we were interested in identifying which selective pressures (if any) were shaping innate immunity genes evolution, hampering the long standing idea of very high conservation of innate immunity.

Secondly, our objective was elucidating whether selection was acting homogeneously across all innate immunity studied genes or, otherwise, there were any gene or group of genes more prone to be under selective pressures exerted by invader organisms.

Finally, we were interested in determining whether these selective pressures were likely to affect some geographic regions at a global scale and thus describing which human populations presented signatures of selection.

**Genes related to glycan biosynthesis**

The evolutionary study of this genetic category was undertaken through three different approximations.

Initially we analyzed the global pattern of variation of a single glycosylation gene; *FUT2*, which had presented previous evidences of selection in diverse human populations to: 1) state the nature of the selection forces driving this locus

evolution, 2) determine whether there existed more alleles than the ones previously described in literature that could also explain its particular secretor/non-secretor phenotype and 3) determine the geographic distribution of the non-secretor phenotype.

Furthermore, we studied a group of pseudogenes belonging to a specific glycosylation family to evaluate whether they were under any selective force other than neutrality and to disentangle their possible biological role in the human genome, where they have been proposed to be inactivated.

Finally, we were interested in determining the possible selective pressures acting on a large set of genes (70) involved in four glycan biosynthesis pathways with the specific purposes of: 1) disentangling the importance that the fact of belonging to a specific pathway can have for the evolution of a gene and 2) elucidating if evolution operates differently on a gene in relation to the step (within a biosynthetic pathway) in which it plays its role. Moreover, we also were interested in understanding which populations were/have been more constrained by selective forces acting on these genes and which possible agents could be responsible of that.

The last and most general objective was evaluating whether the description of evolutionary signatures accordingly to biological context (in functional pathway terms) could potentially contribute to a better understanding of evolutionary events and link it to the process of differential adaptation among human populations.

## 4 RESULTS

---



**4.1 Chapter 1**

Balancing Selection Is The Main Force Shaping the Evolution Of  
Innate Immunity Genes

- This work was published at The Journal of immunology on July 2008 -



Ferrer-Admetlla A, Bosch E, Sikora M, Marquès-Bonet T, Ramírez-Soriano A, Muntasell A, Navarro A, Lazarus R, Calafell F, Bertranpetit J, Casals F. *Balancing selection is the main force shaping the evolution of innate immunity genes.* J Immunol. 2008 Jul 15;181(2):1315-22.

## 4.2 Chapter 2

### A natural history of *FUT2* polymorphism in humans

- This work was submitted for consideration at Journal of Molecular Biology and Evolution on November 2008 –





## A natural history of *FUT2* polymorphism in humans

Ferrer-Admetlla, A.<sup>1</sup>, M. Sikora<sup>1</sup>; H. Laayouni<sup>1,2</sup>, A. Esteve<sup>1</sup>, F. Roubinet<sup>3</sup>, A. Blancher<sup>4</sup>, F. Calafell<sup>1,2</sup>, J Bertranpetit<sup>1,2</sup> and F. Casals<sup>1,5</sup>

1 Institut de Biologia Evolutiva (CSIC-UPF), CEXS-UPF-PRBB,

Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain

2 CIBER Epidemiología y Salud Pública (CIBERESP)

3 Etablissement Français du sang Centre Atlantique, BP 52009, 37020 Tours, Cedex 1, France

4 Laboratoire d'Immunogénétique Moléculaire (EA3034, IFR30) ; Université de Toulouse (UPS) ; Faculté de Médecine de Rangueil ; Bâtiment A2, 31062 Toulouse Cedex 4, France

4 Laboratoire d'Immunogénétique Moléculaire, Faculté de Médecine de Rangueil, Université Paul Sabatier, Bâtiment A2, 31062 Toulouse Cedex 4, France

5 Present address: Centre de Recherche, CHU Sainte-Justine, Université de Montréal, Montréal, Québec H3T 1C5, Canada

Author for correspondence:

Jaume Bertranpetit, IBE, Institut de Biologia Evolutiva (UPF-CSIC), CEXS-UPF-PRBB, Doctor Aiguader, 88, 08003 Barcelona, Catalonia, Spain, Telef. +34.93.316.08.40, Fax. +34.316.09.01, jaume.bertranpetit@upf.edu

*FUT2*, balancing selection, secretor phenotype, non-secretor phenotype, humans, global diversity

Abstract

As pathogens are powerful selective agents, host cell surface molecules used by pathogens as identification signals can be object of adaptation. Most of them are oligosaccharides, synthesized by glycosyltransferases. One known example is balancing selection shaping *ABO* evolution as a consequence of both, A and B antigens being recognized as receptors by some pathogens and anti-A and/or anti-B natural antibodies produced by hosts conferring protection against the numerous infectious agents expressing A and B motifs. These antigens can also be found in tissues other than blood if there is activity of another enzyme, *FUT2*, a fucosyltransferase responsible of *ABO* biosynthesis in body fluids. Homozygotes for null variants in this locus give place to non-secretor phenotype (*se*), and they do not express *ABO* antigens in secretions. Various independent mutations have been shown to be responsible of the non-secretor phenotype, which is coexisting with the secretor phenotype in most populations. Here we resequenced the coding region of *FUT2* in 39 worldwide human populations. We report a complex pattern of natural selection acting on the gene. While frequencies of secretor and non secretor phenotypes are similar, the point mutations at the base of the phenotypes are different, with some variants showing a long history of balancing selection among Eurasian and African, and another a recent and fast spread in East Asian. Thus a convergent phenotype composition has been achieved through different molecular bases with different evolutionary histories. Only the evolutionary molecular analysis allows tracing the natural history of the *FUT2* gene back.

---

## Introduction

*FUT2* gene codes for the  $\alpha(1,2)$ fucosyltransferase responsible for the synthesis of the H antigen, which is the precursor of the ABO histo-blood group antigens in body fluids and on the intestinal mucosa. Several studies have determined that individuals that are homozygous for any non-functional *FUT2* allele fail to present ABO antigens in secretions and on epithelial cells, and they are called non-secretors or *se* individuals, while those individuals carrying at least one functional allele of *FUT2* can express ABO on secretions (Secretors or *Se* individuals). Around 20% of individuals in various populations in the world fail to secrete ABO in body fluids (Soejima et al. 2007).

*FUT2* is 9,980 bp in total length, and it is constituted by 2 exons (of 118 bp and 2,995 bp, respectively) separated by a 6,865 bp intron. While the whole first exon constitutes an untranslated coding region (UTR), the second exon codes for a 343-aa protein that has been extensively studied. Many allelic variants with secretor phenotype have been found across *FUT2* (Hubbard et al. 2007; Koda et al. 2001). The most frequent ones are *Se*<sup>40</sup>, *Se*<sup>375</sup> and *Se*<sup>481</sup> in Xhosa, South-Africa (Liu et al. 1998) and *Se*<sup>357</sup> and *Se*<sup>480</sup> in Xhosas, Ghanaians and Caucasians (Kelly et al. 1995; Liu et al. 1998; Soejima et al. 2007). In total 19 different single nucleotide polymorphisms (SNPs) have been described.

Although many polymorphisms in *FUT2* are population-specific, non-secretor phenotypes are present in most populations (Soejima et al. 2007). Non-secretor phenotypes are caused by mutations in the second exon of *FUT2* gene, with two alleles being the most common cause of the non-secretor status: i) the non-functional allele *se*<sup>428</sup>, which codes for a stop codon at position 143 (Trp-Ter) and is responsible of the non-secretor phenotype in Europeans, Iranians and Africans (Kelly et al. 1995; Liu et al. 1998), and ii) *se*<sup>385</sup>, which is the most frequent cause of non-secretor phenotype in South East and East Asians, due to a reduction of the  $\alpha(1,2)$ fucosyltransferase activity caused by a missense mutation at codon 129 (Ile-Phe) (Henry et al. 1996; Koda et al. 1996; Pang et al. 2001; Yu et al. 1995).

Two other non-secretor alleles appear to have a more restricted geographical distribution:  $se^{302}$  in Thai and Bangladeshi (Chang et al. 1999; Pang et al. 2000) and  $se^{571}$  in Samoans (Pang et al. 2000). Additionally, one deletion ( $se^{778}$ ), two complete deletions of the coding region ( $se^{del}$ ,  $se^{del2}$ ) and one fusion gene ( $se^{fus}$ ) have been reported along it (Soejima and Koda 2008). To the best of our knowledge, the molecular description of variation at *FUT2* has been carried out in one or a few populations, and a global, detailed perspective has not been undertaken.

Some studies have reported the action of balancing selection at *FUT2* on African populations. Neutrality tests based on 18 SNPs in the *FUT2* coding sequence in 121 Ghanaian samples showed an excess of intermediate frequencies, which is indicative of balancing selection (Soejima et al. 2007). Another study, based on genotyping SNPs across 168 genes related to immune function in 3 populations (CEPH Europeans, Han Chinese and Yoruba Nigerians), showed that the allele frequency spectra of SNPs at the *FUT2* gene are skewed towards intermediate frequencies in Yoruba, which is considered to be the result of balancing selection (Walsh et al. 2006). Recently, a putative promoter region of the gene has also been proposed to be under balancing selection in the Yoruba population (Fumagalli et al. 2008). Koda et al. (2001) (Soejima and Koda 2008) estimated a very ancient divergence time between *Se* and  $se^{428}$ , at 3.1 million years (MYA). This divergence time for *FUT2* is in the same range as the estimated for human *ABO* locus (2.7-4.7 MYA); for the latter gene, balancing selection has been claimed to be responsible for its ancient coalescence time (Calafell et al. 2008; Fry et al. 2008; Roubinet et al. 2004; Saitou and Yamamoto 1997).

The possible relationship between *FUT2* alleles and susceptibility to disease has also been extensively studied. The null allele ( $se^{428}$ ) has been shown to confer protection to GGII noroviruses (Norwalk-like virus) infection, which is a major cause of acute gastroenteritis worldwide and has been associated with nosocomial infections and food-borne outbreaks (Larsson et al. 2006; Thorven et al. 2005). It has also been claimed that heterozygous (*Se*/ $se^{428}$ ) individuals are more prone to be infected by Norwalk-like viruses than secretor homozygotes

---

(Se/Se), while non-secretor individuals ( $se^{428}/se^{428}$ ) are relatively resistant to the infection (Marionneau et al. 2005). The null allele  $se^{428}$  has also been strongly associated with slow progression of HIV-1 infection (Kindberg et al. 2006).

In the present work we have resequenced the second exon of *FUT2* on a large number of samples covering most of human variation worldwide. The aim of this work is to describe the geographic variation in sequence, which will allow us to determine the presence of null alleles in a global scenario and to analyze which selective pressures, probably related with the different pathogenic environments existing in the different geographic areas, have acted on this locus throughout human history. To reach this goal, we searched for signatures of natural selection on *FUT2* through: a) the analysis of interpopulation differentiation, b) the phylogenetic relationships among the inferred haplotypes at the continental group level, c) the distribution pattern of the most common secretor and non-secretor haplotypes at the population level and d) the variability at intrapopulation level to check for significant decreases or increases of diversity values from those expected under a neutral evolution model. The aim of this study is to elucidate the evolutionary forces that shaped the genetic variation and function in the *FUT2* gene, including positive or balancing selection in relation to the different pathogenic environments existing in the different geographic areas around the world.

## Materials and methods

### Samples

We sequenced 732 non-related samples from the Human Genome Diversity Panel (HGDP-CEHP) (Cann et al. 2002), after excluding all duplicated individuals and first degree relatives (Cann et al. 2002; Rosenberg 2006). These samples were grouped according to their geographical and ethnical origin into 39 populations to avoid very low sample size and regrouped into 7 continental regions (Europe, Middle East and North Africa, Central and South Asia, East Asia, Oceania, America and Subsaharan Africa) as in (Gardner et al. 2006).

## Sequencing

The second exon of *FUT2* (1,032 bp) was resequenced. The amplification primers (5'-ACACACCCACACTATGCCTGCAC-3' and 5'-ACTTGCAGCCCAACGCATCTT-3') was located at 100bp from both ends of the coding region. A second internal pair of primers (5'-CCAGCTAACGTGTCCCGTTTTCC-3' and 5'-TGCCTCCCTCAAGATGAGTGCC-3'), was located at 13bp downstream and 35bp upstream of the coding region respectively, and were used to sequence the 1,032bp segment. DNA purification was performed with Biomek FX (Beckman Coulter) using Montage Seq 96 Kit from Millipore, and ABI3100 sequencer (Applied Biosystems) was used to read all fragments. Sequences were aligned with SeqMan program of the Lasergene v7.1.0.44 package and revised manually by two independent investigators in order to detect heterozygous positions. Polymorphic positions for all sequences are given in Supplementary Table S1.

## Statistical analysis

The less frequent allele was determined based on the less common allele across all populations (Table S1). Haplotypes were inferred in each population using the Bayesian algorithm in Phase v2.1 software (Stephens et al. 2001) performing 1000 iterations. Diversity statistics and neutrality test were calculated using DnaSP v 4.50.3 (Rozas et al. 2003). For Fu and Li D, Fu and Li F and Fay and Wu tests, the chimpanzee sequence from Ensembl was used as an outgroup. The significance of neutrality tests was calculated by means of coalescent simulations with COSI software (Schaffner et al. 2005), using a model that takes into consideration the demographic history of humans for three reference populations used in HapMap: CEU (Europeans of North and Central Europe), JPT (Japanese from Tokio) and YRI (Yoruba from Nigeria). We performed 10,000 iterations using the local recombination rate estimate obtained from HapMap (<http://www.hapmap.org/>). Significance in Europe, Middle East and North Africa and Central and South Asia was obtained comparing to the simulations for CEU population; in the case of East Asia, Oceania and America to JPT and Sub-

---

Saharan Africa values to YRI. A median-joining network establishing possible genealogical relationships among haplotypes based on number of substitutions was performed with Network v4.5.0.0 (Bandelt et al. 1999). Also with this program, we estimated the time to the MRCA for the *FUT2* coding region. The substitution rate needed for that calculation was estimated as follows: we used the divergence between the human and chimpanzee sequences ( $K=0.01029$ , with a Jukes-Cantor model), which considering that the separation of the human and chimpanzee lineages dates to ~6 MYA, translates to a substitution rate of  $8.575 \times 10^{-10}$  per site and year. Population differentiation statistics ( $F_{ST}$ ) were calculated with Arlequin v3.11 (Excoffier 2005). Extended Haplotype Homozygosity decay was computed with Sweep software package v1.1 (<http://www.broad.mit.edu/mpg/sweep/index.html>) (Sabeti et al. 2002). For Sweep analysis we used publicly genotype data from whole genome scans obtained in the HGDP samples (Jakobsson et al. 2008; Li et al. 2008).

## Results

### Nucleotide variation

We have sequenced the *FUT2* coding region in 732 individuals belonging to 39 human populations covering most human population diversity. Full sequence results are shown in Table S1 (Supplementary material) and the summary information is shown in Table 1. We found a total of 55 single nucleotide polymorphisms (Table S1) in the second exon of *FUT2* (1,032bp). Previous studies had described 19 SNPs and 1 deletion along the region analyzed here (Koda et al. 2001). In this work we report 37 new substitutions. Out of these 37 SNPs, 27 have low minor allele frequency (MAF) < 0.05, most of them (23) being population specific. One of them, the SNP at position 342, is specific of San population with a high MAF (0.20) (Table S1). With the set of samples used here, we could not detect the substitution  $se^{628}$  or the deletion  $se^{778}$  previously described in the literature (Birney et al. 2007; Koda et al. 1996; Liu et al. 1998).



### Interpopulation differentiation analysis

The  $F_{ST}$  statistic was used to calculate the allele frequency differentiation among populations, be it among the 39 populations ( $F_{ST}$ ), among the seven continental groups ( $F_{CT}$ ), and among the populations included in the same continental group ( $F_{SC}$ ). Table 1 shows the values for the 26 SNPs with minor allele frequency over 0.05. As expected, most of the variability is explained by differences among continents ( $F_{CT}$ ) (Barbujani et al. 1997). Results reveal high  $F_{ST}$  values (over 0.20) for eight SNPs. For one of these eight cases, the high  $F_{ST}$  value is mainly explained by allele frequency differences between Sub-Saharan Africa versus the rest of continental groups (SNPs at position 40). In six cases high  $F_{ST}$  values are due to the difference between Europe, Middle East and North Africa, Central-South Asia and Sub-Saharan Africa in relation to the rest of continental groups, mainly East Asia (SNPs at position 171, 216, 385, 428, 739 and 960), a pattern that will be discussed below. And finally, there is one SNP ( $Se^{375}$ ) showing a high MAF in Oceania and not in the rest of groups. Of the eight SNPs presenting the highest  $F_{ST}$  values, four are non-synonymous variants, three of them presenting the highest  $F_{ST}$ , an interesting finding discussed below. Worth to highlight is one of these non-synonymous SNPs,  $se^{385}$ , which presents a very high  $F_{ST}$  (0.39) due to its high minor allele frequency in East Asia (0.44). Furthermore, it is interesting to note the presence of 4 contiguous SNPs with high  $F_{ST}$ , from position 342 to 385.

### Sequence variation

Diversity indexes of the sequence data for each of the 39 populations and results for seven neutrality tests are shown in Table 2. For some populations results are significantly different from those expected under a model of neutral evolution, thus indicating a possible footprint of selection; results for Tajima's D are especially interesting. In particular, Basque and North Italy populations are showing positive significant values for five out of the seven tests, indicating an excess of intermediate frequencies which may have been originated by balancing selection. It is interesting to note that this trend is shown by many of the populations in West Eurasia (and North Africa), even if they are not reaching statistical significance in

---

many cases. Sub-Saharan African presents a more complex pattern: on one hand, populations in West and Central Africa present high and significant values for the neutrality tests (six out seven in Mandenka and four in Biaka Pygmies), while, on the other hand, neutrality tests are negative in other populations (reaching significance in the San but not in Bantu and Mbuti Pygmies), indicating an excess of rare alleles. Thus negative values are found in the eastern and southern African populations.

The empirical distributions of four neutrality tests distributions of 132 genes in European-American and African-American populations (Akey et al. 2004) allow for a comparison with the present results. Considering French and Yoruba populations as proxies, their neutrality tests values fall close to the 95<sup>th</sup> percentile in three distributions (Tajima's D, Fu and Li D\* and Fu and Li F\*). When considering Basques and Mandeka, with higher significance for neutrality tests, *FUT2* values are greater than the 95<sup>th</sup> percentile of the distributions and even exceed the values for *ABO*, a gene already proposed to be under balancing selection (Calafell et al. 2008).

#### Genealogical relationship among haplotypes

Using all sequenced individuals we identified a total of 96 haplotypes in the *FUT2* coding region (Supplementary Table S2, Table 3). To determine the relationship among them we constructed a median-joining network; the chimpanzee sequence was used to root the network. Figure 1 shows the network with relative frequencies and geographic origin. To interpret it according to non-secretor or secretor phenotypes, Figure 2 was constructed, with the same topology as Figure 1 but with all known inactivating substitutions shown instead of geographic origin.

Figure 1 shows that the haplotype structure of *FUT2* is divided into two main groups, and that *se*<sup>428</sup> is one of the polymorphisms that define such groups. This is both a functional and a geographical clustering: the left-hand side of the network contains only non-functional alleles, and chromosomes from the continents where signals of balancing selection were found (namely, West Eurasia and Africa); on

the right-hand side, a cosmopolitan assortment of both functional and non-functional haplotypes can be found. Non-functional haplotypes in the right-hand cluster are the frequent H8 and the much rarer H51 and H53, which are defined by the  $se^{385}$  substitution, and are found in East Asia and derived populations (Oceania and the Americas). Notice the contrast in haplotype diversity between the non-functional carriers of  $se^{428}$  and of  $se^{385}$ ; the latter contrasts not only with the former, but with the star-like structure of the network around its neighbors, H3 and H10. Carriers of  $se^{302}$  (H12 and its derivatives: H33 and H39) are much rare and practically restricted to South and Central Asia.

Figure 3 shows the worldwide distribution of the 4 groups of haplotypes carrying non-secretor alleles ( $se^{302}$ ,  $se^{385}$ ,  $se^{428}$  and  $se^{571}$ ) and the 4 major haplotypes (together with their derived haplotypes) carrying secretor mutations. The worldwide distribution of secretor alleles is mainly explained by haplotypes related H3 (red) and H10 (orange) which are ubiquitous, whereas secretor haplotypes related to H2 (brown) are specific of Europe and Central and South Asia and those related to H17 (yellow) are exclusive of Sub-Saharan Africa. On the other hand, non-secretor haplotypes are frequent in Eurasia and Africa, even if they are produced by different substitutions. Specifically H7, the haplotype carrying the  $se^{428}$  null allele, is the most common null haplotype, being present in half of the Caucasoids and nearly half of most Africans. Notice that haplotypes carrying  $se^{385}$  allele (H8-H51-H53 in dark blue) are exclusive of East Asia, except for the two more eastern Central and South Asian populations (Burusho and North West China, which have received East Asian gene flow) and Melanesians. Haplotypes carrying  $se^{302}$  allele (H12, green) seem to be specific of Central and South Asian populations, although they are slightly represented in Cambodians. Finally, the haplotype carrying  $se^{571}$  allele (H40, cyan) seems to be particular of Cambodian population.

#### Long-range haplotype analysis

Since neutrality tests results tended to indicate the action of positive selection in some East Asian populations, we examined the *FUT2* region for signs of positive

---

selection applying the long range haplotype (LRH) test. For these purpose we analyzed public available SNP data (<http://shgc.stanford.edu/hgdp/index.html>) for the *FUT2* region including SNPs up to 400Kb in both directions from the gene. Data from this locus was compared to data from the same database covering 69 regions related to glycosylation processes (mainly sialylation, fucosylation and galactose tranfering) (Ferrer-Admetlla Aa unpublished data). To detect signals of positive selection on the *FUT2* region, we measured the Extended Haplotype Homozygosity (EHH) versus core haplotype frequency at a fixed length of 0.3cM in both directions from the core haploptype (Sabeti et al. 2002). P-values were significant (<0.05) for 17 core SNPs in four East Asian populations (Yakut, Han, Cambodian and North East China) however none of them remain significant after applying multiple testing correction (q-value = 0.20). Thus, results do not support positive selection acting on this gene in any population.

#### Discussion

Our results indicate that diversity patterns at *FUT2* cannot be explained by neutrality and human demography. The phylogenetic structure of variation, the geographic distribution of variants, the amount of population differentiation and the neutrality tests for widespread human populations show a complex picture that includes natural selection and its different action according to geography. Neutrality tests show highly significant positive Tajima's D values for 12 out of the 39 populations, which mainly belong to four continental groups in West Eurasia and Africa (Europe, Middle East and North Africa, Central and South Asia and Sub-Saharan Africa). Moreover, some of these populations present high significant values for other neutrality test (Table 2). This would suggest balancing selection could be the force governing *FUT2* evolution in these regions, and is in agreement with previous studies (Fumagalli et al. 2008; Soejima et al. 2007; Walsh et al. 2006). The estimation of the time depth of the phylogeny (see materials and methods) gives an age of the MRCA of 2.61-5.27 MYA, which is similar to that estimated for *ABO* (2.7-4.7 MYA), a locus extensively proposed to be under balancing selection (Calafell et al. 2008; Saitou and Yamamoto 1997), and much older than most of the human genome variation.

The interpopulation differentiation statistics ( $F_{ST}$ ) provides more evidence supporting balancing selection in *FUT2*, as well as the existence of different evolutionary forces acting in different continental groups. Although not conclusive (Gardner et al. 2007), high  $F_{ST}$  has also been taken as an indicator of local specific selective pressures, leading to positive selection (Barreiro et al. 2008; Weir et al. 2005). On the other hand, the effect of balancing selection on  $F_{ST}$  is less clear. It has been proposed that balancing selection should decrease the interpopulation differentiation levels (Akey et al. 2002; Weir et al. 2005), as expected if the same allelic variants are maintained in different populations as reported in some innate immunity receptors, and in the *IL10* and *CCR5* genes (Bamshad et al. 2002; Ferrer-Admetlla et al. 2008; Wilson et al. 2006). Our obtained  $F_{ST}$  values reflect the action of different selective forces in different geographic areas. As stated above, the high  $F_{ST}$  values are mainly produced by the very different allele frequencies reported in the East Asia populations. In contrast, if this continent together with America and Oceania are excluded from the analysis, the  $F_{ST}$  values decrease drastically. As an example, the amount of differentiation between continents ( $F_{CT}$ ) for the 428 position, which encodes for the most frequent secretor and non secretor variants, is only 0.044.

In fact, East Asian populations follow a different pattern. It is not trivial to understand why the inactivating African-West Eurasian allele is not present in Asia and thus how a new variant appeared and increased in frequency. The non-secretor phenotype is achieved by another mutation found only in H8 and in two derived haplotypes. Although their frequency is very high in these populations the amount of non-secretor haplotypes is achieved by a different mutation, much more recent than the one inactivating of the H7 and neighbors. It seems as that very similar frequency in all Old World populations has been achieved by two different ways: by a very ancient inactivating mutation that gave rise to a wide range of newly produced neutral haplotypes (within the non-secretor phenotype) spreading in Africa and West Eurasia and a newly produced, recent mutation, that from H3 produced H8 and only two other haplotypes. This inactivating mutation has had a recent origin and a drastic increase in frequency, accounting for around 50% of

---

chromosomes and generating very little haplotype diversity (just two other haplotypes with a single substitution each). Thus, positive selection has to be invoked to explain the increase of the Asian allele. However, our analyses have failed to detect significant signatures of positive selection considering the total *FUT2* variation; in fact this footprint would only affect the parts of the tree as selection has shaped the variation at different times. Finally, this lack of selection signature may be related with first stages of balancing selection in these populations, which would be difficult to recognize.

The main question is the possible meaning of balancing selection for a set of haplotype variation with just two phenotypes and dominance of the secretor one. A plausible explanation for balancing selection might be the already reported beneficial effects for homozygous null-allele individuals. Some works have demonstrated  $se^{428}$  (the null allele carried by H7 haplotype) confers protection against certain pathogens, as the Norwalk-like virus, or that they play a role in slowing the progression of HIV-1 infection (Kindberg et al. 2006; Marionneau et al. 2005). In a recent work this variant has been demonstrated to be in strong linkage disequilibrium with the G allele of  $Se^{171}$  and that women homozygous for the latter had higher  $B_{12}$  levels, suggesting that the non-secretor allele  $se^{428}$  is a plausible mechanism for altered  $B_{12}$  absorption and plasma levels. Recently, several new and old examples of balancing selection exerted by infectious disease have been published, including innate immunity receptors (Ferrer-Admetlla et al. 2008), blood group antigen genes (Fumagalli et al. 2008), or the human major histocompatibility complex (Solberg et al. 2008).

Evolutionary forces have changed in space and time and they have to accommodate adaptation with the already existing variants. An overall human picture would not have given any clue on the natural history of *FUT2*; for example overall Tajima's *D* is low and with a non-significant value (-1.436), whereas it reaches significance (actually, in the opposite direction) for quite a few single populations. Nonetheless, the final adaptation in all Africa and Eurasia seems to have followed a common general pattern through different basic variants. A detailed history of times and place for selection acting on a wide panoply of

genetic elements may be difficult to achieve, but the general trends seem possible to be recovered through the understanding of the molecular variation; at the end it is the fundamental piece of the natural history of the gene and of its phenotypic effects.

#### Acknowledgements

This research was funded by grants BFU2005-00243 and SAF-2007-63171 awarded by Ministerio de Educación y Ciencia (Spain), by the Direcció General de Recerca of Generalitat de Catalunya (Grup de Recerca Consolidat 2005SGR/00608). Funds were also from the Etablissement Français du Sang (EFS) Centre Atlantique, and from the Ministère Français de la Recherche (EA3034). All the sequencing was done at the Genomic Service, Universitat Pompeu Fabra; we thank Stéphanie Plaza and Roger Anglada for their help. Computation was helped by the National Institute for Bioinformatics ([www.inab.org](http://www.inab.org)), and SNP genotyping services were provided by the Spanish "Centro Nacional de Genotipado" (CEGEN; [www.cegen.org](http://www.cegen.org)); both are platforms of Genoma España. A F-A is supported by a PhD fellowship from UPF and MS from the Programa de becas FPU del Ministerio de Educación y Ciencia, Spain (AP2005-3982).

**Table 1.** Minor allele frequency (MAF) and molecular fixation indices (FST) for each polymorphic position with MAF>0.05

Position <sup>a</sup>	Change	Amino acid change	Phenotype <sup>b</sup>	Minor Allele Frequency							Interpopulation Differentiation						
				EUR (n=119) 43.28%	MENA (n=124) 53.63%	CSASIA (n=183) 42.08%	EASIA (n=144) 45.49%	OCE (n=23) 6.52%	AME (n=58) 2.59%	SSAFR (n=81) 26.54%	Total	F <sub>ST</sub> <sup>c</sup>	P-value	F <sub>ST</sub> <sup>d</sup>	P-value	F <sub>ST</sub> <sup>e</sup>	P-value
1	non-syn	M→V	NA	0.008	0.000	0.003	0.000	0.000	0.009	0.012	0.005	0.033	0.003	0.040	0.013	-0.007	0.616
40*	non-syn	I→G	Se	0.000	0.004	0.005	0.007	0.000	0.009	0.288	0.045	0.419	0.000	0.220	0.000	0.255	0.002
107	non-syn	V→G	NA	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.001	0.076	0.000	0.084	0.046	-0.009	0.178
113	non-syn	V→I	NA	0.000	0.000	0.000	0.000	0.000	0.009	0.036	0.006	0.040	0.002	0.005	0.009	0.035	0.005
171*	syn	A→A	Pr, Se	0.378	0.548	0.321	0.007	0.000	0.026	0.220	0.214	0.246	0.000	0.055	0.000	0.203	0.000
216*	syn	Y→Y	Pr, Se	0.437	0.560	0.326	0.007	0.000	0.026	0.220	0.225	0.259	0.000	0.053	0.000	0.218	0.000
221	non-syn	L→P	NA	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.029	0.059	0.038	0.121	-0.009	0.647
278	non-syn	A→V	NA	0.004	0.004	0.016	0.000	0.000	0.000	0.000	0.003	0.012	0.080	0.010	0.094	0.003	0.283
302*	non-syn	L→P	se	0.000	0.000	0.064	0.003	0.000	0.000	0.000	0.010	0.072	0.000	0.026	0.000	0.046	0.004
315*	syn	S→S	Pr, Se	0.000	0.004	0.003	0.000	0.000	0.000	0.012	0.003	0.026	0.021	0.018	0.045	0.008	0.028
342	syn	Q→Q	Pr, Se	0.000	0.000	0.000	0.000	0.000	0.000	0.012	0.002	0.178	0.000	0.206	0.003	-0.035	0.172
357*	syn	N→N	Se	0.479	0.649	0.475	0.201	0.630	0.345	0.687	0.495	0.171	0.000	0.021	0.000	0.154	0.000
375*	syn	E→E	Se	0.000	0.004	0.000	0.000	0.283	0.000	0.012	0.043	0.283	0.000	0.047	0.000	0.247	0.000
385*	non-syn	I→F	se	0.004	0.000	0.041	0.438	0.065	0.000	0.000	0.078	0.391	0.000	0.062	0.000	0.350	0.000
400*	non-syn	V→I	Se	0.000	0.000	0.000	0.000	0.065	0.000	0.000	0.009	0.101	0.001	0.040	0.006	0.064	0.010
428*	non-syn	W→stop	se	0.429	0.560	0.329	0.007	0.000	0.026	0.220	0.224	0.256	0.000	0.050	0.000	0.217	0.000
480*	syn	H→H	Se	0.105	0.085	0.105	0.000	0.000	0.000	0.000	0.042	0.070	0.000	0.027	0.000	0.045	0.002
481*	non-syn	D→N	Se	0.000	0.032	0.003	0.000	0.000	0.000	0.149	0.026	0.147	0.000	0.030	0.000	0.121	0.000
558*	non-syn	G→G	Pr, Se	0.000	0.000	0.000	0.000	0.000	0.000	0.011	0.002	0.034	0.032	0.032	0.035	0.002	0.158
571*	non-syn	R→stop	se	0.000	0.000	0.000	0.007	0.000	0.000	0.000	0.001	0.006	0.213	0.005	0.397	0.001	0.148
616	non-syn	V→I	NA	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.001	0.024	0.097	0.028	0.366	-0.004	0.215
681	syn	I→I	Pr, Se	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.033	0.028	0.037	0.022	-0.004	0.495
739*	non-syn	G→S	NA	0.454	0.637	0.331	0.003	0.000	0.043	0.220	0.241	0.294	0.000	0.055	0.000	0.253	0.000
855	syn	A→A	Pr, Se	0.008	0.000	0.003	0.000	0.000	0.000	0.000	0.002	0.025	0.038	0.031	0.041	-0.006	0.695
954	non-syn	E→D	NA	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.001	0.029	0.036	0.036	0.038	-0.007	0.876
960*	syn	T→T	Pr, Se	0.487	0.452	0.671	0.910	0.957	0.983	0.653	0.730	0.224	0.000	0.053	0.000	0.181	0.000

<sup>a</sup> Positions accordingly to bibliography (Koda et al. 2001); \* SNPs previously described in literature (Koda et al. 2001; Liu et al. 1998; Kelly et al. 1995; Chang et al. 1999; Pang et al. 2000; Yu et al. 1995; Koda et al. 1996; Liu et al. 1999; Peng et al. 1999; Yu et al. 1999). <sup>b</sup> Phenotype defined as Se indicates functional allele, Pr, Se, presumably functional allele, se non-functional allele and NA not available information. <sup>c</sup> F<sub>ST</sub> calculated between the 39 populations, <sup>d</sup> within continents and <sup>e</sup> between continental groups. Frequencies in the header indicate the percentage of null-alleles within each continental region.

SSAFR = Sub-Saharan Africa; MENA = Middle East-North Africa; EUR = Europe; CSASIA = Central-South Asia; EASIA = East Asia; OCE = Oceania; AME = America



# Human Genetic Diversity In Genes Related To Host-Pathogen Interactions

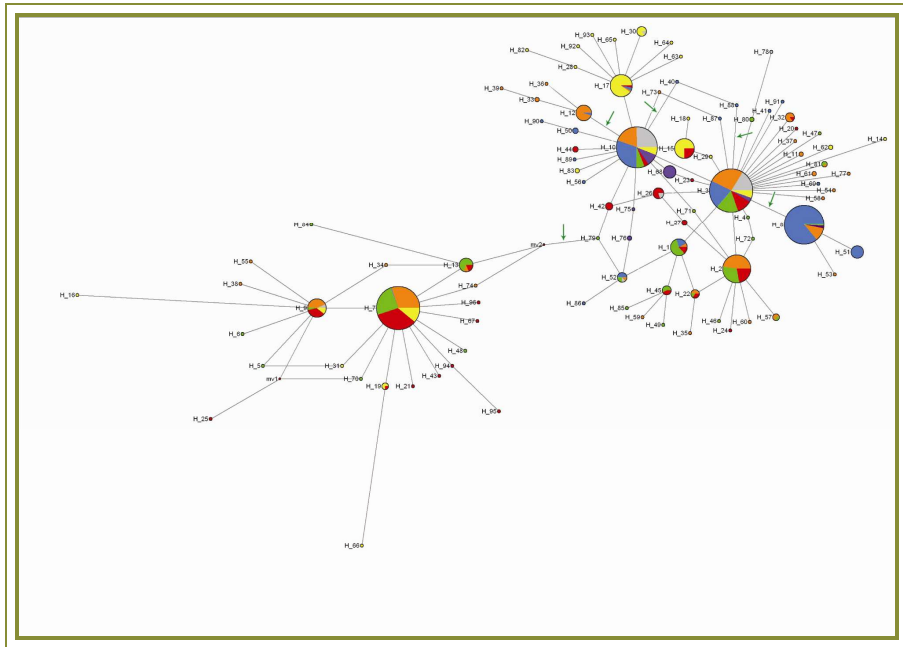
**Table 2.** Diversity statistics and neutrality tests for the 39 populations analyzed.

Continent	Population	n	S	Hd	$\pi$	$\theta$	Tajima's D	Fu&Li D*	Fu&Li F*	Fu&F	Fu&Li D	Fu&Li F	Fay&Wu F
AME	Cdombian	6	6	0.530	0.0013	0.0019	-1.28	-1.71	-1.82	1.28	-0.01	-0.42	-5.18
AME	Naya	21	8	0.813	0.0011	0.0018	-1.17	-0.75	-1.04	-0.97	-0.10	-0.53	-5.33
AME	Sunui	8	1	0.325	0.0003	0.0003	0.16	0.69	0.63	0.55	0.66	0.62	-1.08
AME	Kartiana	11	1	0.485	0.0005	0.0003	1.33	0.64	0.94*	1.39	0.61	0.93	-0.36
AME	Pina	12	3	0.518	0.0006	0.0008	-0.54	-1.36	-1.30	0.37	-0.72	-0.59	-0.38
CSASIA	Balochi	24	9	0.755	0.0032	0.0020	1.79*	0.70	1.24	1.39	0.71	1.27	-0.14
CSASIA	Brakui	23	11	0.853	0.0031	0.0024	0.79	0.31	0.55	-1.63	0.30	0.56	-1.26
CSASIA	Bunusho	18	12	0.797	0.0029	0.0028	0.12	-0.01	0.04	-1.75	0.38	0.44	-1.58
CSASIA	Hazara	21	9	0.747	0.0031	0.0020	1.48*	0.72	1.13	-0.47	0.74	1.17	-0.26
CSASIA	Kalash	17	8	0.702	0.0032	0.0019	1.99*	0.66	1.26	2.37	0.67	1.30	-0.54
CSASIA	Makrani	23	14	0.824	0.0036	0.0031	0.53	-1.30	-0.81	-0.92	-1.45	-0.91	0.49
CSASIA	North_West_China	19	11	0.747	0.0016	0.0025	1.10	-0.70	-0.98	-2.36	-0.81	-1.08	-5.18
CSASIA	Pathan	17	10	0.763	0.0030	0.0024	0.82	0.29	0.54	1.24	0.27	0.54	-0.89
CSASIA	Sindhi	21	11	0.703	0.0033	0.0025	1.04	-0.21	0.23	1.27	-0.26	0.21	0.40
EASIA	Canbodian	9	5	0.621	0.0012	0.0014	-0.42	-1.14	-1.08	-1.75	-1.33	-1.27	-0.65
EASIA	Han	34	4	0.669	0.0009	0.0008	0.27	-0.18	-0.04	-0.01	0.98	0.89	-2.71
EASIA	Japanese	19	4	0.802	0.0012	0.0009	0.79	-0.02	0.25	-0.66	-0.05	0.24	-1.75
EASIA	North_East_China	30	8	0.649	0.0009	0.0017	-1.14	-2.41	-2.35*	-0.93	-0.99	-1.24	-6.39
EASIA	South_China	48	11	0.730	0.0012	0.0021	-1.11	-2.52*	-2.41*	-7.44*	-2.63*	-2.49*	-2.10
EASIA	Yakut	4	7	0.643	0.0019	0.0026	-1.36	-1.36	-1.50	0.08	-0.17	-0.56	-4.71
EUR	Adygei	14	9	0.791	0.0033	0.0022	1.47*	0.23	0.71	0.34	0.20	0.71	-0.03
EUR	Basque	20	7	0.779	0.0030	0.0016	2.51**	1.26*	1.89**	2.54	1.30*	2.00**	-0.13
EUR	French	17	10	0.786	0.0031	0.0024	0.93	0.29	0.58	-0.19	0.27	0.58	-0.48
EUR	North_Italy	20	7	0.688	0.0030	0.0016	2.43**	1.26*	1.90**	1.55	1.30*	1.87**	-0.43
EUR	Orcaadian	8	7	0.442	0.0024	0.0020	0.57	0.73	0.79	2.00	0.74	0.82	-0.43
EUR	Russian	15	10	0.784	0.0027	0.0025	0.30	0.87	0.81	-0.97	0.81	0.85	-2.86
EUR	Sardinian	25	9	0.745	0.0033	0.0020	1.90*	0.03	0.76	-0.50	0.01	0.76	0.47
MENA	Bedouin	32	13	0.792	0.0034	0.0027	0.83	-0.11	0.25	-2.99	-0.14	0.24	0.29
MENA	Druze	40	9	0.646	0.0028	0.0018	1.55*	0.61	1.10	-0.21	0.61	1.12	-0.08
MENA	Mozabite	17	9	0.759	0.0033	0.0021	1.61*	0.16	0.72	1.55	0.13	0.72	-0.04
MENA	Palestinian	35	11	0.705	0.0031	0.0022	1.14	-0.42	0.13	-0.61	-0.47	0.12	0.49
OCE	NAN_Melanesian	10	3	0.663	0.0087	0.0030	0.42	1.01**	0.97**	-0.06	1.01**	1.00	0.69
OCE	Papuan	13	4	0.686	0.0013	0.0010	0.79	1.07*	1.15*	0.02	1.09*	1.18	-1.12
SSAFR	Bantu	12	15	0.822	0.0038	0.0039	-0.07	-1.02	-0.86	-0.30	-1.26	-1.05	0.93
SSAFR	Blaka_Pygmies	25	10	0.793	0.0027	0.0022	0.65*	0.78	0.87*	0.72	0.80*	0.90*	-2.31
SSAFR	Mandenka	17	10	0.770	0.0035	0.0024	1.45**	1.40**	1.66**	1.81*	1.49**	1.76**	0.35
SSAFR	Ntsui_Pygmies	12	12	0.863	0.0019	0.0031	-1.35	-0.28	-0.70	-2.86	-0.56	-0.96	-4.59
SSAFR	San	5	10	0.844	0.0022	0.0034	-1.53*	-1.51	-1.70	-1.46	-0.68	-1.09	-3.38
SSAFR	Yoruba	10	12	0.863	0.0038	0.0033	0.56	0.23	0.38	-1.72	0.18	0.36	0.98

n = number of SNPs; S = segregating sites; Hd = haplotype diversity  $\pi$  = average number of nucleotide differences per site;  $\theta$  = Watterson estimator; \* P < 0.05; \*\* P < 0.01; \*\*\* P < 0.001

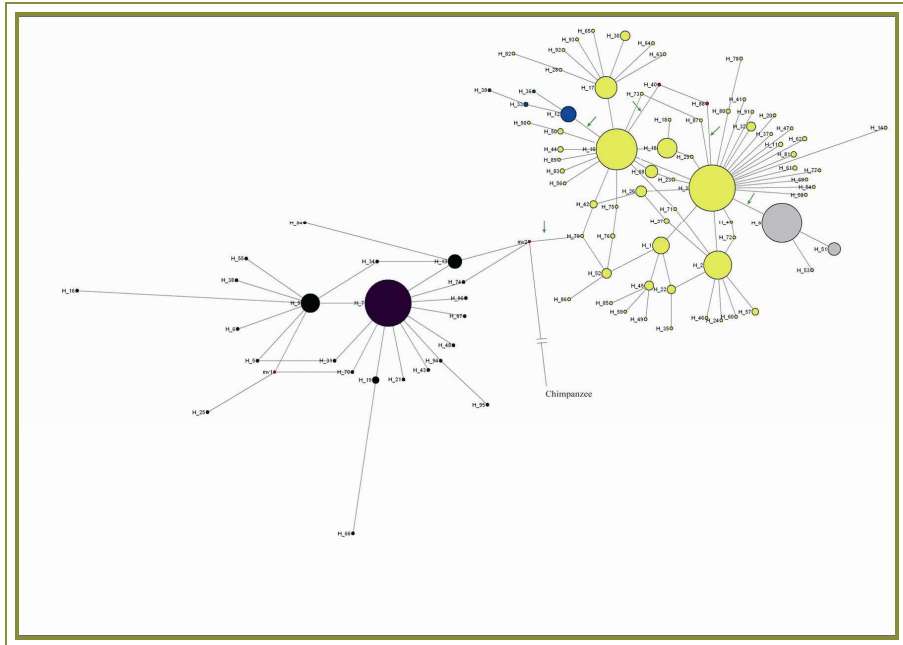


Figure 1. Median-joining network of *FUT2* haplotypes in seven continental regions.



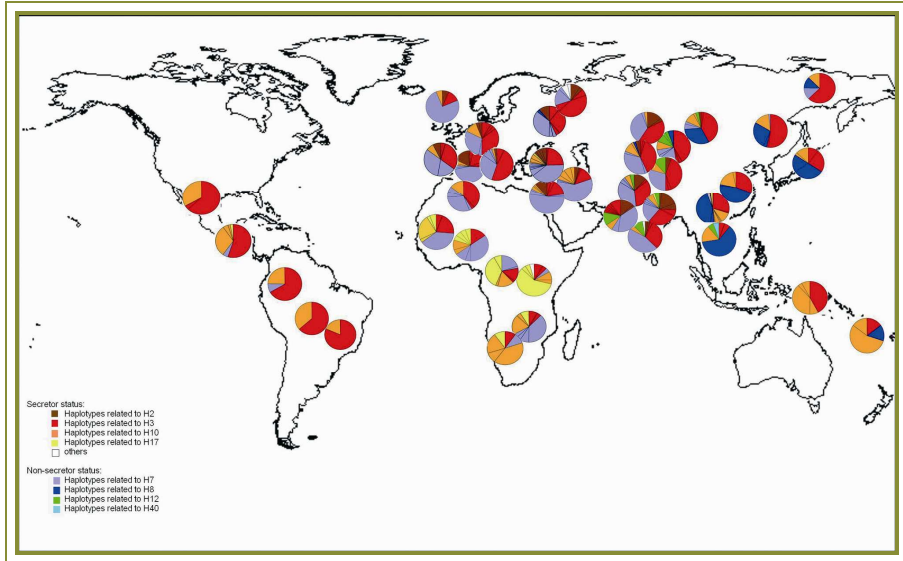
America (grey), Central-South Asia (orange), East-Asia (blue), Europe (green), Middle-East and North Africa (red), Oceania (purple) and Sub-Saharan Africa (yellow). The circle areas are proportional to the frequency of the haplotypes. Haplotype number is shown next to circles. Inactivating mutations have been indicated by an arrow.

Figure 2. Median-joining network of *FUT2* haplotypes according to phenotypes.



. The topology is the same as in Figure 1; here colors indicate the haplotypes carrying different secretor/non-secretor variants; se<sup>302</sup> (blue), se<sup>385</sup> (grey), se<sup>428</sup> (black), se<sup>571</sup> (red) and Se (yellow). Inactivating mutations have been indicated by an arrow.

Figure 3. Geographical distribution of haplotypes carrying variants conferring secretor and non-secretor phenotypes.



Color represents a frequent haplotype and those haplotypes phylogenetically close to it. Each pie corresponds to a different population in each geographical location.

### **4.3 Chapter 3**

Evolutionary analysis of human pseudogenes of the ABO family show  
a complex picture in their dynamics and function loss

Manuscript in preparation



Evolutionary analysis of human pseudogenes of the ABO family show a complex picture in their dynamics and function loss

Ferran Casals<sup>1,2</sup>, Anna Ferrer-Admetlla<sup>1</sup>, Martin Sikora<sup>1</sup>, Anna Ramírez-Soriano<sup>1</sup>, Tomàs Marquès-Bonet<sup>1,3</sup>, Stéphanie Despiau<sup>4</sup>, Francis Roubinet<sup>5</sup>, Francesc Calafell<sup>1,6</sup>, Jaume Bertranpetit<sup>1,6,\*</sup> and Antoine Blancher<sup>4</sup>

<sup>1</sup> Institut de Biologia Evolutiva (CSIC-UPF), CEXS-UPF-PRBB,

Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain.

<sup>2</sup> Present address: Ste Justine Hospital Research Centre, Department of Pediatric, Faculty of Medicine, University of Montreal, Montreal, Quebec H3T 1C5, Canada

<sup>3</sup> Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA

<sup>4</sup> Laboratoire d'Immunogénétique Moléculaire, Faculté de Médecine de Rangueil, Université Paul Sabatier, Bâtiment A2, 31062 Toulouse Cedex 4, France

<sup>5</sup> Etablissement Français du sang Centre Atlantique, BP 52009, 37020 Tours, Cedex 1, France

<sup>6</sup> CIBER Epidemiología y Salud Pública (CIBERESP);

\* Correspondence: [jaume.bertranpetit@upf.edu](mailto:jaume.bertranpetit@upf.edu)



Abstract

The GT6 glycosyltransferases gene family, that includes the ABO blood group, shows a complex phylogenetic pattern, with several events of gain and loss among the different primate species. Most of the members of this family in humans have been proposed to be not functional, and the ABO gene is considered the sole functional member. Here, we analyze these sequences from an evolutionary perspective, by the study of i) their levels of population genetic diversity through the resequencing analysis of European and African individuals, including neutrality tests; ii) the interpopulation differentiation, with genotyping data from a survey of populations covering most of human genetic diversity and iii) the interspecific divergence, by the comparison of the human and other species sequences. Since pseudogenes are expected to evolve under neutrality, they should show an evolutionary pattern different to that of functional sequences, with higher levels of diversity as well as a ratio of synonymous to non synonymous changes close to 1. We describe some departures from these expectations, suggesting that some of these pseudogenes may still be functional, in some case even at the protein level; these includes: i) neutral behaviour for the two processed pseudogenes (HGT2 and LOC401913); ii) no gene conversion has been detected as an homogenizing process; iii) two processed pseudogenes (GGAT1 and IGB3) show signs of selection for inactivation, with likely pressures of unknown pathogen factors; and iv) a special behaviour in FS (Forssman) with a shift in its initial function that put it apart from a pure neutral pseudogene.

## Introduction

The classical definition of pseudogenes as non-functional sequences requires to be reviewed because of the recent advances in the transcriptome and regulatory roles of non-coding RNA. Indeed, recent transcriptome analyses have showed that not only coding regions are transcribed into RNA, but also a high proportion of non-coding regions. This proportion ranges from ~ 15 % of the sequence obtained in the analysis of ten human chromosomes (CHENG *et al.* 2005) to the 93 % described at some ENCODE regions (Hubbard *et al.* 2007), where at least a fifth of the 201 annotated pseudogenes have been demonstrated to be transcribed (ZHENG *et al.* 2007). Although the transcription of a pseudogene is not a direct demonstration that the transcripts have biological function, this has been unequivocally demonstrated in some cases, such as in the suppression of the nNOS synthesis by an antisense mRNA encoded by a pseudogene in *Lymnaea stagnalis* (KORNEEV *et al.* 1999), or the competition of pseudogene and functional *Mkrn1* sequences both for transcription repressors and RNA-digesting enzymes in mouse (HIROTSUNE *et al.* 2003; LEE 2003). Besides of such direct functional roles, if pseudogene sequence is conserved they may be involved in gene conversion or recombination events with functional genes, thus acting as a genetic diversity reservoir (BALAKIREV and AYALA 2003; VARGAS-MADRAZO *et al.* 1995). Finally, pseudogenes were proposed to be considered as *potogenes* (BROSIOUS and GOULD 1992) due to their potentiality for becoming new genes (BALAKIREV and AYALA 2003).

In this work, we analyze the intra and interspecific variability (polymorphisms and divergence) pattern of a set of six human pseudogenes belonging to the glycosyltransferase 6 (GT6) family. This family includes enzymes contributing to the synthesis of histo-blood group-related antigens by transferring a galactose or an N-acetylgalactosamine to different substrates (TURCOT-DUBOIS *et al.* 2007). The ABO gene, the most representative gene of this family, has two functional alleles called A and B responsible for the expression of the A or B blood groups, respectively. Even much before of the molecular characterization of the ABO gene

(Saitou and Yamamoto 1997), phenotypic variation shown by this locus was used in population genetics analysis, revealing a high level of polymorphism (Mourant 1954). Resequencing analysis has confirmed this, showing that this locus exhibits high levels of variation, requiring other forces than neutrality to be explained (AKEY *et al.* 2004; CALAFELL *et al.* 2008; STAJICH and HAHN 2005). Although balancing selection is the best explanation for this gene diversity, the selective agent favouring this pattern remains unknown. Several possible causative agents have been proposed, because of significant associations between some alleles with protection to infectious diseases (Boren *et al.* 1993; Fry *et al.* 2008; Lindesmith *et al.* 2003; Marionneau *et al.* 2005; Rowe *et al.* 2007; Ruiz-Palacios *et al.* 2003; Swerdlow *et al.* 1994), or consistent geographical distributions among ABO alleles and some infectious disease causative elements, as has been proposed to occur with the malaria parasite *Plasmodium falciparum* (CSERTI and DZIK 2007) and confirmed in a case-control study (Fry *et al.*, 2008).

Apart of the ABO gene, other members of the GT6 family are present in all vertebrates, and an expansion occurred in the placental mammals (TURCOT-DUBOIS *et al.* 2007). In this clade, seven members have been identified and classified into the GT6 family: ABO, Forssman (FS), IGB3, GGTA1, GT6m5, GT6m6, and GT6m7. To the best of our knowledge, no much is known on the function of GT6m5, GT6m6, and GT6m7 genes. These genes show a complex phylogenetic pattern of activity / inactivity with several episodes of gain and loss among the different species (TURCOT-DUBOIS *et al.* 2007). In particular, in humans ABO is the sole gene to have functional alleles. In humans, FS and GGTA1 are inactivated by several mutations within the coding region including the catalytic domain, while IGB3 displays promoter alteration and aberrant splicing; GT6m5 and GT6m6 are deleted (MILLAND *et al.* 2006; TURCOT-DUBOIS *et al.* 2007; XU *et al.* 1999), and GT6m7 is probably inactivated by a premature stop codon in the last exon suggesting that the human protein (if it exists) could be not functional. Because carbohydrates at the cell surface of mammals are used by various pathogens as receptors, the inactivation of some glycosyltransferase genes could have been directly selected as resistance factors (ELLIOTT *et al.* 2003). On the

other hand, various pathogens share carbohydrate antigens with mammals, and the inactivation of glycosyltransferase genes could have been indirectly selected because in absence of the cognate carbohydrate antigen allows the production of antibodies protective against infectious agents expressing cross-reacting carbohydrate antigens (GAGNEUX and VARKI 1999; SHARON 1996).

In this work we analyze all non-functional members of the GT6 family found in humans including the four non-processed pseudogenes FS, IGB3, GGTA1 and GT6m7, and two processed pseudogenes: LOC401913, a processed ABO copy on chromosome 19 (Saitou and Yamamoto 1997), and HGT2, a retrotranscript of GGTA1 present in all primates except in prosimians (KOIKE *et al.* 2007). The evolutionary analysis of all members of the GT6 family in humans allow to have an insight in their functional history, a process that can be unravelled through the footprints left by natural selection and possible to read in the extant genome variation. To this end, most methods of unravelling the footprint of selection have been used, retrieving signals from very ancient or recent times in the different analysis. Thus this analysis intends to reconstruct the functional history of a gene family composed mostly of pseudogenes. Evolutionary analysis is thus used as a tool in the understanding of biological function.

## Methods

There are two kinds of data used: resequencing data, obtained in a sample of European and African individuals and another of SNPs along the regions of interest obtained in a world-wide sample.

## Samples

For sequencing, nineteen African samples (ten Baoulés and nine defined as Ivorians) and 26 European samples (from Toulouse, France) were sequenced for the region homologous to ABO exon 7 of FS, GGTA1, GT6m7, and IGB3

pseudogenes. Samples used for each pseudogene are shown in Table S1 (Supplementary material). SNP genotyping was performed on the 1,064 samples included in the Human Genome Diversity Panel of cell lines (HGDP-CEPH), which belong to 51 populations globally distributed (CANN *et al.* 2002). Following the recommendation of Rosenberg (2006) we have used for further analysis the H971 subset, obtained after removing the atypical and duplicated samples, and first-degree relatives. Samples with more than 50% of genotyping missing data have been also excluded from the analysis. The 51 original populations were reorganized in 39 populations, following geographic and ethnic criteria (Gardner *et al.* 2006).

#### Sequencing and Sequence Analysis

A list of PCR and sequencing primers is provided in Table S2 (Supplementary material). These primers were used to amplify both human and other primate DNA. Resulting amplified products were separated on agarose gels, purified using QIAquick PCR Purification Kit (Quiagen, Courtaboeuf, France) and directly sequenced on both strands by using either the fluorescent Dye Terminator method and an ABI 373 automatd sequencer (Applied-Biosystems, Courtaboeuf, France) or the Beckman Coulter sequencing kit (CEQ DTCS-Quick Start) and a CEQ 8000 sequencing apparatus (Beckman Coulter).

Multiple sequence alignments were performed with ClustalW (THOMPSON *et al.* 1994). Although the analyzed sequences are considered pseudogenes, the divergence among the paralogous sequences was not a difficulty and totally unambiguous alignments were produced for the five sequences (the four processed pseudogenes and the *ABO* gene) with no doubts. Diversity statistics and neutrality tests to detect signals of natural selection (Tajima's D, Fu and Li's  $F^*$  and  $D^*$ , Fu and Li's F and D, Fay and Wu's H, Fu's  $F_s$ ) have been calculated using DNAsp 4.00 (ROZAS *et al.* 2003). The significance of these tests has been estimated by means of coalescent simulations performed using the COSI software, under a realistic human past demography (SCHAFFNER *et al.* 2005). 10,000 replicates have been performed. Simulations were performed by

---

considering the recombination rate  $r$ , obtained for each region from HapMap ([www.hapmap.org](http://www.hapmap.org)). Different individuals have been sequenced for different genome regions (Table S1) and thus comparisons may be based on different individual sets. The time to the most recent ancestor (TMRCA) was estimated by means of the maximum likelihood coalescent-based method implemented in the Genetree program (TAVARE *et al.* 1997).

Orthologous sequences of ABO, FS, GGTA1, GT6m7 and IGB3 were obtained from the GenBank (<http://www.ncbi.nlm.nih.gov/>) and Ensembl databases (<http://www.ensembl.org/index.html>). Phylogenetic analyses were performed using the PAML software package (YANG 1997).

## SNPs

We genotyped a total of 43 SNPs in five pseudogene regions (FS, GT6m7, HGT2, IGB3 and LOC401913). Valid genotypes were obtained for 40 of the 43 SNPs. The exact position of each SNP is shown in Table S3 (Supplementary material). SNPs were selected from HapMap Phase I and dbSNP databases, following a common rationale. We selected SNPs in the genic and 5' and 3' flanking regions (up to 30 kb), with a density of 1SNP / 5 Kb approximately and giving priority to those with a minimum allele frequency (MAF) over 0.1. Ancestral alleles were obtained from the chimpanzee genome sequence (panTro2). SNPs were typed using the SNPlex Genotyping System from Applied Biosystems, following the manufacturer's standard protocol. Allele separation was performed on an Applied Biosystems 3730 analyzer and besides the automated allele calling and quality metrics provided by GeneMapper Software 3.5, allele calling was always reviewed manually. Additional publicly available data for 75 SNPs in these five pseudogene regions and the GGTA1 region was retrieved from <http://shgc.stanford.edu/hgdp/>, which includes genotyping in the Human Genome Diversity Panel samples (HGDP-CEPH) at more than 650,000 SNP loci, obtained with the Illumina BeadStation technology (LI *et al.* 2008).

### SNP Data Analysis

Genotype data was stored in SNPator (MORCILLO-SUAREZ *et al.* 2008) (<http://bioinformatica.cegen.upf.es/public/principal/index.php>), that allows storage, quality control and analysis. It allowed for coherence control among the samples duplicated at the same or different plates, as well as basic analysis including allele and genotype frequencies estimation, expected heterozygosity and Hardy-Weinberg equilibrium calculations. All SNPs were in Hardy-Weinberg equilibrium after Bonferroni correction.  $F_{ST}$  values were estimated with a locus by locus Analysis of Molecular Variance (AMOVA) with the Arlequin software (Excoffier 2005), using the 39 populations and the seven continental groups defined.

### Haplotype Estimation and Linkage Disequilibrium Analysis

Haplotype frequencies were estimated using the bayesian algorithm implemented in the PHASE 2 software using the default parameters and 1,000 iterations (STEPHENS *et al.* 2001). Network 4.2.0.1 software was used to generate median-joining networks describing possible genealogical relationships among haplotypes in terms of mutational differences (BANDELT *et al.* 1999). Haploview software (BARRETT *et al.* 2005) was used to represent and calculate LD, measured as  $D'$  or  $r^2$ , between the different SNPs in the different regions.

### Haplotype Extension Analysis

The Extended Haplotype Homozygosity (EHH) and the Relative Extended Haplotype Homozygosity (REHH) for core haplotypes included in the analyzed regions were calculated both for data presented in this work and HapMap data at different genetic distances using the Sweep software (SABETI *et al.* 2002) and considering the default core definition parameters. The integrated haplotype score (iHS) has been calculated from Haplotter (<http://hg-wen.uchicago.edu/selecion/haplotter.htm>) (VOIGHT *et al.* 2006).

---

## Results

### Nucleotide sequence variation

The region homologous to ABO exon 7 of four non-processed human pseudogenes (FS, GGTA1, GT6m7 and IGB3) was sequenced in African and European human samples. We detected six polymorphic sites in FS, four in GGTA1, five in GT6m7, and two in IGB3, some previously unreported (Table 1). Only one deletion of 1bp has been found at position 166 of GT6m7 locus, and the other 16 polymorphisms are single nucleotide substitutions, with transitions (10) being slightly more frequent than transversions (6), and six of them being singletons (38 %) and four of them doubletons (25 %). This abundance of low frequency variants leads to a high level of homozygosity in all four pseudogenes, with a high proportion of individuals homozygous for the major haplotype for the whole tract analyzed: 67% for FS, 79% for GGTA1, 72% for GT6m7, and 93% for IGB3. Haplotypes for each pseudogene have been inferred (see Methods) and are shown in Table 1. Ancestral haplotypes have been deduced from the chimpanzee sequence. In the case of IGB3, we have obtained the sequence through PCR amplification with the same primers used in humans, since no homologous sequences were found in public databases.

### Nucleotide and Haplotype Diversity

The variation observed in the four pseudogene regions resulted in low levels of nucleotide and haplotype diversity (Table 2), with IGB3 showing the most extreme values with only two singletons described in the African population and no variation at all in Europeans. All genes, except GT6m7, show a higher level of variation in Africa than in Europe, as expected in the out-of-Africa model of modern humans origin (CAVALLI-SFORZA and FELDMAN 2003) (Table 2). In the four cases the variability is much lower than that exhibited by the exon 7 of ABO, whose high level of polymorphism has been proposed to be explained by the action of balancing selection (CALAFELL *et al.* 2008; FRY *et al.* 2008). In general the nucleotide diversity values for these pseudogene regions are close or lower than



the average obtained for the 322 resequenced genes included in the SeattleSNPs database (<http://pga.mbt.washington.edu/>), with average heterozygosity ( $\pi$ ) values of 0.0009 in the African population and 0.0007 in the European population.

We have tested whether the amount of variation in these regions show polymorphism levels in accordance with those expected under neutrality or in fact are lower and due to purifying selection acting on them. We performed coalescent simulations with recombination (see Methods) for each region, considering the number of segregating sites (Table 2). Simulations have been performed considering a demographic scenario that simulates the human population history (SCHAFFNER *et al.* 2005). Only in the case of GGTA1 in Africans and in the case of FS in the Europeans the observed  $\pi$  value are lower than expected, although results do not reach statistical significance ( $P = 0.068$  and  $P = 0.095$ , respectively). No simulations could be performed for IGB3 in Europeans, due to the lack of polymorphic positions. In order to gain power, as the four regions show low levels of polymorphism, we have also performed the analysis considering the four regions pooled together. The observed nucleotide diversity level in Africans and Europeans are lower than expected, although the differences are again not significant ( $P = 0.15$  and  $P = 0.25$ ).

#### Neutrality tests

In order to elucidate if this low variability pattern could be expected under a neutral model or is due to natural selection, which can only act if there is a function, several neutrality tests based on allele frequencies (Tajima's  $D$ , Fu and Li's  $D^*$ ,  $F^*$ ,  $D$  and  $F$ , Fay and Wu's  $H$ ) and haplotype diversity (Fu's  $F_s$ ) were performed, with significance calculated by coalescence under a demographic model (see Methods). Results (Table 3) suggest that some of these genome regions have not evolved under neutrality, with differences among them. In the GT6m7 region Fay and Wu's  $H$  values are significantly less than zero in Africans and Europeans. In the FS region, the obtained negative value of Tajima's  $D$  in Europeans is also significant. Finally the observed values for IGB3 (for Tajima's  $D$ , Fu and Li's  $D^*$ ,  $F^*$ ,  $D$ , and  $F$ , and Fu's  $F_s$  in the African population) are the most extreme values

---

and close to the significance level ( $P < 0.088$ ). Due to the low number of segregating sites the power of the test is low and the statistical significance should be interpreted with caution. Neutrality tests could not be performed in IGB3 in Europeans due to the absence of polymorphic sites.

#### Population differentiation

Using the resequencing data, we have measured the population differentiation between Africans and Europeans for all described polymorphisms by means of  $F_{ST}$  values. Only in the case of one polymorphic site described in FS (position 336) the differences are statistically different from zero ( $F_{ST} = 0.34$ ). Nonetheless, it is an isolated case, with the rest of SNPs in the region being much lower.

Genotyping data for 115 SNPs in the six pseudogene regions (see Methods) was used to calculate genetic differentiation of these regions across human populations.  $F_{ST}$  was calculated among the 39 populations on one hand and among and within the seven continental groups on the other. Mean values for each region are shown in Table 4. Results show a low level of population differentiation, with similar values, except in the case of HGT2. The obtained  $F_{ST}$  values were compared to those obtained in an independent ongoing work on 3,058 SNPs located in 69 glycosylation gene regions (unpublished data) that had been selected following the same criteria, making the two data sets comparable. Apart of HGT2, the average  $F_{ST}$  values obtained on the pseudogenes are close to the average  $F_{ST}$  in the glycosylation dataset ( $F_{ST} = 0.095$ ).

#### Linkage disequilibrium structure

SNP genotype data in the six pseudogene regions also provide the opportunity of analyzing the linkage disequilibrium (LD) pattern in every population or continental group. In absence of selective pressures, recombination rate is the main force determining the extent of LD at given region, and thus, we do not expect to find important differences between the LD patterns in different populations. On the other hand, the existence of local selective forces may increase the amount of LD

through a selective sweep, leading to different LD patterns between the populations where the selective event has occurred and those where it has not. However, the role in shaping LD of population history, genetic drift and stochastic factors should also be considered. We used the extended haplotype homozygosity test (EHH), which allows detecting recent selective sweeps by comparing the breakdown of the LD in different haplotypes (SABETI *et al.* 2002). We have performed the EHH test using the data obtained here both considering the 39 populations and the seven continental groups, and no significant departures from neutrality have been found in any of them. In order to analyze larger regions we also performed the iHS tests (VOIGHT *et al.* 2006) to the six pseudogene regions using HapMap data, and again no evidence of recent positive selection events have been found in any of them. The analysis of LD does not show any significant pattern, meaning that there has not been, in recent times, a differential action of positive selection that would act on some populations and not in others.

#### Interspecific divergence

We have calculated the rate of synonymous to non-synonymous changes in the lineage leading to humans after the split with chimpanzee for the four non-processed pseudogenes: FS, GGTA1, GT6m7 and IGB3. Rates are expected to be greater than 1 in case of positive selection, close to 1 in case of neutral evolution (what is expected to occur in a pseudogene), and less than 1 in case of purifying selection. The estimated  $d_N / d_S$  ( $\omega$ ) ratios in the human branch both for the complete pseudogenes and for the exon homologous to ABO exon 7 are shown in Table 5. We tested the likelihood of these sequences evolving under neutrality, by comparison of the likelihood of two nested models, a model estimating a free  $d_N / d_S$  value for the human branch against the likelihood of a model assuming that this branch is evolving with  $d_N / d_S = 1$  (Table 5).

The two models are compared using the likelihood ratio test with as much degrees of freedom as the number of differences on the parameters estimated (that fits to a  $\chi^2$  distribution). With the exception of FS, although the estimated omega values in the human branch are small, they are not statistically different than 1. In the case

of FS, we can reject the hypothesis of this sequence evolving neutrally ( $\omega = 1$ ), suggesting functional constraint in the human branch due to the action of purifying selection.

## Discussion

We have performed an evolutionary analysis of six human loci related to the ABO gene. According to different types of evidence, these six loci have been previously proposed to be pseudogenes in humans. Because of their lack of function, pseudogenes are expected to evolve under neutrality, since no functional constraints should be acting on them. In this case, the absence of purifying selection should be reflected at different evolutionary levels, which could have been acting at different evolutionary time scales. First, since divergence from chimpanzee, the rate of synonymous to non-synonymous changes should be close to 1, since putative amino acid changes will not have any consequence; second, interpopulation differences should also be higher as populations would differentiate by drift alone without common functional constraints; third, sequence variation should fit the theoretical neutral models, tested through many different methods and parameters; and finally, nucleotide diversity levels should be higher in pseudogenes than in functional regions because of lack of purifying selection in the ancestors of extant humans. In this work, we describe that the analyzed pseudogenes do not fully fit to these predictions, suggesting that purifying selection has been or is still acting on some of these sequences.

Processed pseudogenes, those coming from a duplication event, seem to be less prone to show signatures of evolutionary conservation, since they have been originated by retrotransposition, which includes more cases of protein-coding ability loss (ZHENG *et al.* 2007). The two processed pseudogenes included in this work, do not show special signatures of conservation. Both LOC401913 (that has been poorly studied till now) and HGT2 clearly behave as real pseudogenes, with no evolutionary constraints. HGT2 is a pseudogene containing multiple frameshift mutations processed from GGTA1 after the split between haplorrhines and

strepsirrhines (JOZIASSE *et al.* 1991; KOIKE *et al.* 2007), which has been estimated to occur 77.5 MYA (STEIPER and YOUNG 2006).

For the four non-processed pseudogenes, there is a low level of nucleotide diversity and interpopulation differentiation. This conservation could be related with the retention of some function, not necessarily at the protein level. One possible explanation for the conservation of a pseudogene originated by duplication is that it may play a role in the generation of diversity of the original and functional gene, through gene conversion events. However, in all four cases, the pseudogene regions analyzed here show high levels of nucleotide divergence with ABO exon 7 (0.39 for FS, 0.47 for GGTA1, 0.48 for GT6m7, and 0.37 for IGB3). A high level of identity along regions with a minimum size of 200 bp approximately is needed to exist gene conversion between two sequences (LUKACSOVICH and WALDMAN 1999). We have performed a sliding window calculation to check if there are regions of at least 200 bp showing a high level of interespecific conservation, but none has been found, making these hypotheses unreliable. Thus gene conversion have not played a main role in shaping the paralogous ABO sequences

If any function is being retained at these loci, it should be related with some RNA function or even with the protein product (see below), more than with the full primary sequence. Low nucleotide diversity values can also be explained through the action of positive selection. After a positive selection event, the genetic variation is reduced in this region, and this will persist until new neutral mutations will reach high frequencies. This time has been estimated to be up to 1 million years (SABETI *et al.* 2006). However, we have not found evidence of positive selection in the regions analyzed. Only in two cases the neutrality statistics points at this direction, GT6m7 in both Africans and Europeans for Fay and Wu H and, at a lower level, for FS in Europeans for Tajima's D (Table 3). The negative value in the Fay and Wu's H test for GT6m7 in the two populations is originated because the majority of individuals carry the derived allele at two of the five polymorphic positions (Table 1), which, given this low number of polymorphisms in the locus, may have been produced randomly by drift. In the case of the Tajima's D value for

---

FS in European population, the fact that no other neutrality test detects positive selection signatures makes this prediction improbable. In addition, the analysis on LD structure and haplotype extension also fails to detect any signature of positive selection. The synonymous to non-synonymous changes ratio ( $d_N/d_S$ ) analysis performed at the interspecies level also fails to detect any signature of positive selection (that would be more ancient than that detected by intraspecific tests) in any of the pseudogenes.

In fact, the analysis of the interespecific divergence suggests different evolutionary histories for the four non-processed pseudogenes. First, GGTA1 and IGB3 do not show signatures of conservation of the protein sequence. GGTA1 has been suggested to be under strong purifying selection in primates except in catarrhines, where has been demonstrated to be inactive (KOIKE *et al.* 2002; KOIKE *et al.* 2007). The obtained  $d_N/d_S$  value in the human branch seems to corroborate its inactivity, and with the present evidence purifying selection can be rejected in this locus (Table 5). Given the functional constraints on this gene in the noncatarrhine species, and the deleterious effect of its absence demonstrated in knock out mice and pigs, this inactivation in catarrhines (including humans) should have been compensated by yielding important different kind of benefits, as may be an enhanced defence against alphaGal-positive pathogens (KOIKE *et al.* 2007). IGB3 also shows also a relatively high  $d_N/d_S$  (Table 5), and has previously been shown not to be functional in humans (KEUSCH *et al.* 2000), then producing high circulatory levels of anti-Gal $\alpha$ 1,3Gal antibodies (GALILI *et al.* 1993). In both cases the inactivation seem to have produced a defective phenotype that makes that their status as pseudogenes (with inactivation) can only be understood if here is some selective advantage in their inactivation, likely related to pathogen resistance. Thus pathogen resistance would not only explain the functional genetic variation but the lack of functionality.

On the other hand, GT6m7 shows low level of intraspecific variation, low levels in the Fay and Wu test and high levels of divergence with AB0, and thus the model assuming neutrality in the human branch can not be statistically rejected. This is in

agreement with the existing evidence of inactivity in this locus, with a stop codons disrupting the last exon having been reported in humans (TURCOT-DUBOIS *et al.* 2007).

Finally, the case of FS is especially interesting. The presence or absence of Forssman glycolipid has been related to different levels of susceptibility for several infectious diseases (ELLIOTT *et al.* 2003; XU *et al.* 1999). While several mutations in its catalytic domain do not allow the production of the Forssman glycolipid in humans, FS mRNA has been detected in several tissues, suggesting that this gene is transcribed and may be retaining some biological function. Our results would support functional-related findings, because unexpectedly low  $d_N/d_S$  ratio has been described in the human lineage. These results could be compatible with the differentiation of the human FS gene with the acquirement of a function differing from the N-acetyl Galactosyltransferase activity of the classical FS synthase reported in other mammals such as dog and mouse. Interestingly, proteins encoded by human-dog hybrid cDNA demonstrated that the two amino acid substitutions in human are sufficient to inactivate the dog FS activity (XU *et al.* 1999). These two mutations are observed in chimpanzee suggesting that the FS loss of enzyme activity preceded the human-chimpanzee split. Therefore it seems reasonable to evoke the possibility that the new acquired function of the FS protein appeared in the common ancestor of human and chimpanzee.

In conclusion, the evolutionary analysis of one set of human pseudogenes belonging to the GT6 family suggests neutrality in some of them but footprints of selection in others, including selection for inactivation and putative acquisition of new functions. The proportion of proposed non functional members of this family in humans is similar or even higher to that reported in other well described gene families as the olfactory and bitter taste receptors families (FISCHER *et al.* 2005; GILAD *et al.* 2005). In the case of the olfactory receptors, the high proportion of 56 % of the members being non functional in humans, has been explained by a relaxation of the functional constraints in this species related with reduced olfactory needs (GILAD *et al.* 2005; GILAD *et al.* 2003). The GT6 family shows, in addition to the high number of non functional members, a complex pattern of birth-

and-death evolution (TURCOT-DUBOIS *et al.* 2007). These features can be originated by the implication of these proteins in interactions with pathogens, with different and changing environments favouring the absence of some pathogen receptors (GAGNEUX and VARKI 1999). Our analysis suggests that some of these putative human pseudogenes while unable to perform the initial function of AB0, are likely to have been selected for inactivation and in the FS locus maybe used as pathogen target, and thus still retaining a biological function.

#### Acknowledgments

This research was funded by Genoma España (Proyectos Piloto del CEGEN), by grants BFU2004-02002, BFU2005-00243 and SAF-2007-63171 awarded by Ministerio de Educación y Ciencia (Spain), by the Direcció General de Recerca of Generalitat de Catalunya (Grup de Recerca Consolidat 2005SGR/00608), and by the National Institute for Bioinformatics ([www.inab.org](http://www.inab.org)), a platform of Genoma España. SNP genotyping services were provided by the Spanish "Centro Nacional de Genotipado" (CEGEN; [www.cegen.org](http://www.cegen.org)). MS is supported by a PhD fellowship from the Programa de becas FPU del Ministerio de Educación y Ciencia, Spain (AP2005-3982).



# Human Genetic Diversity In Genes Related To Host-Pathogen Interactions

**Table 1.** Haplotypes found in exon homologous to ABO exon 7 in four human pseudogenes

FS	3 <sup>a</sup>	174	336 <sup>f</sup>	353 <sup>g</sup>	510 <sup>h</sup>	603	Africa	Europe	Total
H1	C	G	C	A	C	C	12	31	43
H2	.	.	T	.	.	.	5		5
H3	.	.	T	C	.	.	1		1
H4	.	A	.	.	.	.	2		2
H5	.	.	.	.	.	T		1	1
H6	A	.	.	.	T	.		2	2
Chimpanzee*	.	.	.	.	.	.			
GGTA1	146 <sup>i</sup>	232	331	488			Africa	Europe	Total
H1	G	C	C	C			17	25	42
H2	.	T	.	.			1		1
H3	T	.	.	.			2		2
H4	.	.	.	G				1	1
H5	.	.	T	.			.	2	2
Chimpanzee*	.	.	.	.					
GT6m7	166	323 <sup>a</sup>	394 <sup>b</sup>	522 <sup>c</sup>	584 <sup>d</sup>		Africa	Europe	Total
H1	T	A	C	T	C		31	21	52
H2	.	.	.	C	.		1	1	2
H3	.	.	.	C	A		4		4
H4	.	G	.	C	A		1	1	2
H5	.	.	.	.	.		2		2
H6	.	G	.	C	A		1		1
H7	.	G	T	.	A			2	2
H8	.	.	.	C	A			1	1
Chimpanzee*	.	.	.	C	A				
IGB3	38	471					Africa	Europe	Total
H1	C	G					18	34	52
H2	T	.					1		1
H3	.	T					1		1
Chimpanzee*	.	.							

<sup>a</sup> rs35762223; <sup>b</sup> rs17040344; <sup>c</sup> rs12351198; <sup>d</sup> rs12336965; <sup>e</sup> rs35898523; <sup>f</sup> rs34691037; <sup>g</sup> rs35366884; <sup>h</sup> rs35902535; <sup>i</sup> rs10985245.

For the chimpanzee haplotypes only those positions polymorphic in humans are shown.

**Table 2.** Summary of diversity statistics of four human pseudogenes in African and European populations.

All					
	L	n	S	$\pi$	$\theta$
FS	686	56	6	0.0007	0.0019
GGTA1	685	48	4	0.0004	0.0013
GT6m7	694	66	4	0.0010	0.0012
IGB3	690	54	2	0.0001	0.0006
ABO <sup>d</sup>	687	94	15	0.0047	0.0042
Africa					
	m	n	S	$\pi$	$\theta$
FS	686	20	3	0.0011	0.0012
GGTA1	685	20	2	0.0004	0.0008
GT6m7	694	40	3	0.0010	0.0010
IGB3	690	20	2	0.0003	0.0008
All	2755	18	9	0.0006	0.0010
ABO <sup>d</sup>	687	48	15	0.0056	0.0049
Europe					
	m	n	S	$\pi$	$\theta$
FS	686	34	3	0.0004	0.0011
GGTA1	685	28	2	0.0003	0.0008
GT6m7	694	26	4	0.0012	0.0015
IGB3	690	34	0	0	0
All	2755	24	9	0.0005	0.0009
ABO <sup>d</sup>	687	46	8	0.0034	0.0027

**Table 3.** Neutrality tests for four human pseudogenes in African and European populations

Africa							
	Tajima's D	Fu and Li's D*	Fu and Li's F*	Fu and Li's D	Fu and Li's F	Fay and Wu's H	Fu's Fs
FS	-0.36	-0.12	-0.22	-0.18	-0.27	0.52	-0.72
GGTA1	-1.14	-0.59	-0.85	-0.66	-0.92	0.26	-1.21
GT6m7	-0.16	0.92	0.70	0.92	0.70	-2.23*	-0.37
IGB3	-1.51	-2.05	-2.19	-2.18	-2.33	0.19	-1.86
All	-1.24	-1.02	-1.24	-0.59	-0.93	-0.73	-4.37
Europe							
	Tajima's D	Fu and Li's D*	Fu and Li's F*	Fu and Li's D	Fu and Li's F	Fay and Wu's H	Fu's Fs
FS	-1.38*	-0.30	-0.72	-0.34	-0.76	0.27	0.91
GGTA1	-1.24	-0.71	-0.99	-0.76	-1.04	0.20	-1.59
GT6m7	-0.52	1.07	0.71	1.09	0.73	-2.31**	-1.24
IGB3	ND	ND	ND	ND	ND	ND	ND
All	-1.32	-0.79	-1.09	-0.93	-1.25	-1.80	-2.61

\* P &lt; 0.05; \*\* P &lt; 0.01; \*\*\* P &lt; 0.001.

# Human Genetic Diversity In Genes Related To Host-Pathogen Interactions

**Table 4.** Population differentiation in seven human pseudogenes regions. Each value is the average for all SNPs in each genome region.

	All region <sup>a</sup>			Gene region <sup>b</sup>				
	SNPs	F <sub>ST</sub> <sup>c</sup>	F <sub>CT</sub> <sup>d</sup>	F <sub>SC</sub> <sup>e</sup>	SNPs	F <sub>ST</sub> <sup>c</sup>	F <sub>CT</sub> <sup>d</sup>	F <sub>SC</sub> <sup>e</sup>
FS	19	0.095	0.082	0.026	3	0.108	0.093	0.029
GGTA1	15	0.086	0.079	0.019	11	0.091	0.079	0.025
GT6m7	34	0.098	0.085	0.026	13	0.090	0.073	0.029
IGB3	14	0.086	0.074	0.023	2	0.085	0.066	0.030
HGT2	12	0.124	0.118	0.024	6	0.143	0.134	0.031
LOC401913	14	0.099	0.088	0.025	7	0.097	0.089	0.022

<sup>a</sup> Includes SNPs from 30 Kb upstream to 30 Kb downstream from the gene region. <sup>b</sup> Includes SNPs only in the region between the homologous sequences to the first and last exon of ABO. <sup>c</sup> among the 39 populations <sup>d</sup> among seven continental groups <sup>e</sup> average value of F<sub>ST</sub> calculated among populations within the same continental group

**Table 5.** Interspecific evolutionary analysis of four human pseudogenes.

		Likelihood	Np	2* lnL	diff Np	χ <sup>2</sup> (P value)
FS (complete gene)						
H1	Free (ω = 0.0001)	-2829.92	15	5.07	1	0.024
H0	ω = 1	-2832.46	14			
FS (exon 7)						
H1	Free (ω = 0.0001)	-1887.02	15	5.26	1	0.022
H0	ω = 1	-1889.65	14			
GT6m7 (complete gene)						
H1	Free (ω = 0.24)	-2537.40	19	1.34	1	0.247
H0	ω = 1	-2538.07	18			
GT6m7 (exon 7)						
H1	Free (ω = 0.23)	-2055.30	19	0.80	1	0.371
H0	ω = 1	-2055.70	18			
GGTA1 (complete gene)						
H1	Free (ω = 999) <sup>a</sup>	-2148.47	19	1.85	1	0.173
H0	ω = 1	-2149.40	18			
IGB3 (complete gene)						
H1	Free (ω = 0.50)	-2771.38	15	1.98	1	0.159
H0	ω = 1	-2772.37	14			
IGB3 (exon 7)						
H1	Free (ω = 0.41)	-1770.87	15	1.82	1	0.177
H0	ω = 1	-1771.78	14			

<sup>a</sup> no synonymous differences

**4.4 Chapter 4**

Worldwide human genetic diversity in four major pathways of glycan biosynthesis

Manuscript in preparation



Worldwide human genetic diversity in four major pathways of glycan biosynthesis

Ferrer-Admetlla, A., Sikora, M., Laayouni, H., Bosch, E., Casals, F. and Bertranpetit, J.

<sup>1</sup> Institut de Biologia Evolutiva (CSIC-UPF), CEXS-UPF-PRBB,

Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain.

<sup>2</sup> CIBER Epidemiología y Salud Pública (CIBERESP);

<sup>3</sup> Present address: Ste Justine Hospital Research Centre, Department of Pediatric, Faculty of Medicine, University of Montreal, Montreal, Quebec H3T 1C5, Canada

\* Correspondence: [jaume.bertranpetit@upf.edu](mailto:jaume.bertranpetit@upf.edu)

## Introduction

Glycans have an important biological role in multicellular organs and organisms, since they are mediating the interaction between cells and the surrounding matrix. Specifically, secreted or outer cell surface glycans can modulate or mediate cell-cell, cell-matrix or cell-molecule interactions, and they can also mediate the interaction between different organisms (for example between host and parasite). Surfaces of all cell types are carpeted by a layer of different kinds of glycoconjugates named glycocalyx. These glycoconjugates are built by glycosylation enzymes, which are responsible of adding glycan moieties to the proteins or lipids that will lay on the cell surface. Such glycosylated structures confer protection to the cell against pathogen infections, but it occurs that they may occasionally be recognized by pathogens as target molecules. Therefore, these genes are good candidates to show signatures of genetic adaptation (Bishop and Gagneux 2007).

In the present work we study a total of 70 genes related to glycosylation biosynthesis on a set of human samples worldwide distributed. The genes of study fall in four major glycosylation processes: O-glycosylation biosynthesis, lactoseries biosynthesis, neo-lactoseries biosynthesis and ganglioseries biosynthesis. The former biosynthesis pathway determines the production of certain molecules, as mucines, that contribute to protection against pathogenic infections. Major components of O-glycan biosynthesis pathway are galactosyltransferases. These enzymes, also named ppGalNAcT, determine the range of O-glycans synthesized by a cell. The specific function of this enzymes is adding an N-acetylgalactosamine (GalNAc) moiety to the -OH of serine or threonine residue of polypeptides through different kind of linkages and different type of acceptors. As an example, proteins that are heavily O-glycosylated via  $\alpha$ -linked O-GalNAc are the above cited mucins. These proteins are found in mucous secretions, on epithelial cell surfaces and in body fluids making them a shield against physical and chemical damage and protecting cells against pathogen infections.

---

In turn, lactoseries biosynthesis pathway is responsible of Lewis antigen ( $Le^x$ ,  $Le^b$ , B  $Le^x$ , B  $Le^b$  and sialyl $Le^x$ ) formation, which some pathogens mimic to infect cells. Lacoseries are cell surface carbohydrates implicated in cell-cell interactions. Besides its role in Lewis antigen synthesis, lactoseries have many important biological roles not related to host-pathogen interaction: 1) embryonic development at early stages of preimplantation (Dodd and Jessell 1985), or 2) sorting of axons from olfactory nerves as they converge towards their target (Storan et al. 2004). Regarding again to Lewis antigens synthesis its worth noticing that Neu5Ac, a sialic acid exclusive of human lineage and added at the last step of lactoseries pathways to form sialyl $Le^x$ , has been shown to confer protection to *Plasmodium reichenovi*, but at the same time susceptibility to *Plasmodium falciparum* which recognizes and uses it to infect red blood cells (Lieberman 2004)

Neo-lactoseries biosynthesis pathway is also involved in the synthesis of Lewis antigens. In contrast to lactoseries biosynthesis pathway, which involves a variety of enzymes, neo-lactoseries is mainly constituted by fucosyltransferases (FUT1-FUT9, with the exception of FUT8 which is involved in Keratan sulfate biosynthesis). These molecules are typically transferring fucose from a GDP-donor to N-linked type complex glycopeptides. So, fucose is frequently a terminal moiety of neo-lactoseries branched chain glycoconjugates. By the fact of being terminal molecules, fucosyltransferases could be susceptible to play some role in the contact to pathogens. So far, two fucosyltransferases (*FUT1* and *FUT2*), responsible of the ABO antigen production, have been shown to confer protection/susceptibility towards certain infections (Fry et al. 2008; Marionneau et al. 2005).

The last biosynthesis pathway (ganglioseries biosynthesis) comprises several sialyltransferases. These enzymes are responsible of the addition of sialic acids on the terminating branches of N-glycans, O-glycans and glycosphingolipids. Sialic acids are abundant on outer cell membranes, on the interior of lysosomal membranes and on secreted glycoproteins suggesting they have an important role in the stabilization of molecules and membranes and in modulating the interaction



with environment. As an example, it has been demonstrated that alpha-2,3-sialyltransferase contributes to *Neisseria gonorrhoeae* pathogenesis in an in vivo murine model.

The interaction of pathogens to some of the gene products belonging to these four biosynthesis pathways has been demonstrated in single gene studies, especially the most widely known blood groups. As pathogens can be regarded as powerful selective agents, then, the enzymes responsible of glycan structure formation could possibly be under selective pressures mediated by pathogens. Likely, as a consequence of the existence of area-specific pathogens we would possibly detect local adaptation events on our candidate genes, as has been demonstrated for *ABO* and *FUT2* both genes showing the footprint of balancing footprint as a consequence of pathogen pressures. The present work attempts to detect the footprint of positive selection on our set of candidate genes through; 1) the analysis of allele frequencies in different populations, 2) the estimation of the level of population differentiation and, 3) the analysis of the linkage disequilibrium decay. The final purpose of this study is the understanding of the selective pressures acting on these genes (if any) within their functional and geographical context. It is worth to note that there is an intention of being exhaustive both in studying populations spread all over the world and of the genes related to a unique and complex process, glycosylation.

## Material and methods

### Samples

The study was conducted using human DNA samples from HGDP-CEPH Human Genome Diversity Cell Line Panel (Cann et al. 2002). According to geographic and ethnic criteria, samples were grouped into 39 different populations and regrouped into seven continental regions (Gonzalez-Neira et al. 2004). From the original sample panel containing 1,064 samples, only a subset of 971 samples were considered for this study. 78 individuals first and second degree related, 13

---

duplicated samples and two individuals with uncertain population origin (HGDP00770 and HGDP00980) were removed (Rosenberg 2006).

### Genes

Using GO and KEGG databases (<http://www.geneontology.org/>; <http://www.genome.ad.jp/kegg/pathway.html>), 70 genes were selected. Specifically, we choose all known genes involved in the four major glycosylation pathways: O-glycans biosynthesis, lacroseries biosynthesis, neo-lactoseries biosynthesis and ganglioseries biosynthesis. Most of the genes are part of three glycosylation gene families: fucosyltransferases, sialyltransferases and galactosydases (Table S1 and pathways); nonetheless, the inclusion was not for genetic or evolutionary criteria, but exclusively by functional: genes were included if they were part of the biosynthetic pathways described.

### SNP selection

SNPs were selected from HapMap Phase I (<http://www.hapmap.org>) and dbSNP build 126 (<http://www.ncbi.nlm.nih.gov/SNP>) databases using PupaSNP (<http://pupasnp.bioinfo.ocha.fib.es>) and SNPbrowser (Applied Biosystems). SNPs were chosen from HapMap preferentially with minor allele frequency (MAF) over 10%. SNPs in the coding region were selected at approximately 1 SNP per 5 Kb density. We also covered 30 Kb from the 5' and 3' gene extremes by placing one SNP at 5 Kb, 10 Kb, 20 Kb and 30 Kb from both gene extremes, adding one SNP up to 100-200 bp at 5' end and, another up to 1 Kb at 3' end when possible. Contiguous genes were grouped for analysis in single regions: SIAT7A and SIAT7B were clustered into the SIAT7A\_B region; FUT3, FUT5 and FUT6 into the FUT3\_5\_6 region; FUT1 and FUT2 into the FUT1\_2 region; SIAT7C and SIAT7E into SIAT7C\_E, and SIAT7D and SIAT7F into SIAT7D\_F. All analyzed regions were independent from each other.

### Public available genotyping data

Gene coverage was enriched with 13437 SNPs from publicly available data from CEPH-HGDP samples typed with the Illumina 650k chip (Jakobsson et al. 2008; Li et al. 2008) in order to increase gene coverage and, specially, fulfill flanking gene regions up to 400Kb upstream and downstream each gene, an information that allows further analysis on these genes.

### Genotyping

Genotyping was performed using the ABI SNPLex technology (Applied Biosystems) based on an oligonucleotide ligation assay (OLA). Alleles are detected using two competitive primers (one for each allele) joined to their ZIPcode (labeled nucleotidic fragment of particular length) and using a universal PCR primer. The combination of different lengths and fluorescent dyes for each ZIPcode allows allele discrimination in multiplex reactions (up to 48 SNPs in a single nucleotide reaction). Samples were typed in three assays of 384 samples (one plate). One control sample (CEPH001) was added twice per plate to check for reproducibility and genotyping quality. Allele detection was achieved by a capillary electrophoresis on ABI3730 (Applied Biosystems) and genotypes were assigned with GeneMapper software v3.5 (Applied Biosystems). 764 SNPs produced reliable genotypes. When combining our genotypes with those retrieved from HGDP-CEPH human genome diversity panel database (<http://www.cephb.fr/en/hgdp/main.phpand>) we excluded 140 SNP due to discordant genotypes.

### Analysis

SNPator web application package (<http://bioinformatica.cegen.upf.es>) (Morcillo-Suarez et al. 2008) had been used for data storing, handling and for performing part of the analysis. Hardy-Weinberg equilibrium was tested and 125 SNPs were removed from the analysis because they did not reach Hardy-Weinberg equilibrium.

---

### Allele frequency test

The distribution of allele frequencies was analyzed within 100 Kb sliding windows with a 30 kb overlap across every gene region in each population. We did not consider those windows with less than five SNPs. Based on the method proposed by Walsh et al. (2006) we computed the proportion of SNPs presenting a minor allele frequency (MAF) lower than 0.1 and the proportion of SNPs with MAF higher than 0.4 within each window. To perform derived allele frequency (DAF) analysis, we retrieved the ancestral allele for all SNPs analyzed by querying the orthologous nucleotide of chimpanzee at UCSC genome browser (*snp129OrthoPt2Pa2Rm2*). we excluded those SNPs (1,076) being discordant among the different primate species (macaque, chimpanzee and human) or having available sequence information for only one primate species that did not match with any of the human alleles. Subsequently, we calculated the proportion of SNPs presenting a derived allele frequency lower than 0.2 and higher than 0.8 (Walsh et al. 2006) per each window.

### $F_{ST}$ statistic

Population differentiation statistics  $F_{ST}$ ,  $F_{CT}$  and  $F_{SC}$  were computed using Arlequin software v3.11 (Excoffier 2005). Statistical significance for  $F_{ST}$  values were estimated by performing an Analysis of molecular variance (AMOVA). To perform population differentiation statistics we did not use those genes on the X chromosome (*OGT* and *C1GALT1C1*) as their particular population genetics give place to higher  $F_{ST}$  values which can not be compared to autosomic data.

### Long Range Haplotype analysis (LRH)

We used fastPHASE software (Scheet and Stephens 2006) to phase genotypes for all SNPs at population level. LD decay was analyzed with two different approximations: the EHH method (Sabeti et al. 2002) and another method inspired on the iHS (Voight et al. 2006) and implemented in our group (Sikora et al. 2008). For both tests we calculated the extended haplotype homozygosity (EHH) using

Sweep software version 1.1 (<http://www.broad.mit.edu/mpg/sweep/index.html>). For the first test, we measured the EHH, and the relative EHH (REHH) at a fixed length of 0.3 cM in both directions from each of the haplotype cores defined in each region (Sabeti et al. 2002). Then, we plotted such haplotype distance versus its respective core haplotype frequency. Core haplotypes were distributed in 20 bins of 5% frequency and their EHH and REHH measures were log transformed to obtain a normal distribution that then we used to check for cores with extreme values. For the second test we took each SNP as a core. Then we measured the extended haplotype length (EHL), which is the physical distance at which EHH reaches a fixed value (0.25) for both, the minor and the major alleles of the core analyzed. After that, we calculated a score for each SNP based on the log of the ratio between the EHL of minor and major alleles. To compare values across SNPs, we then normalized each score value using the mean and the standard deviation of all core SNPs falling in the same frequency bin. For LHR tests we excluded the two genes on X chromosome in order to limit comparisons within autosomic data.

To calculate statistical significance we used the mean and standard deviation of the empirical distribution of the respective scores. *P*-values were adjusted based on the concept of false discovery rate (Benjamini and Hochberg 1995). If no SNPs are truly significant the *p* values will follow a *U* (0,1), where *U* stands for 'uniform distribution'. The so-called Mixture Distribution Partitioning (MDP) methodology assumes that the distribution of *p*-values consists of a set of null *p*<sub>0</sub> and alternative *p*<sub>1</sub> components. This partition forms the basis for estimating various quantities as for example the *q*-values, which were obtained here with the QVALUE software (Storey and Tibshirani 2003). We used a *q*-value cut-off  $\leq 0.05$ , meaning that the maximum expected proportion of false positives incurred when calling a particular gene as significant is 5%.

## Results

In this work we have analyzed genotype data for 14,036 SNPs in 971 samples covering most of the human genetic diversity. The SNPs analyzed cover 70 gene

---

regions (Table 1) with a density of 3.4Kb/SNP, being lower the density of SNPs outside gene regions (4.3Kb/SNP).

#### Minor and derived allele frequency analysis (MAF and DAF)

Analysis of the allele frequency may provide information on the selective forces that have existed on a given gene. These analyses are based on detecting an excess of alleles with extreme frequencies. Thus, a high proportion of rare MAF (<0.1) could indicate positive selection, since selective sweep events lead to a reduction of the variation. In contrast, a high proportions of high MAF (>0.4) could be considered an indicator of balancing selection, since this kind of selection maintains allele frequencies most of the times at intermediate levels. In the case of DAF, an excess of high frequency derived alleles could indicate positive selection. On the contrary, excesses of low frequency derived alleles could be indicative of purifying selection. Figures 1a and 1b show the proportion of MAF or DAF, respectively, of all 100Kb windows within the same continent. In turn, colored diamonds indicate windows falling in or overlapping to candidate regions. All continental regions represented in Figure 1a present some windows outside the 99% confidence interval of the full distribution. More specifically, in lower MAF analysis we find seven regions standing out from the full distribution (Table 2a) corresponding to various populations from the same continent. This is the case of *GALNT7\_17*, which shows three overlapping windows in North West China and a different window shared in Balochi and Pathan, both of them different than the 100Kb window presented in five European populations (Basque, French, North Italy, Orcadian and Russian). It is also interesting the case of *GALNT5* which outstands in Brahui, Sindhi, Balochi (Central and South Asia) and Adygei, Basque and Sardinian (Europe). When considering upper DAF analysis, we found six genes with high proportion of DAF>0.8 (Table 2b). Out of these genes, two overlap with the ones detected in lower MAF analysis: *GALNT5* and *GALNT7\_17* and *GALNTL1*. There are two genes *SIAT8A* and *B4GALT3* presenting an excess of high DAF in nearly all populations of the same continent (America and Sub-Saharan Africa, respectively). Particularly interesting is *GALNT7\_17* which did not show up for MAF analysis in American populations, whereas in DAF analysis

points to two (Karitiana and Pima). Also notable are *GALNT5*, outstanding in two Central and South Asian populations and *SIAT7D\_F* doing so in two European populations.

For lower DAF analysis, we detect eight outstanding genes. The striking results are presented by *B3GNT3* which shows an excess of lower DAF (0.75) in exactly the same 100Kb window in five out of the seven Central and South Asia populations (Balochi, Brahui, Hazara, Burusho and Sindhi) although also presenting an excess of low DAF in Russian. Moreover, *GALNT2* and *SIAT6* outstands in four European populations. Notice the case of *FUT8* which, within a 310 Kb range, shows a high proportion of low DAF from 0.75 to 0.90 in several populations of different continental origin (Biaka Pygmies, Cambodian, Hazara, Maya and North West China; Table 2c). In conclusion, the genes that shows the highest proportion of lower DAFs overall are *B3GNT3* in more than half of Central and South Asia populations, and *GALNT2* and *SIAT6* in four European populations.

Table 2d shows six genes with an excess of high MAFs in more than one population from the same continental region. The most interesting cases are presented by a single *FUT10* window in North East China, South China and Yakut that overlaps with another outstanding window in Pathan, Sindhi and Hazara and, *GALNT7\_17* which shows 4 non-overlapping windows with a high proportion of high MAFs (0.73-0.83). The first window comprises 160 Kb and outstands in Adygei, Karitiana, Orcadian and Druze; the second comprises 130 Kb in Colombian; the third is 130Kb in length and stands out in Adygei, Bedouin, Druze and NAN Melanesian; and, the last one is a single window in Colombian and Karitiana. By analyzing regions with high proportions of high MAF a preliminary sign from those genes possibly under balancing selection would be obtained.

#### Population differentiation statistic

We have calculated the population differentiation in allele frequency for each SNP included in the study among the 39 populations ( $F_{ST}$ ), among the seven

---

continental regions ( $F_{CT}$ ) and among the populations within continental regions ( $F_{SC}$ ). Table 3 shows  $F_{ST}$ ,  $F_{CT}$  and  $F_{SC}$  mean values per each gene or gene cluster. It is of interest to note that most of the variation observed is due to differences among continental regions. The average  $F_{ST}$  across all SNPs analyzed is 0.107. This value is at the same range as that obtained when considering the whole set of 650000 SNPs from Illumina 650K array typed in HGDP-CEPH samples ( $F_{ST}=0.096$ ). Also similar are the upper 95% and 99% percentile of our  $F_{ST}$  distribution (0.24 and 0.34 respectively) to the whole 650K data set distribution (0.22 at the 95% percentile and 0.31 at the 99% percentile). Notice that the genes presenting the highest mean  $F_{ST}$  in our study are *GALNT5* and *FUT1\_2*. Since  $F_{ST}$  varies a lot even for very close SNPs, we just considered for this analysis those genes whose mean has been calculated using at least 20 SNPs. According to this condition we end up with seven genes showing a  $F_{ST}$  value over 0.096 (the overall 650000  $F_{ST}$  mean) (Table 3). When we calculate the averaged mean of each gene family, we find out glycosyltransferases are presenting the more extreme values. Thus, N-acetylgalactosaminyltransferase subfamily is showing the highest  $F_{ST}$  mean (0.096), whereas the subfamily of galactosyltransferases is presenting the lowest (0.079). The former case is due to five members of this subfamily (Table 3). Interestingly, three of these genes correspond with the gene regions outstanding in MAF and DAF analysis: *GALNT5*, *GALNT7\_17* and *GALNT11\_15*.

#### LRH tests

As a last strategy to detect the signature of recent selection, we applied two methods based on the Extended Haplotype Homozygosity (EHH). These tests focus on the detection of long range haplotypes of high frequency since selective sweeps rise the frequency of the selected allele before recombination breaks down its association with nearby markers (See Methods).

For one of the tests, we have calculated the normalized Haplotype Length Score (nHLS) based on cores of one SNP along the 63 studied regions (see material and methods). Tables 4a to 4e show the significant results for each continental region after correcting for multiple testing. Table 4a shows results for Sub-Saharan



continental group. Eight genes stand out in this test because of showing a significant nHLS after multiple testing correction. Four of them have significant nHLS scores in more than half of the populations: *GALNT12* (4), *GALNT13* (3), *GALNT7\_17* (4) and *SIAT7C\_E* (3). Table 4b summarizes results for European continental group. This table reveals nine genes standing out of nHLS tests. From these genes, three (*GALNT13*, *GALNT11\_15* and *SIAT6*) have significant nHLS in four, three and four European populations respectively, and three more genes (*FUT8*, *SIAT7C\_E* and *WBSCR17*) present high scores in two populations. In the case of Central and South Asia (Table 4c) seven genes show significant nHLS scores. Three of them have significant core SNPs in various populations: *GALNT13* (5), *SIAT7C\_E* (5), and *GALNT7\_17* (3).

East Asia is the continental group presenting more genes with high HLS values (13) (Table 4d). Seven of these genes present significant scores in more than one population. This is the case of *GALNT11\_15* (2), *GALNT12* (2), *GALNT13* (3), *GALNT5* (3), *GALNT7\_17* (2) and *SIAT7C\_E* (5). Notice that *GALNT11\_15* and *SIAT7C\_E* are the most interesting cases since both of them are outstanding in five out of the six East Asian populations. Finally, when focusing on results concerning the Americas (Table 4e) we detect ten genes with significant nHLS scores, two of them (*GALNT12* and *WBSCR17*) show up in two populations whereas the majority (*GALNT1*, *GALNT10*, *GALNT11\_15*, *GALNT8*, *GALNT13*, *GALNT7\_17* and *GALNT18*) show up in single populations. The most interesting case is *SIAT7C\_E* which stands out in three out of the five American populations (Colombian, Pima and Surui). Significant results in the case of Middle East and North-Africa and Oceania continental regions are scant and are not significant after multiple testing correction.

Interestingly, there are genes showing up in various populations in only one continental region. This is the case of *GALNT5* which is presenting significant nHLS in East Asia and *SIAT6* which stands out in four European populations. Also notice, that some genes coincide in various continental regions. This is the case of *WBSCR17* appearing in Europe and America; *GALNT12* standing out in Sub-Saharan Africa and in two American populations; *GALNT11\_15* which is present in

---

European and East Asia, *GALNT13* which presents high nHLS scores in all continental region except in America; and *GALNT7\_17* which stands out in various populations from Central and South Asia, East Asia and Sub-Saharan Africa.

To detect local adaptation we have considered all populations separately and then we have looked for adjacent SNPs with an nHLS over |2.5|. By doing so we have detect 4 SNPs within *GALNTL1* presenting high scores in Central and South Asia and East Asia, which would be in agreement with results obtained form MAF and DAF analysis.

By making use of FatiGO (Al-Shahrour et al. 2004) web tool we check whether genes presenting significant nHL scores fell in the same KEGG pathway results indicate that within genes presenting significant nHLS there is a significant over-representation of genes belonging to O-glycan biosynthesis. (p-value=2.28e<sup>-3</sup> and q-value=1.57e<sup>-2</sup>)

With the second approach, based on EHH, we identify three genes (*GALNT1*, *SIAT4A* and *SIAT7C\_E*) outstanding in East Asian populations (Figure 2); one gene in Middle East and North Africa (*GCNT2*), and two genes in Europeans (*GCNT2* and *GALNT18*). Although these core haplotypes show p-values lower than 0.05 none of them remain significant after applying multiple testing correction.

## Discussion

The purpose of the present study was to look for signatures of selection in all known genes to be involved in four metabolic pathways related to glycosylation biosynthesis because of their likeliness to be under pathogen-mediated selection. As it is well known that pathogens are highly structured with geography, it is expected that the specific adaptation to different sets of pathogenic environments will have left a different footprint in the genome in populations inhabiting different parts of the planet.

To detect the possible signatures of adaptive evolution on candidate genes we undertook an analysis based on three different methods: allele frequencies, population differentiation and long range haplotype tests. As these methods give sight about adaptive selection acting on a locus in different time scales (Sabeti et al. 2006), have different strength and are differently affected by ascertainment bias (Sabeti et al. 2007) as many as these tests give a signature of selection at the same region more reliable the detected signal will be. In this sense our results reveal four genes presenting evidences of positive selection for at least two of these tests. To take into account the possible demographic factors affecting results we performed calculations at population level and we interpreted results within a continental context. Finally, we considered trustable those genes presenting different evidences for positive selection in at least half of the populations of the same continental region. Thus, *GALNT7\_17* shows significant nHL scores and a high proportion of low MAF in Central and South Asia. Also notice that in this continental region some core SNPs with significant nHLS within *GALNT13* and *SIAT7E\_C* are accompanied by  $F_{ST}$  over 95% CI, although they do not present outstanding results for allele frequency tests. *GALNT5* shows significant nHL scores and present high proportions of high DAFs in half of East Asia populations. in this continental region *GALNT13* and *GALNT5* also shows some SNPs with significant nHLS and simultaneously  $F_{ST}$  values over the 95% of the distribution. Two SNPs (one in each gene) behaving so are synonymous. Notice the curious behavior of *SIAT6* which outstands in nHLS test in four European populations for some of which it shows an excess of lower MAFs.

Interestingly, three of the genes presenting evidences of positive selection are members of the galactosyltransferase family. As stated at results section the genes in which we detect positive selection signatures coincide to participate in the same biosynthesis pathway. Strikingly, the three genes encode enzymes that catalyze the first step of the O-glycan biosynthesis pathway, so they transfer a N-acetylgalactosamine (GalNAc) moiety to the -OH of serine or threonine residue of polypeptide. This pathway determines the production of molecules protecting cells towards pathogen infections as mucins. It is worth noticing that the different

---

galactosyltransferases we report are presenting signals of selection in different continental regions. Thus, *GALNT7\_17* shows evidences of selection in Central and South Asia populations, whereas *GALNT13* and *GALNT5* present the footprint of positive selection in East Asia populations.

It is also remarkable that the other two genes where we detect some signatures of positive selection: *SIAT6* and *SIAT7C\_E*, also present such selection signatures in particular continental regions which could indicate that pathogens are able to finely discriminate among different sialylated structures.

In conclusion, it seems that different continental groups present different positively selected genes in the first step of O-glycosylation pathway, although *GALNT13* is showing some evidences of selection in nearly all continental groups. Sialyltransferases are also susceptible to pathogen-mediated selection, specially when considering together the outstanding  $F_{ST}$  mean values and the top 95%CI values of allele frequency analysis distributions since in this case six sialyltransferases (*SIAT4A*, *SIAT7C\_E*, *SIAT8A*, *SIAT8B*, *SIAT8C* and *SIAT8F*) and four galactosyltransferases (*GALNT5*, *GALNTL1*, *GALNT11\_15* and *GALNT13*) present outstanding results.

#### Acknowledgements

This research was funded by grants BFU2005-00243 and SAF-2007-63171 awarded by Ministerio de Educación y Ciencia (Spain), by the Direcció General de Recerca of Generalitat de Catalunya (Grup de Recerca Consolidat 2005SGR/00608). Computation was helped by the National Institute for Bioinformatics ([www.inab.org](http://www.inab.org)), and SNP genotyping services were provided by the Spanish "Centro Nacional de Genotipado" (CEGEN; [www.cegen.org](http://www.cegen.org)); both are platforms of Genoma España. A F-A is supported by a PhD fellowship from UPF and MS from the Programa de becas FPU del Ministerio de Educación y Ciencia, Spain (AP2005-3982).

## Human Genetic Diversity In Genes Related To Host-Pathogen Interactions

**Table 1.** Summary of all genes and gene clusters analyzed

Gene	Chromoso	Initial	Final	Pathway
ABO	9	135120384	135140451	Glycosphingolipid biosynthesis - lactoseries/-neo-
B3GALT1	2	168383428	168435612	Glycosphingolipid biosynthesis - lactoseries
B3GALT2	1	191414799	191422347	Glycosphingolipid biosynthesis - lactoseries
B3GALT5	21	39850239	39956685	Glycosphingolipid biosynthesis - lactoseries
B3GNT1	11	65869419	65871737	Glycosphingolipid biosynthesis - neo-lactoseries
B3GNT2	2	62276766	62305370	Glycosphingolipid biosynthesis - neo-lactoseries
B3GNT3	19	17766658	17785385	Glycosphingolipid biosynthesis - neo-lactoseries
B3GNT4	12	121254181	121258037	Glycosphingolipid biosynthesis - neo-lactoseries
B3GNT5	3	184453726	184473873	Glycosphingolipid biosynthesis - lactoseries/ - neo-
B3GNT6	11	76423083	76430653	O-Glycan biosynthesis
B4GALT1	9	33100642	33157231	Glycosphingolipid biosynthesis - neo-lactoseries
B4GALT3	1	159407725	159413938	Glycosphingolipid biosynthesis - neo-lactoseries
B4GALT4	3	120413277	120442442	Glycosphingolipid biosynthesis - neo-lactoseries
C1GALT1	7	7240414	7250506	O-Glycan biosynthesis
C1GALT1C	X	119643562	119647969	O-Glycan biosynthesis
FUT1_2	19	53891050	53950459	Glycosphingolipid biosynthesis - lactoseries/ - neo-
FUT10	8	33365956	33450206	-
FUT11	10	75202055	75205982	-
FUT12	20	30259357	30290128	-
FUT13	21	45508271	45532239	-
FUT4	11	93916775	93922712	Glycosphingolipid biosynthesis - neo-lactoseries
FUT5_6	19	5781637	5821551	Glycosphingolipid biosynthesis - lactoseries(FUT3)/ - neo-
FUT7	9	139044447	139047283	Glycosphingolipid biosynthesis - neo-lactoseries
FUT8	14	64947593	65279715	-
FUT9	6	96570590	96760477	Glycosphingolipid biosynthesis - neo-lactoseries
GALNT1	18	31488594	31545792	O-Glycan biosynthesis
GALNT10	5	153550488	153780003	O-Glycan biosynthesis
GALNT11_	7	151284444	151347945	O-Glycan biosynthesis
GALNT12	9	100609802	100652180	O-Glycan biosynthesis
GALNT13	2	154436712	155018734	O-Glycan biosynthesis
GALNT14	2	30986837	31205994	O-Glycan biosynthesis
GALNT18	11	11248999	11600128	O-Glycan biosynthesis
GALNT2	1	228464239	228482595	O-Glycan biosynthesis
GALNT3	2	166311570	166335456	O-Glycan biosynthesis
GALNT4	12	88437321	88442666	O-Glycan biosynthesis
GALNT5	2	157822586	157876159	O-Glycan biosynthesis
GALNT6	12	50032100	50071467	O-Glycan biosynthesis
GALNT7_1	4	172971232	174481692	O-Glycan biosynthesis
GALNT8	12	4700013	4752153	O-Glycan biosynthesis
GALNT9	12	131246870	131415862	O-Glycan biosynthesis
GALNTL1	14	68796668	68890936	O-Glycan biosynthesis
GCNT1	9	78263966	78312152	O-Glycan biosynthesis
GCNT2	6	10663935	10737587	Glycosphingolipid biosynthesis - lactoseries/ - neo-
GCNT3	15	57691275	57699494	O-Glycan biosynthesis
GCNT4	5	74359045	74362480	O-Glycan biosynthesis
OGT	X	70682292	70712465	O-Glycan biosynthesis
SIAT1	3	188131210	188279035	-

## RESULTS

---

SIAT10	3	99933842	99995926	Glycosphingolipid biosynthesis - neo-lactoseries
SIAT4A	8	134540325	134653344	O-Glycan biosynthesis/ - ganglioseries
SIAT4B	16	68972810	68992042	O-Glycan biosynthesis/ - ganglioseries
SIAT4C	11	125731306	125789743	Glycosphingolipid biosynthesis - lactoseries
SIAT6	1	43974491	44169418	Glycosphingolipid biosynthesis - lactoseries
SIAT7A_B	17	72073077	72151489	O-Glycan biosynthesis(SIAT7A) - ganglioseries(SIAT7B)
SIAT7C_E	1	76312992	77302325	Glycosphingolipid biosynthesis - ganglioseries
SIAT7D_F	9	129687422	129719126	Glycosphingolipid biosynthesis - ganglioseries
SIAT8A	12	22245200	22378915	Glycosphingolipid biosynthesis/ - neo-lactoseries/ -
SIAT8B	15	90738144	90812962	Glycosphingolipid biosynthesis - ganglioseries
SIAT8C	18	53170719	53187159	Glycosphingolipid biosynthesis - ganglioseries
SIAT8D	5	100170803	100266869	Glycosphingolipid biosynthesis - ganglioseries
SIAT8E	18	42513079	42591037	Glycosphingolipid biosynthesis - ganglioseries
SIAT8F	10	17402682	17536260	-
SIAT9	2	85919782	85969668	Glycosphingolipid biosynthesis - ganglioseries
WBSCR17	7	70235725	70816520	O-Glycan biosynthesis

---

\*Positions correspond to the human genome positions from UCSC hg18 assembly. Hyphen stays when any of the studied pathways contain the gene.

## Human Genetic Diversity In Genes Related To Host-Pathogen Interactions

**Table 2a to 2d.** Outlier regions of lower and upper MAF/DAF analysis concerning candidate genes.

Gene	Position	Population	maf.snps	maf.lower
GALNT5	157902879	Adygei	7	0.86
GALNT5	157872879	Adygei	11	0.64
GALNT5	157902879	Balochi	8	0.75
GALNT5	157872879	Balochi	13	0.69
GALNT5	157902879	Basque	11	0.91
GALNT5	157872879	Basque	14	0.71
GALNT5	157902879	Brahui	11	0.82
GALNT5	157872879	Brahui	15	0.73
GALNT5	157902879	Palestinian	9	0.67
GALNT5	157902879	Sardinian	7	0.86
GALNT5	157902879	Sindhi	10	0.90
GALNT5	157872879	Sindhi	13	0.77
GALNT7_17	173593751	Balochi	14	0.64
GALNT7_17	172933751	Basque	10	0.80
GALNT7_17	174463751	Bedouin	15	0.67
GALNT7_17	172933751	Druze	9	0.78
GALNT7_17	172933751	French	10	0.80
GALNT7_17	173833751	French	17	0.65
GALNT7_17	172933751	Hazara	8	0.75
GALNT7_17	173293751	Mbuti.Pygmies	12	0.58
GALNT7_17	174463751	North.East.China	15	0.80
GALNT7_17	172933751	North.Italy	8	0.63
GALNT7_17	174433751	North.West.China	10	0.70
GALNT7_17	174463751	North.West.China	10	0.70
GALNT7_17	174493751	North.West.China	9	0.67
GALNT7_17	172933751	Orcadian	9	0.78
GALNT7_17	173353751	Papuan	9	0.89
GALNT7_17	173593751	Pathan	14	0.64
GALNT7_17	172933751	Russian	9	0.78
GALNT7_17	173593751	Russian	14	0.64
GALNT7_17	173563751	Russian	16	0.63
GALNT7_17	174283751	Surui	5	1.00
GALNT7_17	174313751	Surui	10	1.00
GALNT7_17	174373751	Surui	11	1.00
GALNT7_17	174343751	Surui	16	1.00
GALNT7_17	174403751	Surui	7	0.86

\* The number of SNPs and the proportion lower or upper MAF/DAF are given per window analyzed (100Kb). Position corresponds to the center of each 100kb window.

Table 2b.

Gene	Position	Population	daf.snps	upper.daf
B4GALT3	159361585	Yoruba	5	0.60
B4GALT3	159361585	San	5	0.40
B4GALT3	159361585	Biaka.Pygmies	5	0.40
B4GALT3	159361585	Mandenka	5	0.40
GALNT5	157902879	Orcadian	10	0.60
GALNT5	157902879	Palestinian	10	0.60
GALNT5	157902879	Papuan	7	0.86
GALNT5	157902879	Kalash	8	0.75
GALNT5	157902879	NAN.Melanesian	7	0.86
GALNT5	157902879	North.West.China	10	0.60
GALNT5	157902879	Sindhi	10	0.60
GALNT5	157902879	South.China	9	0.67
GALNT5	157902879	Yakut	9	0.67
GALNT5	157902879	Pathan	9	0.67
GALNT5	157902879	Russian	10	0.60
GALNT5	157902879	Sardinian	9	0.67
GALNT5	157902879	Burusho	10	0.60
GALNT5	157902879	French	10	0.60
GALNT5	157902879	Adygei	9	0.67
GALNT5	157902879	Balochi	9	0.67
GALNT5	157902879	Druze	10	0.60
GALNT5	157902879	Bedouin	10	0.50
GALNT5	157902879	Han	9	0.67
GALNT5	157872879	Pathan	13	0.62
GALNT5	157872879	Balochi	13	0.62
GALNT5	157872879	Yakut	12	0.67
GALNT5	157872879	South.China	12	0.67
GALNT5	157872879	Adygei	13	0.62
GALNT5	157872879	Druze	13	0.62
GALNT5	157872879	North.West.China	13	0.62
GALNT5	157872879	Kalash	12	0.67
GALNT5	157872879	Han	12	0.67
GALNT5	157872879	Bedouin	14	0.50
GALNT5	157872879	Sindhi	13	0.62
GALNT5	157872879	Palestinian	14	0.50
GALNT5	157872879	Papuan	10	0.80
GALNT5	157782879	Mbuti.Pygmies	9	0.44
GALNT7_17	172963751	Pima	5	1.00
GALNT7_17	172933751	Palestinian	10	0.50
GALNT7_17	172933751	North.Italy	7	0.71
GALNT7_17	172933751	Pima	5	1.00
GALNT7_17	172933751	Karitiana	5	1.00
GALNT7_17	172933751	Japanese	7	0.71
GALNT7_17	172933751	Sardinian	6	0.83
GALNT7_17	172933751	Druze	8	0.50
GALNT7_17	172933751	North.East.China	8	0.63
GALNTL1	68747775	San	6	0.50
SIAT8A	22417968	Mozabite	10	0.50
SIAT8A	22417968	Orcadian	8	0.63
SIAT8A	22417968	Palestinian	9	0.56
SIAT8A	22417968	Pathan	8	0.63
SIAT8A	22417968	North.East.China	8	0.75
SIAT8A	22417968	North.Italy	8	0.63
SIAT8A	22417968	North.West.China	8	0.63



## Human Genetic Diversity In Genes Related To Host-Pathogen Interactions

---

SIAT8A	22417968	South.China	8	0.75
SIAT8A	22417968	Surui	6	1.00
SIAT8A	22417968	Yakut	8	0.75
SIAT8A	22417968	Pima	6	1.00
SIAT8A	22417968	Russian	7	0.71
SIAT8A	22417968	Sardinian	8	0.63
SIAT8A	22417968	French	8	0.63
SIAT8A	22417968	Han	9	0.67
SIAT8A	22417968	Bedouin	10	0.50
SIAT8A	22417968	Cambodian	8	0.75
SIAT8A	22417968	Colombian	6	1.00
SIAT8A	22417968	Basque	8	0.63
SIAT8A	22417968	Karitiana	6	1.00
SIAT8A	22417968	Druze	9	0.56
SIAT8A	22417968	Japanese	8	0.75
SIAT8A	22387968	Yakut	11	0.64
SIAT8A	22387968	South.China	10	0.70
SIAT8A	22387968	Surui	8	1.00
SIAT8A	22387968	Japanese	10	0.70
SIAT8A	22387968	North.East.China	10	0.70
SIAT8A	22387968	Karitiana	8	1.00
SIAT8A	22387968	Han	11	0.64
SIAT8A	22387968	Colombian	8	1.00
SIAT8A	22357968	Surui	8	1.00
SIAT8A	22327968	Surui	9	1.00
SIAT8A	22297968	Surui	10	1.00

Table 2c.

Gene	Position	Population	daf.snps	lower.daf
B3GNT3	17834449	Hazara	12	0.75
B3GNT3	17834449	Russian	12	0.75
B3GNT3	17834449	Sindhi	10	0.70
B3GNT3	17834449	Balochi	12	0.75
B3GNT3	17834449	Burusho	12	0.75
B3GNT3	17834449	Brahui	12	0.75
FUT8	65018701	Makrani	10	0.70
FUT8	65018701	Japanese	7	0.86
FUT8	65018701	Hazara	8	0.88
FUT8	65018701	Han	7	0.86
FUT8	65018701	Pima	5	0.80
FUT8	65018701	Pathan	8	0.88
FUT8	65018701	Papuan	6	0.83
FUT8	65018701	Maya	9	0.78
FUT8	65018701	Cambodian	8	0.88
FUT8	65018701	Balochi	10	0.70
FUT8	65048701	Pathan	9	1.00
FUT8	65048701	Papuan	5	1.00
FUT8	65048701	Japanese	6	1.00
FUT8	65048701	Basque	10	0.70
FUT8	65048701	North.West.China	7	0.71
FUT8	65048701	Makrani	11	0.73
FUT8	65048701	Biaka.Pygmys	10	0.90
FUT8	65048701	Bedouin	10	0.70
FUT8	65048701	Maya	10	0.80
FUT8	65048701	Han	5	1.00
FUT8	65048701	Balochi	11	0.73
FUT8	65048701	Pima	5	1.00
FUT8	65048701	Hazara	9	1.00
FUT8	65048701	Cambodian	7	1.00
FUT8	65078701	Pima	7	0.71
FUT8	65078701	Pathan	10	0.70
FUT8	65078701	Papuan	5	0.80
FUT8	65078701	Hazara	10	0.80
FUT8	65078701	Japanese	7	0.71
FUT8	65078701	Cambodian	8	0.88
FUT8	65108701	Cambodian	10	0.90
FUT8	65108701	North.West.China	12	0.75
FUT8	65108701	Hazara	13	0.77
FUT8	65108701	Papuan	7	0.86
FUT8	65108701	Japanese	10	0.70
FUT8	65108701	NAN.Melanesian	5	0.80
FUT8	65108701	Biaka.Pygmys	10	0.90
FUT8	65138701	Cambodian	11	0.82
FUT8	65138701	Hazara	14	0.71
FUT8	65138701	Maya	15	0.73
FUT8	65138701	North.West.China	14	0.71
FUT8	65138701	Papuan	8	0.75
FUT8	65168701	San	5	1.00
FUT8	65168701	Cambodian	9	0.78
FUT8	65168701	Maya	12	0.83
FUT8	65168701	North.West.China	12	0.75
FUT8	65168701	Papuan	7	0.71
FUT8	65198701	Maya	9	0.89
FUT8	65228701	Cambodian	7	0.71

## Human Genetic Diversity In Genes Related To Host-Pathogen Interactions

---

FUT8	65228701	Japanese	7	0.71
FUT8	65228701	Maya	8	0.75
FUT9	96763842	Makrani	18	0.72
FUT9	96793842	Mozabite	18	0.72
FUT9	96793842	South.China	14	0.71
GALNT2	228420183	North.Italy	23	0.74
GALNT2	228420183	Orcadian	25	0.72
GALNT2	228420183	Palestinian	24	0.75
GALNT2	228420183	Basque	23	0.70
GALNT2	228420183	Burusho	24	0.71
GALNT2	228420183	Russian	23	0.74
GALNT2	228420183	Makrani	24	0.75
SIAT6	44165707	Orcadian	10	0.70
SIAT6	44165707	Sardinian	11	0.73
SIAT6	44165707	South.China	11	0.73
SIAT6	44165707	North.Italy	11	0.73
SIAT6	44165707	French	12	0.75
SIAT6	44165707	Han	11	0.73
SIAT6	44195707	North.Italy	13	0.69
SIAT6	44195707	French	15	0.73
SIAT6	44195707	Orcadian	13	0.69

Table 2d.

Gene	Position	Population	maf.snps	maf.upper
FUT10	33329383	Colombian	11	0.91
FUT10	33449383	Hazara	12	0.67
FUT10	33449383	North.East.China	11	0.82
FUT10	33479383	North.East.China	10	0.70
FUT10	33389383	North.East.China	15	0.67
FUT10	33359383	Pathan	15	0.67
FUT10	33449383	Pathan	14	0.64
FUT10	33359383	Sindhi	15	0.73
FUT10	33329383	Sindhi	14	0.71
FUT10	33389383	Sindhi	14	0.71
FUT10	33449383	South.China	11	0.73
FUT10	33449383	Yakut	10	0.90
FUT10	33479383	Yakut	10	0.70
FUT10	33419383	Yakut	11	0.64
GALNT7_17	173773751	Adygei	11	0.82
GALNT7_17	173353751	Adygei	9	0.78
GALNT7_17	173413751	Adygei	13	0.77
GALNT7_17	173383751	Adygei	13	0.77
GALNT7_17	173323751	Adygei	8	0.63
GALNT7_17	173773751	Bedouin	11	0.82
GALNT7_17	173743751	Bedouin	13	0.62
GALNT7_17	173833751	Cambodian	15	0.67
GALNT7_17	173893751	Cambodian	19	0.63
GALNT7_17	173413751	Druze	13	0.77
GALNT7_17	173773751	Druze	12	0.75
GALNT7_17	173353751	Druze	10	0.70
GALNT7_17	173743751	Druze	13	0.62
GALNT7_17	173353751	French	9	0.78
GALNT7_17	173263751	French	15	0.67
GALNT7_17	173293751	French	11	0.64
GALNT7_17	173323751	French	8	0.63
GALNT7_17	174223751	Karitiana	5	1.00
GALNT7_17	174163751	Karitiana	9	0.89
GALNT7_17	173353751	Karitiana	8	0.88
GALNT7_17	174193751	Karitiana	8	0.88
GALNT7_17	173353751	Maya	8	0.88
GALNT7_17	174373751	Mbuti.Pygmies	17	0.41
GALNT7_17	173803751	NAN.Melanesian	11	0.82
GALNT7_17	173773751	NAN.Melanesian	11	0.82
GALNT7_17	173383751	North.Italy	13	0.69
GALNT7_17	173413751	North.Italy	13	0.69
GALNT7_17	173353751	Orcadian	9	0.78
GALNT7_17	173383751	Orcadian	13	0.77
GALNT7_17	173413751	Orcadian	13	0.69
GALNT7_17	173323751	Orcadian	8	0.63
GALNT7_17	173773751	Palestinian	11	0.64
GALNT7_17	173383751	Palestinian	13	0.62
GALNT7_17	174523751	San	6	0.50
GALNT7_17	173773751	San	12	0.42
GALNT7_17	173353751	San	5	0.40
GALNT7_17	172963751	Sardinian	10	0.70
GALNT7_17	173563751	Surui	5	1.00
GALNT7_17	173533751	Surui	6	1.00
GALNT7_17	174163751	Yoruba	12	0.42
GALNT7_17	174193751	Yoruba	12	0.42

## Human Genetic Diversity In Genes Related To Host-Pathogen Interactions

**Table 3.** Population differentiation in the regions analyzed.

Region	F <sub>ST</sub> <sup>a</sup>	F <sub>CT</sub> <sup>b</sup>	F <sub>SC</sub> <sup>c</sup>	Gene length	N° of SNPs
GALNT5	0,188	0,181	0,033	53573	10
FUT1_2	0,168	0,163	0,030	59409	19
GCNT3	0,146	0,116	0,034	8219	3
SIAT7D_F	0,139	0,147	0,012	31704	5
GALNTL1*	0,136	0,126	0,031	94268	35
B3GNT5	0,133	0,131	0,020	20147	4
SIAT8F*	0,127	0,117	0,028	133578	47
B3GALT2	0,114	0,084	0,033	7548	1
GCNT4	0,113	0,092	0,023	3435	2
SIAT8C	0,109	0,087	0,037	16440	9
SIAT4C	0,108	0,098	0,026	58437	19
GALNT13*	0,106	0,090	0,030	582022	130
FUT12	0,103	0,090	0,026	30771	6
GALNT11_15*	0,103	0,087	0,029	63501	25
SIAT8B*	0,101	0,085	0,030	74818	34
SIAT8A*	0,100	0,085	0,030	133715	42
B4GALT1	0,099	0,080	0,032	56589	16
GALNT18*	0,098	0,090	0,022	351129	249
B3GNT3	0,095	0,075	0,032	18727	4
WBSCR17	0,094	0,081	0,026	580795	184
GALNT7_17	0,094	0,084	0,022	1510460	293
SIAT7C_E	0,093	0,082	0,024	989333	303
GALNT10	0,092	0,083	0,022	229515	106
B3GALT5	0,092	0,073	0,030	106446	43
SIAT7A_B	0,090	0,080	0,023	78412	58
FUT9	0,089	0,078	0,023	189887	52
B3GALT1	0,088	0,072	0,028	52184	10
SIAT10	0,087	0,068	0,031	62084	15
SIAT8D	0,087	0,073	0,027	96066	18
SIAT1	0,084	0,072	0,023	147825	61
FUT8	0,084	0,063	0,031	332122	43
GALNT14	0,083	0,075	0,020	219157	75
SIAT4A	0,083	0,072	0,022	113019	57
GALNT12	0,082	0,065	0,028	42378	20
GCNT2	0,081	0,067	0,025	73652	22
GALNT2	0,079	0,062	0,028	18356	11
SIAT6	0,079	0,068	0,021	194927	38
GCNT1	0,078	0,072	0,016	48186	14
ABO	0,085	0,072	0,025	20067	17
GALNT6	0,077	0,064	0,023	39367	19
FUT5_6	0,076	0,064	0,023	39914	13
B3GNT6	0,076	0,060	0,018	7570	2
SIAT8E	0,076	0,067	0,020	77958	34
GALNT1	0,073	0,057	0,025	57198	12
GALNT9	0,072	0,049	0,031	168992	17
GALNT3	0,069	0,073	0,007	23886	3
SIAT9	0,068	0,060	0,018	49886	16
FUT10	0,068	0,054	0,022	84250	14
B4GALT3	0,064	0,025	0,041	6213	2
SIAT4B	0,063	0,067	0,006	19232	1
B3GNT2	0,062	0,054	0,016	28604	8
GALNT8	0,056	0,035	0,026	52140	14
B4GALT4	0,050	0,043	0,014	29165	10
FUT13	0,046	0,044	0,009	23968	4

<sup>a</sup> between the 39 populations; <sup>b</sup> between continental groups; <sup>c</sup> within continents. \* indicates those genes with more than 20 SNPs and in bold, genes presenting with a F<sub>ST</sub> value higher than the mean obtained overall 650K chip typed in the same samples (F<sub>ST</sub>=0.096)

**Table 4.** Results of nHLS test for footprint of positive selection. All significant core SNPs falling within genes are shown and divided by continental group a, Sub-Saharan Africa; b, Europe; c, Central and South Asia; d, East Asia; e, America

a)

SNPs in core	SNP position	Gene	Population	Haplotype	Sequence	Frequency bin	Score (adj)	p-value	q-value	logP	Minor Allele	Ancestral Allele	Function UCSC	Alleles
rs4021	53945073	FUT1_2	San	2	T	0.2	-4.05	5.12E-05	3.00E-02	0.20	T	C	untranslated-3	C/T
rs7808728	151340172	GALNT11_15	Bantu	2	T	0.15	-4.04	5.40E-05	3.06E-02	N/A	N/A	C	intron	C/T
rs1019173	151341480	GALNT11_15	Bantu	2	G	0.25	-3.93	8.41E-05	3.67E-02	0.38	A	A	intron	A/G
rs998471	100641466	GALNT12	Bantu	2	G	0.3	4.41	1.04E-05	2.22E-02	0.29	A	A	intron	A/G
rs2277189	100641466	GALNT12	Mbuti_Pygmies	2	G	0.5	4.34	1.44E-05	2.22E-02	0.50	G	G	intron	C/G
rs1407503	100638430	GALNT12	San	2	A	0.2	-5.43	5.51E-08	2.22E-02	0.20	G	G	intron	A/G
rs4743285	100639579	GALNT12	San	2	G	0.2	-4.43	9.28E-06	2.22E-02	0.20	A	A	intron	A/G
rs1407506	100640073	GALNT12	San	2	C	0.2	-4.43	9.28E-06	2.22E-02	0.20	T	C	intron	C/T
rs998471	100641466	GALNT12	Yoruba	2	G	0.15	4.86	1.19E-06	2.22E-02	0.12	A	A	intron	A/G
rs3856384	154882137	GALNT13	Bantu	2	A	0.25	-4.03	5.55E-05	3.06E-02	0.23	A	C	intron	A/C
rs7566164	154879103	GALNT13	Mbuti_Pygmies	2	G	0.35	-4.69	2.73E-06	2.22E-02	0.31	G	G	intron	G/T
rs707081	155015863	GALNT13	Mbuti_Pygmies	2	G	0.3	-3.64	1.25E-04	4.68E-02	0.27	G	A	untranslated-3	A/G
rs16836453	154873487	GALNT13	Mbuti_Pygmies	2	A	0.1	-3.81	1.38E-04	4.77E-02	0.08	A	A	intron	A/G
rs10185861	154652950	GALNT13	San	2	G	0.5	-3.82	1.35E-04	4.72E-02	0.50	G	A	intron	A/G
rs7845821	112944474	GALNT18	San	2	A	0.4	-3.99	6.74E-05	3.23E-02	0.40	A	G	intron	A/G
rs12042442	173921448	GALNT7_17	Bantu	2	T	0.1	-4.40	1.10E-05	2.22E-02	0.05	T	G	intron	G/T
rs1696946	173105520	GALNT7_17	Mandenka	2	A	0.45	4.75	2.18E-06	2.22E-02	0.44	A	G	intron	A/G
rs8845010	173412506	GALNT7_17	San	2	T	0.2	-4.95	7.51E-07	2.22E-02	0.20	T	T	intron	C/T
rs7690646	173105520	GALNT7_17	Yoruba	2	A	0.4	4.51	6.51E-06	2.22E-02	0.44	A	G	intron	A/G
rs2828505	99945988	SIAT10	San	2	A	0.2	-4.15	3.32E-05	2.43E-02	0.20	A	C	intron	A/C
rs2828604	99989874	SIAT10	San	2	C	0.4	-3.83	1.28E-04	4.68E-02	0.40	C	C	intron	C/T
rs4949824	76505205	SIAT7C_E	Bantu	2	G	0.15	-4.50	6.68E-06	2.22E-02	0.11	G	A	intron	A/G
rs199722	77301014	SIAT7C_E	Bantu	2	C	0.4	-4.31	1.64E-05	2.22E-02	0.50	T	N/A	intron	C/T
rs199722	77301014	SIAT7C_E	Biaka_Pygmies	2	T	0.5	-4.31	1.64E-05	2.22E-02	0.46	C	N/A	intron	C/T
rs1474752	77249448	SIAT7C_E	Mandenka	2	G	0.3	-4.01	6.14E-05	3.15E-02	0.24	G	C	intron	C/G
rs1933398	77237260	SIAT7C_E	Yoruba	2	C	0.4	-3.78	1.55E-04	5.12E-02	0.38	G	G	intron	C/G

b)

SNPs in core	SNP position	Gene	Population	Haplotype	Sequence	Frequency bin	Score (adj)	p-value	q-value	logP	Minor Allele	Ancestral Allele	Function Illumina	Alleles
rs743285	65137886	FUT8	Basque	2	T	0.45	-4.017	5.89E-05	2.55E-02	0.44	T	T	intron	C/T
rs2411822	64948148	FUT8	Russian	2	T	0.45	-4.117	3.83E-05	2.55E-02	0.44	T	T	untranslated-5	C/T
rs1953416	64948560	FUT8	Russian	2	C	0.45	-4.117	3.83E-05	2.55E-02	0.44	C	C	near-gene-5	C/T
rs4549697	151322751	GALNT11_15	North_Italy	2	A	0.35	-4.413	1.02E-05	2.55E-02	0.35	G	A	intron	A/G
rs11770365	151320606	GALNT11_15	North_Italy	2	C	0.45	-4.306	1.66E-05	2.55E-02	0.45	T	T	intron	C/T
rs11765654	151327965	GALNT11_15	Orcadian	2	A	0.5	5.031	4.89E-07	2.55E-02	0.47	G	G	intron	A/G
rs4549697	151322751	GALNT11_15	Orcadian	2	A	0.45	3.964	7.36E-05	2.55E-02	0.43	G	A	intron	A/G
rs11765654	151327965	GALNT11_15	Russian	2	G	0.4	-4.114	3.88E-05	2.55E-02	0.38	A	G	intron	A/G
rs1919522	151313315	GALNT11_15	Sardinian	2	A	0.15	4.420	9.99E-06	2.71E-02	0.14	G	G	intron	A/G
rs6960270	151311005	GALNT11_15	Sardinian	2	T	0.35	3.966	9.39E-05	2.71E-02	N/A	N/A	T	missense	C/T
rs1364557	154989524	GALNT13	Adygei	2	C	0.2	3.802	1.44E-04	2.71E-02	0.18	C	C	intron	A/C
rs7568164	154879103	GALNT13	French	2	T	0.45	3.654	7.67E-05	2.71E-02	0.31	G	G	intron	G/T
rs6435070	154774452	GALNT13	French	2	G	0.1	3.791	1.50E-04	2.71E-02	0.22	G	A	intron	A/G
rs1020848	154719378	GALNT13	Orcadian	2	G	0.25	4.078	4.58E-05	2.95E-02	0.23	G	A	intron	A/G
rs17204619	154890096	GALNT13	Orcadian	2	G	0.3	3.830	1.28E-04	3.04E-02	0.30	G	T	intron	G/T
rs768187	157856980	GAI_NTR	North_Italy	2	T	0.15	-3.639	1.40E-04	3.17E-02	0.13	C	T	intron	C/T
rs17305821	173675840	GALNT7_17	French	2	T	0.45	-3.901	9.57E-05	3.17E-02	0.41	T	T	intron	C/T
rs7546818	44018257	SIAT6	French	2	G	0.35	4.128	3.66E-05	3.59E-02	0.34	G	A	intron	A/G
rs7546818	44018257	SIAT6	North_Italy	2	G	0.35	4.165	3.12E-05	3.61E-02	0.34	G	A	intron	A/G
rs7546818	44018257	SIAT6	Orcadian	2	G	0.25	3.808	1.40E-04	4.06E-02	0.34	G	A	intron	A/G
rs7546818	44018257	SIAT6	Sardinian	2	G	0.35	4.624	3.76E-06	4.08E-02	0.34	G	A	intron	A/G
rs1436087	76748025	SIAT7C_E	North_Italy	2	T	0.2	-4.433	9.28E-06	4.77E-02	0.18	T	T	intron	C/T
rs620381	76922048	SIAT7C_E	Orcadian	2	C	0.35	-4.255	2.09E-05	4.89E-02	N/A	N/A	C	unknown	A/C
rs1023043	17445523	SIAT8F	Orcadian	2	A	0.4	-4.021	5.80E-05	4.89E-02	0.37	A	A	intron	A/C
rs1202640	70380969	WBSOR17	Basque	2	A	0.4	-4.051	4.88E-05	4.91E-02	0.40	A	A	intron	A/C
rs1202640	70380969	WBSOR17	Orcadian	2	C	0.4	4.336	1.45E-05	5.01E-02	0.40	A	A	intron	A/C

# Human Genetic Diversity In Genes Related To Host-Pathogen Interactions

c)

SNP in core	SNP position	Gene	Population	Haplotype	Sequence	Frequency bin	Scores (sd)	p-value	q-value	MAF	Minor Allele	Ancestral Allele	Function UCSC	Aliases
rs145329	153631919	GALNT10	Pathan	2	G	0.1	4.275	0.000	0.021	0.34	G	G	intron	A/G
rs882166	153631796	GALNT10	Pathan	2	T	0.25	4.041	0.000	0.037	0.36	T	G	intron	G/T
rs753685	153630456	GALNT10	Pathan	2	C	0.25	3.983	0.000	0.040	0.07	C	T	intron	C/T
rs1919522	151313315	GALNT11_15	Kalash	2	A	0.25	4.401	0.000	0.017	0.23	G	G	intron	A/G
rs4546997	151322751	GALNT11_15	Sindhi	2	G	0.3	-3.969	0.000	0.039	0.50	G	A	intron	A/G
rs2277189	100641992	GALNT12	Makrani	2	C	0.3	-0.398	0.000	0.041	0.24	G	G	intron	C/G
rs6706444	154674722	GALNT13	Balochi	2	G	0.35	3.888	0.000	0.049	0.27	G	N/A	intron	A/G/T
rs10208448	154719378	GALNT13	Brahui	2	G	0.3	4.072	0.000	0.035	0.23	G	A	intron	A/G
rs9435070	154774452	GALNT13	Drahu	2	G	0.15	4.307	0.000	0.020	0.22	G	A	intron	A/G
rs959872	154778813	GALNT13	Brahui	2	A	0.2	3.995	0.000	0.039	0.14	A	NA	intron	A/T
rs10208448	154719378	GALNT13	Kalash	2	G	0.1	4.126	0.000	0.031	0.23	G	A	intron	A/G
rs1364537	154698524	GALNT13	Kalash	2	C	0.2	3.939	0.000	0.043	0.18	C	C	intron	A/C
rs9435070	154774452	GALNT13	Makrani	2	G	0.25	4.096	0.000	0.033	0.22	G	A	intron	A/G
rs6706444	154674722	GALNT13	Pathan	2	G	0.2	4.055	0.000	0.036	0.27	G	N/A	intron	A/G/T
rs17252939	174032057	GALNT7_17	Makrani	2	G	0.45	0.410	0.000	0.033	0.10	G	A	intron	A/G
rs17057585	173095187	GALNT7_17	North_West_China	2	C	0.4	-3.997	0.000	0.039	0.10	C	A	intron	A/C
rs10001933	173994843	GALNT7_17	Pathan	2	C	0.35	-4.293	0.000	0.020	0.09	C	T	intron	C/T
rs12042595	174000429	GALNT7_17	Pathan	2	G	0.15	-0.417	0.000	0.027	0.11	G	A	intron	A/G
rs1546918	44010257	SIAT7C	Sindhi	2	G	0.15	3.919	0.000	0.045	0.34	G	A	intron	A/G
rs302809	76384590	SIAT7C_E	Burusho	2	G	0.35	-3.995	0.000	0.039	0.10	G	NA	intron	A/G
rs4245647	76383845	SIAT7C_E	Burusho	2	A	0.5	-4.575	0.000	0.011	0.10	A	A	intron	A/G
rs1513970	76814282	SIAT7C_E	Burusho	2	A	0.1	-4.175	0.000	0.027	0.28	A	NA	intron	A/G
rs7548549	763838660	SIAT7C_E	Burusho	2	T	0.25	3.985	0.000	0.042	0.21	C	C	intron	C/T
rs4245647	76383845	SIAT7C_E	Kalash	2	A	0.35	-4.388	0.000	0.017	0.10	A	A	intron	A/G
rs484299	76871317	SIAT7C_E	Makrani	2	T	0.3	-4.026	0.000	0.039	N/A	N/A	T	unknown	G/T
rs1436088	76748096	SIAT7C_E	North_West_China	2	T	0.4	-4.056	0.000	0.036	0.16	T	T	intron	C/T
rs289696	76381079	SIAT7C_E	Pathan	2	G	0.3	-5.055	0.000	0.003	0.18	G	G	intron	A/G

d)

SNP in core	SNP position	Gene	Population	Haplotype	Sequence	Frequency bin	Scores (sd)	p-value	q-value	MAF	Minor Allele	Ancestral Allele	Function UCSC	Aliases
rs4672482	62301903	B3GNT2	Cambodian	2	G	0.45	-4.472	7.73E-06	1.27E-02	0.45	G	NA	intron	A/G
rs2092914	65148868	FUT8	Cambodian	2	C	0.1	-3.771	1.63E-04	3.34E-02	0.10	C	T	intron	C/T
rs549804	31518055	GALNT1	North_East_China	2	A	0.3	3.647	2.65E-04	4.42E-02	0.26	A	A	intron	A/G
rs625212	153624405	GALNT10	Cambodian	2	C	0.1	-4.260	2.05E-05	1.31E-02	0.10	C	C	intron	C/T
rs11785654	151327965	GALNT11_15	Cambodian	2	G	0.4	-4.869	1.12E-06	1.27E-02	0.40	A	G	intron	A/G
rs10235029	151325334	GALNT11_15	Han	2	A	0.35	5.538	3.06E-08	1.27E-02	0.34	C	A	intron	A/C
rs1919522	151313315	GALNT11_15	Han	2	A	0.2	4.279	1.88E-05	1.30E-02	0.19	G	G	intron	A/G
rs10246261	161316706	GALNT11_16	Japanoco	2	A	0.4	4.610	6.21E-06	1.27E-02	0.36	C	A	intron	A/G
rs1919522	151313315	GALNT11_15	North_East_China	2	A	0.15	0.435	1.34E-05	1.27E-02	0.11	G	G	intron	A/G
rs10246261	151313706	GALNT11_15	North_East_China	2	A	0.35	4.514	6.35E-06	1.27E-02	0.31	C	A	intron	A/C
rs1919522	151313315	GALNT11_15	South_China	2	A	0.2	4.333	1.47E-05	1.27E-02	0.15	G	G	intron	A/G
rs1407508	100640073	GALNT12	Cambodian	2	C	0.35	5.680	1.35E-08	1.27E-02	0.15	C	C	intron	C/T
rs999471	100641466	GALNT12	Cambodian	2	G	0.35	5.680	1.35E-08	1.27E-02	0.15	A	A	intron	A/G
rs2277189	100641992	GALNT12	Cambodian	2	C	0.4	3.957	7.60E-05	2.19E-02	0.40	G	G	intron	C/G
rs1407508	100640073	GALNT12	North_East_China	2	C	0.4	0.508	3.71E-07	1.27E-02	0.38	T	C	intron	C/T
rs999471	100641466	GALNT12	North_East_China	2	G	0.4	0.508	3.71E-07	1.27E-02	0.38	A	A	intron	A/G
rs10931862	164639000	GALNT13	Han	2	A	0.3	4.346	1.39E-06	1.27E-02	0.30	A	A	intron	A/G
rs10931862	154539000	GALNT13	Japanese	2	A	0.3	-0.456	5.09E-06	1.27E-02	0.21	G	A	intron	A/G
rs6739800	154528029	GALNT13	Japanese	2	G	0.25	-4.116	3.85E-05	1.61E-02	0.21	G	NA	intron	A/G
rs7558032	154734701	GALNT13	Japanese	2	A	0.25	-3.923	8.73E-05	2.39E-02	0.25	A	A	intron	A/G
rs10931862	154539000	GALNT13	South_China	2	A	0.45	-4.407	1.05E-05	1.27E-02	0.30	A	A	intron	A/G
rs7558032	154734701	GALNT13	South_China	2	A	0.4	-3.732	1.90E-04	3.67E-02	0.25	A	A	intron	A/G
rs3214040	157865607	GALNT5	Cambodian	2	G	0.2	-4.690	2.73E-06	1.27E-02	0.20	A	A	synon.cds-re	A/G
rs2353292	157866760	GALNT5	Cambodian	2	G	0.2	-3.723	1.97E-04	3.68E-02	0.20	A	A	intron	A/G
rs2166488	157868082	GALNT5	Cambodian	2	A	0.2	-3.723	1.97E-04	3.68E-02	0.20	G	A	intron	A/G
rs7566187	157859960	GALNT5	Cambodian	2	T	0.2	-3.590	3.19E-04	5.16E-02	0.20	C	T	intron	C/T
rs3214040	157865607	GALNT5	Japanese	2	G	0.25	-3.618	2.96E-04	4.83E-02	0.20	A	A	synon.cds-re	A/G
rs7566187	157859960	GALNT5	North_East_China	2	T	0.25	-3.735	1.88E-04	3.66E-02	0.21	C	T	intron	C/T

rs17080691	175056603	GALNT7_17	Japanese	2	G	0.16	3.037	8.28E-06	2.28E-03	0.11	C	C	intron	A/G
rs8116534	174225430	GALNT7_17	North_East_China	2	C	0.45	-4.022	5.54E-05	1.93E-02	0.43	C	C	intron	C/T
rs1670349	7632603	SIAT7C_E	Cambodian	2	C	0.1	-4.506	6.61E-06	1.27E-02	0.10	C	T	intron	C/T
rs4245647	7632645	SIAT7C_E	Cambodian	2	A	0.1	-4.506	6.61E-06	1.27E-02	0.10	A	A	intron	A/G
rs322809	76324590	SIAT7C_E	Cambodian	2	G	0.1	-4.506	6.61E-06	1.27E-02	0.10	G	NA	intron	A/G
rs12064337	76381279	SIAT7C_E	Cambodian	2	T	0.15	-0.659	2.30E-11	1.27E-02	0.15	T	G	intron	G/T
rs184095	76381365	SIAT7C_E	Cambodian	2	C	0.15	-5.291	1.22E-07	1.27E-02	0.15	C	T	intron	C/T
rs2152388	76380939	SIAT7C_E	Cambodian	2	T	0.15	-5.022	5.10E-07	1.27E-02	0.15	T	T	intron	A/T
rs12064337	76381279	SIAT7C_E	Han	2	T	0.1	-4.785	1.71E-06	1.27E-02	0.15	T	G	intron	G/T
rs226906	76361079	SIAT7C_E	Han	2	G	0.2	-5.670	1.43E-05	1.27E-02	0.18	G	G	intron	A/G
rs4245647	7632645	SIAT7C_E	Han	2	A	0.2	-4.084	4.42E-05	1.72E-02	0.10	A	A	intron	A/G
rs2029256	76380606	SIAT7C_E	Japanese	2	T	0.15	-4.448	8.68E-06	1.27E-02	0.14	T	NA	intron	C/T
rs2152388	76380939	SIAT7C_E	Japanese	2	T	0.2	-4.798	1.60E-06	1.27E-02	0.15	T	T	intron	A/T
rs7546049	76380660	SIAT7C_E	Japanese	2	C	0.25	-5.076	3.85E-07	1.27E-02	0.21	C	C	intron	C/T
rs1474752	77246448	SIAT7C_E	Japanese	2	G	0.25	-4.119	3.81E-05	1.61E-02	0.24	G	C	intron	C/G
rs11162066	76373802	SIAT7C_E	Japanese	2	C	0.25	-4.044	5.27E-05	1.88E-02	0.25	C	C	intron	C/T
rs289697	76370797	SIAT7C_E	Japanese	2	C	0.25	-3.845	1.21E-04	2.83E-02	0.23	C	T	intron	C/T
rs289686	76381079	SIAT7C_E	South_China	2	G	0.3	-5.597	2.18E-08	1.27E-02	0.18	G	G	intron	A/G
rs7546049	76380660	SIAT7C_E	South_China	2	C	0.25	-4.188	2.81E-05	1.43E-02	0.21	C	C	intron	C/T
rs7546049	76380660	SIAT7C_E	Yakut	2	C	0.35	-0.434	1.45E-05	1.27E-02	0.21	C	C	intron	C/T
rs11162113	76522115	SIAT7C_E	Yakut	2	C	0.45	3.742	1.83E-04	3.62E-02	0.44	C	T	intron	C/T
rs1889751	76365646	SIAT7C_E	Yakut	2	C	0.25	-3.697	2.18E-04	3.88E-02	0.24	C	T	intron	C/T
rs727248	22956340	SIAT8A	Japanese	2	C	0.45	-4.205	2.61E-05	1.38E-02	0.45	C	C	intron	C/T
rs9804230	17421670	SIAT8F	Cambodian	2	C	0.1	-4.406	5.01E-05	1.84E-02	0.10	C	T	intron	C/T
rs6721125	85956240	SIAT9	Cambodian	2	C	0.2	-3.833	1.27E-04	2.91E-02	0.20	C	NA	intron	C/T

e)

SNPs in core	SNP position	Gene	Population	Haplotype	Sequence	Frequency bin	Score (rd)	p-value	q-value	MAF	Mirror Allele	Ancestral Allele	Function UCSC	Aliases
rs12959837	31490999	GALNT1	Colombian	2	C	0.4	4.523	6.1E-06	3.3E-06	0.36	T	C	intron	C/T
rs11662739	31495598	GALNT1	Colombian	2	G	0.4	4.523	6.1E-06	3.3E-06	0.36	C	T	intron	C/T
rs149329	153631919	GALNT10	Colombian	2	G	0.3	-0.534	9.4E-08	2.9E-04	0.34	G	G	intron	A/G
rs625212	153624405	GALNT10	Colombian	2	T	0.2	-0.376	1.7E-04	3.0E-04	0.10	C	C	intron	C/T
rs1919522	151313315	GALNT11_15	Pima	2	A	0.3	4.958	7.1E-07	9.7E-04	N/A	N/A	A	intron	A/G
rs10245251	151315706	GALNT11_15	Pima	2	A	0.5	-0.375	1.7E-04	1.7E-03	0.43	A	A	intron	A/C
rs4743285	100639579	GALNT12	Colombian	2	T	0.5	-0.431	1.6E-05	4.3E-03	0.50	G	G	intron	A/G
rs1407506	100640073	GALNT12	Karitiana	2	C	0.3	6.190	6.0E-10	4.3E-03	0.25	T	C	intron	C/T
rs990471	100641466	GALNT12	Karitiana	2	T	0.3	6.190	6.0E-10	7.9E-03	0.25	A	A	intron	A/G
rs799790	154947881	GALNT13	Surui	2	G	0.4	-4.265	2.0E-05	8.2E-03	0.33	C	N/A	intron	C/T
rs12417042	11473723	GALNT18	Colombian	2	C	0.2	-3.811	1.4E-04	9.9E-03	0.14	A	C	intron	A/C
rs11737748	174367078	GALNT7_17	Colombian	2	T	0.3	-5.197	2.0E-07	1.0E-02	0.29	T	T	intron	C/T
rs230	173133048	GALNT7_17	Colombian	2	T	0.3	-3.768	1.6E-04	1.1E-02	0.29	A	A	intron	A/G
rs1094699	174379449	GALNT7_17	Colombian	2	T	0.3	-3.710	2.1E-04	2.8E-02	0.29	A	A	intron	A/G
rs10001933	173994843	GALNT7_17	Colombian	2	A	0.3	-3.660	2.5E-04	2.8E-02	0.09	C	T	intron	C/T
rs1860347	4715984	GALNT8	Colombian	2	G	0.5	3.857	1.1E-04	2.8E-02	0.43	G	A	intron	A/G
rs1436087	76748025	SIAT7C_E	Colombian	2	C	0.5	-0.433	1.5E-05	3.1E-02	0.18	T	T	intron	C/T
rs4949764	77235250	SIAT7C_E	Colombian	2	A	0.5	4.210	2.6E-05	3.2E-02	0.50	T	C	intron	C/T
rs199722	77301014	SIAT7C_E	Colombian	2	T	0.5	3.870	1.1E-04	3.2E-02	0.50	T	N/A	intron	C/T
rs209256	76380606	SIAT7C_E	Pima	2	A	0.4	-3.836	1.3E-04	3.2E-02	0.14	T	NA	intron	C/T
rs10873890	76748305	SIAT7C_E	Surui	2	A	0.4	4.821	1.4E-06	3.5E-02	0.39	A	A	intron	A/G
rs199722	77301014	SIAT7C_E	Surui	2	C	0.4	-0.419	2.8E-05	3.6E-02	0.39	C	N/A	intron	C/T
rs1544442	70470492	WBSOR17	Colombian	2	C	0.5	3.805	1.4E-04	3.6E-02	0.43	T	C	intron	C/T
rs1544443	70470551	WBSOR17	Colombian	2	A	0.5	3.805	1.4E-04	4.1E-02	0.43	A	G	intron	A/G
rs1202640	70380969	WBSOR17	Karitiana	2	C	0.4	3.866	1.1E-04	4.6E-02	0.40	A	A	intron	A/C



**Figure 1a.** Proportion of MAF per each 100Kb window

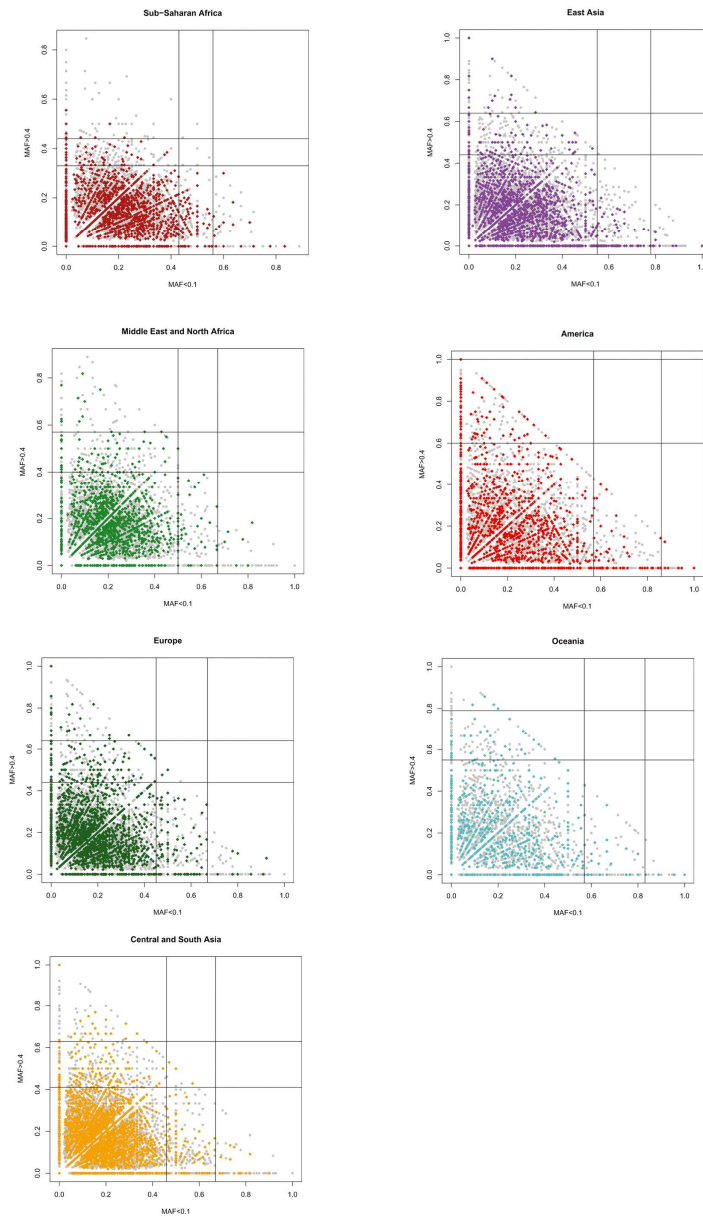


Figure 1b. Proportion of DAF per each 100Kb window

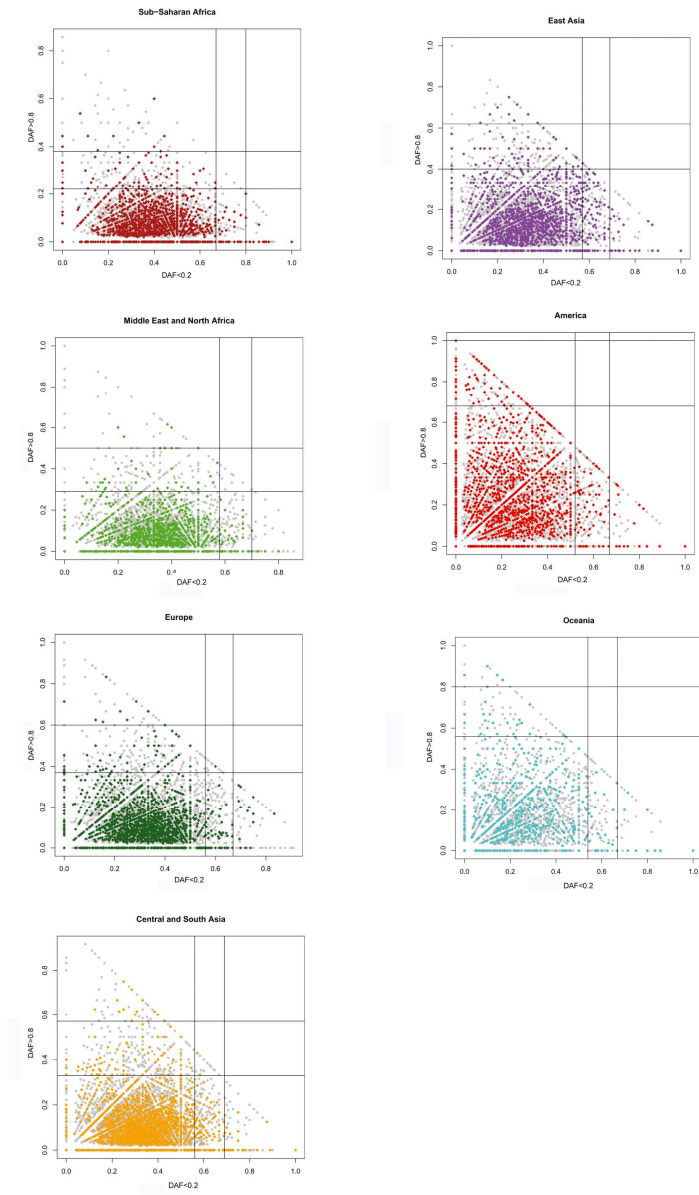
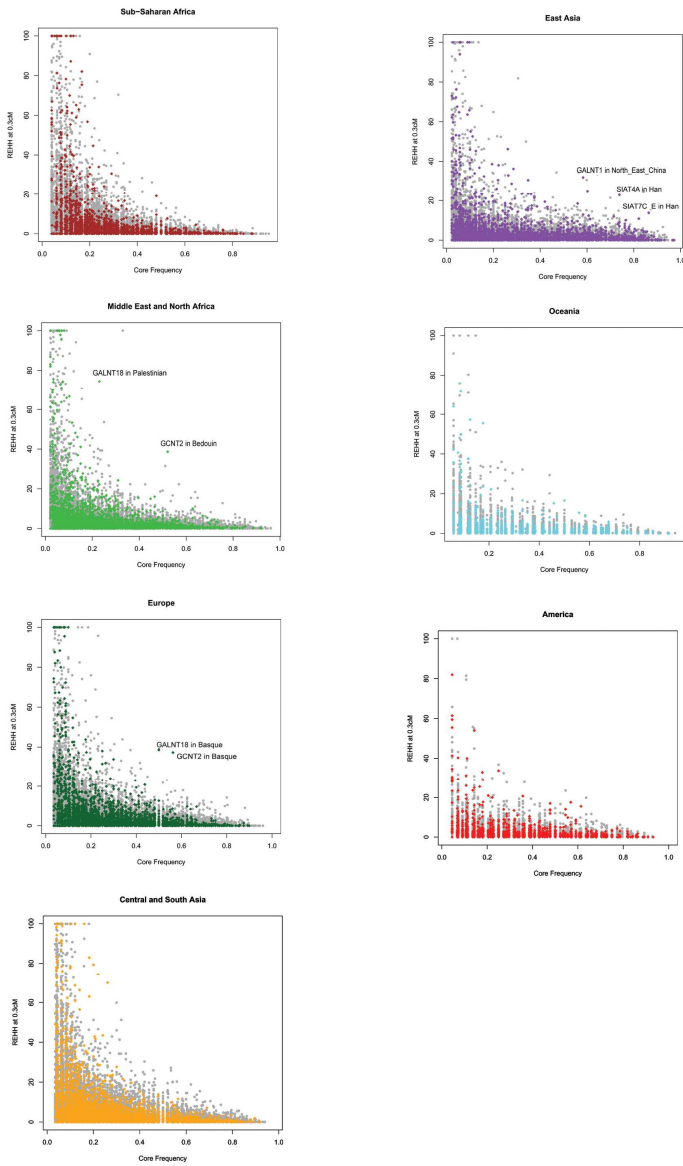


Figure 2. Distribution of REHH versus frequency



**5 DISCUSSION**

---



### **Understanding the effect of evolution in particular ascertained protein pathways in human populations**

Studying the genetic diversity of genes involved in the interaction between pathogen and host in human populations can help us better understand: a) where pathogens and humans encountered, b) to which extent pathogens have exerted a selective force on the human genome and c) how this interaction has shaped human genome evolution (Semple et al. 2006).

In the development of the current thesis we tackle this topic through four different works (chapters 1 to 4 from results section).

In the first chapter we analyze the genetic diversity of 15 genes participating in the innate immunity response at different stages. The biological function of innate immunity genes, which directly relates to pathogen recognition and clearance, make this gene category a good candidate where looking for pathogen-mediated selection. Moreover, the fact of studying a variety of genes participating in different stages of such response allows us to understand their evolution in a contextualized manner.

In chapters two to four from results section we present three works on genes related to glycosylation processes. Such genes are responsible of the biosynthesis of glycan structures attached to cell surface proteins. Among others, one of the functions of these glycan structures is protecting cells against infectious agents, although in some occasions these glycan structures are used as targets by some pathogens. Many glycans show remarkably discontinuous distribution across evolutionary lineages. And, as reported by Bishop and Gagneux in 2007, pathogens have evolved to exploit host lineage-specific glycans and are constantly shaping the glycomes of their hosts (Bishop and Gagneux 2007). This reason stimulated our interest on studying the possible departures from neutrality of genes involved in glycan biosynthesis. Below we briefly describe the three strategies we carried out to incur in the evolution of glycosylation field.

In chapter two (from results section) we focus on the study of one fucosyltransferase (*FUT2*) participating in the biosynthesis pathway of the ABO antigen. The study of this gene was motivated by the fact this locus determines the ability of individuals to express ABO antigens in secretions. Failing on expressing *FUT2* exposes individuals to certain infections but at the same time protects them from others. Previous studies on *FUT2* had described evidences for selection at this locus in single or few populations. Our interest on studying this locus in a extensive manner to obtain a detailed picture of its genetic diversity pattern across the globe was the reason impelling us to perform a single locus analysis.

In chapter three we have studied the genetic variation pattern of seven members from the glycosyltransferase 6 (GT6) family. This set of genes includes six non-functional loci described in humans plus *ABO* gene. Both intraspecific and interspecific variability patterns have been considered to assess if these genes, not functional in humans, are evolving under neutrality in human lineage because of lacking from a functional role while being constrained, when functional, in the non-human primate lineage.

For having a broader point of view about the evolution of glycosylation genes, our last work has been centred on the study of 70 loci involved in four glycan biosynthesis pathways in HGDP-CEPH samples (see chapter four). The main purposes of this work were elucidating whether evolution affects differently some glycosylation biosynthesis pathways than others and elucidating if evolution operates differently on a gene accordingly to the step (within a pathway) in which it plays its role.

Works included in this manuscript have been done under the assumption that detecting signatures of positive selection is a useful tool to identify functionally relevant genomic regions since selection locally shapes the functional variation. Until now the genetic diversity pattern of human populations has been studied using a variety of approaches. Some works (Tang et al. 2007; Voight et al. 2006) focuses on scanning the whole human genome to detect genes showing the

footprint of positive selection. Other strategies focus on particular genes previously described as good candidates for disease resistance or susceptibility (Sabeti et al. 2006). An example of that would be malaria disease that appears as a pathogen-host model where looking for the effects of selection. In this context some important findings are *HBB* alleles conferring resistance to malaria (see introduction, section 1.2.2.2): *Glu6Val* (HbS), *Glu6Lys* (HbC) and *Glu26Lys* (HbE).

Our research project moves over the strategies like classical single-gene studies or whole genome scans to the study of particular gene categories and protein networks presumably related to selection mediated by pathogens. We have proceeded so with the aim to understand specific gene evolutionary signatures in the meaningful way as possible (taking into account the evolutionary signature of the rest of genes interacting with it, its function...).

### **Innate immunity**

In the case of genes related to the immune function evolutionary studies can contribute to identify genes involved in the resistance to disease, but they can as well contribute to the understanding of the molecular mechanisms of pathogenesis. Nevertheless, despite its putative importance, the majority of the works focused on immune response (Sabeti et al. 2006), have been carried out on genes related to the acquired immunity, and just a small number have considered innate immunity genes

As mentioned in the introduction (see introduction, section 1.4.41.) innate immunity confers protection to our organism against infections by recognizing molecular patterns common to a wide variety of pathogens (PAMPs). This part of the immune system constitutes a first-line defence barrier that responds rapidly to infection. To respond in such a rapid manner innate immunity needs from receptors already fixed on the genome ready to recognize these general molecular structures. Therefore, to keep untouched the ability of recognizing general antigens one would expect innate immunity to show high levels of conservation. However, our results indicate that innate immunity has been evolving under other



forces than demography and purifying selection. Specifically, we have found signatures of balancing selection in six loci (*CD14*, *TLR1*, *TLR6*, *TLR10*, *IL6* and *IL10*). Three of them (*IL6*, *IL10* and *TLR10*) show balancing selection signatures in Africa-American population (IIPGA; Innate Immunity Programme in Genomics Applications) and four (*CD14*, *TLR6*, *TLR10* and *IL10*) do so in European-American population. On the other hand, we have detected signatures of positive selection in two genes (*TLR9* and *TLR10*) in European-American population and in one gene (*TLR1*) in African-Americans.

The most striking result concerns *CD14*. Balancing selection detected on this locus suggests the presence of more than one functional form of CD14. So far one mutation (6043) has been related to higher concentration of CD14 as a consequence of increased transcription (Baldini et al. 1999; LeVan et al. 2001) although there is controversy about it (Liang et al. 2006; von Aulock et al. 2005). This variant is located at a potential regulatory region. Besides, this regulatory region we have predicted two more potential regulatory regions in the 5' extreme of *CD14* by detecting conserved regions from the alignment of 5 mammal species. Two nucleotide substitutions located at positions 3748 and 5391 fall within these regions and could also be related to functional changes of this locus. Moreover, each of these two substitutions fall in a different branch of the CD14 phylogenetic structure, which for Europeans shows three groups of haplotypes separated by a large number of mutations. CD14 is an adaptor molecule that binds LPS and promotes innate immunity to Gram-negative bacteria. It is important to point out that the protein encoded by CD14 is required in order to trigger the TLR4 signalling cascade. CD14 exists as a membrane-bound and soluble protein that functions by binding LPS and presents it to the signalling complex TLR4/MD2 (Miyake 2006; Miyake 2006). Depending on the functional and clinical setting in which CD14 function has been analyzed, certain alleles have been associated with susceptibility to infections (Yuan et al. 2008), or with the development of autoimmunity and allergies (LeVan et al. 2001).

In European population *TLR1*, *TLR6* and *IL10* are also showing an excess of intermediate alleles, though for these loci statistical significance is less conclusive

than for *CD14*. The hypothesis of balancing selection acting on *IL10* is supported by literature (Wilson et al. 2006). Taking a look on their function these three genes have an impact on the inflammatory response. Specifically, *TLR1* and *TLR6* are first-line sensors for pathogens capable of signalling through NF-kappa B to trigger an inflammatory response (Papadimitraki et al. 2007) and *IL10* is a cytokine with a strong ability to modulate the development of the inflammatory response.

On the other hand *TLR9* and *TLR10* showed an excess of rare alleles in European population. The high LD amount detected at *TLR10* region in this population gives one more sight towards the hypothesis of positive selection acting on this gene. The case of *TLR9* is much less clear since it shows a complex LD pattern and an unusually low interpopulation differentiation value ( $F_{ST}=0.05$ ) which do not give support to the hypothesis of positive selection. Moreover, *TLR9* falls in a high LD region and thus it can not be concluded whether the detected diversity pattern is due to forces acting on it or to forces acting on nearby genes which can be creating a hitchhiking effect.

In African population *TLR10*, *IL10* and *TLR1* showed up in neutrality tests because of an excess of intermediate frequencies. Despite of this fact, it is important to remark they show weaker significance than the rest of studied genes.

In conclusion, our results show more cases of balancing selection and with more statistical significance in Europeans than in the African population which do not fit with the expectation since balancing selection needs a long time, probably a number of generations greater to those needed to separate two populations, to produce a detectable signal, except in cases of strong selective pressures (Barton and Navarro 2002) as those of *HBB* and thalasemias related with malaria are (Kwiatkowski 2005).

Our results can be explained by the fact that two out of the six genes presenting signals of balancing selection (*IL10* and *TLR10*) outstand in both sample sets which suggest widespread balancing selection with recurrent episodes of balancing selection and/or partial selective sweeps in response, likely, to

epidemics in European populations. This could account for the general positization of the values of most statistics in most populations (explained by more ancestral selective episodes), and the very significant and positive values in certain populations (explained by balancing selection acting upon a frequency spectrum that was already biased by older episodes).

Interestingly, all genes with evidences of balancing selection encode proteins with an essential role in the magnitude of the inflammatory immune response. Our results seem to point at a tight and fine tuning of the inflammatory response as the force determining the balanced selection of alleles observed in these genes, since too strong inflammatory responses are causing undesirable effects like autoimmune diseases (Anders et al. 2005) while weak inflammatory responses attenuates the subsequent immune response (Misch and Hawn 2008).

## **Glycosylation**

### **Study of a single glycosylation gene; *FUT2***

The study of *FUT2* global diversity has revealed this locus is under a complex evolutionary pattern since it suggests the possibility of different selective signatures (balancing and positive selection) operating in different geographic areas.

Results from the initial study we undertook indicate *FUT2* evolution is mainly driven by balancing selection. Evidences come from neutrality tests which show significant positive Tajima's D values for West Eurasia and Africa populations (Europe, Middle East and North Africa, Central and South Asia and Sub-Saharan Africa). Specifically, 12 out of the 39 populations analyzed, all of them belong to the previous mention continental regions, give support to these fact. Moreover, the old age we estimated for *FUT2* MRCA (2.61-5.27MYA) also points to this direction.

Despite balancing selection seems to be governing *FUT2* evolution in most of the studied populations it turned out not to be the unique force. Positive selection could also be contributing to shaping *FUT2* evolution, specifically in East Asia, since four out of the six studied East Asian populations show negative Tajima's D values. Although in general the mentioned neutrality test values are not significant, the case of South China presents significant negative values for five out of the seven statistics.

There exist different null-alleles truncating *FUT2* protein or dramatically decreasing its expression levels. With this work we state that *FUT2* null-alleles present in West Eurasia and Africa populations are different than those present in East Asia. Interestingly, although the alleles from both areas are different, their frequencies are very similar (~50%). The maintenance of similar allele frequencies at so high levels could be explained by the fact of null-alleles conferring possible beneficial effects to their carriers.

In the case of  $se^{428}$ , the most common null-allele within West Eurasia and Africa, beneficial effects when in homozygosis have been described in literature: protection against Norwalk virus, slowing HIV-1 infection or regulation of B<sub>12</sub> plasma levels (Hazra et al. 2008; Kindberg et al. 2006; Marionneau et al. 2005). Although the East Asia null-allele ( $se^{385}$ ) has been not related yet to any beneficial effect, it is likely it would have similar consequences which would maintain it at so high frequencies.

The old *FUT2* MRCA age and the presence of the  $se^{428}$  null-allele in Sub-Saharan African populations seem to indicate this allele existed before the out-of-Africa, which would suggest it should be present in East Asia. But this allele has not been found in the present study in East Asian populations which may be due to two possibilities: 1) either this allele disappeared because suddenly conferring low fitness to its carriers, which do not seem plausible because no traces of such null-allele are detected in any East Asian population (except in Yakut which is geographically close to north Europeans) and as being recessive it should easily be maintained at low frequencies or 2) individuals carrying this null-allele never

reached East Asia (which seem extremely improbable) and thus a new null-allele giving place to the same phenotype latter appeared. Any the reason was, it still remains unsolved. The issue of who the responsible was for a new null-allele playing the same role as the old and high frequency null-allele  $se^{428}$ , which did not re-appear when necessary in East Asia, is of interest and it is open for further research in East Asian populations using a case-control setting.

### **Study of a glycosylation pseudogenes family**

The evolutionary analysis of a set of putative human pseudogenes from *ABO* family has revealed that some of these sequences may be still retaining some function in the human lineage.

Since glycosylation genes are responsible of the formation of molecules used as targets for some pathogens the high proportion of pseudogenized members of the *ABO* family might be explained by the exposure of humans to environments favouring the absence of some pathogen receptors. In fact other gene categories, as olfactory genes, have already been demonstrated to have an elevated proportion (56%) of pseudogenized members due to a reduction of, in this case olfactory, needs.

The set of genes studied in this work is composed by two processed pseudogenes (*LOC401913* and *HGT2*) and four non-processed pseudogenes (*FS*, *GGTA1*, *GT6m7* and *IGB3*). When studying processed pseudogenes we found no traces of conservation, which indicate these pseudogenes are behaving as expected normal pseudogenes. In contrast, non-processed pseudogenes show a complex evolutionary pattern. Both *GGTA1* and *IGB3* show high  $d_N/d_S$  values suggesting they are not evolving under purifying selection. In the particular case of *GGTA1* statistical significance allows to reject conservation whereas for *IGB3* it can not be stated. Purifying selection signatures on these loci would have indicated that despite of being pseudogenes, these sequences would have been playing a necessary role. Clearly this is not the case for *GGTA1* neither *IGB3*. The former is inactive in catarrhines which have an enhanced defence against alphaGal positive

bacteria, possibly as a compensative mechanism. The latter is inactive in humans, which then produce high anti-Gal antibodies.

In the case of *FS* and *GT6m7* the low  $d_N/d_S$  values detected indicate purifying selection could be operating on these genes. Despite of this fact, for the case of *GT6m7* the neutrality model can not be rejected. From the set of pseudogenes studied here, *FS* outstands because of showing conclusive conservation signatures.

Forssman glycolipid (FG), the product of Forssman synthetase (FS), is widely expressed among non-primate mammalian species. It has been proven that, despite their high degree of sequence identity, murine and canine *FS* genes express a functional enzyme whereas the human *FS* expresses a protein that lacks FS activity. Two mutations present in both chimpanzee and human lineages have been shown to inactivate FS enzyme (Xu et al. 1999). Interestingly, the expression of a functionally active *FS* modifies shiga toxin (Stx) receptor glycolipids to FG and results in markedly decreased susceptibility to toxin (Elliott et al. 2003). Therefore, it has been speculated that the inactivation of the *FS* gene during primate evolution may account, at least in part, for susceptibility of human cells to Stx. The substitutions responsible for FS inactivation could, then, possibly confer a new function to *FS* pseudogene in human and chimpanzee lineages which would have been important and therefore conserved.

In general, catalytical pseudogenitization of the analyzed sequences seem to be the result of resistance to infection as the silencing of some of these sequences allows the expression of certain antibodies.

It should not escape notice the fact of studying the entire gene could give more clues on the possible selective regimes operating on these loci. Moreover, it would contribute to the finding of more inactivating mutations. Nevertheless, the study we have undertaken is considering the seventh exon of this pseudogene family, the exon encoding the catalytical region and, thus the one presumably more prone to show the footprint of selection.

### **Study of glycan biosynthesis pathways**

When considering the evolutionary pattern of all genes playing a role in four biosynthesis pathways related to glycosylation we detect the signature of adaptive selection in three regions concerning galactosyltransferases (*GALNT5*, *GALNT13* and *GALNT7\_17*) and two concerning sialyltransferases (*SIAT6* and *SIAT7C\_E*).

Especially interesting is that these galactosyltransferases which participate in the same step of the O-glycan biosynthesis pathway show selection signatures in different geographical regions. In accordance to our hypothesis it seems certain steps of a biosynthetic pathway are more prone to be under selective pressures than others. This explanation would also account for galactosyltransferases, which all participate in the same step of O-glycosylation, and sialyltransferases, which despite of not belonging to the same pathway both participate in the last steps of its respective biosynthesis pathways.

Moreover, FatiGO web tool showed that the genes presenting positive selection signatures (the ones with significant nHL scores) are an over-representation of genes belonging to O-glycan biosynthesis ( $p\text{-value}=2.28e^{-3}$  and  $q\text{-value}=1.57e^{-2}$ ) and thus this would suggest that indeed certain metabolic pathways are more susceptible to evolution forces than others.

*GALNT5*, *GALNT13* and *GALNT7\_17* enzymes catalyze the first key step of mucin-type O-glycosylation adding a N-acetylgalactosamine to an -OH serine or threonine of a polypeptide. Despite playing a role in the same step of O-glycan biosynthesis these enzymes differ on the kind of linkages ( $\alpha$  or  $\beta$ ) they perform and the donors they use.

So far two studies have been performed concerning the gene regions in which we detect signatures of positive selection. Both studies have been motivated by the fact that glycosylation process is involved in the progression of primary tumor into metastatic disease. One of these study has revealed *GALNT13* is up-regulated in patients diagnosed of stage 4 of bone marrow neuroblastoma (NB) (Berois et al.

---

2006). The other study was conducted on genes from *GALNT1* to *GALNT 9* in order to particularly study the first seven ppGalNAc-T (-T1, -T2, -T3, -T6 and -T7) because of being expressed in breast tumours (Freire et al. 2006). The latter did not find evidences for cancer progression and expression of any of the genes we report under positive selection

Studies addressing the relationship of ppGalNAc-Ts to infection disease are scant. One of the really few studies performed in this direction suggest that Eg-ppGalNAc-T1, a *Echinococcus granulosus* ppGalNac-T expressed in the hydatid cyst wall and the subtegumental region of larval worms, could participate in the biosynthesis of O-glycosylated parasite proteins exposed at the interface between *Echinococcus granulosus* and its hosts (canids) (Freire et al. 2004). It is important to remark, since this thesis is quite centred on humans, that livestock and humans are important *E. granulosus* intermediate hosts. Therefore, it would occur that human galactosyltransferases evolution could have been driven by other pathogens, similar to *E granulosu*, probably area-specific, which would account for different ppGalNAc-Ts being selected at different geographic locations.

In the case of sialyltransferases, the positive selection signatures have been found on *SIAT6* and *SIAT7C\_E* regions. Both enzymes belong to the same family and are responsible of attaching sialic acid moieties to the terminal positions of N-glycans, O-glycans and glycosphingolipids. In contrast to galactosyltransferases the two genes presenting signatures of positive selection participate in different biosynthesis pathways.

*SIAT6* is a member of the lactoseries biosynthesis pathway, where it forms the precursor molecules for *FUT3*, the enzyme responsible for the last modification of Lewis antigen formation. This locus specifically acts in the formation process of Lewis sialylated antigen (sialylLe(x)). This Lewis antigen (sialylLe(x)) carries the Neu5Ac sialic acid, exclusive of human lineage, which has already been demonstrated to be related to unique human vulnerabilities as Alzheimer's disease and multiple sclerosis Interestingly, our results indicate *SIAT6* is under adaptive evolution in Europeans. It has been previously reported that north Americans and



Europeans predominantly expresses type 2 Lewis antigens (Lex and Ley epitopes) (Monteiro et al. 1998). One example of an organism taking advantage of Lewis antigen to infect cell is *Helicobacter pylori*. Its cell envelope contains a lipopolysaccharide (LPS) the O-chain of which frequently expresses type 2 Lewis x (Lex) and Lewis y (Ley) blood group antigens that mimic human gastric mucosal cell-surface glycoconjugates. Thanks mimiquing Lewis antigens *Helicobacter pylori* evades the immune response and thus enhance bacterial adherence to gastric epithelium (Monteiro et al. 2000). Thus, as *Helicobacter pylori*, other pathogens present in Europe taking advantage of sialylLe(x) antigen could also account for the positive selection signatures found at this locus in Europeans.

Our results show *SIAT7C\_E* region present signatures of positive selection in America and East Asia. *SIAT7C* seem not to be explored yet, since there is no literature available about it. That makes difficult to speculate about the possible agents driving this gene evolution. In the case of *SIAT7E*, a gene encoding a type II membrane sialyltransferase, two studies has been done both on them assessing this locus expression profile which have been correlated: 1) to rheumatoid arthritis (RA) (Galligan et al. 2007), and 2) with a reduction of cell adhesion (Jaluria et al. 2008).

Due to the lack of information about these genes, one can hypothesise about the possible agents exerting positive selection pressures in America and East Asia. Although being pure speculation dengue virus type III (DENV-3 virus) could be a candidate for driving this loci evolution since the current diversity of main DENV-3 genotypes appeared between the middle 1960s and the middle 1970s, coinciding with human population growth, urbanization, and massive human movement, and with the description of the first cases of DENV-3 hemorrhagic fever in Asia. Furthermore, migration patterns of the DENV-3 genotype III spread across Asia, East Africa and into the Americas (Araujo et al. 2008).

**As a general conclusion**

In 1999 Gagneux and Varki suggested that glycans, the oligosaccharide chains of glycoproteins, could be subjected to diversifying selection mediated by bacteria, parasite and viruses recognizing them. Based on this premise genes which modify the extracellular glycoprotein scenery would constitute a target group in which to search for the footprint of selection.

From the expansion of modern humans out of Africa to the occupation of the Old World pandemics must not have occurred frequently and thus selection could have acted in a regional manner. As a consequence of this, the effects of the selection impelled by coadaptation of humans to their pathogens could be detected at a regional context and translated as an increase of the interpopulational differentiation of genetic parameters such as  $F_{ST}$ .

Therefore, genes responsible of the formation of pathogen receptors are first line candidates to be susceptible to selection, likely, as a result of pathogen pressures. As described in this thesis this is so, at least, for some genes related to immunity and glycan biosynthesis which act as receptors or are involved to their synthesis.

During the present discussion it has been shown that the fact of studying genes within their functional context or in relation to their partners is a helpful strategy to better understand where selective pressures are acting and the possible biological meaning it could have. A good example of this is the innate immunity study we performed which showed that those genes involved in the magnitude of the innate immune response are the ones under balancing selection pressures.

Working on scenarios as complete as possible (as metabolic networks, signalling cascades, biosynthesis pathways...) could be the most optimal way to understand how selection operates and why it does on certain genes. In short, working on the whole system could be the best manner to understand its evolution. And thus, from such a need the evolutionary systems biology field might be born



## **6 REFERENCES**

---



- 
- Ackerman HC, Ribas G, Jallow M, Mott R, Neville M, Sisay-Joof F, Pinder M, Campbell RD, Kwiatkowski DP (2003) Complex haplotypic structure of the central MHC region flanking TNF in a West African population. *Genes Immun* 4: 476-86
- Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* 2: e286
- Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 12: 1805-14
- Al-Shahrour F, Diaz-Uriarte R, Dopazo J (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20: 578-80
- Anders HJ, Zecher D, Pawar RD, Patole PS (2005) Molecular mechanisms of autoimmunity triggered by microbial infection. *Arthritis Res Ther* 7: 215-24
- Araujo JM, Nogueira RM, Schatzmayr HG, Zanotto PM, Bello G (2008) Phylogeography and evolutionary history of dengue virus type 3. *Infect Genet Evol*
- Balakirev ES, Ayala FJ (2003) Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet* 37: 123-51
- Balaresque PL, Ballereau SJ, Jobling MA (2007) Challenges in human genetic diversity: demographic history and adaptation. *Hum Mol Genet* 16 Spec No. 2: R134-9
- Baldini M, Lohman IC, Halonen M, Erickson RP, Holt PG, Martinez FD (1999) A Polymorphism\* in the 5' flanking region of the CD14 gene is associated with circulating soluble CD14 levels and with total serum immunoglobulin E. *Am J Respir Cell Mol Biol* 20: 976-83

Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. *Nat Rev Genet* 4: 99-111

Bamshad MJ, Mummidi S, Gonzalez E, Ahuja SS, Dunn DM, Watkins WS, Wooding S, Stone AC, Jorde LB, Weiss RB, Ahuja SK (2002) A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. *Proc Natl Acad Sci U S A* 99: 10539-44

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: 4516-9

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. *Nat Genet* 40: 340-5

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263-5

Barton NH, Navarro A (2002) Extending the coalescent to multilocus systems: the case of balancing selection. *Genet Res* 79: 129-39

Berois N, Blanc E, Ripoche H, Mergui X, Trajtenberg F, Cantais S, Barrois M, Dessen P, Kagedal B, Benard J, Osinaga E, Raguenez G (2006) ppGalINAc-T13: a new molecular marker of bone marrow involvement in neuroblastoma. *Clin Chem* 52: 1701-12

Biosystems A SOLiD technology. <http://solid.appliedbiosystems.com/>

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM,

---

Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799-816

Bishop JR, Gagneux P (2007) Evolution of carbohydrate antigens--microbial forces shaping host glycomes? *Glycobiology* 17: 23R-34R

Biswas S, Akey JM (2006) Genomic insights into positive selection. *Trends Genet* 22: 437-46

Boren T, Falk P, Roth KA, Larson G, Normark S (1993) Attachment of *Helicobacter pylori* to human gastric epithelium mediated by blood group antigens. *Science* 262: 1892-5

Braslavsky I, Hebert B, Kartalov E, Quake SR (2003) Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci U S A* 100: 3960-4

Brosius J, Gould SJ (1992) On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proc Natl Acad Sci U S A* 89: 10706-10



Calafell F, Roubinet F, Ramirez-Soriano A, Saitou N, Bertranpetit J, Blancher A (2008) Evolutionary dynamics of the human ABO gene. *Hum Genet* 124: 123-35

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL (2002) A human genome diversity cell line panel. *Science* 296: 261-2

Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. *Nat Genet* 33 Suppl: 266-75

Chang JG, Yang TY, Liu TC, Lin TP, Hu CJ, Kao MC, Wang NM, Tsai FJ, Peng CT, Tsai CH (1999) Molecular analysis of secretor type alpha(1,2)-fucosyltransferase gene mutations in the Chinese and Thai populations. *Transfusion* 39: 1013-7

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308: 1149-54

Consortium H (2005) A haplotype map of the human genome. *Nature* 437: 1299-320

Cserti CM, Dzik WH (2007) The ABO blood group system and *Plasmodium falciparum* malaria. *Blood* 110: 2250-8

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D (2005) Efficiency and power in genetic association studies. *Nat Genet* 37: 1217-23

- 
- Dodd J, Jessell TM (1985) Lactoseries carbohydrates specify subsets of dorsal root ganglion neurons projecting to the superficial dorsal horn of rat spinal cord. *J Neurosci* 5: 3278-94
- Elliott SP, Yu M, Xu H, Haslam DB (2003) Forssman synthetase expression results in diminished shiga toxin susceptibility: a role for glycolipids in determining host-microbe interactions. *Infect Immun* 71: 6543-52
- Excoffier LGL, and S. Schneider (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics*, vol Online pp 47-50
- Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L (2007) Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci U S A* 104: 17614-9
- Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405-13
- Feng Z, Smith DL, McKenzie FE, Levin SA (2004) Coupling ecology and evolution: malaria and the S-gene across time scales. *Math Biosci* 189: 1-19
- Ferrer-Admetlla A, Bosch E, Sikora M, Marques-Bonet T, Ramirez-Soriano A, Muntasell A, Navarro A, Lazarus R, Calafell F, Bertranpetit J, Casals F (2008) Balancing selection is the main force shaping the evolution of innate immunity genes. *J Immunol* 181: 1315-22
- Fischer A, Gilad Y, Man O, Paabo S (2005) Evolution of bitter taste receptors in humans and apes. *Mol Biol Evol* 22: 432-6
- Flint J, Harding RM, Boyce AJ, Clegg JB (1998) The population genetics of the haemoglobinopathies. *Baillieres Clin Haematol* 11: 1-51

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-61

Freire T, Berois N, Sonora C, Varangot M, Barrios E, Osinaga E (2006) UDP-N-acetyl-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 6 (ppGalNAc-T6) mRNA as a potential new marker for detection of bone marrow-disseminated breast cancer cells. *Int J Cancer* 119: 1383-8

Freire T, Fernandez C, Chalar C, Maizels RM, Alzari P, Osinaga E, Robello C (2004) Characterization of a UDP-N-acetyl-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase with an unusual lectin domain from the platyhelminth parasite *Echinococcus granulosus*. *Biochem J* 382: 501-10

Fry AE, Griffiths MJ, Auburn S, Diakite M, Forton JT, Green A, Richardson A, Wilson J, Jallow M, Sisay-Joof F, Pinder M, Peshu N, Williams TN, Marsh K, Molyneux ME, Taylor TE, Rockett KA, Kwiatkowski DP (2008) Common variation in the ABO glycosyltransferase is associated with susceptibility to severe *Plasmodium falciparum* malaria. *Hum Mol Genet* 17: 567-76

- 
- Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133: 693-709
- Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M (2008) Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res*
- Gagneux P, Varki A (1999) Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology* 9: 747-55
- Galili U, Anaraki F, Thall A, Hill-Black C, Radic M (1993) One percent of human circulating B lymphocytes are capable of producing the natural anti-Gal antibody. *Blood* 82: 2485-93
- Galligan CL, Baig E, Bykerk V, Keystone EC, Fish EN (2007) Distinctive gene expression signatures in rheumatoid arthritis synovial tissue fibroblast cells: correlates with disease activity. *Genes Immun* 8: 480-91
- Galvani A-P, Slatkin M (2003) Evaluating plague and smallpox as historical selective pressures for the CCR5-Delta 32 HIV-resistance allele. *Proc Natl Acad Sci U S A* 100: 15276-9
- Gardner M, Bertranpetit J, Comas D (2008) Worldwide genetic variation in dopamine and serotonin pathway genes: Implications for association studies. *Am J Med Genet B Neuropsychiatr Genet*
- Gardner M, Gonzalez-Neira A, Lao O, Calafell F, Bertranpetit J, Comas D (2006) Extreme population differences across Neuregulin 1 gene, with implications for association studies. *Mol Psychiatry* 11: 66-75
- Gardner M, Williamson S, Casals F, Bosch E, Navarro A, Calafell F, Bertranpetit J, Comas D (2007) Extreme individual marker F(ST) values do not imply population-specific selection in humans: the NRG1 example. *Hum Genet* 121: 759-62

Gerber-Lemaire S, Juillerat-Jeanneret L (2006) Glycosylation pathways as drug targets for cancer: glycosidase inhibitors. *Mini Rev Med Chem* 6: 1043-52

Gilad Y, Man O, Glusman G (2005) A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res* 15: 224-30

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

Gonzalez-Neira A, Calafell F, Navarro A, Lao O, Cann H, Comas D, Bertranpetit J (2004) Geographic stratification of linkage disequilibrium: a worldwide population study in a region of chromosome 22. *Hum Genomics* 1: 399-409

Hall N (2007) Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol* 210: 1518-25

Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* 66: 1669-79

Hanchard NA, Rockett KA, Spencer C, Coop G, Pinder M, Jallow M, Kimber M, McVean G, Mott R, Kwiatkowski DP (2006) Screening for recently selected alleles by analysis of human haplotype similarity. *Am J Hum Genet* 78: 153-9

Harding RM, Healy E, Ray AJ, Ellis NS, Flanagan N, Todd C, Dixon C, Sajantila A, Jackson IJ, Birch-Machin MA, Rees JL (2000) Evidence for variable selective pressures at MC1R. *Am J Hum Genet* 66: 1351-61

Harpending H (1993) The Human Scenario Beclouded. *Science* 260: 1176-1178

Harris EE, Meyer D (2006) The molecular signature of selection underlying human adaptations. *Am J Phys Anthropol Suppl* 43: 89-130

---

Hastbacka J, de la Chapelle A, Mahtani MM, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, Kusumi K, Trivedi B, Weaver A, et al. (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell* 78: 1073-87

Hazra A, Kraft P, Selhub J, Giovannucci EL, Thomas G, Hoover RN, Chanock SJ, Hunter DJ (2008) Common variants of FUT2 are associated with plasma vitamin B12 levels. *Nat Genet* 40: 1160-2

Hedrick P (2004) Estimation of relative fitnesses from relative risk data and the predicted future of haemoglobin alleles S and C. *J Evol Biol* 17: 221-4

Henry S, Mollicone R, Fernandez P, Samuelsson B, Oriol R, Larson G (1996) Molecular basis for erythrocyte Le(a+ b+) and salivary ABH partial-secretor phenotypes: expression of a FUT2 secretor allele with an A-->T mutation at nucleotide 385 correlates with reduced alpha(1,2) fucosyltransferase activity. *Glycoconj J* 13: 985-93

Hill AV (2006) Aspects of genetic susceptibility to human infectious diseases. *Annu Rev Genet* 40: 469-86

Hill AV, Allsopp CE, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA, Bennett S, Brewster D, McMichael AJ, Greenwood BM (1991) Common west African HLA antigens are associated with protection from severe malaria. *Nature* 352: 595-600

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: 1072-9

Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-Boris A, Yoshiki A (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* 423: 91-6

Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E (2007) *Ensembl* 2007. *Nucleic Acids Res* 35: D610-7

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153-9

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335: 167-70

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451: 998-1003

Jaluria P, Chu C, Betenbaugh M, Shiloach J (2008) Cells by design: a mini-review of targeting cell engineering using DNA microarrays. *Mol Biotechnol* 39: 105-11

Janeway CA, Jr., Medzhitov R (2002) Innate immune recognition. *Annu Rev Immunol* 20: 197-216

Jobling MA, Hurler M, Tyler-Smith C (2004) *Human evolutionary genetics: origins, peoples & disease*. Garland Science, New York

---

Joziase DH, Shaper JH, Jabs EW, Shaper NL (1991) Characterization of an alpha 1----3-galactosyltransferase homologue on human chromosome 12 that is organized as a processed pseudogene. *J Biol Chem* 266: 6991-8

Kelly RJ, Rouquier S, Giorgi D, Lennon GG, Lowe JB (1995) Sequence and expression of a candidate for the human Secretor blood group alpha(1,2)fucosyltransferase gene (FUT2). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. *J Biol Chem* 270: 4640-9

Keusch JJ, Manzella SM, Nyame KA, Cummings RD, Baenziger JU (2000) Expression cloning of a new member of the ABO blood group glycosyltransferases, iGb3 synthase, that directs the synthesis of isoglybosphingolipids. *J Biol Chem* 275: 25308-14

Khor CC, Chapman SJ, Vannberg FO, Dunne A, Murphy C, Ling EY, Frodsham AJ, Walley AJ, Kyrieleis O, Khan A, Aucan C, Segal S, Moore CE, Knox K, Campbell SJ, Lienhardt C, Scott A, Aaby P, Sow OY, Grignani RT, Sillah J, Sirugo G, Peshu N, Williams TN, Maitland K, Davies RJ, Kwiatkowski DP, Day NP, Yala D, Crook DW, Marsh K, Berkley JA, O'Neill LA, Hill AV (2007) A Mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. *Nat Genet* 39: 523-8

Kim Y, Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167: 1513-24

Kindberg E, Hejdeman B, Bratt G, Wahren B, Lindblom B, Hinkula J, Svensson L (2006) A nonsense mutation (428G-->A) in the fucosyltransferase FUT2 gene affects the progression of HIV-1 infection. *Aids* 20: 685-9

Koda Y, Soejima M, Kimura H (2001) The polymorphisms of fucosyltransferases. *Leg Med (Tokyo)* 3: 2-14



Koda Y, Soejima M, Liu Y, Kimura H (1996) Molecular basis for secretor type alpha(1,2)-fucosyltransferase gene deficiency in a Japanese population: a fusion gene generated by unequal crossover responsible for the enzyme deficiency. *Am J Hum Genet* 59: 343-50

Koike C, Fung JJ, Geller DA, Kannagi R, Libert T, Luppi P, Nakashima I, Profozich J, Rudert W, Sharma SB, Starzl TE, Trucco M (2002) Molecular basis of evolutionary loss of the alpha 1,3-galactosyltransferase gene in higher primates. *J Biol Chem* 277: 10114-20

Koike C, Uddin M, Wildman DE, Gray EA, Trucco M, Starzl TE, Goodman M (2007) Functionally important glycosyltransferase gain and loss during catarrhine primate emergence. *Proc Natl Acad Sci U S A* 104: 559-64

Korneev SA, Park JH, O'Shea M (1999) Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J Neurosci* 19: 7711-20

Kwiatkowski DP (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet* 77: 171-92

Landsteiner K (1900) Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. *Zbl. Bakt.* 27: 357-362

Larsson MM, Rydell GE, Grahn A, Rodriguez-Diaz J, Akerlind B, Hutson AM, Estes MK, Larson G, Svensson L (2006) Antibody prevalence and titer to norovirus (genogroup II) correlate with secretor (FUT2) but not with ABO phenotype or Lewis (FUT3) genotype. *J Infect Dis* 194: 1422-7

Lee JT (2003) Molecular biology: Complicity of gene and pseudogene. *Nature* 423: 26-8

---

LeVan TD, Bloom JW, Bailey TJ, Karp CL, Halonen M, Martinez FD, Vercelli D (2001) A common single nucleotide polymorphism in the CD14 promoter decreases the affinity of Sp protein binding and enhances transcriptional activity. *J Immunol* 167: 5838-44

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74: 175-95

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319: 1100-4

Li WH, Wu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* 2: 150-74

Liang XH, Cheung W, Heng CK, Liu JJ, Li CW, Lim B, Wang de Y (2006) CD14 promoter polymorphisms have no functional significance and are not associated with atopic phenotypes. *Pharmacogenet Genomics* 16: 229-36

Liberman B (2004) Details of being human. *Nature* 454

Lindesmith L, Moe C, Marionneau S, Ruvoen N, Jiang X, Lindblad L, Stewart P, LePendu J, Baric R (2003) Human susceptibility and resistance to Norwalk virus infection. *Nat Med* 9: 548-53

Liu Y, Koda Y, Soejima M, Pang H, Schlaphoff T, du Toit ED, Kimura H (1998) Extensive polymorphism of the FUT2 gene in an African (Xhosa) population of South Africa. *Hum Genet* 103: 204-10

Lukacsovich T, Waldman AS (1999) Suppression of intrachromosomal gene conversion in mammalian cells by small degrees of sequence divergence. *Genetics* 151: 1559-68

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-80

Marionneau S, Airaud F, Bovin NV, Le Pendu J, Ruvoen-Clouet N (2005) Influence of the combined ABO, FUT2, and FUT3 polymorphism on susceptibility to Norwalk virus attachment. *J Infect Dis* 192: 1071-7

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351: 652-4

Mecas J, Franklin G, Kuziel WA, Brubaker RR, Falkow S, Mosier DE (2004) Evolutionary genetics: CCR5 mutation and plague protection. *Nature* 427: 606

Medzhitov R, Janeway C, Jr. (2000a) Innate immunity. *N Engl J Med* 343: 338-44

Medzhitov R, Janeway C, Jr. (2000b) The Toll receptor family and microbial recognition. *Trends Microbiol* 8: 452-6

Milland J, Christiansen D, Lazarus BD, Taylor SG, Xing PX, Sandrin MS (2006) The molecular basis for gal $\alpha$ (1,3)gal expression in animals with a deletion of the  $\alpha$ 1,3galactosyltransferase gene. *J Immunol* 176: 2448-54

- 
- Misch EA, Hawn TR (2008) Toll-like receptor polymorphisms and susceptibility to human disease. *Clin Sci (Lond)* 114: 347-60
- Mivake K (2006) Links Roles for accessory molecules in microbial recognition by Toll-like receptors. *J Endotoxin Res* 12: 195-204
- Miyake K (2006) Roles for accessory molecules in microbial recognition by Toll-like receptors. *J Endotoxin Res* 12: 195-204
- Monteiro MA, Chan KH, Rasko DA, Taylor DE, Zheng PY, Appelmek BJ, Wirth HP, Yang M, Blaser MJ, Hynes SO, Moran AP, Perry MB (1998) Simultaneous expression of type 1 and type 2 Lewis blood group antigens by *Helicobacter pylori* lipopolysaccharides. Molecular mimicry between *h. pylori* lipopolysaccharides and human gastric epithelial cell surface glycoforms. *J Biol Chem* 273: 11533-43
- Monteiro MA, Zheng P, Ho B, Yokota S, Amano K, Pan Z, Berg DE, Chan KH, MacLean LL, Perry MB (2000) Expression of histo-blood group antigens by lipopolysaccharides of *Helicobacter pylori* strains from asian hosts: the propensity to express type 1 blood-group antigens. *Glycobiology* 10: 701-13
- Morcillo-Suarez C, Alegre J, Sangros R, Gazave E, de Cid R, Milne R, Amigo J, Ferrer-Admetlla A, Moreno-Estrada A, Gardner M, Casals F, Perez-Lezaun A, Comas D, Bosch E, Calafell F, Bertranpetit J, Navarro A (2008) SNP analysis of results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data. *Bioinformatics* 24: 1643-4
- Mourant AE (1954) Blood grouping. *Br Med J* 1: 37-9
- National Institutes of H, National Institute of General Medical S, National Institute on A, Institute for Medical R, Coriell Institute for Medical R (1982) Catalog of cell lines. Catalog of cell lines.

Nei M (1986) Stochastic errors in DNA evolution and molecular phylogeny. *Prog Clin Biol Res* 218: 133-47

Nielsen R (1998) Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theor Popul Biol* 53: 143-51

Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Brittenham G, Looareesuwan S, Clark AG, Tokunaga K (2004) Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. *Am J Hum Genet* 74: 1198-208

Pang H, Fujitani N, Soejima M, Koda Y, Islam MN, Islam AK, Kimura H (2000) Two distinct Alu-mediated deletions of the human ABO-secretor (FUT2) locus in Samoan and Bangladeshi populations. *Hum Mutat* 16: 274

Pang H, Koda Y, Soejima M, Fujitani N, Ogaki T, Saito A, Kawasaki T, Kimura H (2001) Polymorphism of the human ABO-Secretor locus (FUT2) in four populations in Asia: indication of distinct Asian subpopulations. *Ann Hum Genet* 65: 429-37

Papadimitraki ED, Bertias GK, Boumpas DT (2007) Toll like receptors and autoimmunity: a critical appraisal. *J Autoimmun* 29: 310-8

Rana BK, Hewett-Emmett D, Jin L, Chang BH, Sambuughin N, Lin M, Watkins S, Bamshad M, Jorde LB, Ramsay M, Jenkins T, Li WH (1999) High polymorphism at the human melanocortin 1 receptor locus. *Genetics* 151: 1547-57

Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet* 70: 841-7

- 
- Roubinet F, Despiau S, Calafell F, Jin F, Bertranpetit J, Saitou N, Blancher A (2004) Evolution of the O alleles of the human ABO blood group gene. *Transfusion* 44: 707-15
- Rowe JA, Handel IG, Thera MA, Deans AM, Lyke KE, Kone A, Diallo DA, Raza A, Kai O, Marsh K, Plowe CV, Doumbo OK, Moulds JM (2007) Blood group O protects against severe *Plasmodium falciparum* malaria through the mechanism of reduced rosetting. *Proc Natl Acad Sci U S A* 104: 17471-6
- Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496-7
- Ruiz-Palacios GM, Cervantes LE, Ramos P, Chavez-Munguia B, Newburg DS (2003) *Campylobacter jejuni* binds intestinal H(O) antigen (Fuc alpha 1, 2Gal beta 1, 4GlcNAc), and fucosyloligosaccharides of human milk inhibit its binding and infection. *J Biol Chem* 278: 14112-20
- Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832-7
- Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) Positive natural selection in the human lineage. *Science* 312: 1614-20
- Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo

A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449: 913-8

Saitou N, Yamamoto F (1997) Evolution of primate ABO blood group genes and their homologous genes. *Mol Biol Evol* 14: 399-411

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 94: 441-8

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74: 5463-7

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15: 1576-83

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78: 629-44

Semple CA, Gautier P, Taylor K, Dorin JR (2006) The changing of the guard: Molecular diversity and rapid evolution of beta-defensins. *Mol Divers* 10: 575-84

Sharon N (1996) Carbohydrate-lectin interactions in infectious disease. *Adv Exp Med Biol* 408: 1-8

- 
- Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309: 1728-32
- Sikora M, Ferrer-Admetlla A, Mayor A, Bertranpetit J, Casals F (2008) Evolutionary analysis of genes of two pathways involved in placental malaria infection. *Hum Genet* 123: 343-57
- Soejima M, Koda Y (2008) Distinct single nucleotide polymorphism pattern at the FUT2 promoter among human populations. *Ann Hematol* 87: 19-25
- Soejima M, Pang H, Koda Y (2007) Genetic variation of FUT2 in a Ghanaian population: identification of four novel mutations and inference of balancing selection. *Ann Hematol* 86: 199-204
- Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, Thomson G (2008) Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum Immunol* 69: 443-64
- Stajich JE, Hahn MW (2005) Disentangling the effects of demography and selection in human history. *Mol Biol Evol* 22: 63-73
- Steiper ME, Young NM (2006) Primate molecular divergence dates. *Mol Phylogenet Evol* 41: 384-94
- Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68: 978-89
- Storan MJ, Magnaldo T, Biol-N'Garagba MC, Zick Y, Key B (2004) Expression and putative role of lactoseries carbohydrates present on NCAM in the rat primary olfactory pathway. *J Comp Neurol* 475: 289-302



Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100: 9440-5

Swerdlow DL, Mintz ED, Rodriguez M, Tejada E, Ocampo C, Espejo L, Barrett TJ, Petzelt J, Bean NH, Seminario L, et al. (1994) Severe life-threatening cholera associated with blood group O in Peru: implications for the Latin American epidemic. *J Infect Dis* 170: 468-72

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585-95

Tang K, Thornton KR, Stoneking M (2007) A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. *PLoS Biol* 5: e171

Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics* 145: 505-18

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22: 4673-80

Thorven M, Grahn A, Hedlund KO, Johansson H, Wahlfrid C, Larson G, Svensson L (2005) A homozygous nonsense mutation (428G-->A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections. *J Virol* 79: 15351-5

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* 293: 455-62

- 
- Toomajian C, Kreitman M (2002) Sequence variation and haplotype structure at the human HFE locus. *Genetics* 161: 1609-23
- Turcot-Dubois AL, Le Moullac-Vaidye B, Despiau S, Roubinet F, Bovin N, Le Pendu J, Blancher A (2007) Long-term evolution of the CAZY glycosyltransferase 6 (ABO) gene family from fishes to mammals--a birth-and-death evolution model. *Glycobiology* 17: 516-28
- Vargas-Madrado E, Almagro JC, Lara-Ochoa F (1995) Structural repertoire in VH pseudogenes of immunoglobulins: comparison with human germline genes and human amino acid sequences. *J Mol Biol* 246: 74-81
- Varki A, Cummings R, Esko J, Freeze H, Hart G, Marth J (1999) *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York
- Voight BF, Kudravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4: e72
- von Aulock S, Rupp J, Gueinzus K, Maass M, Hermann C (2005) Critical investigation of the CD14 promoter polymorphism: lack of a role for in vitro cytokine response and membrane CD14 expression. *Clin Diagn Lab Immunol* 12: 1254-6
- Wallace RB, Shaffer J, Murphy RF, Bonner J, Hirose T, Itakura K (1979) Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res* 6: 3543-57
- Walsh EC, Sabeti P, Hutcheson HB, Fry B, Schaffner SF, de Bakker PI, Varilly P, Palma AA, Roy J, Cooper R, Winkler C, Zeng Y, de The G, Lander ES, O'Brien S, Altshuler D (2006) Searching for signals of evolutionary selection in 168 genes related to immune function. *Hum Genet* 119: 92-102

Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proc Natl Acad Sci U S A* 103: 135-40

Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Res* 15: 1468-76

Wilson JN, Rockett K, Keating B, Jallow M, Pinder M, Sisay-Joof F, Newport M, Kwiatkowski D (2006) A hallmark of balancing selection is present at the promoter region of interleukin 10. *Genes Immun* 7: 680-3

Wood ET, Stover DA, Slatkin M, Nachman MW, Hammer MF (2005) The beta - globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria. *Am J Hum Genet* 77: 637-42

Wright S (1931) Evolution in Mendelian Populations. *Genetics* 16: 97-159

Wright S (1950) Genetical structure of populations. *Nature* 166: 247-9

Xu H, Storch T, Yu M, Elliott SP, Haslam DB (1999) Characterization of the human Forssman synthetase gene. An evolving association between glycolipid synthesis and host-microbial interactions. *J Biol Chem* 274: 29390-8

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13: 555-6

Yu LC, Yang YH, Broadberry RE, Chen YH, Chan YS, Lin M (1995) Correlation of a missense mutation in the human Secretor alpha 1,2-fucosyltransferase gene with the Lewis(a+b+) phenotype: a potential molecular basis for the weak Secretor allele (Sew). *Biochem J* 312 ( Pt 2): 329-32

Yuan FF, Marks K, Wong M, Watson S, de Leon E, McIntyre PB, Sullivan JS (2008) Clinical relevance of TLR2, TLR4, CD14 and FcγRIIA gene polymorphisms in *Streptococcus pneumoniae* infection. *Immunol Cell Biol* 86: 268-70

Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M, Ruan Y, Wei CL, Gingeras TR, Guigo R, Harrow J, Gerstein MB (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res* 17: 839-51



<sup>7</sup> **APPENDIX**





**7.1 Chapter 5**

Is There Selection for the Pace of Successive Inactivation of the *arpAT*  
Gene in Primates?

- This work was published at the Journal of Molecular Evolution on October 2007-





Casals F, Ferrer-Admetlla A, Chillarón J,  
Torrents D, Palacín M, Bertranpetit J.

*Is there selection for the pace of successive  
inactivation of the arpAT gene in primates?*

J Mol Evol. 2008 Jul;67(1):23-8. Epub 2008 Jun  
20.

**7.2 Chapter 6**

Neuropathologic findings in an Aged Albino Gorilla

- This work was published at the Veterinarian Pathology on March 2008-



Márquez M, Serafin A, Fernández-Bellon H,  
Serrat S, Ferrer-Admetlla A, Bertranpetit J,  
Ferrer I, Pumarola M.

*Neuropathologic findings in an aged albino  
gorilla.*

Vet Pathol. 2008 Jul;45(4):531-7.