Universidad Autónoma de Madrid

Facultad de Ciencias

Departamento de Biología Molecular

Tesis Doctoral

A Comparative Genomic Study of Human and Chimpanzee Evolution: Natural Selection, Function, and Disease

Leonardo D. Arbiza Brustin

Octubre - 2008

Director:
Dr. Hernán J. Dopazo

Contents

Resumen

Introducción

Entender lo que nos hace humanos ha sido una de las búsquedas más antiguas e importantes en la historia de la humanidad. Sin embargo, ha sido tan solo en las últimas décadas que hemos adquirido la capacidad de abordar esta pregunta desde los múltiples campos y disciplinas científicas que podrían comenzar a dar lugar al enfoque multidisciplinario requerido para lograr avances importantes en la materia. El conocimiento de la proximidad y parentesco evolutivo entre humanos y otros primates, ha existido tan solo desde la formulación de la teoría de selección natural. Desde hace aproximadamente poco más de un siglo, se han comenzado a realizar estudios desde un punto de vista comparativo utilizando nuestro pariente evolutivo más cercano, el chimpancé, con el propósito de buscar posibles pistas sobre la respuesta a esta importante pregunta.

Los primeros estudios naturalmente tuvieron lugar a nivel anatómico, seguidos por comparaciones a nivel de comportamiento, patología, y posteriormente a nivel molecular basados en estudios bioquímicos y patológicos. Desde aquél entonces, se han realizado múltiples estudios profundizando sobre las diferencias moleculares, pero incluso, aun en periodos tan recientes como el de los últimos 5 años donde hemos podido abordar una comparación de sus genomas, nos hemos maravillado ante la paradoja de las pocas diferencias que hemos logrado encontrar y las diferencias tan marcadas que podemos observar entre nosotros y los chimpancés.

Sin lugar a duda, existen varias diferencias conocidas entre ambas especies dentro de varios aspectos de nuestra fisiología, nuestras habilidades innatas, nuestro comportamiento, y nuestra capacidades lingüísticas y cognitivas. Por esta razón pareciera casi increíble que las comparaciones genómicas muestren tan pocas diferencias. Las estimas de variación entre los genes de ambas especies muestran diferencias sorprendentemente bajas: de entre el 1% y 4%.

Aun a pesar de la poca diferencia, resulta natural preguntarse qué procesos evolutivos que han dado lugar a estas diferencias. En la última década se han desarrollado varios métodos mejorados para la detección de la señales moleculares indicativas de procesos de selección natural. Considerando la reciente explosión en la disponibilidad de datos moleculares, nos encontramos situados en el escenario más favorable que ha existido para poder intentar responder a esta pregunta. Claramente, en los últimos años se han realizado muchos estudios orientados a este propósito. El crecimiento también en la disponibilidad de otros tipos datos a nivel molecular, por ejemplo datos de expresión, junto con la observación de las diferencias en habilidades lingüísticas y cognitivas, entre otras, ha motivado estudios de la relación entre los procesos evolutivos y sus posibles efectos al nivel de los órganos y tejidos del organismo.

Sin embargo es muy importante destacar, que la consideración de la pregunta sobre ¿qué es lo que nos hace humanos?, puede ser engañosa. Por un lado pareciera dar extrema importancia a las funciones relacionadas con el cerebro, lo cual es importante evitar dado que podrían existir diferencias en muchísimas otras funciones relevantes, como por ejemplo, en el metabolismo, la

reproducción, y la habilidad de prevenir ciertas enfermedades, entre muchísimas otras. La pregunta también pareciera enfocar solo sobre las diferencias que existen entre nosotros y otros organismos. Sin embargo, la relación de parentesco evolutivo que tenemos como especies derivadas de un ancestro en común, también lleva a la consideración tanto, de las funciones que se han mantenido a lo largo de la historia evolutiva, como de las innovaciones y susceptibilidades que podemos haber heredado de forma común. Estas consideraciones cobran mayor importancia cuando se considera que la selección natural es hasta hoy en día, la mejor explicación de la existencia de la adaptación, mediante la cual los organismos han desarrollado sus funciones, en las cuales diferentes fallos o desarrollos dañinos pueden dar lugar a procesos de enfermedad.

La importancia de la relación entre la selección natural, la funcionalidad, y la enfermedad se puede observar en el gran numero de estudios realizados, y en publicaciones recientes, buscando profundizar nuestro conocimiento en la materia. En algunos casos, aproximaciones indirectas a las estimación de presiones selectivas han dado lugar a varios métodos para la detección de variaciones genéticas asociadas a enfermedad. En otros, se ha estudiado la relación de eventos de selección positiva y funcionalidad dando conclusiones en algunos casos controvertidas, y en otros, contradictorias. Sin embargo, el desarrollo de métodos probabilísticos robustos para la medición de tasas evolutivas y la detección de genes y residuos bajo la influencia de la selección natural, nos permiten abordar de forma más acertada la relación entre selección natural, función, y enfermedad.

Desde el desarrollo de la teoría neutralista de la evolución molecular se han desarrollado varios métodos para testar la presencia de eventos de PS utilizando el neutralismo como hipótesis nula. La teoría neutralista postula que la mayoría las diferencias observables entre organismos y especies se deben a cambios selectivamente neutros, o casi neutros, que se fijan por deriva génica en poblaciones de tamaños finitos. Existen varios métodos para la detección de selección positiva, cada uno con sus ventajas y desventajas, pero todos ellos se engloban bajo la clases de pruebas llamadas test de neutralidad. Entre la diversidad de métodos disponibles, existen dos tipos principales. Los que se basan en la comparación de variación genética dentro de las especies, y los que comparan la variación genética entre especies. Mientras los test basados en la comparación de variación dentro de las especies se ven afectados por parámetros demográficos, aquellos basados en la comparación entre especies, particularmente aquellos que además comparan distintas tasas de mutación dentro de las mismas regiones génicas, evitan este problema.

El test de Ka/Ks (también conocido como dN/dS), basado en la comparación de tasas de sustitución no sinónimas por sitio no sinónimo (Ka) frente a la estima análoga para la tasa sinónima (Ks), es uno de los métodos que evitan este tipo de problemas. Además tiene un comportamiento que permite obtener evidencia certera de casos de selección positiva. En sus implementaciones más recientes dentro de un marco probabilístico, se han desarrollado modelos de codones de máxima verosimilitud (ML) a partir del modelado se sustituciones

como procesos de cadena de Markov continuos. El desarrollo de estos modelos a permitido la implementación de modelos específicos que permiten estimar por métodos de optimización numérica, las tasas de sustitución a lo largo de linajes, para diferentes tipos de sitios, o para ambos en los modelos conocidos como modelos de rama-y-sitio ("branch-site"). La implementación de estos modelos ML, también tiene la ventaja que permite contrastar que modelo se ajusta mejor a los datos mediante test estadísticos. De esta forma se pueden codificar y contrastar hipótesis neutras y variantes que permiten la presencia de linajes, sitios, o sitios dentro de linajes evolucionando bajo selección positiva.

Métodos y Resultados

En el primer capitulo de los resultados se aborda la detección de genes bajo procesos de aceleración de tasas, relajación de presiones selectivas (RSC), y selección positiva (PS) en los genomas de humano y chimpancé.

El análisis parte de los 30,709 genes humanos en la base de datos Ensembl (versión 30), y las anotaciones de ortología contenidas en el Ensembl Compara. Tras filtrar los genes humanos con ortólogos en chimpancé, rata, ratón, y perro, que a la vez habían sido confirmados por su presencia en la base de datos SwissProt, se obtuvieron, 14,185 genes con sus correspondientes ortólogos en cada una de las especies antes mencionadas. Después de alinear cada una de las secuencias de cada gen mediante el programa de alineamientos múltiples ClustalW, se realizó una medición de tasas de sustitución sinónimas (Ks) y no sinónimas (Ka) con el programa RRTree. Aquellas secuencias que mostraron tasas de Ka>0.5 o Ks>1 fueron descartadas, dando lugar a un conjunto de 13,197 genes que se mantuvieron para análisis posteriores.

El análisis mediante el test de tasas relativas (RRT) implementado en el programa RRTree, mostró que habían más genes acelerados en chimpancé que en humano. El número de genes acelerados en tasas no sinónimas (Ka) fue casi cuatro veces mayor a el de los genes acelerados en tasas sinónimas (Ks). La comparación mediante el test de Kolmogorov-Smirnov de las medias de Ka vs. Ks, como la tasa de Ka/Ks entre humano y chimpancé, no mostró diferencias significativas en ninguno de los casos. La comparación de las estimas para ambas tasas en ambas especies mostró ser muy similar a los resultados obtenidos por el Consorcio de Secuenciación y Análisis del Genoma de Chimpancé cuando publicaron el genoma del chimpancé.

Para el análisis de selección positiva (PS) los 13,197 genes utilizados para la detección de genes con aceleración de tasas, se filtraron descartando aquellos con menos de 3 nucleótidos de diferencia entre las secuencias de humano y chimpancé. Los 9.674 genes restantes se analizaron con los test I y II de PS, implementados en el programa CodeML del paquete de programas PAML. Ambos test comparan la bondad de ajuste de un modelo neutro y otro que permite la presencia de sitios bajo selección positiva en un linaje seleccionado. La diferencia entre ambos test es que mientras el test II es robusto en cuanto a la detección

de eventos de PS, el test I tiende a incluir falsos positivos debido a su ineficacia en la diferenciación de casos de relajación de presiones selectivas (RSC) de verdaderos eventos de selección positiva. De esta forma, la comparación de los resultados de ambos tests, permite discernir aquellos genes que probablemente han evolucionado bajo RSC.

Tras la aplicación de ambos test a los linajes de humano, chimpancé, y el linaje ancestral de ambos desde el punto de divergencia de los roedores, se pudieron discernir varios patrones interesantes. Los genes bajo selección positiva (test II) en humanos (108) son bastante menos que en chimpancé (577), y ambos son menores que los encontrados en el linaje ancestral (750). Comparando los resultados con los genes encontrados bajo RSC (exclusivamente en el test I), se observa que la RSC ha sido un proceso más frecuente que la selección positiva en todos los linajes.

La comparación de los grupos de genes entre procesos de selección positiva y aceleración muestra que ambos grupos son bastante diferentes. Mediante una exploración del cruce de estos grupos y su relación con el valor estadístico de corte utilizado en el RRT, se demuestra que hay genes que muestran una tasa de $Ka/Ks > 1$, sin mostrar resultados positivos en el test II, y que también ocurre la situación inversa. Sin embargo aquellos genes que se encuentran bajo el primer caso, se deben a que las estimas en ambos Ka y Ks son mínimas, y el valor de $Ka/Ks>1$ ocurre principalmente debido a variaciones aleatorias permitidas dentro de un proceso de evolución neutra. El caso opuesto, demuestra el poder del test II, que pudiendo detectar eventos de selección en pocos codones del alineamiento, es mucho más sensitivo que un enfoque basado en la detección de una elevación media en Ka con respecto al Ks a lo largo de toda la secuencia del gen.

En el segundo capitulo de los resultados, se pasa a analizar la posible implicación funcional de los genes observados bajo aceleración de tasas, RSC, y PS. Utilizando la anotación funcional por términos de Ontología Génica (GO), se aplica el programa FatiGO para estudiar si existen clases funcionales sobreo sub-representadas dentro de los grupos de genes bajo cada uno de estos procesos evolutivos, y en la comparación bajo los mismos procesos entre humano y chimpancé. El test estadístico implementado en Fatigo es el test exacto de Fisher (de dos colas) con corrección de test múltiples.

El análisis de los genes acelerados en Ka y Ks del análisis de RRT no muestra diferencias en la representación de términos GO, ni en la comparación de cada grupo entre humano y chimpancé, ni en la comparación de los grupos acelerados frente al resto de los genes del genoma.

Al analizar los genes bajo selección positiva entre especies y en comparación con el resto de los genes del genoma, se obtienen resultados similares. Es decir, los genes bajo PS y RRT no muestran diferencias en términos funcionales entre humano y chimpancé, ni representan un grupo particular en términos funcionales en comparación con el resto de los genes en sus respectivos genomas. Las clases funcionales que muestran la mayor representación bajo PS son las mismas en ambas especies : metabolismo proteico celular, receptor acoplado

a proteína G, percepción sensorial, transcripción y regulación de la transcripción, y respuesta inmune, entre otras. Sin embargo, al comparar los genes bajo PS dentro de las mismas clases funcionales, se observa que mayoritariamente y a pesar de encontrarse agrupados bajo el mismo tipo de función, son genes diferentes. Al comparar la representación funcional de humano y chimpancé con la del linaje ancestral se obtienen algunos patrones interesantes. Se observa que en comparación con la representación porcentual observada en el el linaje ancestral: algunas clases incrementan en humano mientas disminuyen en chimpancé (receptor acoplado a proteína G, percepción sensorial, metabolismo celular de carbohidratos) y otras hacen lo opuesto (metabolismo proteico celular, transcripción y su regulación, regulación del metabolismo de nucleobases, nucleótidos, y nucleósidos). Desde el punto de vista de la tendencia ancestral, se observan variaciones que no se podían observar al comparar entre especies de forma directa. Las mayores diferencias desde este punto de vista relativo se observan en las clases GO: metabolismo proteico celular, favorecida por PS en chimpancé, y tanto receptor acoplado a proteína G como percepción sensorial, favorecidas bajo PS en humano.

En el análisis de cambios en tendencia funcional bajo los procesos de RSC y PS entre el linaje ancestral y los derivados, se observan dos clases que muestran diferencias significativas. La clase de receptor acoplado a proteína G muestra un aumento significativo bajo RSC en humano comparado al linaje ancestral. La clase de percepción sensorial muestra una mayor representación bajo RSC que bajo PS en chimpancé. La comparación de los genes bajo RSC contra el resto de los genes en el genoma muestra un sesgo significativo en ambas clases con mayor representación bajo RSC. Finalmente, se emplea un análisis de regresión lineal entre la diferencia en representación porcentual de clases funcionales en humano y chimpancé, frente a quella en el linaje ancestral. Los cambios a partir de la tendencia ancestral bajo RSC tienden a incrementar o disminuir conjuntamente (p-valor=3.16e-15) en las especies derivadas, mientras no existe correlación alguna entre los cambios en tendencia bajo PS. Esta diferencia demuestra que los grupos de genes deducidos bajo ambos procesos, cuado menos desde un punto de vista funcional, tienen ciertas diferencias.

En la ultima sección del segundo capitulo de los resultados se aborda la exploración de las posibles implicaciones funcionales de PS desde otro punto de vista. En la primera parte, se había aplicado un enfoque de dos pasos basado en testar la asociación funcional de genes deducidos en un paso previo, a partir de los resultados obtenidos mediante la utilización del valor de corte estadístico en el test II. Mientras este enfoque es el que se ha empleado mayoritariamente en estudios anteriores para estudiar la relación de PS y función, se puede observar que el enfoque, es cuando menos, ineficiente. Las dos rondas de test múltiples implican la necesidad de dos correcciones independientes. La utilización de un valor de corte ha sido criticada en base a la posible inclusión de falsos positivos que solo por llegar a sobrepasar un valor determinado, independiente de la fuerza de la señal, se consideran como resultados positivos. Por estos motivos se aplica un test de particiones donde el estadístico antes utilizado para

comparar con el valor de corte en el test II de selección positiva, es utilizado ahora para ordenar la lista de todos los genes analizados independiente de su resultado. Sobre estos genes, ahora ordenados de mayor a menor evidencia de selección positiva, se aplica el FatiScan que prueba la existencia de de posibles distribuciones asimétricas de clases funcionales hacía los extremos de la lista. Mediante esta metodología se obtienen varias clases funcionales que muestran valores elevados de PS (27 en humano y 54 en chimpancé, de las cuales solo 16 son comunes en ambas especies). Algunas de estas clases habían sido propuestas en estudios previos, donde carecían de significación estadística robusta, pero también se observan otras no descritas de forma previa.

En el tercer capitulo de los resultados, se exploran los patrones de selección natural al nivel de los órganos y tejidos del organismo. Para lograr una definición robusta de los genes específicos de tejidos a ser utilizados, se emplean tres bases de datos. La primera y principal (GNF) cuenta con 840 genes específicos de tejido (TSG) pertenecientes a 8 agrupaciones de órganos (categorías TSG) y 859 genes "housekeeping" (HK) obtenidos del Tissue Atlas Database. GNF está basada en la medición de patrones de expresión por microarrays, y utiliza la determinación de presencia / ausencia de un gen en diferentes tejidos basadas en el algoritmo de análisis MAS5. Cada una de las 8 categorías TSG agrupa varios tejidos bajo órganos similares. De esta forma solo aquellos genes detectados con presencia en cualquier numero de tejidos, pero dentro de una única categoría TSG, se incluyeron en el análisis. Los genes HK se definieron como aquellos que mostraban expresión en más de 60 tejidos (de 83 tejidos totales en GNF). El segundo grupo de TSGs se obtuvo a partir de la base de datos GeneNote que analiza 12 tejidos humanos mediante un procedimiento de normalización desarrollado por sus autores, el cual permite la identificación de TSGs expresados de forma elevada en uno de los 12 tejidos que a la vez muestran un expresión disminuida en todos los otros. La tercera base de datos utilizada fue la TissueDistributions Database (TDDB), basada en la anotación de datos de ESTs de Unigene, donde ambos, los niveles de expresión, y la representación porcentual de genes dentro de agrupaciones de órganos, se utilizan para definir TSGs y HK.

La comparación de las mismas clases TSG derivadas de las tres bases de datos de expresión diferentes mostró un grado bajo de coincidencia. Por ejemplo, en la categoría TSG de cerebro que contaba con entre 100 y 200 genes en cada grupo, solo 4 genes se encontraron definidos en grupos derivados de las tres bases de datos. Este desacuerdo entre las definiciones de categorías TSG obtenidas a partir de cada una de las bases de datos de expresión, demuestra que tanto los métodos experimentales, como el método utilizado para detectar y definir genes de expresión específicaen tejidos, son factores importantes cuando se pretenden sacar conclusiones de aplicabilidad generalizada sobre TSGs.

Para el análisis evolutivo, se empleó un procedimiento similar al detallado en el primer capitulo de los resultados, utilizando una versión más reciente de la base de datos Ensembl (v42) y descartando alineamientos que mostraron errores tras una inspección manual de todos aquellos genes, que además de tener

datos de expresión, tenían ortólogos anotados en las 5 especies (human, chimpancé, ratón, rata, y perro.) Para la medición de tasas Ka/Ks a lo largo de los linajes, humano, chimpancé, ratón, y rata, se utilizó el ajuste de modelos de máxima verosimilitud que permiten que cada uno de los linajes tenga una tasa media independiente (modelo "free branch" en PAML). Para la comparación de tasas entre linajes se utilizo la media del Ka/Ks pesada por el error estandard para cada linaje, y p-valores derivados de un tests de permutación (10,000 permutaciones) entre grupos.

Al comparar las estimas entre los linajes de primates y roedores, se observaron diferencias altamente significativas ($p<<0.001$) para la mayoría de las clases incluyendo los genes HK. La comparación entre los linajes de primates solo mostró diferencias significativas, para los genes de cerebro, con tasas mayores en chimpancé. Tres definiciones adicionales de genes HK, publicados en estudios previos, fueron evaluadas, y tanto estas como la comparación de estimas de todos los genes en el genoma, mostraron diferencias significativas. Este patrón, demuestra que existen efectos de linaje que han de ser tomados en cuenta en la comparación de tasas entre linajes. Al dividir las estimas de cada categoría TSG y HK por las estimas genómicas, desaparecieron todas las diferencias salvo en los TSG de cerebro y otros (la TSG "otros," se deriva de la agrupación de todas las TSG iniciales con menos de 20 genes que ante esta falta de datos no se tomaron en cuenta como categorías independientes). En ambos casos las diferencias tanto, entre primates y roedores, como las diferencias entre linajes de primates, se deben a variaciones en las tasas de los genes TSG de chimpancé. Las mismas son más elevadas que el resto de los linajes en la TSG de cerebro, y similarmente menores en la TSG otros. Aunque en la mayoría de los casos no se encontraron diferencias significativas, de forma general se observa que las tasas en chimpancé tienden a ser más elevadas que las de los otros linajes.

La comparación entre categorías TSG dentro de los linajes mostraron que las TSG de testículos, páncreas, e hígado son aquellas con tasas mas elevadas dentro de cada linaje. De forma similar pero opuesta, la TSG de cerebro y los genes HK son las categorías que muestran las tasas mas conservadas de todas la categorías. Al comparar la categorías mediante tests de permutación de todas contra todas las caetorias, se observan dos patrones interesantes. La TSG de de cerebro es la que muestra mayores diferencias, siendo menor que muchas de las otras en todas las especies, salvo en chimpancé. En los roedores, se observa adicionalmente, que las clases con tasas elevadas, mencionadas anteriormente, también muestran una elevación significativa en comparación con la mayoría de las otras clases TSG.

La TSG de cerebro, en todos los linajes, es la categoría mas conservada, con estimas inclusive menores a las de los genes HK en humano. Tras realizar un análisis de regresión lineal entre, la tasa Ka/Ks y el numero de tejidos en los que se expresa un gen, se obtienen dos perspectivas interesantes. Los resultados de la regresión demuestran una correlación negativa fuerte ($0.45<R^2<0.81$) y altamente significativa ($2.4e\text{-}05<p<0.01$) en todos los linajes. Estos resultados demuestran que el Ka/Ks disminuye de forma aditiva mientras aumenta

el número de tejidos en los que se expresa un gen. Dado que los genes HK, por definición en base a su función requieren una expresión constitutiva en la mayoría de los tejidos, este resultado muestra que su utilización como punto de comparación al realzar estudios sobre el comportamiento de los TSG, no es indicado. Al contrario, es un grupo con sesgos importantes que no debería ser utilizado como punto de referencia único para entender el comportamiento de genes TSG en general. La utilización de los mismos en publicaciones anteriores, como la de Dorus et al., donde se concluye que existen tasas elevadas en cerebro, se observa como un fallo importante en los mismos. Lo que es particularmente interesante, es que las tasas de la TSG de cerebro, no estando sujetas a la acumulación de presiones selectivas negativas impuestas por expresión en más de un tejido, muestran tasas aun menores que la de los genes HK. Lo cual sirve para demostrar, que lejos de haber evidencias de una elevación de tasas en genes específicos de cerebro, uno de los patrones más llamativos al comparar entre todos los TSG, es precisamente las tasas particularmente bajas que se observan en esta categoría. Abordar un entendimiento de la razones por las cuales los genes del cerebro han evolucionado con tasas tan bajas, podría resultar de interes para lograr un mayor concimiento de los mecanismos de evolución que han dado lugar a este órgano tan particular.

Finalmente, se estudiaron los resultados del test II especifico para la detección de selección positiva, donde se observa que los genes del cerebro humano no han evolucionado mediante eventos masivos de selección positiva. El numero de casos de PS en es menor en humanos que en chimpancé, y la TSG de cerebro no muestra un numero particular de eventos de PS en comparación con el resto de los tejidos analizados.

En el quinto capitulo de los resultados se aborda el estudio de las relación entre la selección natural y la enfermedad. Para ello se utilizan métodos similares a los estudios anteriores utilizando 11 species de vertebrados (humano, chimpancé, ratón, rata, perro, monodelphis, pollo, sapo, y tres especies de peces), y modelos de máxima verosimilitud que estiman las presiones selectivas (Ka/Ks) por sitio. Se obtienen datos de frecuencias de mutación asociadas a enfermedad de tres conjuntos de bases de datos para realizar distintos estudios. El primer conjunto se obtiene de la base de datos IARC, la cual es el recurso mas completo con anotación de frecuencias de mutación asociadas a enfermedad en la proteína supresora de tumores p53. El segundo conjunto, se obtiene de las bases de datos IDR, RettBase y Cosmic que en conjunto contienen información de 43 genes sobre la frecuencia de mutaciones asociadas a enfermedades inmunes y cáncer. Finalmente, se analiza la anotación de polimorfismos causantes y no causantes de enfermedad en SwissProt que permite el análisis de más de 8.000 mutaciones puntuales distribuidas a lo largo de 1.434 proteínas.

Primero se estudia la distribución de Ka/Ks en los sitios y dominios de la proteína p53. La proteína p53 tiene 5 dominios principales: un dominio de transactivación N-terminal (TA), un dominio rico en prolina (PR), un domino de unión al ADN (DB), uno dominio de tetramerización (TR), y un dominio C-terminal (CO). Las mutaciones causantes de enfermedad más frecuentes se

encuentran a lo largo de los dominios DB y TR los cuales muestran medias y medianas de Ka/Ks significativamente más bajas que los otros dominios ($p<0.05$ test de Kolmogorov-Smirnov). Mientras ambos muestran medias de Ka/Ks de 0.1, el dominio DB muestra un distribución de valores menores que la del dominio TR. Al mapear las presiones selectivas de los codones a lo largo de la proteína sobre las estructuras de la misma unida al ADN o tetramerizada, se observa que los valores más pequeños de Ka/Ks se encuentran en los residuos específicos que interaccionan con el sitio de unión al ADN, y los que interaccionan con otros residuos durante la tetramerización. Adicionalmente, todos los sitios con frecuencias elevadas de mutación asociados a enfermedad se observan con un valor de Ka/Ks igual o menor a 0.1. Algunos residuos adicionales sin datos asociados a enfermedad también se observan bajo presiones selectivas fuertes, pero coincidían de forma notable con sitios estructurales importantes y asociaciónes a enfermedad destacados en publicaciones no incluidas en la base de datos IARC. Los residuos con $Ka/Ks > 0.3$ se observan solamente en la periferia de la proteína y aquellos con $Ka/Ks > 0.2$ no coinciden en ningún caso con sitios estructurales importantes ni con aquellos que muestran frecuencias altas de mutaciones relacionadas a enfermedad.

Partiendo de la observación de que todos los residuos asociados significativamente a enfermedad han evolucionado con un $Ka/Ks < 0.1$, se tomo este valor como un valor de corte representativo presiones selectivas fuertes (selección purificadora). Utilizando los datos del segundo conjunto de bases de datos, se comprobó la frecuencia de mutaciones asociadas a procesos de enfermedad frente a estimas de presiones selectivas mayores o menores al valor de corte. La diferencia en la distribución de frecuencias obtenida resultó altamente significativa ($p=3.03e\text{-}05$). La exploración de valores de cortes alternativos no dieron lugar a particiones más significativas entre las distribuciones.

Finalmente, para testar la aplicabilidad del poder predictivo de esta consideración sobre la asociación de presiones selectivas elevadas y sitios de importancia en procesos de enfermedad, se analizaron los datos de la tercera bases de datos (SwissProt). De nuevo el la utilización del valor de corte, mostró una separación significativa entre los polimorfismos asociados a enfermedad ($Ka/Ks<1$) y los polimorfismos neutros sin un efecto fenotípico detectable ($Ka/Ks>1$).

En el ultimo capitulo de los resultados, se describe la contribución realizada a la comunidad científica, mediante la incorporación, tanto de las herramientas y predicciones desarrollados en este trabajo, como otros métodos no accesibles en la web, dentro de dos plataformas integradas para el análisis de datos: el PupaSuite y el Phylemon. La incorporación de los métodos dentro de estos entornos ha representado una contribución importante para la publicación de estas herramientas.

La inclusión del predictor desarrollado en este trabajo en el servidor web integrado PupaSuite logra varios propósitos importantes. Se permite el acceso publico de los resultados de este método mediante una interfase web derivada de herramientas ya conocidas y altamente utilizadas. También se hace dentro de un entorno que facilita la interpretación de los resultados tanto por las propiedades

de visualización gráfica del servidor, como la posibilidad de obtener resultados de estimas sobre poblaciones y patrones de variación asociados a la misma. Es decir, los resultados de los otros métodos en el PupaSuite pueden servir como resultados complementarios para el estudio de varias hipótesis evolutivas. De forma similar, la inclusión del predictor complementa la funcionalidad del servidor haciendo disponible un método adicional para selección de polimorfismos con posibles efectos fenotípicos, el cual es el principal propósito del PupaSuite como plataforma para la selección y análisis de polimorfismos importantes en el marco de proyectos de genotipado a gran escala.

La inclusión de todos los programas utilizados en este trabajo, más otros adicionales en el servidor Phylemon, también logra varios propósitos importantes. Permite la reproducción de los métodos utilizados aquí. Abre un amplio rango de posibilidades tanto, para la obtención y preparación de datos, como en el análisis de hipótesis de evolución molecular diversas. La integración de todos los programas en una misma plataforma, respaldada por un entorno computacional de alta capacidad, y provista de ejemplos que muestran las configuraciones necesarias para correr los análisis principales para los cuales han sido diseñados los programas, también representa una contribución de un valor incalculable para los usuarios. Sean usuarios principiantes o expertos, la integración de todos estos programas bajo una misma suite de herramientas permite lanzar análisis donde se requieren pasos múltiples, varios programas, y en muchos casos un poder de computación considerable.

Discusión

Desde hace años los biólogos evolutivos han deseado conocer el grado con el cual la selección positiva y la deriva génica han contribuido a la variación genética de las poblaciones y las especies. Diferentes pruebas de neutralidad han dado lugar a herramientas valiosas para el desarrollo de hipótesis en esta área de la biología. El primer objetivo de los estudios en este campo se había enfocado en obtener inferencias generales sobre las causas de la evolución molecular, y en búsqueda de dicho objetivo, se han llevado acabo numerosos estudios para rastrear posibles desviaciones de la hipótesis del reloj molecular. Sin embargo, durante la ultima década el enfoque ha cambiado a la búsqueda de eventos de selección positiva (PS).

Varios estudios de PS a nivel genómico en humano y chimpancé se han realizado. Sin embargo, estos han dado lugar a conclusiones algunas veces controvertidas, y en otros, contradictorias. La disponibilidad de métodos nuevos y más sensibles para la detección de eventos de selección positiva se puede utilizar para dar respuesta a preguntas que anteriormente solo han recibido respuestas parciales. En la primera parte de los resultados, se presenta un estudio genómico completo de los genes bajo PS, relajación de presiones selectivas (RSC), y aceleración de tasas evolutivas (RRT). Es el primer estudio donde se distinguen eventos de PS de posibles casos de RSC utilizando modelos de ramas-y-sitios

("branch-site" tests) de máxima verosimilitud (ML), para agregar robustez y poder comparar la frecuencia de ambos procesos.

En total y después de corregir ante los efectos de test múltiples, aproximadamente el 5% de los genes analizados en humano y el 10% en chimpancé, mostraron pertenecer a alguno de estos procesos desviándose de una hipótesis de evolución neutra. A partir de estas cifras se puede concluir que las desviaciones de neutralidad no han sido procesos frecuentes en la evolución de ambos genomas. Sin embargo, es importante resaltar que los grupos obtenidos tras la corrección necesaria de los efectos de test múltiples, solo puede ser considerada como el grupo mínimo y más significativo de genes bajo estos procesos.

La comparación de genes acelerados en tasas de sustitución sinónimas (Ks) y no sinónimas (Ka) no mostró diferencias significativas entre ambas especies. El numero de genes acelerados bajo Ka es mayor que en Ks, y en ambos casos, muestran más eventos de aceleración en humano que en chimpancé. Esto correlaciona con la mayor frecuencia de de eventos de PS en chimpancé y puede ser debido al tamaño de la pobalción, menor que el de chimpancé, que ha marcado la evolución de la especie humana. Sin embargo la diferencia es importante y sugiere que chimpancé ha experimentado una proporción mayor de eventos de PS que humano. Este resultado se confirma por duplicado cuando se utiliza una versión posterior de Ensembl para mirar los casos de PS asociados a tejidos. Esta observación, introducida por primera vez en este estudio, es citada y verificada posteriormente por Bakewell y colaboradores en la revista PNAS.

La comparación de eventos de PS con los de RRT mostró que la comparación de tasas mediante el método "pairwise" (donde se comparan los linajes por pares) de Li y los resultados del test II de ramas-y-sitios para la detección de PS, son significativamente distintos. Mientras las razones de este comportamiento ya se han discutido en la sección de resultados, estos resultados demuestran tanto la importancia de la utilización de los métodos ML más sensibles, como los posibles errores asociados a la utilización de métodos simples para intentar concentrar eventos de PS.

Las posibles implicaciones a nivel funcional de los genes encontrados bajo cada uno de estos procesos también fue estudiada. A partir de estos resultados podemos observar como tanto, los métodos para la detección de selección positiva, como el enfoque para el análisis funcional, son importantes y pueden afectar los resultados. En ellos podemos ver como la distinción de posibles falsos positivos debidos a casos de RSC, los cuales muestran sobre-representación en algunas clases funcionales, es de extremada importancia a la hora de obtener conclusiones robustas sobre los genes de PS. Se observa también que cuando se aplica la corrección necesaria ante los efectos de test múltiples, ninguna de las clases funcionales, particularmente algunas reportadas en otros estudios, muestran representaciones significativamente mayores bajo eventos de PS.

La comparación de clases funcionales mediante el enfoque de comparación entre linajes derivados frente al linaje ancestral también ha dado resultado a perspectivas nuevas. Se observa que el proceso de RSC es mucho más frecuente en ambos linajes derivados. También se observan cambios de tendencia, cor-

respondientes un aumento o disminución opuesto en humano y chimpancé, y se obtienen mayores evidencias de que los genes bajo RSC y PS son grupos distintos con tendencias diferentes a nivel funcional.

También podemos observar que la aproximación al estudio funcional partiendo de un enfoque nuevo en su aplicación para el estudio de procesos evolutivos, nos muestra que existen asociaciones significativas entre PS y clases funcionales cuya relación había sido previamente controvertida y en otras que hasta el momento no se habían descrito.

Del análisis funcional de los procesos de selección natural, se pasa a un estudio de selección natural al nivel de órganos y tejidos del del órganismo. El propósito del mismo es ver si existen evidencias de selección positiva en conjuntos de genes agrupados por funcionalidad desde una perspectiva del organismo como sistema. Particularmente en la búsqueda de las diferencias únicas de humano, varios estudios sobre la posible acción de PS sobre genes específicos de cerebro se han publicado. Algunos, mediante enfoques indirectos basado en la comparación de medias entre tasas, habían concluido que la acción de PS había sido tan extensa como para haber producido una elevación general de tasas en genes asociados al desarrollo del cerebro humano. En los resultados podemos observar que tanto, los métodos experimentales, como aquellos utilizados para la definición de genes específicos de tejidos (TSG), puede afectar los resultados obtenidos. Por lo cual es necesario abordar los anáslisis de TSG mediante la consideración de varios métodos alternativos. Sin embargo, independiente del método utilizado, se observa que las diferencias en la tasa de Ka/Ks entre linajes, se debe principalmente a efectos de linaje, y sesgos en chimpancé. La comparación de las tasas evolutivas entre las categorías TSG dentro de cada linaje mostraron que mientras existen algunas clases con tasas Ka/Ks significativamente elevadas, los TSG de cerebro muestran tasas evolutivas bajas. Aun más, destacan entre todas las otras clases de TSG, al ser la que muestra el mayor numero de diferencias significativas en humano siendo las más bajas de todas. Esta observación, es importante en cuanto a que lejos de ser una clase de genes con posibles tasas elevadas, sería interesante poder entender porqué los genes de cerebro han evolucionado bajo presiones selectivas particularmente fuertes. Los resultados de la aplicación del test II de selección positiva confirma que no existen evidencias de que los TSG de cerebro hayan evolucionado de forma distinta en comparación con chimpancé y otros tejidos. De hecho la mayor parte de los eventos de PS se observan en genes expresados en más de un solo tejido.

En el estudio de la relación de la selección natural y procesos de enfermedad, se utilizan métodos de rama-por-sitio para poder discernir si la medición de presiones selectivas a nivel de codones permite distinguir sitios cuya funcionalidad es importante para el organismo, y en el cual las mutaciones tendrán una mayor probabilidad de causar efectos fenotípicos y enfermedad. Los métodos anteriores han usado aproximaciones simples a las estimas de presiones selectivas junto con medidas de parámetros estructurales, o físico químicos, para obtener predictores de este tipo. Sin embargo, este es el primer estudio donde se utilizan métodos precisos para la estimación directa de presiones selectivas a nivel de codones,

y se emplean para explorar la relación entre presiones selectivas y enfermedad. En los resultados podemos observar como estimas de presiones selectivas fuertes pueden distinguir dominios y sitios funcionalmente importantes y en los cuales las mutaciones generalmente llevan al desarrollo de enfermedad en la proteína modelo, p53. La p53 es utilizada debido a que ha sido una de las proteínas más estudiadas y existen datos y estudios numerosos reportando las mutaciones de p53 que se observan en procesos de enfermedad. Partiendo de la observación que todos los residuos asociados a enfermedad en esta proteína, se encuentran bajo valores de Ka/Ks < 0.1, se utiliza este valor de corte como parámetro predictor del posible papel funcional de residuos proteicos y su asociacion con procesos de enfermedad. La hipótesis se confirma tras un análisis extensivo de bases de datos con anotación de mutaciones en proteínas que causan o no enfermedad, y consecuentemente se comprueba que las presiones selectivas medidas a nivel de codones es un estimador capaz de predecir sitios frecuentemente asociados a enfermedad. Este método, fue citado y utilizado en el análisis de variación genética y polimorfismos en rata por el consorcio STAR, en un estudio genómico publicado en la revista Nature Genetics.

Una característica interesante de este estudio es que representa un abordaje directo al análisis de la conexión de selección natural y con efectos fenotípicos. La limitación en este caso, se encuentra en que mientras algunas de las enfermedades estudiadas representan enfermedades complejas, la gran mayoría de las mutaciones analizadas están relacionadas a enfermedades mendelianas. Mientras el estudio de las enfermedades complejas está limitado por la dificultad de obtención de datos sobre la localización precisa de los genes y sitios causantes, se comienzan a ver bases de datos que colectan anotaciones relacionadas. Quizás en un futuro estas podrían permitir un abordaje más completo de este "eslabón perdido," de la biología evolutiva: el de la conexión entre eventos de selección natural y sus efectos fenotípicos.

Conclusiones

1. La aplicación de metodologías nuevas, mas sensitivas, y robustas para la detección de eventos de selección positiva, junto con el enfoque estadístico apropiado, da lugar a diferencias importantes en los resultados relacionados a la frecuencia y posibles implicaciones funcionales asociados a este proceso evolutivo. Notablemente, no se observan diferencias funcionales asociados a genes de selección positiva, cuando se toma en cuenta las correcciones necesarias ante los efectos de test múltiples.

2. La diferenciación de eventos de relajación de presiones selectivas de aquellos de selección positiva, da lugar a diferencias importantes en cuanto a la frecuencia, identidad, e implicaciones funcionales de los procesos evolutivos fuera de la neutralidad que han marcado la evolución de humano y chimpancé. La relajación de presiones selectivas ocurre con mayor frecuencia que la selección positiva en ambos genomas, y puede ser responsable de

conclusiones controvertidas encontradas en estudios previos acerca de la sobre-representación de algunas clases funcionales bajo el procesos de selección positiva. La selección positiva ha sido mas frecuente en el genoma de chimpancé que en el de humano.

3. El abordar el estudio de posibles implicaciones funcionales de genes bajo selección positiva desde metodologías distintas a las empleadas mayoritariamente en estudios previos, muestra que existen clases funcionales que muestran sesgos es su distribución hacia una evidencia elevada de selección positiva.

4. El análisis de las presiones selectivas entre diferentes grupos de genes pertenecientes a distintos tejidos, muestra que mientras las diferencias entre linajes son pocas, el efecto de la selección natural al nivel de tejidos u órganos del sistema no es uniforme, resultando en diferencias significativas entre distintas categorías. Los genes específicos de cerebro se destacan por tener las tasas significativamente más bajas entre las distintas clases, y no muestran evidencia de haber evolucionado de forma particular mediante eventos de selección positiva.

5. La utilización de estimas de presiones selectivas a nivel de codones se puede utilizar para la predicción de sitios en los cuales las mutaciones tienen alta probabilidad de causar efectos fenotípicos resultando en enfermedad.

Part I

Introduction

Chapter 1

Understanding Ourselves From a Comparative Standpoint

1.1 An Age Old Pursuit

Understanding what makes us human has been one of the greatest pursuits throughout human history. However it is only relatively recently that we have started to acquire the ability to address such a question from the wide array of disciplines and understandings that may provide the integrative scientific insight necessary to make progress in this respect. The knowledge of the evolutionary relatedness and proximity of humans and other primates has only been around since the times of the development the theory of natural selection (Varki *et al.* , 2008), and efforts in the past century or so had begun searching to provide clues using chimpanzee, our closest evolutionary relative, from a comparative approach.

The first studies focused naturally on anatomical differences, and were followed by behavioral, pathological, and later, by molecular approaches based on biochemical and immunological studies in the 60's (Varki *et al.* , 2008). From then up until today, many more detailed and more numerous analyses have been conducted at different and deeper molecular levels, but even in as little as the last decade, when the genomes of both species were sequenced, we have marveled at the paradox of how relatively small differences have been found to set us apart and how markedly different we really are.

Certainly there are many disparities that span numerous aspects of our morphology, physiology, behavior, and both, general, and cognitive abilities, but it seems incredible to note that molecular comparisons reveal surprisingly low differences. Estimates of differences among protein coding genes range from initial estimates of approximately 1%, to estimates of approximately 4% when

additional regions consisting of duplications and deletions have been analyzed (Varki *et al.* , 2008).

In spite of these small differences, it is inevitable to wonder about the evolutionary processes that have given rise to them, as well as the possible impact and implications they may have in different aspects of our organism. Both of these considerations are tied hand in hand. Distinctly, an appropriate understanding of the possible functional impact of this variation cannot be properly achieved without an understanding of what part of it is due to changes that are neutral, having little or no effect on the organism's fitness, and those which do have an effect and are acted upon by natural selection. It is for this reason that many studies have focused on the functional roles of genes showing signatures of natural selection; be it those that have highly conserved regions, where mutations are deleterious and are removed by the action of negative or purifying selection, or those where beneficial mutations have been fixed through positive selection (PS), leading to adaptation.

In the past decade, the development of more powerful methods for detecting signatures of positive and negative selection in contrast to neutral evolution, together with the explosion in the amount of genomic sequences, have provided a the most favorable scenario as of yet for addressing these questions. Additionally, the increasing availability of a plethora of other types of molecular data – i.e. expression data, together with the remarkable differences we observe in cognitive abilities between human and chimp, have prompted studies considering the potential contribution of genomic variation to differences at the organ system level, and in particular, in the human brain. In this light many other types of integrative studies have been addressed in order to approach and understanding ourselves from a comparative standpoint

However, given the evolutionary relatedness as species descending from a common ancestor, it is important not to forget that the information that can be gained from a comparative perspective also encompasses a consideration of the functions and features which have been maintained throughout evolutionary history, as well as the innovations or susceptibilities we may have inherited in common with our close relatives. Those which have been conserved do not only serve as an adequate point of comparison for gaging what has changed, but it is also in itself an indicator of those that have been important enough to resist change throughout millions of years. These considerations gain particular importance when considering that natural selection is to date the best explanation for the mechanism leading to the development and maintenance of function, where different failures or harmful developments can give rise to disease.

1.2 A Review of Some of the Most Relevant Studies

The importance of understanding the relationship between natural selection, function, and disease can be seen from the great number of studies that have been conducted and published relatively recently, seeking to deepen our understanding of the subject. The work presented here takes place precisely in the time period just before the publication of the chimpanzee genome to date, and being based on a comparative genomic approach, deals specifically with human-chimp and human-mammal comparisons for exploring patterns change and conservation resulting from natural selection on protein coding genes. More precisely it is specifically concerned with the application of maximum likelihood codon models for the estimation and testing of the different selective pressures that have acted on human protein coding sequences compared to those in chimp, their relation with function, their potential differences at the organ-system level, and using the deeper phylogenetic framework of vertebrates for their estimation at a codon level, their relation with disease. In order to provide a better understanding of the importance of the methods and approaches used here, a brief review of the previous works that have addressed these subjects are crucial.

Earlier efforts at a large or genomic scale employing a similar methodology had been conducted to elucidate the intricacies of human evolution by means of comparing rate differences and PS against other fully sequenced species. The work of Dorus *et al.* , 2004 found significantly higher rates of gene evolution in the primate nervous system when comparing against housekeeping and among subsets of brain specific genes. From this data they proposed natural selection as the underlying mechanism. Other efforts have focused on finding direct molecular evidence of PS. Clark *et al.* (2003a), using more than 7,600 homologous sequences, found 1,547 human and 1,534 chimp genes as likely candidates to have been acted upon by PS. In a later study, Nielsen *et al.* (2005), using more than 13,000 orthologous sequences, found that 733 genes deviated from strict neutrality, showing evidences of PS. In the latest genomic study the Chimpanzee Sequencing and Analysis Consortium (CSAC) had found 585 out of 13,454 human–chimp orthologous genes as potential candidates to have been acted upon by PS (Sequencing & Consortium, 2005).

Indeed, while these three publications have been hallmarks in the genomic-scale search for events showing PS and have provided much insight into the subject, the combination of methods used have produced certain disagreements and have left some important considerations unaccounted for. As noted in the CSAC publication, the set of 585 genes observed may only be enriched for cases of PS given that, for example, the statistic used could be greater than the cutoff by chance in almost half of these genes if purifying selection is allowed to act non uniformly (Sequencing & Consortium, 2005). In Clark *et al.* , 2003a, the branch-site test used for PS allowed distinguishing of lineage-specific cases of selection in the branches of human and of chimp, but has been criticized by

other authors given that it may have suffered from the inclusion of false positives originating from the lack of power of the test to distinguish true cases of PS from cases of relaxation of selective constraints (RSC) (Sequencing & Consortium, 2005; Zhang, 2004). The study by Nielsen *et al.* , 2005, with the exception of a small subset of 50 analyzed genes, was based on pair-wise comparisons that make it impossible to know in which of these lineages selection had occurred. In addition, in all of these studies, differentiation of the sets of genes under PS from the sets that are likely cases of RSC has not been done nor used specifically for study.

It is also important to note that likelihood ratio tests like those used here (see Chapter 3) and in some previous studies are sensitive to model assumptions (Zhang, 2004; Zhang *et al.* , 2005). While the tests used in this study have been shown to have a good performance under a variety of conditions (Zhang *et al.* , 2005), the definition of a genomic set of genes under PS is addressed from a conservative standpoint. Thus, while some of these studies have disregarded or considered multiple testing corrections only for case-specific observations after comparisons, here corrections for multiple testing are employed as the norm for all comparisons, while considering the uncorrected sets for confirmation of specific results where appropriate.

Most of the above mentioned studies have also addressed the possible functional impact of different evolutionary processes with a particular emphasis on those genes that have shown evidence of positive selection (PS). However, all of the above mentioned considerations, that are unavoidable in order to deduce PS events robustly, also play an important role on the inference of their possible functional impact. Particularly being able to distinguish PS from likely cases of RSC, as we shall observe in the results, proves fundamental. The consideration of corrections for multiple testing also plays an immense role. On the one hand, the consideration of results that are not corrected for multiple testing may lead to false functional associations. On the other, and quite importantly it also acts as an obstacle in the most frequent used approaches for functional inference. The necessity of correcting for multiple testing twice, first when deducing genes under particular evolutionary processes, and again when testing for a possible enrichment of functional classes within the deduced sets of genes (a two step approach), makes the approach rather inefficient (Al-Shahrour *et al.* , 2005b). Instead of disregarding the necessary corrections when running tests in favor of positive but less robust results, a different method is brought into consideration here. The use of a threshold free approach to detect skews in the distribution of genes belonging to functional classes ranked by their evidence of positive selection is employed, circumventing the inefficiency (Al-Shahrour *et al.* , 2005b) of multiple corrections for multiple testing and the dangers of threshold based approaches (Hughes, 2007).

Other studies had also looked at the possible action of natural selection at the organ system's level by comparing rates of evolution in genes expressed in different tissues. Special attention had naturally focused on human evolution and human brain specific genes. The conclusions from these studies had varied

and at times been contradictory and controversial. (Dorus *et al.* , 2004) found that the average dN/dS ratio was higher in brain specific genes in primates (humans and chimpanzees) than in rodents (mouse and rat), and in comparison to housekeeping genes. They also argued that among the genes showing the largest disparity in rates between primates and rodents, there was an excess of genes with human specific acceleration. Based on these observations they argued that genes in the human nervous system had experienced accelerated evolution caused by positive selection during human evolution. Khaitovich *et al.* (2005) found a slightly elevated rate of substitutions in brain specific genes in humans compared to chimpanzees. While the elevation was not significantly larger than what is observed for genes not expressed in the brain, it was found to be significantly larger than the ratio of human to chimpanzee rates of evolution in other tissue specific genes ($p < 0.05$). The comparisons in Khaitovich *et al.* , 2005 were based on comparing the rate of amino acid substitution in coding regions with the rate of nucleotide substitution in non-coding regions. In the study by Khaitovich *et al.* , 2005 it was found that the rate of amino acid evolution, as measured by dN/dS, or by any other measure, was overall much slower in brain specific genes than in genes expressed in other tissues. Studies specifically examining evidence for positive selection on the human lineage had found very little evidence for positive selection on the human linage (e.g. Clark *et al.* 2003b; Sequencing & Consortium 2005; Arbiza *et al.* 2006, although several genes relating to brain size show evidence for positive selection (Mekel-Bobrov *et al.* , 2005). In the light of these somewhat contradictory conclusions, a more thorough examination of rates of evolution in brain specific genes in the human lineage was needed. Here the evolution of genes at an organ systems level is addressed as a whole using various definitions and methods for defining tissue specificity, using other tissues as a point of comparison, testing the possible influence of different methods used to estimate and combine rates for groups of genes, and finally using a direct approach where positive selection is tested for specifically.

Previous studies focusing on the effect of genetic variation and natural selection on disease had also been conducted. Some of them focused primarily on the study of structural parameters of proteins such as stability, types of bonds, and the chemical properties of the amino acids and their environment, among others, together with sequence similarity cutoff values, and conservation profiles along alignments of homologous sequences from different species. Among these we can find the studies Sunyaev et al. (Sunyaev *et al.* , 2000, 2001), Chasman & Adams (2001), Wang & Moult (2001). Other studies have employed more direct estimates of evolutionary rates including the important consideration of the phylogenetic framework necessary in order to asses and properly deal with the non-random structure of dependence defined in the historical relationships among species (Felsenstein, 1985) which can have an important influence in the estimation of selective pressures (Miller & Kumar, 2001). In this light, studies such as those of Miller & Kumar (2001), Koref *et al.* (2003), and Saunders & Baker (2002) were conducted, which although based on single or few proteins,

provided much insight to the importance of using more direct estimations of evolutionary estimates and phylogenetically minded approaches. In particular, the study of Saunders & Baker (2002) concludes that the combination of both structural and evolutionary parameters is better than any one on its own, and suggested that when high numbers of sequences are available, evolutionary parameters are more informative than structural parameters to predict deleterious mutations. In addition to these studies, many methods have been developed to use structural, conservation, and evolutionary data to predict the possible phenotypic effect of single point mutations and single nucleotide polymorphisms (SNPs). All of them make use of homologous sequences, in many cases without differentiating relationships of orthology and paralogy, given that sequences and structural conservation scores are generally derived from blast based approaches to define protein families or sequences for study. An extensive review of these methods can be found in (Dopazo, 2008). Here a direct and phylogenetically minded approach for the estimation selective pressures at a codon level is used both, to study the relationships of selective pressures and disease, and to build a predictor of the phenotypic effects of nonsynonymous mutations and polymorphism in protein sequences that is applied genomically in human.

Chapter 2

Natural Selection and the Neutral Theory of Molecular Evolution

2.1 A Brief History of Molecular Evolution

One of the main aims of evolutionary biology is the understanding of forces shaping the evolution of populations and species. Under this premise, and particularly within the field of molecular evolution, clarifying the roles that drift and selection have played in molding the patterns of genetic variation observed in extant sequences has become one of the central points of focus. However, in spite that our ability to assess their relative contributions has benefited enormously from the availability of data and considerable developments in methodology, even today, it still remains as a broadly open question.

The earliest discussions of Darwin's theory of natural selection, in a point in time where molecular data was scarce, gave origin to the consideration of diverse theoretically based hypotheses on the relationship of natural selection and the observable variety in traits among organisms. Indeed, early notions tended closer to a panselectionist perspective where natural selection was postulated as the main force driving observed differences in traits and variation. With the advent of molecular biology and an increase in the availability of data to test hypotheses against observations, alternative considerations became increasingly important. In particular several experiments revealed that variation in the composition of proteins was far too great and evenly spaced in time to be explained by natural selection alone.

Zuckerkandl and Pauling's observation that the number of amino acid substitutions in hemoglobin was proportional to the phylogenetic divergence times of sequences through calibration with fossil records, provided the basis for the molecular clock hypothesis (Zuckerkandl & Pauling, 1965). The molecular clock

hypothesis states that the rate of evolutionary change of proteins is approximately constant over time and across lineages. Smith and Margoliash extended the analysis to Cytochrome-C (Smith & Margoliash, 1964). Both of these studies agreed fairly well with established divergence times based on paleontological evidence.

Taking advantage of the development of gel electrophoresis, studies such as those pioneered by Hubby and Lewontin (Hubby & Lewontin, 1966), also revealed that even within populations, the amount of polymorphism observed within proteins were far to great to be explained by either side of a purely selectionist view: it could certainly not be explained by the classical view of selection, where natural selection reduces the amount of variation by favoring optimal genes, and was too high even for the theory of balancing selection acting to maintain variation by overdominance.

In 1968, Motoo Kimura, based on previous studies of the molecular clock, calculated the average rate of DNA substitution per 100 amino acids per million years (Kimura, 1968). The observed rate was of approximately one substitution per 1.8 years in mammalian sequences. A rate that he proposed was impossibly high to explain without the existence of selectively neutral mutations. As such, he introduced the neutral theory of molecular evolution which postulates that substitutions which do not have an impact on an organisms fitness occur frequently and could come to fixation in populations of finite size in the absence of natural selection due to genetic drift. The study was followed in the next year by more extensive work by King and Jukes (King & Jukes, 1969) that also supported the relevance of selectively neutral mutations, and together provided both, a theoretical basis to explain the molecular clock hypothesis, and for delving deeper into an understanding of the forces shaping the patterns of evolution among populations and species.

Finally, Tomoko Ohta, focusing on the interplay of natural selection and genetic drift, particularly for mutations that were not strictly, but nearly neutral, established the notion of a competition between these forces, where slightly deleterious mutations (later extended to include slightly advantageous mutations as well) could be eliminated by natural selection or chance, or they could come to fixation by genetic drift (Ohta, 1973). Therefore, given the inverse relation of population size and generation time, the slow rate of protein evolution in comparison with that of non-coding DNA could be explained by the stronger action of natural selection in populations of larger size weeding out slightly deleterious mutations. More specifically, the neutral theory holds that slightly deleterious mutations are frequent and are mainly governed by the stochastic effects of genetic drift in populations of finite size: where negative selection acts to eliminate deleterious mutations (the majority of mutations), and directional or balancing selection are considered as being infrequent events. As such, the effective population size (N_e), and the neutral mutation rate, μ_0, determine the levels of polymorphism (H) within species, and the rates of divergence among species, respectively.

Nowadays, it is clear that both neutral evolution and natural selection have

played important roles throughout evolutionary history, and the question is centered instead, not on which view is correct, but to what proportion these processes have played a role in the evolution of populations and species. Importantly, both theories yield exclusive propositions, and the theory of neutral evolution has become the hallmark null hypothesis against which to test for natural selection. Indeed, with the recent explosion in the availability of molecular data we are now better suited to enlighten the roles each of these processes have played throughout evolutionary history.

2.2 Testing for Deviations from Neutrality

The neutral theory of molecular evolution states that the levels of polymorphism in the population are mainly the results of mutations that are neutral or nearly neutral with regard to their effect on fitness and reproductive success. On the other hand, natural selection is the mechanism through which the relative genotype frequencies vary according to their reproductive success in a population. Mutations that confer a greater reproductive success to the individual, increase in successive generations tending towards fixation through the action of positive selection as the underlying mechanism for adaptation. Mutations that confer a reproductive disadvantage decrease in successive generations due to the action of negative or purifying selection by which previous adaptations are maintained.

Accordingly, both positive and negative selection perturb the patterns of genetic variation relative to what can be expected from a neutral model. Thus, the neutral theory has become an important point of reference acting as the null hypothesis against which natural selection can be tested for. In this form, a wide array of tests for the signatures of positive and negative selection have been devised. Most of them fall under two important perspectives: 1) those looking at recent selection in populations based on an analysis of polymorphism within species 2) those looking at ancestral selection based on the patterns of divergence between species. In some cases, given that the neutral theory predicts a positive correlation between the levels of polymorphism within species and divergence between species, both of these perspectives are employed for analysis. Box 2.1, provides a summary of some examples of each of these methods.

Methods based on the analysis of polymorphisms within species can shed light on recent and ongoing selection in a population. They can consist in the detection of skews in the allele frequency distributions, reduced levels of genetic variation, or elevated levels of linkage disequilibrium (LD) relative to the neutral expectation (Biswas & Akey, 2006). However, it is important to note that the presence of confounding effects from population demographic history can cause similar signatures as those of natural selection. The problem is that the neutral population model employed includes assumptions such as those of constant population size and lack of population structure which are violated partially or severely in many if not most real scenarios (Nielsen, 2001). For

Tests based on polymorphisms within species

Tajima's D: measures the difference between two estimators of the population mutation rate, θw and π. Under neutrality, their means should be approximately equal and significant deviations from zero indicate a skew in the allele frequency distribution relative to neutral expectations.

Fu and Li's D and F: tests for a skew in the allele frequency spectrum, but makes the distinction between old and recent mutations as determined by where they occur on the branches of genealogies. The D and F statistics compare the population mutation rate based on the number of derived variants seen only once in a sample with θw or π, respectively. Similarly,significant deviations from zero in D and F indicate a departure from the neutral expectation.

Fay and Wu's H test: is used to detect the presence of an excess of high frequency derived alleles in a sample, which is a hallmark of positive selection.

Long range haplotype (LRH) test: examines the relationship between allele frequency and the extent of linkage disequilibrium (LD). Positive selection is expected to accelerate the frequency of an advantageous allele faster than recombination can break down LD at the selected haplotype. Thus, a hallmark of recent positive selection is an allele that has greater long-range LD given its frequency in the population relative to neutral expectations..

iHS: is applied to individual SNPs and is based on a standardized and log normalized ratio of the estimates of the integral of the decay in homozygosity as a function of distance from the SNP (iHH) for ancestral and derived alleles. Large positive and negative values of iHS indicate unusually long haplotypes carrying the ancestral and derived allele, respectively.

LD decay (LDD): detects large differences in the extent of LD between two alleles at a particular locus SNPs with high ALnLH values imply that the decay in LD for one allele is unusual compared with that of the alternative allele, in which the pattern of LD decay is within an a priori defined bound of the genome-wide average.

FST: quantifies levels of differentiation between subpopulations. Under neutrality, levels of FST are largely determined by genetic drift and migration, but local adaptation can accentuate levels of population differentiation at particular loci thus resulting in large FST values.

Tests based on polymorphisms within species and the divergence between species

Hudson–Kreitman–Aguade (HKA) test: is used to determine if levels of nucleotide variation within and between species at two or more loci conform to the neutral expectation of a positive correlation between levels of polymorphism within species and divergence between species.

McDonald Kreitman (MK) test: also uses polymorphism and divergence data, but compares synonymous versus nonsynonymous sites at a specific locus. In the MK test, a 2x2 contingency table is formed to compare the number of nonsynonymous and synonymous sites that are polymorphic within a species (PN and PS) and fixed between species (DN and DS).

Tests between species

dn/ds test: the ratio (ω) of nonsynonymous substitutions per nonsynonymous site (dN) is compared to that of synonymous ones (dS) in protein coding loci. Values that deviate significantly from the neutral expectation of ω= dN/dS = 1, are subject to functional constraint and purifying selection on dN (ω<1), or under positive selection (ω>1).

Box 2.1: Methods and approaches to Detect Positive Selection. Reproduced from Biswas & Akey, 2006

example, population expansions or bottlenecks can cause patterns of genetic variation that would depart significantly from the neutral expectation, leading to possible false inferences (Biswas & Akey, 2006), and estimators such as Tajima's D (Box 2.1) can actually be used to detect not only selective sweeps, but also population bottlenecks and subdivision (Simonsen *et al.*, 1995). As a response to this inherent problem, when employing these tests, some authors have suggested, it important to use strategies such as the comparison of signatures present in various loci in order to search for patterns that are particularly striking above the confounding effects of population demographic history that would affect most loci in the genome (Biswas & Akey, 2006). However, even such measures may prove inefficient to detect cases of weak selection which are likely to have only a mild effect on the genealogy (Nielsen, 2001).

Other tests are based on the comparison of both on the polymorphism between species and the divergence between species. Examples of these are the HKA and MK tests described in Box 2.1. The HKA test has the additional advantage of drawing power from combining data from multiple loci. However, the confounding effects of population demography are still inherent since the variance in the number of segregating sites within a single population may be affected by migration (Nielsen, 2001). The MK test conversely, uses quite a different strategy based on the comparison of variability among different classes of mutations: synonymous and nonsynonymous mutations. Since the test is based in the comparison of different types of variability within the same loci, the effects of the demographic model is the same for both types of variation making them independent of these kinds of effects. Thus the results of the MK test are quite appropriate for the obtention of unambiguous signals of natural selection. Unfortunately, it is not always evident which type of selection may have taken place and it is thus limited in certain cases for the deduction of positive selection (Nielsen, 2001).

Tests based on comparison between species have also been devised. In particular the dN/dS method, which compares two different rates of substitution –synonymous and nonsynonymous rates, has become one of the most direct methods for the detection of positive selection. It uses both rates to look at information under the same loci and is thus free from demographic assumptions. It can distinguish among different types of selection, and has been one of the most successful approaches in providing non-ambiguous evidence for the presence of positive selection (Nielsen, 2001). However, this method is not free of assumptions, and important considerations must be taken into account in order to achieve robust results. In the next section, we shall delve deeper into this method which is the underlying base of the work presented here, where signals of positive and negative selection are examined from a comparative genomic stand point.

Certainly the methods mentioned in this section are not the only ones available. However, they provide an example of the range of available forms to test selection against neutrality, from a wide array of perspectives that are based on different approaches and properties that underlie the theoretical bases of these

evolutionary processes. As such, they serve to provide a brief introduction and a picture of the framework within which the methods used in this work are situated; highlighting, in particular, some of the important considerations with regard to model assumptions and confounding factors, that must be taken into account when considering tests of neutrality in general.

Chapter 3

ML Codon Based Methods

3.1 A Probabilistic Maximum Likelihood Framework

As mentioned in the last section, the dN / dS ratio method (also known ω = dN/dS = Ka/Ks) is based on the comparison of two different rates of substitution within a given protein coding locus. More precisely, the rate of nonsynonymous substitution per nonsynonymous site (dN) is compared to that of synonymous substitutions per synonymous site (dS). As mutations in nonsynonymous sites can more directly affect protein function, they are more likely to affect the fitness of an organism than changes which would leave the protein intact. Under negative or purifying selection, nonsynonymous substitutions that would confer a reproductive disadvantage accumulate more slowly than synonymous mutations yielding estimates of $\omega<1$. Under positive selection, nonsynonymous mutations that are advantageous would become fixed at a faster rate than synonymous mutations leading to an $\omega>1$. In the case were nonsynonymous mutations did not have an effect on fitness, they would accumulate at a rate approximately equal to that of synonymous mutation which are in this case the proxy used to estimate the neutral expectation. Thus, under a scenario of neutral evolution we would expect an $\omega \simeq 1$. Therefore, testing for a dN that is significantly different from a dS constitutes a test of neutrality that can be used to infer different selective processes, and has become a standard measure of selective pressure (Pond *et al.* , 2009).

A fitting implementation of this concept requires several important considerations. First it is important to consider that the genetic code forces a consideration of codons as the basic evolutionary unit. It is under this consideration that both synonymous and nonsynonymous substitutions may be defined. Initial implementations of this method addressed pairwise comparisons among sequences providing estimates of synonymous and nonsynonymous rates based on averages of the numbers of each type of substitution that would be required

for the conversion of one codon to another based on all the shortest possible paths that between both codons (i.e. the methods of Li (1993) and Nei and Gojobori (1989). However, it is clear that these earlier *ad hoc* methods lacked the consideration of important aspects. For example, they could not account for cases where all alternative paths among codons involved both types of substitutions, they were difficult to extend to comparisons of more than two lineages, and could not take into account that the variation present in extant sequences can be derived from common ancestry.

However, in the early nineties probabilistic codon based models of nucleotide sequence evolution were introduced independently by Muse and Gaut (MG94) (Muse & Gaut, 1994) and Goldman and Yang (GY) (Goldman & Yang, 1994). Both of them model nucleotide sequences as continuous time Markov Chains within the space of codons available in a particular genetic code. In one parametrization, the instantaneous rate matrix of the process

$$Q = \{q_{ij}\} \tag{3.1}$$

is given by :

$$q_{ij} = \begin{cases} 0 & \text{if i and j differ at two or three nucleotide positions,} \\ \pi_j & \text{if i and j differ by one synonymous transversion,} \\ \kappa\pi_j & \text{if i and j differ by one synonymous transition,} \\ \omega\pi_j & \text{if i and j differ by one nonsynonymous transversion,} \\ \omega\kappa\pi_j & \text{if i and j differ by one synonymous transition.} \end{cases} \tag{3.2}$$

where $\pi_j$is the stationary frequency of codon j, $\kappa$is the transition/transversion rate ratio, and ω (= dN/dS) is the nonsynonymous / synonymous rate ratio. Variations in the parametrization allow coding for nucleotide substitution models that can consider different rates of nucleotide changes. $\pi_j$can be calculated in various ways. For example, as the frequency of the target nucleotides in each of the three codon positions as in the MG94 model, or as the frequency of the target codon in the GY model. The transition probability matrix is calculated as :

$$P(t) = e^{Qt} \tag{3.3}$$

where time or branch-length t is measured as the expected number of nucleotide substitutions per codon averaged over all sites (Li and Goldman, 1998). Using this model and Felsensteins's Pruning Algorithm (Felsenstein, 1981), parameter estimates can be obtained by numerical optimization under maximum likelihood (ML) taking into account relationships among sequences by modeling parameters and substitutions probabilistically across all of the nodes in a phylogeny. That is to say, given a phylogeny, a probabilistic model for the internal nodes

relating them can be included without the need of assuming star phylogenies, as in pairwise methods. As mentioned earlier, different parametrizations also allow for the incorporation of complex mutational models, also taking the structure of the genetic code into account, which permits correction for codon and nucleotide substitution biases. It is also important to note that the behavior of ML estimates, given enough data, tend to converge to the true values and have minimum variance among all unbiased estimators for any reversible substitution model with a finite number of states (Pond *et al.* , 2009).

3.2 Different ML Models: branch, site, and branch-site methods

From the initial implementations of ML models, many developments have taken place allowing computationally accessible means of extending how different model parameters can be distributed along and across the different phylogenetic partitions that provide the basic framework for their estimation. The first and simplest models considered parameter distributions –such as those of dN and dS, that were global and were taken as the sum of all fractions occurring across all sites and lineages in the phylogenetic tree. These initial models had interesting applications in scenarios where the purpose was a rough approximation of of parameters from small or limited data sets, when the estimation phylogenetic relationships or ancestral sequence reconstruction where the focus and substitution rates could be labeled of as nuisance parameters, or for use as a null model against which to test non-homogeneous distributions of parameters along and across different lineages of the phylogenetic tree. However, it is clear that these global models are not realistic enough to be of much use when estimating dN and dS for biological sequences. More precisely, selective pressures are likely to vary among sites of different functional importance along proteins, and it is likely that they may vary with time among different lineages of the phylogeny. For example, under a global model, widely different selective regimes varying across sites or among lineages may produce very similar estimates of substitution rates when averaged.

Indeed various different local models have been implemented, as extensions of the basic probabilistic framework described in the previous section, that allow parameters to vary among lineages, different sites, or both lineages and sites. Under these local models, different independent partitions or groups can be specified, under which particular parameters can be averaged. Branch models average estimates of dN/dS across lineages in a phylogeny. Partitions ranging from one independent lineage or lineages to the most complex model where all lineages have independent averages can be coded. However, it is important to consider that as a means to detect positive selection, estimates based on whole lineage averages are quite conservative since it is expected that purifying selection should be quite extense among functional genes and dN to be much smaller than dS when averaging along all sites.

Site models, allow for the variation of rates averaging for codons across a phylogeny instead of along each one of the sequences, and given enough data, permit testing for the presence of a few sites under positive selection in protein coding genes. Several implementations are available, but two main approaches will be discussed. Not being able to distinguish in which of the branches selection has occurred, the use of these models is generally geared to the detection of the actual sites that are functionally important or have experienced positive selection with a protein.

One approach is modeling heterogeneous rates across sites (as in Yang & Nielsen, 2000) with a simple distribution –usually the gamma distribution. In practice, the shape parameter(s) of the distribution are estimated by numerical optimization from the data under ML, and the sum of the conditional probabilities for the rate parameters at each site are used for computing the likelihood from the expectation over the distribution of rates. Using one, or a combination of different free or bound distributions or free parameters, different models can be coded allowing for different different groups of sites with different mean estimates of ω (Pond *et al.* , 2009). Empirical Bayesian approaches can then be used to compute the posterior probability of of each site belonging to each one of the rate classes and to provide a post estimate of the mean dN / dS at each site. In particular this method benefits from the ability to pool information across multiple sites, and can be useful both, for the estimation of selective pressures at individual codon sites, and in the detection of cases of weak selection, where only a few sites that would be undetectable from an average along the full coding sequences, can be detected. Nevertheless, these models are not useful when searching for lineage specific changes in rates.

A second approach is that implemented in programs such as the site-wise likelihood ratio test (Massingham & Goldman, 2005). Instead of drawing substitution rates from a distribution, they are estimated directly at each site. In particular, the SLR method uses ML methods as described above to estimate important global parameters accounting for nucleotide and codon biases from all of the sites and lineages. Each codon site is then considered independently for each of the branches in the phylogeny. Estimates of site-wise dN/dS ratios can thus be obtained and different models can be coded for each site allowing models of neutral ($dN \simeq dS$), positive ($\omega > 1$), or negative selection ($\omega < 1$). Some of the advantages of this kind of approach is that no assumption needs to be made on the underlying distribution of rates among sites. Some of its disadvantages is that it requires a greater amount of sequences in order to draw enough power for inference.

Finally, a more complex set of methods exist that combine both site and branch specific local models allowing for different site classes in particular branches within the phylogeny. Practically, the most commonly used implementations of these models –such as in Zhang *et al.* (2005), allow two partitions among the lineages of a phylogeny, and different site classes for one or both of these. While the power in the detection of positively selected sites with a particular lineage or group of lineages is hampered from a reduction in the

sampling size making it more difficult to discover sites under selection, these models benefit from a more realistic approach to the detection of positive selection among lineages. More clearly, they allow that the selection of a few sites within a particular lineage, which is more realistic than a regime expecting a mean increase in dN above that of dS along the full protein, be employed in the detection of lineage specific events of positive selection.

Specific examples containing further detail, together with additional references, of these models are provided in Part III, Materials and Methods. As we shall care to cover next, the flexibility of considering different variants and constraints within the different models that have been covered can not only be employed to obtain parameter estimates, but also and importantly, to construct different statistical tests for positive selection.

3.3 Testing Evolutionary Hypotheses Using ML Models

The different types of models described earlier can be considered as evolutionary hypothesis which can be contrasted statistically given that the overall likelihood estimate for the model is a measure of the probability of generating the data given the model. For the purpose of contrasting evolutionary hypothesis under this framework, null (H_0) and alternative (H_A) models can be compared through a likelihood ratio test (LRT) when models are nested; that is, when the null model can be obtained from the alternative by constraining a given number of its parameters. For this test the likelihood ratio test statistic (LR) is given by :

$$2\Delta = 2(log\, L(H_a|D) - log\, (H_0|D) \tag{3.4}$$

where twice the difference in log likelihood of the alternative minus the null model is compared against a given cutoff above which the null hypothesis may be rejected in favor of the alternative one. For nested models, the distribution of $2\Delta$for data generated under the H_0model, follows a χ^2distribution with a number of degrees of freedom equal to the number of constrained parameters, and where the cutoff is taken from this distribution at a given significance level. Thus, through the LRT, when $2\Delta \geq$cutoff, $H_0$is rejected under the accepted false positive rate given the chosen significance level.

Through this approach, different tests can be constructed to test a wide range of hypothesis regarding the values and partitions of parameters that can be coded in the models presented in the previous section. Importantly and specifically relevant for this work, comparison of site, branch, or branch-site models can thus be used to code for different tests for positive selection.

Part II

Objectives

Objectives

1. The application of newer, more sensitive, and more robust maximum likelihood codon based models for the exploration of the relationship between natural selection, function, and disease from a comparative genomic's perspective.

2. The determination of the groups of genes that have undergone rate acceleration, positive selection, and relaxation of selective constraints in the human and chimp genomes, and a comparison of the relative frequencies of events and overlap of these sets amongst themselves, between lineages, and each and all of these sets against the rest of genes in the genome.

3. The determination of possible functional differences observed among the different evolutionary processes, among lineages, including a comparison with the ancestral lineage for the identification of possible relative trends, and in comparison to the rest of the genes in the genome.

4. The comparison of selective pressures and evidence of PS among tissue specific genes, and within and among primate and murid lineages, in order to understand if natural selection has affected different functional groupings at an organ system's level, and to undertake a controlled approach to re-examine the possible elevation of rates in genes expressed in the human brain due the action of positive selection.

5. Examining if the association between estimates of selective pressures at a codon level and functional importance of residues can be used and extended towards a general application in the prediction of sites where mutation is likely to cause disease.

6. Providing greater accessibility of methods used and results obtained to the members of the scientific community by making them available on line within integrated software platforms.

Part III

Materials and Methods

3.4 Test Based Inferences of Natural Selection at a Genomic Scale

3.4.1 Data Obtention and Preparation for Analyses

For the initial study of acceleration, positive selection, and relaxation presented in Chapters 4 and 5, ortholog annotations for the subset of 20,469 "known" Ensembl human protein-coding genes within the full set (30,709 genes) of the Ensembl version 30.35h H. sapiens database (Lander *et al.* , 2001) were retrieved from the Ensembl Compara database version 30 (Hubbard *et al.* , 2005). In this version of Ensembl-Compara, orthology annotations were based on reciprocal best hits corrected by synteny. Coding sequences (CDS) for the proteins represented by the largest transcript of each ortholog were retrieved from the Ensembl databases (Human: version 30.35c, Chimp: version 30.2, Mouse: version 30.33f, Rat: version 30.34, Dog: version 30.1b).

3.4.2 Relative Rate Estimation and Quality Filters

DNA CDS were aligned using ClustalW (Thompson *et al.* , 1994) and parameters by default with translated protein sequences as templates. Codons containing gaps were removed. Alignments smaller than 50 bp were excluded from the analysis. The upper limit for Ka and Ks rates considered were those of the human interferon γ (Ka = 3.06) and the relaxin protein (Ks = 6.39 substitutions per site per 109 years), showing the highest rates in human (Li, 1997). Assuming the human–mouse and human–chimp differentiation times to be about 80 million and 5 million years, respectively (Hedges & Kumar, 2003), all the comparisons with orthologs showing Ks $\geq$ 1 and Ka $\geq$ 0.5 substitutions/site for the RRT estimates, and those showing Ks $\geq$ 0.032 and Ka $\geq$ 0.0152 substitutions/site for ML lineage estimates, were excluded from the analysis. The RRT was performed using Li's method (Li, 1993) as implemented in the RRTree program (Robinson-Rechavi & Huchon, 2000). Sequences of human and of chimp were tested for deviation from a molecular clock using mouse, rat, and dog as the outgroup. Weights for each species in the outgroup were determined according to the topological scheme ((mouse:1/4, rat:1/4), dog:1/2)) as implemented in RRTree. Ka and Ks estimations were made on the CDS alignments of the largest transcripts of genes showing differences in GC content of less than 10%. Only three genes showed a GC content difference greater than 10% and were excluded from the analysis. Differences in human and in chimp rates were assessed using the Kolmogorov–Smirnov two-sample test (Sokal, 1981). ML estimations of Ka and Ks were computed jointly under a branch model for each ortholog using CodeML.

3.4.3 Branch-site Tests of Positive Selection and Relaxation of Selective Constraints

Positive Selection (PS) was evaluated using two different branch-site model Tests (I and II) (Zhang *et al.* , 2005) implemented in the CodeML program of the PAML (3.15) package (Yang, 1997). Branches in the phylogeny were defined *a priori* as foreground and background lineages. Under these models only the foreground lineage may contain events of PS. Human, chimp, and their ancestral lineage, derived from the common ancestor of mouse and rat, were tested independently as the foreground lineage. Sequences with fewer than three unique base pair differences in codons between human and chimp were removed for the analysis of PS.

In contrast to the statistical behavior of previous branch-site tests (Zhang, 2004), Tests I and II, developed and tested by Zhang *et al.* 2005 and employed at a genomic scale in this study, are improved methods of branch-site test models using an ML approach which has proved to be more successful with regard to differentiating PS from RSC. Test I compares model M1a against model A. M1a assumes two site classes, $0 < \omega_0 < 1$ and $\omega_1 = 1$, fixed in all the lineages of the phylogenetic tree. Model A considers four classes of sites. Site class 0 includes codons conserved throughout the tree with $0 < \omega_0 < 1$. Site class 1 includes codons evolving neutrally throughout the tree with $\omega_1 = 1$. Site classes 2a and 2b include codons conserved or evolving neutrally on the background branches, but which become under PS on the foreground branches with $\omega_2 > 1$. The proportion p_i of the site classes (p_0,p_1,p_2,p_3) and the mean value of ω_2 are estimated from the data by ML methods. Test II compares the null model A1 against model A. Parameters in A1 are equal to those of A with the exception that site classes 2a and 2b are fixed in the foreground with $\omega_2 = 1$. As was demonstrated by simulations (Zhang *et al.* , 2005), Test I cannot suitably distinguish cases of RSC from true events of PS. On the other hand, Test II, by allowing selectively constrained sites in the background to become relaxed under the proportion of site classes with $\omega_2 = 1$ set in the foreground of A1, is able to make this distinction, having an acceptable false discovery rate. One can therefore compare the results of both tests to distinguish cases of PS from events of RSC. Since the compared models are nested, likelihood ratio tests were performed and 2Δ values were posteriorly transformed into exact p-values using the pchisq function of the R statistical package (Ihaka & Gentleman, 1996). The chi-squared distribution with d.f. = 2 and d.f. = 1, which have been shown to be conservative under conditions of PS (Zhang *et al.* , 2005), were used to perform Tests I and II, respectively. In all cases, unless otherwise stated, p statistics derived from clock and PS analysis were false discovery rate–adjusted for multiple testing using the method of Benjamini and Hochberg (Benjamini & Hochberg, 1995).

3.5 Natural Selection and Biological Function

Functional characterization of accelerated and positively selected genes (PSG) was carried out by means of the FatiGO and FatisScan programs for functional annotation using GO (Al-Shahrour *et al.* , 2004, 2005a). FatiGO implements an inclusive analysis, where levels correspond to those in the directed acyclic graph hierarchy defining the relationship between GO terms (Eilbeck *et al.* , 2005) which is chosen for the analysis (Al-Shahrour *et al.* , 2004, 2005a). The program computes a Fisher's two-tail exact test in order to statistically define over- or under-represented terms in between two lists of genes considering p-values corrected for multiple testing (false discovery rate–independent adjustment) (Benjamini & Yekutieli, 2001).

FatiScan, uses a ranked list of genes which are subdivided into a prespecified number of partitions (containing the same amount of genes), and runs the test similar to that implemented in FatiGO by testing two lists that grow and shrink respectively as the genes in each partition are successively added to the first list and subtracted from the second list. This simple extension, permits an overview of functional classes, namely blocks of genes that share some functional property, showing a significant asymmetric distribution towards the extremes of a list of ranked genes. Given that multiple functional classes (C) are tested in multiple partitions (P), the unadjusted p-values for a total of C $\times$ P tests are corrected by the widely accepted FDR (Benjamini & Yekutieli, 2001). A fundamental advantage of the FatiScan method is that it does not depend on the original data from which the ranking of the list was derived. The significance of the test depends only on the numerical values used to rank the genes in the list and the strategy used for performing the partitions.

3.6 Natural Selection at the Organ System Level

3.6.1 Data Obtention and Preparation for Analyses

3.6.1.1 Expression Databases and Definition of Tissue Specificity

Expression data was obtained from three separate and independent databases. The first is the human-specific HG_U133A/GNF1H microarray based Tissue Atlas Dataset from the Genomics Institute of the Novartis Research Foundation Su *et al.* (2004). This database provides the results of MAS5 software based Absence / Presence calls for duplicate runs of each measurement of expression in separate preparations of the tissues analyzed. The collection of tissues analyzed were examined to count the original numbers of genes available belonging to each of 6 higher order groupings (tissue specific gene categories) under which the originating tissues are found (*pancreas, testis, respiratory, brain, B-lymphoblasts, and liver*), and into a 7th category, *other*, when they belonged to tissues that had less than 15 members which could not be grouped under one existing higher organ or tissue structure to surpass this number. Tissue

specific genes (TSG) were defined as those which were expressed in any number of tissues in one tissue category but not in any of the tissues belonging to any of the others. Housekeeping genes were defined as those that were expressed in 60 or greater of the original tissues and in all of the seven tissue specific categories defined. All disease related tissues were excluded before the analysis.

A second set was derived from GeneNote Shmueli *et al.* (2003) which is also derived from on replicate analyses microarray analyses using Affymetrix GeneChip HG-U95 on 12 healthy human tissues. Here tissue specificity is determined by analyzing genes highly expressed in one tissue and not others through a normalized expression levels approach developed by the authors. The genes included in each of the categories were thus determined as those which according to their method show high expression in one of the tissues but are lowly expressed in all of the others. As with all of the databases employed, all TSG categories that had less than 15 members were discarded, and those remaining were: *brain, prostate, spinal cord, liver, pancreas, bone marrow, skeletal muscle, kidney,* and *lung.*

The third expression data base used was the human specific portion of TissueDistributionsDB database, a repository of tissue distribution profiles based on Expressed Sequence Tags (ESTs) from UniGene (Wheeler *et al.* , 2006), kindly provided by the authors (Jonnakuty *et al.* , 2006). In this database, expression data for single unigene clusters from different tissue sources are grouped and related through a Tissue Synonym Library, the Brenda Tissue Ontology (Schomburg *et al.* , 2004), and Tissue Slims according to four different levels of organismal organization: the whole-body level (Level 2), the body-system level (Level 3), the organ-system level (Level 4), and the tissue-system level (Level 5). See and Brenda Tissue Ontology (Schomburg *et al.* , 2004) for specific definitions of tissue ontology terms. Tissue specific genes were defined, at the organ-system level, as those genes which were either uniquely expressed in a single tissue, or were expressed in 2-10 tissues with more than 60% of the expression falling in one tissue. House-keeping genes were defined as those genes expressed in 60 or more tissues, at the tissue-system level, where none of the individual tissues accounted for more than 40% of the total level of expression of each gene. Prior to the analysis, all percentages and tissue specificity indices were re-calculated to exclude disease tissues and cell lines. The remaining TSG categories were: *brain, male reproductive gland, germ, kidney, eye*, and *gastrointestinal tract.*

3.6.1.2 DNA sequences and Orthology Relations

Ortholog annotations from human, chimp, mouse, and rat and the corresponding coding sequences (CDS) were retrieved from the Ensembl-Compara and the Ensembl v42 database respectively (Birney *et al.* , 2006). This version of the database includes the second assembly of the chimp genome (PanTro II) and uses instead phylogeny based methods for the determination of orthologous relations of genes among species. For genes with multiple transcripts, the largest transcript was chosen. CDS from orthologous genes of human, macaque, mouse,

and rat reported in the study of Dorus *et al.* (2004), were retrieved from Genbank (Benson *et al.* , 2008). Genes where any of the four orthologous sequences from Genbank lacked coding sequence annotation were discarded. CDSs were aligned with Muscle (Edgar, 2004) using translated protein sequences as templates, with an upper limit of 9999 iterations or 5 hours running time and other parameters by default. Codons containing gaps were removed. Gapless alignments that did not cover at least 50% of the original multiple alignment length and those showing miss-aligned regions upon eye-inspection of the full set of alignments were removed.

3.6.2 Testing for Positive Selection and Branch Model Based Estimates of Evolutionary Rates

Measurements of evolutionary rates were conducted using two programs from the PAML package (Yang, 1997). Analyses are based on both pairwise comparisons and lineage specific analyses using a phylogenetic tree. Pairwise estimation of dN/dS (=ω) were done using the (Li, 1993) method as implemented in yn00 program and using maximum-likelihood (ML) estimates under the F3x4 model with estimated transition/transversion rate ratios. In the lineage specific analyses, the same model is used, but with estimates of ω obtained on each lineage of a five-taxon tree by means of the codeml program's free branch model. To get an overall estimate of the value of ω from all genes in a category, the standard error (SE) weighted mean was used (SEWM):

$$\bar{\omega} = \sum\nolimits_{i=1}^{n} \omega_i h_i, \quad h = \frac{SE\,(\omega_i)^{-1}}{\sum_{j=1}^{n} SE\,(\omega_j)^{-1}} \tag{3.5}$$

where ω_i is the estimate of ω from the ith gene among n genes in the category and SE(ω_i) is the square root of the variance of ω_i approximated using asymptotic likelihood theory. This weighting scheme minimizes the standard error of the estimate. Error bars in figures 1 and 2 represent the standard error of estimated as

$$SE\,(\bar{\omega}) = \left[\frac{1}{n-1} \sum\nolimits_{i=1}^{n} (\,\omega_i h_i - \frac{\bar{\omega}}{n}\,)^2 \right]^{1/2} \tag{3.6}$$

Notice that this method for calculating the standard error takes into account that the true value of ω may vary among loci within a category. Confidence values for the difference among means are presented as two-tailed p-values calculated using a permutation test based on 10,000 samples. Hypothesis testing for difference in the mean between groups is similarly done by permutation of

estimates from different genes. In the permutation procedure comparing primate and murid SE weighted means under the free branch model, the estimates from different species (e.g. human and chimp or mouse and rat) were treated as independent samples. Lineage specific tests for positive selection were conducted on tissue specific genes using branch-site test II (Yang & Nielsen, 2000) which uses relaxation of constraints as the null hypothesis and is described in Subsection 3.4.3 of this Chapter.

3.7 Natural Selection and Disease

3.7.1 Data Obtention and Preparation for Analyses

3.7.1.1 Mutational Frequency Databases

Cancer-associated mutations for the p53 gene were obtained from the International Agency for Research on Cancer (IARC), containing all somatic and inherited mutations that have been reported in published literature since 1989 which is the most extensive database containing related information. It does not include information on human tissues that have been reported as negative with respect to p53 mutation: *in vitro* induced mutations in cell lines and from other animal models are not included. The complete set of codon mutations and frequencies (18,145 single nucleotide mutations) associated with all the cancer forms contained in release 10 (R10) (Olivier *et al.* , 2002) were used for the p53 analysis.

The number of mutations associated with each specific amino acid residue were also analyzed for all of the proteins in the MeCP2 mutation database (RettBASE†), the immunodeficiency resource (IDR) (Valiaho *et al.* , 2002), and the catalog of somatic mutations in cancer (COSMIC) (Bamford *et al.* , 2004). In order to avoid circularity, p53 data derived from the COSMIC dataset were removed from the analysis. Genes observed in more than one database were replaced by the one containing more information (the highest number of associated mutations). A total of 264 genes were collected from these databases to test for the statistical association of ω with the frequency of mutation observed in disease.

Human disease associated or neutral polymorphic protein variants were obtained from SwissProt database release 48 (Wu *et al.* , 2006). SwissProt classifies protein variants as disease related (i.e., with pathological effects), neutral polymorphism (i.e., with no effect on human health), or unclassified. Only variants with single point mutations were kept for analysis leaving 8,987 sites in the dataset which corresponded to 1,434 protein coding genes.

3.7.1.2 Structural Information

The p53 core domain in complex with DNA (Cho *et al.* , 1994) and the tetramerization domain (Lee *et al.* , 1994) structures were downloaded from PDB (Berman

et al. , 2000) (ID codes: 1TSR, 1PES). Figures showing structures were made using the MOLMOL program (Koradi *et al.* , 1996).

3.7.1.3 DNA Sequences and Orthology Relations

The coding sequences of p53 in 15 vertebrate species were retrieved from GenBank: human (Homo sapiens, X02469), green monkey (Cercopithecus aethiops, X16384), common tree shrew (Tupaia belageri, AF175893), mouse (Mus musculus, AF161020), rat (Rattus norvergicus, X13058), rabbit (Orictolagis cuniculus, X90592), dog (Canis familiaris, AF060514), cat (Felis catus, D26608), sheep (Ovis aries, X81705), pig (Sus scrofa, AF098067), chicken (Gallus gallus, X13057), the pipid frog (Xenopus laevis, M36962), trout (Salmo irideus, M75145), zebrafish (Danio rerio, U60804) and catfish (Ictalurus punctatus, AF074967). Using the orthologous relationships from the Ensembl-Compara database, orthologous sequences for the 264 genes were retrieved from version 35 the Ensembl Database Project (Hubbard *et al.* , 2005) in 11 vertebrate species: human (Homo sapiens v.35_35h); chimpanzee (Pan troglodytes v.35_3); mouse (Mus musculus v.35_34c); (Ratus norvegicus v.35_34e); dog (Canis familiaris v.35_1d); opossum (Monodelphis domestica v.35_2); chicken (Gallus gallus v.35_1k); pipid frog (Xenopus tropicalis v.35_1b); pufferfish (Takifugu rubripes v.35_2g); spotted pufferfish (Tetrahodon nigroviridis v.35_1d) and zebrafish (Danio rerio v.35_5b). Phylogenetic relationships were established according to recent advances in molecular studies on mammals (Springer *et al.* , 2004) and fishes (Chen *et al.* , 2004), providing the following topology: ((((((1, 2), (3, 4)), 5), 6), 7), 8), ((9 ,10),11). Multiple alignments of individual sets of orthologous proteins were done using ClustalW (Thompson *et al.* , 1994) with parameters by default. DNA codon sequences were aligned using ClustalW with translated protein sequences as templates. Codons containing gaps were removed before maximum likelihood estimation of selective pressures.

3.7.2 Site Based Tests of Positive Selection and Estimates of Selective Pressures at the Codon Level

The selective pressures acting at a codon level of proteins were evaluated by means of two alternative approaches: (1) tests of positive selection using codon-based maximum likelihood models (M1a, M2a, M7 and M8) (Anisimova *et al.* , 2001, 2002; Wong *et al.* , 2004); and (2) a test of neutrality using the site-wise likelihood-ratio (SLR) method (Massingham & Goldman, 2005). Under the first approximation it is impractical to use one ω parameter for each site of the protein. The standard approach is to use a statistical distribution to describe the variation of ω among sites on models, assuming for example, a number k of different site classes at a proportion p_k in the sequences, with ω_k values to account for purifying selection ($0<\omega_0<1$), neutral evolution ($\omega_1=1$), and positive selection ($\omega_2>1$). The test of positive selection involves two major steps: first to test whether sites exist where $\omega_2>1$, which is achieved by a

likelihood-ratio test (LRT) comparing a model that does not allow for such sites (neutral models) with a more general model that does (positive selection model); and second, to use the Bayes theorem to compute the posterior probability of sites belonging to specific site classes ω_i defined under an *a priori* distribution.

Two different models of neutral evolution (M1a and M7-beta) and positive selection (M2a and M8) were evaluated by maximum likelihood on the sequences. The nearly neutral model (M1a) assumes the existence of two site classes: conserved sites at which nonsynonymous mutations are nearly deleterious ($0<\omega_0<1$), and completely neutral sites ($\omega_1=1$) at proportions p_0 and p_1. The selection model (M2a) has an additional class ω_2 and applies to a proportion of p_2 of sites, with the constraints that $p_0+p_1+p_2=1$. The proportions p_k of the site class i and the mean value of ω_0 and ω_2 are estimated from the data by maximum likelihood. Model M7 (β) assumes a β-distribution for $0<\omega_0<1$. Model M8 ($\beta+\omega$) adds an extra category to M7, with proportion p_1 of sites with ω_1, while the rest of the sites (at frequencies $p_0=1-p_1$) have ω from the β-distribution between 0 and 1. Positive selection was evaluated by means of the corresponding likelihood-ratio tests, comparing twice the log-likelihood difference between M1a versus M2a, and M7 versus M8 with a $\chi 2$ distribution with two degrees of freedom (Yang *et al.* , 2000; Nielsen & Yang, 1998) Empirical Bayes analysis (Nielsen & Yang, 1998; Yang *et al.* , 2005) calculates the posterior probability that each site is from a particular site class (0, 1, 2) in the models M2a and M8, and gives us the posterior ω values, which represent an estimation of the strength of natural selection acting on each residue. Codon-based likelihood models were implemented in the codeml program of the PAML (3.14a) package (Yang, 1997). By default, codeml provides the posterior ω values for those sites with probable evidence of positive selection ($\omega>1$). In order to obtain all the estimations derived from the empirical Bayes analysis, the codeml source code was modified to print all posterior ω estimates, instead of only those showing a posterior probability of $\omega>1$ greater than the 0.5 cutoff, which was set by default. These posterior ω values (ωM2 or ωM8) denoting purifying selection, positive selection or neutral evolution on residues, were the values used for all proteins evaluated under the codon-based likelihood models M2a and M8.

Aside from these codon-based maximum likelihood model estimations, purifying selection, positive selection and neutral evolution were tested using the sitewise likelihood-ratio method (SLR) (Massingham & Goldman, 2005). This method uses a site by site approach to test for neutrality, but in contrast to similar approaches developed previously (Suzuki & Gojobori, 1999), it uses the entire alignment of the sequence to determine quantities common to all sites, such as evolutionary distances. Using this approach, there is no need to specify a model of how ω varies along the sequence, and it also uses a correction for multiple testing in order to obtain statistical confidence for inferences on deviations from neutrality on each site. SLR estimations (ωSLR) were computed by using the code distributed by the authors (Massingham & Goldman, 2005).

3.7.3 Statistical Analysis

The number of mutations observed on each one of the protein residues were transformed into normalized frequencies using the following equation:

$$f_r^p = \frac{m_r^p}{\sum_p \sum_r m_r^p} \tag{3.7}$$

where f represents the frequency of mutations observed in the residues r of the protein p, while m is the number of mutations observed. The denominator is a normalization factor accounting for all of the mutations observed in all residues for all of the proteins. Frequencies were independently calculated for p53 and human disease proteins.

In order to avoid the statistical influence of genes with a low or high number of observed mutations, genes with less than ten or more than 2000 mutations were discarded from the analysis. Two alternative datasets were analyzed using alignments from human, chimp, mouse and rat (mammals), and the 11 vertebrate species mentioned above. A total of 43 genes were used to test for the statistical association between ω values and the mutation frequency of human disease genes. The K–S test used was that implemented in the R statistical package (Ihaka & Gentleman, 1996).

Part IV

Results

Chapter 4

Test Based Inferences of Natural Selection at a genomic Scale

4.1 Testing the Molecular Clock Hypothesis

The analysis begins with the complete set of 30,709 genes in the Ensembl Human Database version 30.35c. These were filtered to remove all genes that had not been confirmed through mapping to SwissProt, RefSeq, or SPTreEMBL, and a total of 20,469 genes, which in this manner had acquired the Ensembl known gene status, remained. Inspection of ortholog annotations for this set of genes in the Ensembl-Compara database (version 30) yielded 14,185 human genes with ortholog predictions in chimp, mouse, rat, and dog, corresponding to 69% of the known Ensembl human genome. After filtering the sequences by length and exceedingly high evolutionary rates, 13,197 genes were analyzed by means of the relative rates test (RRT). Evolutionary differences in rates between human and chimp were evaluated using the rates of nonsynonymous mutation per nonsynonymous site (Ka), as well as the analogous measure for synonymous rates (Ks). Rate saturation was observed for 959 (7.3%) genes. After the RRT analysis, significant deviations from the molecular clock were observed for 844 (6.4%) human genes and for 1,260 (9.5%) chimp genes. After correcting for multiple testing ($p < 0.05$), the number of genes retained for further statistical analysis were 469 in human and 651 in chimp.

A more detailed analysis showed significant deviations in both Ka and Ks tests for 65 (0.5%) genes, out of which 18 evolved relatively faster in human than in chimp (HF), and 47 evolved relatively faster in chimp than in human (ChF). It is important to note that HF and ChF terms represent relative, rather than absolute, rate definitions. The number of genes for which there were significant differences in either, only Ka, or only Ks, was higher for chimp (477 and 99) than

for human (352 and 83), respectively. The RRT performed showed that a higher number of genes have significantly accelerated in nonsynonymous (938) rather than in synonymous changes (247). The ratio of the number of genes showing an acceleration of nonsynonymous to synonymous rates was similar and more than threefold (approximately 3.8) in both species. This bias constitutes an indirect evidence of the already characterized overdispersed clock in mammals, which suggests that protein evolution cannot be explained by a simple model theory of neutral evolution (Gillespie, 1991; Ohta & Ina, 1995).

Table 4.1 shows the mean values obtained from RRT in the group of genes with significant deviations from the molecular clock hypothesis. They are arranged according to mutational changes (Ka and Ks), three ranges of p-values adjusted for multiple testing, and the two alternative directions of acceleration (HF or ChF).

Acceleration	**Ka**	**dKa**	**rKa**	**Na**	**%**	p_r
HF	0.069	5.0E-04	3.067	57	14.77	Low
	0.064	0.008	3.661[a]	48	12.43	Med
	0.083	0.007	9.711[b]	281	72,80	High
ChF	0.059	7.7E-03	3.077	114	20.65	Low
	0.070	0.012	3.607	63	11.41	Med
	0.101	0.019	9.919[b]	375	67.94	High
Acceleration	**Ks**	**dKs**	**rKs**	**Ns**	**%**	p_r
HF	0.498	0.051	3.18	25	24.75	Low
	0.394	0.031	3.772	21	20.79	Med
	0.426	0.034	5.855	55	54.46	High
ChF	0.523	0.071	3.248	35	23.97	Low
	0.431	0.034	3.748	27	18.49	Med
	0.461	0.078	5.796	84	57.54	High

Table 4.1: Relative Rates Test Results Evolutionary rate of genes with significant deviations from the molecular clock hypothesis. Accelerated human (HF) and chimp (ChF) genes are arranged according to three ranges of significance (p_r) of the relative rates test. p_r: low, medium and high correspond to adjusted p–values in the ranges $0.01 < p \leq 0.05$, $0.001 < p \leq 0.01$, and $p \leq 0.001$, respectively. Ki is the mean evolutionary rate value measured in substitutions per site. dKi (Ki_h-Ki_{ch}) is the mean rate difference between species. rKi (Ki_h/sdi_h-Ki_{ch}/sdi_{ch}) is the mean normalized difference in rates, where sd is the standard deviation. Ni is the number of genes analyzed where i is either s (synonymous) or a (nonsynonymous) respectively. [a]The value was significantly higher than that observed in chimp ($p<0.05$, K-S test). [b]The values were influenced by outliers. If median, instead of mean values were calculated, 8.63 and 8.34 are obtained for ChF and HF, respectively. [c]The value was significantly higher than that observed in human ($p<0.05$, K-S test).

The bulk of all genes fall within the category showing the highest rates of evolution changing by nonsynonymous mutations ($p < 0.001$, pr = high in Table 4.1), suggesting a favorable scenario for the presence of positive selection (PS) in human and in chimp. The Kolmogorov–Smirnov (K–S) test performed on mean normalized differences in rates (rKi in Table 4.1) detected significant differences in the distribution of the medium rKa category, favoring human, and in the low rKs category, favoring chimp ($p < 0.05$). These minor differences were not

Data sets	Parameter	Human	Chimp	R	p
RRT significant	Ka	0.079	0.088	1.11	0.13
	sd	0.088	0.094		
	Ks	0.437	0.470	1.08	0.24
	sd	0.160	0.165		
	Ka/Ks	0.181	0.187	1.03	
Genome*	Ka	0.086	0.087	1.01	0.77
	sd	0.156	0.156		
	Ks	0.430	0.432	1.00	0.83
	sd	0.196	0.196		
	Ka/Ks	0.200	0.201	1.01	

Table 4.2: Evolutionary Rates of Human and Chimp Ka and Ks for human and chimpanzee were estimated using the method of Li with mouse, rat, and dog as a weighted outgroup using RRTree. Results for the set of genes evolving by significant differences in RRT and for the complete genome* are shown. Human and chimp columns hold the mean number of substitutions per site. The standard deviation (sd) is shown beneath each estimate. R is the mean relative rates ratio of chimp to human. p corresponds to the two sample K-S test probability value under Ho: Ki_h-Ki_{ch}=0, where i is either a (nonsynonymous) or s (synonymous), respectively. Genome* represents the full set of 13,197 filtered orthologous gene alignments.

sufficient to produce a net significant difference when comparing the full sets of genes without clock-like behavior between both species.

Table 4.2 shows the mean evolutionary rates estimated for human and for chimp using a topologically weighted outgroup, with mouse, rat, and dog as the reference in two alternative data sets. On the one hand, using only the group of genes showing significant RRT differences, the mean estimation of the human nonsynonymous rate of evolution (Ka = 0.079) was slower than that of chimp (Ka = 0.088), although the difference was not significant (p = 0.13). The same occurred for the synonymous rate change (p = 0.24). The relative evolutionary rate of chimp to human (R on Table 4.2) was 1.11 for Ka and 1.08 for Ks. On the other hand, when considering the full set of filtered orthologous genes, mean rates in substitutions per site were Ka = 0.086 and Ks = 0.430 for human, and Ka = 0.087 and Ks = 0.432 for chimp. Rate differences for Ka and Ks between species were again not significant. The mean Ka/Ks rate was similar between species and was slightly higher for the set of genes representing the complete genome than for those showing significant deviations from clock behavior (0.20 versus 0.18). This is due to the relative increase of the mean Ks rate observed on genes with significant deviations from clock (Table 4.2).

ML estimations of evolutionary rates in the human branch and in the chimp branch were calculated using PAML (Yang, 1997) and compared with those obtained by the Chimpanzee Sequencing and Analysis Consortium (CSAC) in the publication of the chimp genome (Sequencing & Consortium, 2005). While our estimations were slightly faster for human (Ka = 0.0014, Ks = 0.0063 versus CSAC: Ka = 0.0013, Ks = 0.0062) and for chimp (Ka = 0.0015, Ks = 0.0066 versus CSAC: Ka = 0.0012, Ks = 0.0060), they were considerably similar to those obtained by the CSAC using a highly curated set of 7,043 orthologous genes (Se-

quencing & Consortium, 2005). The total number of genes with Ka/Ks > 1 was 445 in human and 539 in chimp, representing 5% and 6% of the total number of genes with a measurable ML estimation of the rates ratio, respectively.

4.2 Testing for Positive Selection and Relaxation of Selective Constraints

The set of genes used for clock testing were also analyzed for signals of positive selection (PS). After discarding those with fewer than three unique base pair differences, 9,674 human–chimp–mouse–rat–dog orthologous sequences remained. This set was then analyzed for signals of PS with Tests I and II, which can be used to distinguish relaxation of selective constraints (RSC) from true events of PS when used in conjunction with each other. That is, it is known that most tests of PS are not able to distinguish real events of positive Darwinian selection from cases of RSC (Zhang, 2004). This is the case with Test I used in this study. As has been previously demonstrated by Zhang *et al.* , 2005, the genes observed exclusively in Test I but not in Test II correspond to likely cases of RSC.

Both tests were performed on human and on chimp lineages, and 146 (1.51%) human and 672 (6.95%) chimp genes were obtained when the more restrictive Test II was considered. After correcting for multiple testing ($p < 0.05$), 108 (1.12%), and 577 (5.96%) genes in human and in chimp remained and were considered as true cases of PS occurring in their respective genomes. Chimp thus shows much more evidence of positive selection than human even after correction for multiple testing. Figure 4.1 shows the distribution of total and common genes observed in both tests for the three lineages analyzed. As expected, the great majority of H-PSG and Ch-PSG shown in Test II were also observed in Test I. After correcting for multiple testing, 216, 793, and 941 genes were detected in Test I for human, for chimp, and for the ancestral lineage, respectively. Only 122 human (1.26%), 245 chimp (2.53%), and 287 ancestral (2.97%) genes were found exclusively in Test I. This exclusive set of genes was considered to have likely undergone RSC and used for further analyses of this process.

4.3 PS and Nonsynonymous Rate Acceleration

It is held that genes showing acceleration in nonsynonymous rate are likely to concentrate cases of PS. However, both sets show overlaps that are surprisingly low given this consideration. Moreover, as we shall observe in Sections 5.1 and 5.2 of Chapter 5 the comparison of the functions that have undergone acceleration and positive selection (Tables 5.1 and 5.2) reveal outstanding differences between most of the functional categories represented under both processes. To understand these and other discrepancies in the number of positives observed in

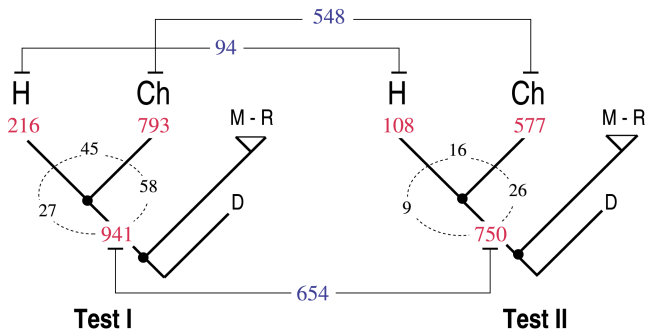


Figure 4.1: Phylogenetic Distribution of Genes Deduced under Tests I and II The differential distribution of genes along tree branches, suggests a different pattern of occurrence of PS (Test II) and RSC (Test I) in derived and ancestral lineages. Numbers in red represent the total number of genes detected in each test exclusively after correcting for multiple testing. Numbers in black are common orthologous genes observed between lineages. Numbers in blue are genes observed in both tests.

Ka rate based approaches and Test II, the relationship between the nonsynonymous rates difference (dKa = Kah−Kach), the mean normalized differences in nonsynonymous rates between the species (rKa = dKa/sd), and the normalized nonsynonymous rate (Ka/Ks), were studied.

Figure 4.2 shows the distribution of rKa versus dKa values for those genes with significant and non significant differences in Ka resulting from the RRT ("molecular clock" in Figure 4.2). Under this distribution, four alternative groups have been labeled: those showing 1) both PS and Ka/Ks > 1 (red circles), 2) PS and Ka/Ks < 1 (blue circles), 3) Ka/Ks > 1 with no evidence of PS (black asterisks), and 4) Ka/Ks < 1 with no evidence of PS (gray circles).

The total number of genes with Ka/Ks > 1 considered in the analysis of Figure 4.2 was 336 in human (437 in chimp), out of which 22 (86) have shown evidence of PS (red circles in Figure 4.2) and only five (18) have shown significant deviations from the molecular clock in Ka rate (circles above the broken line). Alternatively, 58 human (407 chimp) genes with Ka/Ks < 1 were positively selected (blue circles). This shows that 72% of positively selected human genes did not show a Ka/Ks > 1 (82% in chimp). Similarly, 314 (93%) human and 351 (80%) chimp genes showing Ka/Ks > 1 have not shown evidences of PS (black asterisks). Notice that most of these genes have evolved without signs of nonsynonymous deviations from clock behavior, suggesting that these values of Ka/Ks > 1 correspond to variations falling under a neutral model of evolution. The fact that many genes showed evidence of PS under clock-like behavior (red and blue circles below the broken line) highlights the high sensitivity of the branch-site test employed where a few amino acid sites are probably involved in events of PS, without major changes in evolutionary rates between lineages (dKa).

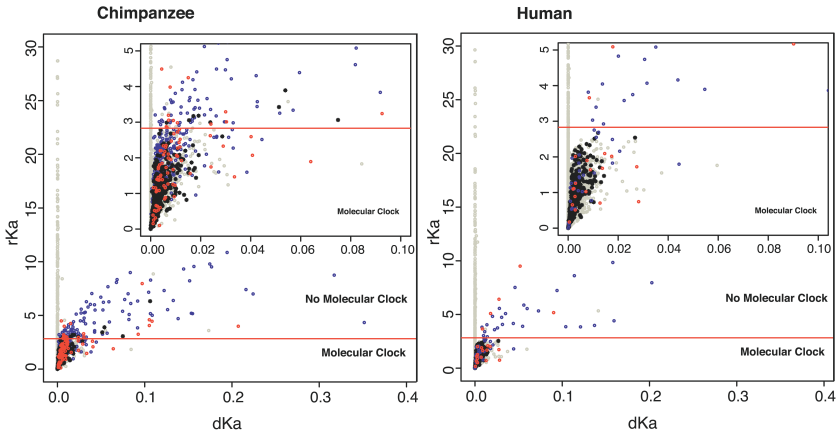


Figure 4.2: Positive Selection and Rates of Evolution A minor proportion of genes with Ka/Ks $>$ 1 match events of PS in human and in chimp (red circles). Many of the genes with Ka/Ks $<$ 1 show evidence of PS (blue circles). Genes with Ka/Ks $>$ 1 without evidence of PS (black asterisks) fall mostly under molecular clock conditions for nonsynonymous changes (circles below the broken red line). Most of the genes without evidence of PS and Ka/Ks $<$ 1 (gray circles) are scattered below the boundary limiting molecular clock like behavior and are observed at dKa $<$ 0.0006 when molecular clock conditions are not fulfilled. Genes outside of clock conditions and dKa $>$ 0.0006 coincide mostly with events of PS in both of the species (red and blue circles above the broken line). dKa and rKa are as defined in Table 1.

In a similar manner, when considering differences in Ka rate instead of Ka/Ks rate ratios, 386 human genes (552 in chimp) have experienced a significant acceleration of nonsynonymous rate, and only approximately 32 of these genes (120 in chimp) have shown a reliable signal of PS. However, when considering genes with a significant acceleration in Ka rate and a dKa $>$ 0.0006, most of them show evidence of PS (81% in human and 94% in chimp). Although it is important to remember that they are still a minority out of all of the genes with a significant deviation in Ka.

In summary, it is found that only those genes with a significant Ka from the RRT and dKa $>$ 0.0006 could possibly be considered as candidates for an enriched probability of having been positively selected. These results serve to highlight one of the downfalls of using elevated normalized Ka rates as a means of concentrating likely cases of PS in an *a priori* fashion.

Chapter 5

Natural Selection and Biological Function

5.1 Functional Analysis of Accelerated Genes in Human and Chimp

Using human Gene Ontology (GO) terms (Ashburner *et al.* , 2000), the focus has been set on seeing if there are any functional differences in the set of genes accelerated within the human genome and in between both lineages. GO terms for chimpanzee were deduced from the corresponding human orthologs.

Table 5.1 shows the main GO terms corresponding to biological processes at GO level 6 associated to human and to chimp genes accelerated in synonymous and nonsynonymous changes. The most significant terms in the analysis of Ka and Ks are shown. The table is arranged according to those terms represented above 5% in the set of human nonsynonymous accelerated genes (column 1). Other terms above 5%, not shown in the table, were indeed observed in other categories. For instance, *cation transport* (6.78%) was observed in the list of genes with coding sequences evolving faster in chimp than in human by means of nonsynonymous changes. Other terms such as *RNA metabolism, DNA metabolism, regulation of protein metabolism, regulation of programmed cell death, protein catabolism,* and *cellular carbohydrate metabolism* correspond to some of the human sequences and the chimp sequences accelerated by synonymous changes above 5%.

To find out if there were any over- or under-represented GO terms in between human and chimp, a Fisher exact test with p-values corrected for multiple testing was run using FatiGO (Al-Shahrour *et al.* , 2004, 2005a). Neither the test applied on HF and ChF genes with coding sequences evolving by means of nonsynonymous changes, nor that for synonymous ones, reported any significant difference for GO-term representation at any level (GO levels 3–6). In conclusion, there are no statistically significant differences in functional GO classes

	Ka			Ks		
Gene Ontology Term	**HF**	**ChF**	**p**	**HF**	**ChF**	**p**
Cellular protein metabolism	34.31 (82)	30.68 (104)	1	21.67 (13)	29.73 (22)	1
Transcription	22.18 (57)	21.83 (79)	1	18.33 (11)	18.92 (14)	1
Regulation of transcription	20.43 (53)	19.17 (74)	1	16.67 (10)	18.92 (14)	1
Regulation of nucleobase ‡	22.18 (53)	22.12 (75)	1	16.67 (10)	18.92 (14)	1
Phosphate metabolism	9.62 (23)	9.14 (31)	1	5.00 (3)	4.05 (3)	1
Cell. macromolecule catabolism	9.62 (23)	7.67 (26)	1	5.00 (3)	8.11 (6)	1
Macromolecule biosynthesis	8.37 (20)	4.72 (16)	1	6.67 (4)	6.76 (5)	1
Protein catabolism	7.53 (18)	6.78 (23)	1	3.33 (2)	8.11 (6)	1
Protein biosynthesis	7.53 (18)	4.72 (16)	1	6.67 (4)	6.76 (5)	1
Apoptosis	7.11 (17)	4.72 (16)	1	1.67 (1)	8.11 (6)	1
Protein transport	6.69 (16)	5.90 (20)	1	5.00 (3)	2.70 (2)	1
DNA metabolism	6.69 (16)	5.01 (17)	1	6.67 (4)	2.70 (2)	1
Immune response	6.28 (15)	6.78 (23)	1	11.67 (7)	10.81 (8)	1
RNA metabolism	6.28 (15)	4.13 (14)	1	1.67 (1)	8.11 (6)	1
Cytoskeleton organization and biogenesis	5.44 (13)	4.72 (16)	1	1.67 (1)	1.35 (1)	1
Total number of genes	386	552		101	146	
Genes with GO terms at level 6	239	339		60	74	
Genes with GO terms at other levels	100	125		20	38	
Genes without GO terms	47	88		21	34	

Table 5.1: Functional Analysis of Genes with Deviations from the Molecular Clock Values in the HF and ChF columns are percentages of the total genes with GO at level 6. The numbers of genes for each term are shown in parenthesis. The list of genes accelerated in human (HF) were used as the query list, while those accelerated in chimp (ChF) were used as the reference for the statistical analysis. Terms were arranged according to a decreasing percent-representation within the subset of HF accelerated in Ka above 5%. ‡ *Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism.*

represented in the sets of the genes without clock-like behavior between the two species. Finally, the hypothesis that accelerated human genes represent an unbiased sample of the human genome in functional terms was tested. Again, no GO terms were found to be significantly over- or under-represented among accelerated human genes when compared with the rest of the genome.

In summary, no GO terms were detected as being differentially distributed between the significantly accelerated genes of human and of chimp. Moreover, the set of functions accelerated in human does not represent a special subset of genes with functional particularities within the human genome.

5.2 Functional Analysis of Positively Selected Genes

Table 5.2 shows the main GO terms associated to the set of PSG detected using Test II in human and in chimp, as well as the difference in representation of GO terms for the sets of genes under PS for both species when compared with their ancestral lineage. As before, terms shown are those represented above 5% in human PSG (H-PSG).

Initially, when comparing representations of terms under human and chimp directly (Table 5.2), it is evident that with minor modifications of frequencies H-PSG show almost the same set of biological functions as those in chimp (Ch-PSG). It is interesting to note that in this comparison the highest differences in representation of genes between both lineages are found under terms such as *G-protein coupled receptor* (GPCR), *sensory perception, electron transport, integrin-mediated signaling pathway, inflammatory response,* and *cellular protein metabolism*, among others. All of these terms were represented to a greater extent in human with the exception of *cellular protein metabolism*, which was higher in chimp. Although the highest differences range from 4% to 15%, they were non significant at any level (GO levels 3–6). Likewise, no term was significantly over- or under-represented in the comparison of H-PSG against the rest of the human genome. However, it is important to note that at least one difference seems evident: only a minor number of orthologous PSG are common between both species (N_o column in Table 5.2). This shows that PS driven evolution of different genes under the same functional classes is the most frequent pattern occurring after speciation. Finally, it is important to note, that while four of the GO categories, each containing more than 50 genes with a significant nonsynonymous rate acceleration (Table 5.1), are within those most highly represented under PS in both species (Table 5.2), the terms *G-protein coupled receptor signaling pathway* and *sensory perception* were absent among those showing a significant acceleration in nonsynonymous rates (see Section 4.3 for a more detailed examination of the differences among PS and RRT results).

A more striking difference becomes noticeable when switching from the perspective of a direct comparison of the functional GO categories under PS for human and for chimp, to that based on the relative differences observed between

	Adaptive Evolution				
Gene Ontology Term	**H**	**Ch**	**N_o**	**H-AH**	**Ch-AH**
Cellular protein metabolism	16.67 (7)	31.00 (102)	3	-10.68	1.66
G-protein coupled receptor †	21.43 (9)	6.08 (20)	0	8.33	-4.56
Sensory perception	16.67 (7)	3.65 (12)	0	9.37	-1.66
Regulation of nucleobase ‡	11.90 (5)	14.29 (47)	0	-3.12	0.49
Transcription	11.90 (5)	15.20 (50)	0	-4.42	0.08
Regulation of transcription	11.90 (5)	13.98 (46)	0	-2.86	0.45
Cellular macromolecule catabolism	9.52 (4)	9.42 (31)	1	2.09	1.27
Immune response	9.52 (4)	4.86 (16)	0	-1.57	-5.31
Protein transport	7.14 (3)	4.86 (16)	1	4.16	0.96
Protein catabolism	7.14 (3)	8.81 (29)	1	1.3	1.99
Intracellular protein transport	7.14 (3)	3.95 (13)	1	5.73	1.64
Cytoskeleton organization and biogenesis	4.76 (2)	4.86 (16)	2	4.68	1.81
Phosphate metabolism	7.14 (3)	7.90 (26)	0	-5.21	-2.06
Cellular carbohydrate metabolism	4.76 (2)	2.74 (9)	1	2.08	-1.24
Resp. to pest, path. & parasite §	7.14 (3)	2.43 (8)	0	-1.82	-4.05
DNA metabolism	4.76 (2)	6.08 (20)	1	0.52	0.59

Table 5.2: Functional Analysis of Genes Under Positive Selection Values in the H and Ch columns are percentages of the total genes with GO at level 6 for human and chimp, respectively. Numbers in parenthesis represent H-PSGs and Ch-PSGs selected exclusively in each lineage. Terms were arranged according to the decreasing percent-representation observed within the subset of H-PSGs above a 5% frequency. Percentages exclude common orthologous genes selected in both lineages (N_o). H-AH and Ch-AH represent the differences in GO representation between derived and ancestral lineages (AH: ancestral hominid lineage). Positive and negative values show an increase or a decrease in relation to the ancestor respectively. Numbers in bold show opposite trends occurring in the derived lineages. Query and reference groups, p, and symbols as in Table 5.1 on page 62. † *G-protein coupled receptor protein signaling pathway.* §*Response to pest, pathogen or parasite.*

the ancestral lineage and each one of the corresponding derived species. The H-AH and Ch-AH columns in Table 5.2 show the difference in representation of GO categories between the derived and ancestral lineages for human and for chimp, respectively. The representation of PSG under *G-protein coupled receptor, sensory perception,* and *cellular carbohydrate metabolism*, increase (+ values) in the human lineage while decreasing (− values) in chimp when compared with the ancestral lineage. In a similar but opposite manner, terms such as *cellular protein metabolism, transcription* and its regulation, *regulation of nucleobase, nucleoside, and nucleotide metabolism*, and *cellular carbohydrate metabolism* show a relative increase in chimp while decreasing in human. From this perspective, differences can be observed that could not be discerned from a direct comparison between derived lineages only: some terms have increased or decreased in relation to the ancestor in both species, others have changed in opposite directions in human and in chimp. The greatest relative differences observed (>10% between H-AH and Ch-AH) in the distribution of functional categories under PS correspond only to three categories: *cellular protein metabolism* which was comparatively favored by natural selection in chimp, and *G-protein coupled receptor singling pathway* and *sensory perception*, comparatively favored in human. Finally, the relative differences observed in the remaining GO categories in Table 5.2 were below 5%.

5.3 Ancestral and Derived Trends of Relaxation and Positive Selection

Figure 5.1 shows the results of the statistical comparisons performed (filled circles) between the representations of genes (numbers on branches) observed under PS and RSC between human, chimp, and the ancestral lineage for four functional GO categories. These categories were among those most represented within both tests, and serve at the same time as examples of the different patterns of differentiation observed between common categories of human and of chimp.

A common pattern observed for all of the functional categories represented in the set of genes under RSC was the absence of functional differentiation between human and chimp (gray-filled circles). However, a highly significant increase (red-filled circles) occurred in the representation of the term *G-protein coupled receptor protein signaling pathway* in the derived lineages in comparison with the ancestral lineage (Figure 5.1). This significant over-representation of genes under RSC was higher for human (+32.68%, $p < 1e\text{-}05$) than for chimp (+18.36%, $p = 0.006$). Considering the time elapsed in each of the branches (approximately 75 Mya in the ancestral lineage against 5 Mya in the evolution of hominids), this suggests that a higher number of genes per unit time have experienced RSC after speciation in both this category and that of *sensory perception* (Figure 5.1). Given that the relative representations of PSG belonging to *G-protein coupled receptor* and *sensory perception* increased in

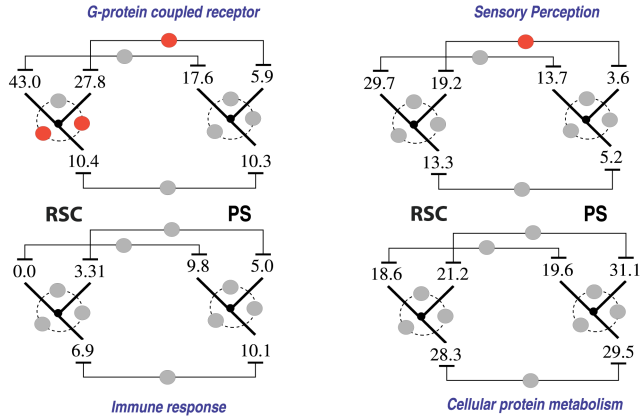


Figure 5.1: Phylogenetic Distribution of Four Representative GO categories under Tests I and II The phylogenetic distribution of four representative GO categories is shown in human, in chimp, and in the ancestral lineage as depicted in the tree defined above. Numbers correspond to the percentage representation of genes under PS and RSC for each term out of the total number of genes with GO annotation. Filled circles show significant (red) and non significant (gray) differences in the comparisons (see text for a detailed explanation).

humans while decreasing in chimp after speciation (Table 5.2, Figure 5.1), it is not surprising that statistically significant differences were only detected in chimp (red-filled circles). Furthermore, *G-protein coupled receptor* and *sensory perception* were statistically over-represented ($p < 1e\text{-}05$) when comparing the set of genes under RSC against the rest of the genes available in our data set as representatives of the human genome. In summary, although both categories have increased in representation in human after speciation, a more frequent process of RSC has occurred under both of these, in both species.

The opposite pattern was observed for the *cellular protein metabolism* category (Figure 5.1). In this case, the representation of genes under RSC decreased after speciation in both species. However, a higher representation of PSG under this category occurs in chimp and is the consequence of a marginal increase relative to the ancestral condition. A more pronounced reduction in the number of genes found under RSC occurred for the *immune response* category. In this case, no genes were observed to be under RSC in human, and considering the relative representation in each lineage, it seems to suggest that human showed little variation and chimp decreased in comparison to the ancestral proportion of PSG, while both species decreased under RSC.

Figure 5.2 shows the evolutionary changes in representations before and after the speciation process for all of the common GO classes deduced under both tests. The difference in representation between human and the ancestral lineage for each functional term (H-AH) is plotted against the difference observed between chimp and the ancestral lineage (CH-AH). Each point represents a functional category, and depending on its location in each one of the quadrants

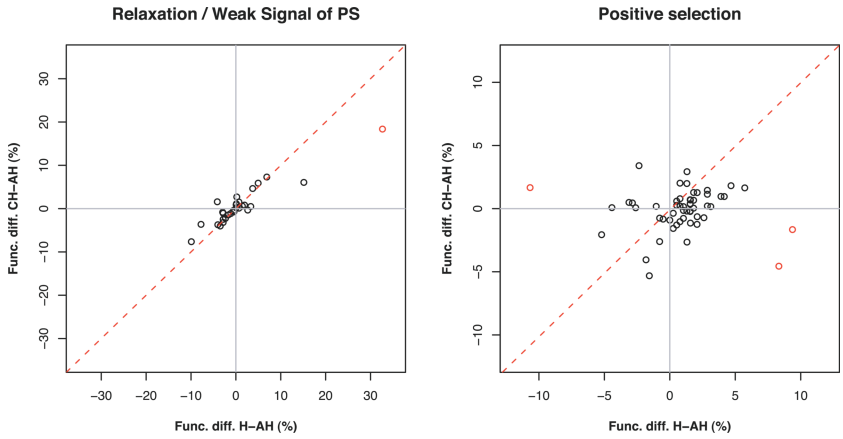


Figure 5.2: Ancestral and Derived Trends in Adaptation and RSC Differences in GO term representation between the sets of the derived and the ancestral lineages (H-AH, human versus ancestral lineage; CH-AH, chimp versus ancestral lineage) are plotted against each other using genes exclusively observed in Test I (RSC) and Test II (PS). Each quadrant represents a particular evolutionary scenario increasing or decreasing in GO representation for each of the lineages after speciation. Terms showing a difference in representation between H-AH and CH-AH >10% were labeled in red: *G-coupled protein receptor* was found in both Test I (14.32%) and Test II (12.89%), and *sensory perception* (11.03%) and *cellular protein metabolism* (−12.34%) in Test II. Only the terms common to all lineages are shown.

(Q) under both graphs, alternative evolutionary scenarios can be deduced. The diagonal represents a homogeneous increase (positive values) or decrease (negative values) in relation to values observed for the ancestral lineage during the evolution of both species.

GO terms with positive differences in representation in both axes correspond to those increasing in both species after the speciation process (Q1). Considering the adaptive evolutionary process, a total of 26 functional categories fit this pattern (PS graph). Most of them (21) showed higher differences in representation in human than in chimp (H-AH%, Ch-AH%), i.e., *synaptic transmission* (1.57, 0.68), *detection of abiotic stimulus* (2.87, 0.21), *intracellular protein transport* (5.73, 1.64), *energy derivation by oxidation of organic components* (3.13, 0.16), and *small GTPase mediated signal transduction* (2.87, 1.14), among others. Another 20 GO terms showed a relative increase in their relative representation in human while decreasing in chimp after speciation (Q4), i.e., *G-protein coupled receptor* and *sensory perception* (differences in Table 5.2 and Figure 5.1), *electron transport* (1.3, −2.65), *male gamete generation* (0.26, −1.57), *blood vessel morphogenesis* (1.04, −0.77) and *wound healing* (1.56, −0.23), among others. The opposite process, favoring the relative increase of PSG in chimp while decreasing in human, was detected for seven GO terms (Q3): *apoptosis* (−2.61, 0.07), *transcription* (−4.42, 0.08), *regulation of transcription* (−2.68, 0.45), and *cellular protein metabolism* (differences in Table 5.2 and Figure 5.1), among others. Finally, a relative decrease from the ancestral representation of PSG was

observed in six GO categories for both species (Q3): *inflammatory response* (−0.78, −2.61), *response to pest, pathogens, and parasites* (−1.82, −4.05), and *immune response* (differences in Table 5.2 and Figure 5.1), among others.

In summary, although Test II detected a higher number of PSG in chimp than in human, and GO term representations between them were not significant, the comparison between ancestral and derived adaptive trends show that out of a total of 59 common GO terms to all lineages, 41 showed a higher proportion of PS events occurring in the human lineage. Only 11 terms showed a higher proportion of PSG in chimp. Additionally, the difference in data distributions between the sets of RSC and that of PS, suggested by Figure 5.2, is persuasive. While differences in the percentage of GO terms are widely distributed between the species, variations in GO representation of genes under RSC are highly correlated between variables (p = 3.6e-15) and fall mostly along the diagonal. The pattern describes a regular increase and decrease of genes undergoing RSC under each GO category at proportional and similar rates in both species after the speciation process. Only two of the GO terms deviated from this general pattern; *G-protein coupled receptor* and *sensory perception* were both located in Q1 below the diagonal, and serve to highlight the high proportion of genes under these categories that are likely cases of RSC in both species.

It is worth noting that the fact that many of the genes found exclusively in Test I have functionally important products, such as homeobox- and polymerase-related proteins among others, seems to suggest that it is highly improbable that all of them have undergone a process of RSC. Probably many of them are genes with a weak yet true signal of PS not sufficient to be detected by Test II (Zhang *et al.* , 2005). It is evident that further statistical methods are necessary to accurately differentiate weak signals of PS from real cases of RSC.

5.4 Functional Roles of PSG in Human and in Chimp

Tables 5.3 and 5.4 show the gene name of some of the genes deduced under PS and RSC belonging to a select few of the more representative GO categories observed in the analyses. In agreement with the estimations based on an acceleration-rate approach (Sequencing & Consortium, 2005), many of the selection events associated to sensory perception in human and in chimp were detected in different genes related to auditory perception. For instance, EDN3 was positively selected in human and is related to sensorineural deafness and hypopigmentation (Aoki *et al.* , 2005). USH1 was positively selected in chimp, and its loss of function produces the most severe form of the Usher's syndrome (Adato *et al.* , 2005). However, PS on genes related to the perception of sound was also found in the ancestral lineage. For instance, the KPTN murine ortholog is a candidate gene for the Nijmegan waltzer mouse mutant, which has vestibular defects and a variable sensorineural hearing loss (Bearer *et al.* , 2000). Other genes related to sensory perception were also found under PS: taste perception

was principally observed in human and the ancestral lineage, visual perception and olfactory receptor genes were found in all of the lineages. Nevertheless, as was previously suggested (Sequencing & Consortium, 2005; Zhang, 2004), most of the events of RSC found under the sensory perception category involved olfactory receptors. RSC in olfactory receptors was abundant in all three lineages. One striking observation was the high number of genes related to visual perception under RSC in the ancestral lineage of hominids. Although further research on this group of genes would be required, the observation could make sense considering the functional change produced by the loss of the nocturnal way of life in higher primates (Wang *et al.* , 2004).

GO Description	Adaptive Evolution		
	Human	Chimpanzee	Ancestral
Sensory Perception	EDN3 GRM6 HKR3 OR2A14 OR51D1 OR5D18 TS1R1	ABCA4 COL1A1 ERCC8 GJA3 MYH9 MYO9A OR52N1OR5I1 ROBO1 USH1C	CCL1 CCL3 COL11A1 CRB2 DSPP GPRC5D IL8RA KPTN MYH14 O10D4 OR10T2OR1B6 OR5A1 OR5P3 PROM1 RP1 TAS1R3 TAS2R38 TAS2R41 TRPA1
G-PCR signaling pathway †	GPR111 HKR3 OR52W1 PTGER4 TS1R1	ADRA1B ADRA2A SORCS1 GRPR AKAP12 TAAR1 TAAR6 PARD3 HTR5A EDG8 OR8D2	AAR5 ADRA1A AS1R3 CCL3 CCR2 CD3 EDRD2 ENPP2 GABBR1 GALR2 GAP43 GLP1R GPCR116 GPR154 GPR43 HTR1D IL8RA MRGPRD OR13A1 PLCE1 PTHR1 RAI3
Immune response ξ	CCL4 ITGAL ITGB1	AFP AMBP CSF1R CSF2RB GABBR1 HFE HLA-G HLA-J IGSF2 IKBKE IL1F10 IL1R1 KLF6 OTUB1 SEMA7A STAT5A TCF7 UBD	AHSG ARTS1 AZGP1 CCL1 CCL3 CCR2 CD72 CD80 CRISP3 CSF3 D3E EXOSC9 FCGR2B FCN2 FTH1 GBP1 GBP2 HLA-DOB HLA-F HLA-G HLA-H ICOS IL18 ITGAL LIRA4 LTB4R LY75 NFX1 S100A9 STAT3 TLR1 TREM1 TRIM22
DNA RNA ¥	CHTF18 NASP	ARID1A ERCC8 FANCG LIG1 MSH4 MUS81 MYST3 POLD3 POLI RAD23B RFC1 SUPT6H TOPBP1 TP73 UBE1 UV39H2 XRCC4	ARSL CASKIN1 CIDEA DCP2DHX15 ELAVL4 EXOSC9 HILS1 NEIL2 NF1B NFIC NFX1 NOLC1 OGG1 PARP2 POLE2 POLG POLM POLN RAD51L3
Trans–cription *	ARNT2 KLF14 NFKB2	ASCL1 CDC5L CEPPZ COA5 EDF1 ERCC8 FLI1 GLI1 HUWE1 KCNH5 KLF12 KLF6 MYEF2 MYST3 NPAS1 PHF15 PHF20 POLR1B POLR2A POLR3B PRDM1 RFC1 RRAGC SNAPC1 ST18 STAT5A SUPT6H TCF7	ATF4 BGALP BLZF1 CD80 CNOT4 DMRTC2 EGR4 ERG ETV2 MEF2B MXD4 NANOG NF1B NFIC NFX1 PHF20 POLR1A POLR2J RBM9 RELA SALF SHPRH SIN3A SOX15 STAT3 TBX6 TF2AY TLE4 TRIM22 TF2AY TLE4 TRIM22 TULP4 ZNF317

continued on next page...

GO Description	Adaptive Evolution		
	Human	Chimpanzee	Ancestral
Cellular protein metabolism	BMP2K GZMB MAPK8 RRBP1 SRCUSP44	ACE ACR ADAM11 ATE1 BIRC4 CAPN6 CASP6 CASP8 CDC2L5 CIT CUL1 DAG1 DNAJC11 EGFR FYN HECW1 HFE HSPCB HUWE1 MAP2 MKNK1 MNAB MTMR1 MYPEP MYST3 NEO1 NEURL NLN PRSS3 PSMB5 RAD23B SRPK1 TFRC TOR1A USP40 USP48	A2M ALG1 APOE ARTS1 BMP2K CAMK2D CDC42BPB CIT CNOT4 COG8 CPB2 CTSB DAPK1 DMPK ELA2A ERG ERO1L HARSL HPT HTRA2 LRP8 NEDD4 NEK11 PARP2 PEN2 PIGQ REN RIPK3 SLMAP STK16 STK3 TGM1 TLR1 USP49 VPS11

Table 5.3: A Sample of Some the Human and Chimp Genes Deduced Under Positive Selection (Test II) Genes deduced to have undergone postive selection in specific lineages are shown grouped by functional categories of Gene Ontology. Immune response ξ, includes *humoral immune response*, and *response to pest pathogens and parasites*. DNA/RNA¥ includes *RNA and DNA metabolism, DNA repair* and *regulation of DNA metabolism*. Transcription* includes its regulation.

GO Description	Relaxation of Selective Constraints / Weak Signal of PS		
	Human	Chimpanzee	Ancestral
Sensory Perception	DFNB31 O2AG1 OR1072 OR10K1 OR10Q1 OR2B6 OR3A3 OR4C16 OR51G1 OR52E5 OR52N1 OR5J2 OR5T1 OR6K2 OR8J3	CNGA2 GUCY2D O10D4 OR13G1 OR2A12 OR2B2 OR2F2 OR2L2 OR4C13 OR4FE OR4K13 OR51B4 OR52E2 OR6P1 OR8I2 OR9A4 PCDH15 TAS2R60 TECTA	BBS2 CNGA1 COL1A2 GUCY2D OPA3 OPN1SW OR11L1 OR2F2 OR4A16 OR4C16 OR4E2 OR51B4OR51I1 OR51V1 OR571 OR5AS1 OR8J1 OR8J3 TNFRSF11A TRPM8 TULP2
G-PCR signaling pathway †	CAP1 IMPG2 OR10T2 OR3A3 Q8NG2 Q8NGU0 Q8NH71 Q8NH88 RBP3 TSHR	ADRA1D CALM1ECE2 ELSR2 OR2B2 OR2T4 OR4C11 OR52M1 OR6K6 PLCE1 R4C13 RAMP3	OR4C15 OR4E2 OR5T2 PYY TSHR
Immune response ξ		CRIP1 ELF4 IVNS1ABP ODZ1 PARP4 STAB1	CAMP CFH C1QG GSR INHA PRF1 PTGS2
DNA RNA ¥	EPRS SUPT6H	ADRA1D CHD5 CHTF18 DARS IVNS1ABP NAP1L5 ORCL3 PAPR4 POLRMT SHPRH SUPT5H	ADARB2 MCM3AP MRE11A MSH2 POLQ POLRMUT RPUSD4 SMN1 CYCS SUPT6H SYV XRCC5
Trans–cription *	AGGF1 CREM FOXI1 ZFP37 ZNF76 POLR3K SUPT6H TITF1 GLIS3 GRIP1 LHX1 MYEF2	CHD5 E2F1 ELF4 GMEB2 HNF4A HOXA1 HOXA3 KLF3 NOC3L PHF19 POLR1A POLRMT SALL2 SIX1 SSRP1 TEAD2 VGLL4	CEBPZ ETS2 PER2 PMFBP1 POLRMT SOX30 SP110 SUPT6H TERF2IP TRIB3 TSC22D4 VCPIP1 YBOX2 ZBTB3
Cellular protein metabolism	DAG1 EPHB6 EPRS GPA11 USP2 USP47 ZNF294 RPL11 TMPRSS2 TRIM50C LAMC1 RANBP2 RNF40	CHD5 COG2 DARS ECE2 GUCY2D HSF1 LPAMARK1 MYLK2 PARP3 PARP4 PTPRC RNF141 RPL31 RPS27 SCPEP1 SHH TIPARP TPSD1 TRIM8 TSTA3	ADAMTS1 BHMT CCRK CTSW DK4 EEF1GF13A1 FGG GALNT6 GRPEL2 KLK15 KLK19 LGMN LMLN LRP2 MMP2 MRE11A MRPL3 MRPS30 OPN1SW PAPPA RNF8 SLC7A11 SMN1 SPTA1 TBB6 TGFBR2 TRIB3 TRIM37 TRPM6 TSSK1

Table 5.4: A Sample of Some the Human and Chimp Genes as Likely Cases of Relaxation of Selective Constraints (Test I) Genes deduced exclusively in Test I in specific lineages are shown grouped by functional categories of Gene Ontology. Symbols as in Table 5.3.

Many other genes with a strong signal of PS in human (H), in chimp (Ch), in human and chimp (H-Ch), and in the ancestral lineage of hominids (AH) were related to: a) nervous system, H: ARNT2 (Stolze *et al.* , 2002), H: GFRalpha-3 (Onochie *et al.* , 2000), Ch: DRP2 (Wrabetz & Feltri, 2001), NES (Ernst & Christie, 2006); b) immune response, H: PTGER4 (Kabashima *et al.* , 2003), CCL4 (Del Cornò *et al.* , 2005), Ch: AFP (Ritter *et al.* , 2004), HLA-G (Sargent, 2005), H-Ch: IGHG3 (Dard *et al.* , 2001), AH: HLA-DOB (Naruse *et al.* , 2002); c) cell cycle, H: VEGFC (Karkkainen *et al.* , 2004), Ch: CCNE2 (Möröy & Geisen, 2004), AH: EXT2 (Pedrini *et al.* , 2005), SEPTIN8 (Nagata *et al.* , 2004); d) metabolism of xenobiotics, H: ARNT2 (Xu *et al.* , 2005); Ch: AKR1C1 (Ciaccio *et al.* , 1994), AH: ABCB4 (Van der Bliek *et al.* , 1987); e) epidermis development, H: KRA58 (Perez *et al.* , 1999), Ch: KRT10 (Zimek & Weber, 2005), COL7A1 (Chen *et al.* , 2002), AH: TGM5 (Cassidy *et al.* , 2005), KTR2A (Mahler *et al.* , 2004); f) inflammatory response, H: ITGAL (Lu *et al.* , 2002), CCL4 (Del Cornò *et al.* , 2005), Ch: IL1F10 (Nicklin *et al.* , 2002), IL1R1 (Tseng *et al.* , 2006), AH: CCL3, CCL1, CCR2 (Sebastiani *et al.* , 2001); g) bone morphogenesis, H-CH-AH: BMP2K (Kearns *et al.* , 2001), Ch: COL1A1 (Pochampally *et al.* , 2005), DCN (Goldberg *et al.* , 2005), AH: BGLAP (Raymond *et al.* , 1999), AHSG (Rittenberg *et al.* , 2005); h) learning and memory, Ch: FYN (Yamada & Nabeshima, 2004), GRIN2A (Adams *et al.* , 2004), AH: APOE (Pfankuch *et al.* , 2005) i) thyroid regulation, Ch: SLC5A5 (Smanik *et al.* , 1997), JMJD1C (Lee *et al.* , 1995); AH: CGA (Vamvakopoulos *et al.* , 1980), PTHR1 (Schipani *et al.* , 1993); and j) reproduction, Ch: CGA (Amato *et al.* , 2002).

These functions are a small sample of those observed in this study and point out the great variety of functions modified by natural selection during hominid evolution.

5.5 Distribution of Functional Classes by Evidence of PS

The most widely used approach when studying the functional implications of g deduced by tests of neutrality involves a threshold based approach, as that used in the previous section, which can be described as a two step approach. However two important criticisms should be taken into account. The first is that all model based approaches are r to their approximations –the model assumptions. Therefore, threshold based inference at a genomic scale may produce more false positives than a rank based approach, considering instead, those cases that show

the most striking deviations among all loci (Hughes, 2007). Second, the use of a two step approach where PSGs are first deduced by testing loci repeatedly for significant deviations, and are then tested repeatedly again to recover possible association with functional classes, is inefficient. It is is inefficient, primarily because it requires a necessary statistical adjustment for multiple testing at each step. While, many studies have skipped this unavoidable consideration and have reported that PSGs show significant associations with certain functional categories, the results in the previous section demonstrate that when proper corrections are implemented, the evidence is lacking.

The Fatiscan program (Al-Shahrour *et al.* , 2005a, 2007), implements a partition test (see Materials and Methods) which searches for functional classes that show significantly asymmetric distributions towards the extremes of a list of ranked genes considering p-values corrected for multiple testing. All genes that were tested for positive selection in in Chapter 4 were ranked by the log base 2 transformed 2Δ statistic derived from PS test II (see Equation 3.4). As such, genes were ranked by the amount of evidence of positive selection that they show. Sixty partitions were used for the analysis. Table 5.5, shows the functional classes deduced as being significantly biased towards or away from either extremes of the list showing higher or lower evidence of PS. The table is ordered by GO level (lvL, levels 3-6 are shown) and the FDR-adjusted p-values corresponding to the most significant partition for each term are provided. Only terms that had at least 20 genes were considered.

The results show a markedly different scenario than those of the previous section. In total 26 categories in human and 54 in chimp show significant biases. In human, 21 categories showed over-representation towards higher evidence of PS, while 5 showed under-representation in low values of PS; both of which are patterns that show more evidence of positive selection among functional categories. In chimp, 39 where over-represented in higher values of PS, while 15 where under-represented in higher values of PS; showing patterns of association to evidence of PS and away from evidence of PS, respectively. In comparison, 15 categories were common to both species. 12 of them are over-represented towards higher evidence of PS (OR in higher PS) and include terms associated with: reproduction, cellular organization and adhesion, development, and anatomy, among others. The remaining three, all related to sensory perception, showed opposing patterns being both under-represented in the opposite extremes of the list (human shows a shift away from lower evidence of PS, and chimp tends towards a higher evidence of PS). While an under representation at either extreme is harder to interpret since functional classes expand the range of values except at either extreme, it is worth noting that categories deduced in other studies and showing high, although not significant, representations in the previous section are observed here as being under represented in low values of PS in human with the highest p-values among categories: *G-protein coupled receptor signaling pathway* and *sensory perception*. The categories showing the most biased distributions towards higher values of PS ($p<0.01$) in human are: *cellular component organization and biogenesis, membrane lipid metabolic*

process, protein kinase cascade, actin filament-based process, organelle organization and biogenesis, cell adhesion, and *multicellular organismal development.* In chimp they are: *system development, positive regulation of signal transduction, regulation of signal transduction, anatomical structure development, multicellular organismal development, regulation of biological process, organ morphogenesis, establishment of localization, protein kinase cascade, cellular component organization and biogenesis, cell adhesion, actin filament-based process, organ development, nervous system development, regulation of a molecular function, transport, anatomical structure morphogenesis,* and *negative regulation of cellular process.*

Result	Lvl	GO Term	L1	L2	G1	G2	*p*-value
		Human					
OR in high PS	3	cellular component organization and biogenesis	1428	2652	55	331	5.7E-04
OR in high PS	3	cell adhesion	1496	2584	22	103	6.2E-03
OR in high PS	3	multicellular organismal development	1224	2856	35	313	7.5E-03
OR in high PS	3	cell cycle	1768	2312	40	95	2.1E-02
OR in high PS	3	anatomical structure development	1224	2856	33	288	2.1E-02
OR in high PS	3	regulation of biological process	1224	2856	54	554	2.1E-02
OR in high PS	4	organelle organization and biogenesis	1428	2652	26	150	4.8E-03
OR in high PS	4	locomotory behavior	1700	2380	11	17	1.9E-02
OR in high PS	4	regulation of multicellular organismal process	1292	2788	10	40	2.1E-02
OR in high PS	4	embryonic development	476	3604	5	48	2.1E-02
OR in high PS	4	system development	1224	2856	25	222	3.2E-02
OR in high PS	4	reproductive process in a multicellular organism	680	3400	3	21	3.4E-02
OR in high PS	4	cell-cell adhesion	544	3536	5	45	3.9E-02
UR in low PS	4	sensory perception	2108	1972	109	63	4.6E-02
UR in low PS	5	sensory perception of chemical stimulus	2108	1972	82	28	4.7E-06
UR in low PS	5	cell surface receptor linked signal transduction	2244	1836	213	106	1.7E-03
OR in high PS	5	cytoskeleton organization and biogenesis	1428	2652	13	70	1.5E-02
OR in high PS	5	embryonic development ending in birth or egg hatching	340	3740	3	17	3.2E-02
OR in high PS	5	immune system development	1224	2856	5	25	4.0E-02
OR in high PS	5	organ development	1224	2856	20	168	4.2E-02
UR in low PS	6	sensory perception of smell	2108	1972	78	27	1.1E-05
UR in low PS	6	G-protein coupled receptor protein signaling pathway	2176	1904	147	71	2.8E-04
OR in high PS	6	membrane lipid metabolic process	2108	1972	4	20	1.4E-03
OR in high PS	6	protein kinase cascade	2176	1904	19	42	1.7E-03
OR in high PS	6	actin filament-based process	1496	2584	7	33	3.7E-03
OR in high PS	6	hemopoietic or lymphoid organ development	1224	2856	5	25	3.4E-02
		Chimp					
OR in high PS	3	anatomical structure development	1404	3276	125	275	1.3E-05
OR in high PS	3	multicellular organismal development	1404	3276	134	309	7.9E-05
OR in high PS	3	regulation of biological process	1404	3276	209	505	8.7E-05
OR in high PS	3	establishment of localization	1248	3432	112	321	6.1E-04
UR in high PS	3	response to biotic stimulus	1170	3510	1	47	1.1E-03

continued on next page ...

Result	Lvl	GO Term	L1	L2	G1	G2	*p*-value
OR in high PS	3	cellular component organization and biogenesis	1170	3510	101	359	1.1E-03
OR in high PS	3	cell adhesion	1248	3432	45	111	1.4E-03
UR in high PS	3	defense response	1326	3354	12	77	3.0E-03
OR in high PS	3	regulation of a molecular function	1404	3276	27	58	3.2E-03
UR in high PS	3	biosynthetic process	936	3744	13	191	1.1E-02
OR in high PS	3	protein localization	1170	3510	31	83	1.9E-02
OR in high PS	3	cellular developmental process	1170	3510	84	292	2.3E-02
OR in high PS	3	response to external stimulus	2574	2106	72	45	2.6E-02
OR in high PS	3	reproductive process	1170	3510	15	49	2.8E-02
UR in high PS	3	response to stress	1560	3120	47	143	3.1E-02
OR in high PS	3	death	468	4212	7	124	3.1E-02
OR in high PS	3	muscle contraction	1482	3198	14	25	3.2E-02
OR in high PS	4	system development	1404	3276	110	221	7.1E-07
OR in high PS	4	transport	1248	3432	110	312	4.2E-03
OR in high PS	4	anatomical structure morphogenesis	1404	3276	71	146	6.7E-03
OR in high PS	4	positive regulation of biological process	1482	3198	69	150	1.2E-02
UR in high PS	4	generation of precursor metabolites and energy	1170	3510	18	120	1.3E-02
UR in high PS	4	response to other organism	1170	3510	1	40	1.3E-02
UR in high PS	4	response to wounding	1638	3042	15	64	1.4E-02
OR in high PS	4	regulation of cellular process	1170	3510	136	495	1.5E-02
UR in high PS	4	sensory perception	1170	3510	14	154	1.8E-02
UR in high PS	4	cellular biosynthetic process	936	3744	10	172	2.0E-02
OR in high PS	4	negative regulation of biological process	858	3822	27	178	2.1E-02
OR in high PS	4	reproductive process in a multicellular organism	1170	3510	9	25	2.9E-02
UR in high PS	4	organic acid metabolic process	1170	3510	12	95	2.9E-02
OR in high PS	4	regulation of catalytic activity	1404	3276	25	56	3.0E-02
OR in high PS	5	regulation of signal transduction	546	4134	10	74	1.2E-05
UR in high PS	5	sensory perception of chemical stimulus	1248	3432	4	98	1.5E-04
OR in high PS	5	organ development	1404	3276	80	169	2.1E-03
OR in high PS	5	nervous system development	1014	3666	28	99	2.5E-03
OR in high PS	5	negative regulation of cellular process	858	3822	27	164	9.4E-03
UR in high PS	5	inflammatory response	1638	3042	10	49	1.5E-02
OR in high PS	5	cytoskeleton organization and biogenesis	1170	3510	27	71	1.6E-02
OR in high PS	5	cell development	1404	3276	93	188	1.9E-02
OR in high PS	5	intracellular signaling cascade	390	4290	12	234	2.0E-02
OR in high PS	5	positive regulation of enzyme activity	1404	3276	13	35	3.6E-02
OR in high PS	5	embryonic development ending in birth or egg hatching	1092	3588	9	13	3.6E-02
OR in high PS	5	transmission of nerve impulse	1326	3354	21	31	4.4E-02
UR in high PS	5	carboxylic acid metabolic process	1170	3510	12	95	4.4E-02
OR in high PS	5	reproductive structure development	1170	3510	9	13	4.4E-02
OR in high PS	6	positive regulation of signal transduction	390	4290	3	20	8.7E-06
OR in high PS	6	organ morphogenesis	1638	3042	36	45	2.4E-04
OR in high PS	6	protein kinase cascade	390	4290	4	68	6.4E-04
UR in high PS	6	sensory perception of smell	1248	3432	4	93	1.0E-03
OR in high PS	6	actin filament-based process	1170	3510	18	33	2.0E-03
OR in high PS	6	cellular morphogenesis during differentiation	1170	3510	9	15	1.0E-02
OR in high PS	6	vascular development	1326	3354	15	25	2.1E-02
UR in high PS	6	glycoprotein metabolic process	1872	2808	4	24	2.4E-02
OR in high PS	6	chordate embryonic development	1092	3588	9	13	3.8E-02

Table 5.5: Functional Classes Showing Biased Distributions According to Evidence of PS
Biological function associated GO terms in levels 3-6 (Lvl) showing significant biases towards (OR under Result) or away (UR under Result) from lower (low PS under Result) or higher (high PS under Result) values when ordered by the 2ΔLRT statistic after correction for multiple testing (p-value) are shown for human and chimp. Only those terms that had at least 20 genes were considered. The total number of genes in each partition (L1 and L2) together with the corresponding genes under the given GO category in each partition (G1, G2) used for the Fisher exact test implemented in Fatiscan are provided.

Chapter 6

Natural Selection at the Organ System Level

6.1 Determination of Tissue Specific Genes

In order to select a robust set of tissue specific genes (TSGs) and to evaluate the possible dependence of results on the methods and sets used to draw observations on the evolutionary patterns of TSGs, 3 databases with sensible definitions of tissue specificity were used. The first, which we shall use as our main point of reference, contained 840 tissue specific genes (TSGs) belonging to 7 different human tissues and 859 housekeeping (HK) genes. Both definitions were deduced from available MAS5 absence / presence calls from the human-specific HG_U133A/GNF1H Tissue Atlas Dataset from the Genomics Institute of the Novartis Research Foundation (GNF hereafter) (Su *et al.* , 2004). A second set was derived from GeneNote (Shmueli *et al.* , 2003) which is based on replicate analyses of Affymetrix GeneChip HG-U95 on 12 healthy human tissues. Here tissue specificity is determined by analyzing genes highly expressed in one tissue and not others through a normalized expression levels approach developed by the authors. The third set was obtained from the human-specific data set of the TissueDistributionsDB database (TDDB hereafter) (Jonnakuty *et al.* , 2006), a repository of tissue distribution profiles based on Expressed Sequence Tags (ESTs) from UniGene (Wheeler *et al.* , 2006). In this last database both the levels of expression and relative percentage of expression in different groups of tissues are provided to asses tissue specificity (see Materials and Methods). Only categories with more than 15 tissue specific genes were considered for further analyses.

An initial inspection of the tissue categories resulting from the GNF, GeneNote, and TDDB databases showed that they were largely different (see Table 6.3 at the end of this chapter for a list). In Figure 6.1, the overlap among genes contained for two categories present in all of the databases is shown. The *brain*

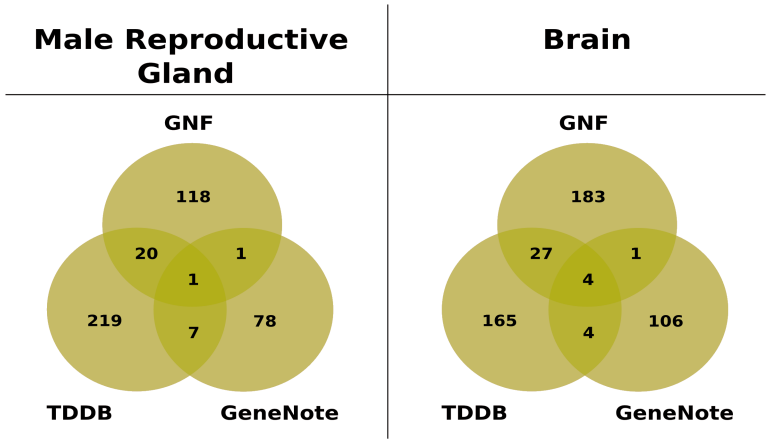


Figure 6.1: Overlap among TSG Categories from Different Databases The number of genes found exclusively and overlapping under categories from the 3 databases used. *Male Reproductive Gland*, is composed from *testis* TSGs in GNF, *prostate* genes in GeneNote, and includes both of these in TDBB. *Brain* is the exact same category in all three databases. Whether similar or exactly the same categories are examined, genes deduced as tissue specific from the different methods used to detect expression and define tissue specificity seem to provide sets with strikingly low overlap. Thus highlighting the importance of a methodologically broad perspective when hoping to draw conclusions about tissue specific genes in general.

category was found as such in each of the databases. The *male reproductive gland* category is composed of both testis and prostate gland tissues, in the GNF and GeneNote sets respectively, and from a combination of both in TDDB. While on the one hand, testis tissue was not analyzed in GeneNote, this lack of overlap among categories is also the result of *prostate* having too few members in the GNF set. This disagreement, even before comparing the identity of genes within categories found in all three databases, provides certain indication that different methods and or experimental procedures should be taken into account when considering a definition of tissue specificity. When the overlap of both GeneNote and GNF with GeneNote is examined for the specific genes deduced in comparison with the *male reproductive gland* category in TDDB, or when observing the overlap among genes in all sets for the *brain* category, the severity of this situation becomes much more evident. Figure 6.1, shows that the datasets themselves barely overlap. Thus the three largely independent sets derived from each of the databases are used for this study.

6.2 Differences Among Primates and Murids

All of the genes that had annotations among the three databases used for determination of tissue specificity, that also had orthologs in all human, chimp, mouse, rat, and dog in the later Ensembl Database v42, were analyzed following the schema used in Chapter 4 (with some modifications, see Materials

and Methods). Evolutionary estimates of the rates of nonsynonymous substitutions to nonsynonymous sites divided by the that for synonymous ones (ω) were derived independently along each lineage from the orthologous sequence alignments for each gene through the ML free branch model in PAML (Yang, 1997). Estimates of ω for each gene of each lineage were then reported as the standard error weighted mean (SEWM, Equation 3.5) for all of the genes belonging to each category. Throughout, results are presented for the analyses conducted on the GNF data set, and specific differences among observations are compared and contrasted with results from TDDB and GeneNote where appropriate.

Figure 6.2A, shows that in accordance with previous results (Ohta, 1993) the value of ω is generally larger in primates than in murids. The differences among these groups are seen for all categories except *pancreas*. This general pattern is also observed throughout all of the databases examined (see Table 6.3 at the end of this Chapter). Similarly, but in direct opposition to the results observed for the set of HK genes analyzed in Dorus *et al.* 2004, HK genes here, also show a strongly significant difference in rates between primates and murids ($p<<0.001$). In order to understand the possible reasons for this discrepancy, two additional, previously published, categorizations of HK genes were evaluated: Velculescu *et al.* 1999 (HK-V) and Eisenberg & Levanon, 2003 (HK-EL), together with those from the analysis of the TissuesDistribution Database (HK-TDDB), as well as those used in Dorus *et al.* 2004 (HK-D), were evaluated. The HK-V, HK-EL, HK-TDDB and the HK-D data sets contain 121, 82, 551, and 55 genes, respectively. In all three of the new data sets (HK-V, HK-EL, HK-TDDB), a significant difference was observed when comparing primates and murids ($p<0.001$) using permutation tests to evaluate the difference between SEWM ω. It is only in the relatively small data set used by Dorus *et al.* 2004 in which no evidence for a larger value of in HK genes in primates than in murids was found. When genomic estimates were drawn from the full set of more than 16,400 genes with orthologs in all of the species available from this version of the Ensembl Database, differences among primates and murids were again highly significant ($p << 0.001$ under SEWM free branch ML for the Genome in Table 6.1). Specifically, genomic SEWM ω was 0.130, 0.145, 0.070, and 0.074 for human, chimp, mouse, and rat respectively. While it is likely that this overall variation in rates observed independent of groups and across genomes may be due in part to stronger purifying selection owing to the larger effective population sizes in murids (Ohta, 1993), what it suggests in general, is that independent from the causes, lineage effects exists that must be taken into account when comparing rates of evolution among these lineages. Indeed, other authors working at the same time in this question proposed that dividing lineage specific estimates by those of their respective genomes is a necessary consideration when comparing among groups of genes from different lineages (Wang *et al.* , 2007).

Figure 6.2 B shows the rates of lineages for each of the categories as in Figure 6.2 A, but normalized by dividing over the genomic estimate for each lineage (ω^*) . After applying this correction for lineage effects, only two categories

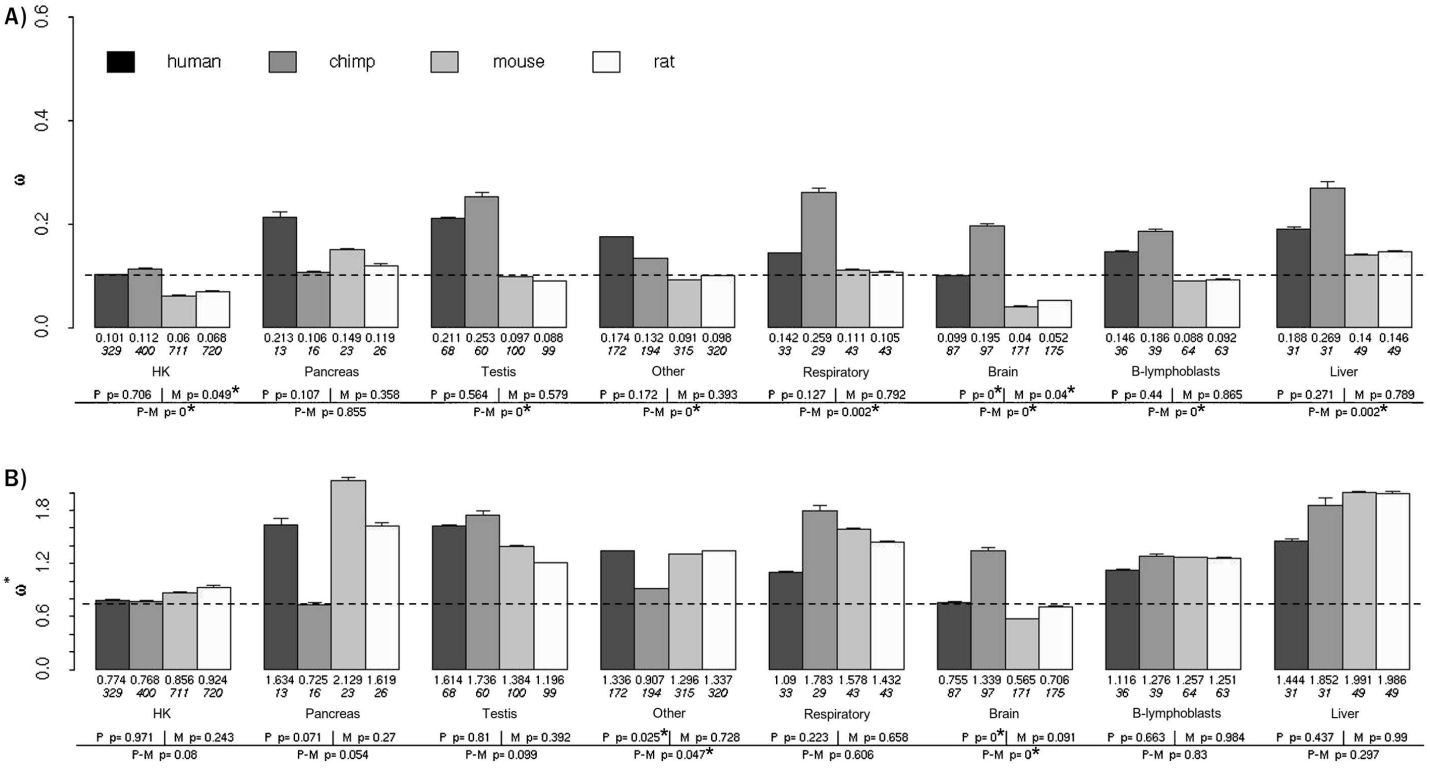


Figure 6.2: Evolutionary Rates of Tissue Specific Gene Categories Before and After Correcting for Lineage Effects The standard error weighted mean (Equation 3.5) is shown for housekeeping genes and 7 tissue specific categories. The value of omega (ω) is shown immediately below each bar followed by the number of genes used for estimation in each case. Below the category labels, the *p*-value for the permutation based comparison of means is shown for the comparison within primates (P), murids (R), and among primates and murids (P-M). Significant differences are marked with an asterisk (*). Whiskers represent the standard error associated to the mean (Equation 3.6) **A)** Results for the branch model based estimates for each lineage in each of the categories. All categories except pancreas, show significantly different means between primates and murids. **B)** Omega values are shown divided by the genomic means (ω*) for each of the four lineages. Only those comparisons among groups within *Brain* and *Other* tissue categories remain significant after correction for lineage effects.

show significant differences in rates among primate and murid lineages: *brain* and *other*. *Other*, is a combination of other categories with fewer than 15 TSGs, and thus contains different tissue sources out of which many are immune related. Whether this difference within the Others category is an effect of the variation of particular TSGs or an effect of the distributions of rates spread among different categories, would require further analysis. With respect to the difference observed for *brain* when comparing primate and murid lineages, as we will see next, a better understanding can be obtained in the light of the differences in rates found specifically within primate lineages. Additionally it is also important to mention, that although the data after correction for lineage effects in other databases is not shown, differences between lineages in brain specific genes are not statistically significant. The only other category showing significant differences among primate and murid lineages, is *pleura* TSGs from the TDDB set ($p = 0.017$). Moreover, the difference observed here is due to similarly elevated or low values within primates ($\omega^* \backsimeq 1.5$) and within murids ($\omega^* \backsimeq 0.8$), and as such, it constitutes the class with most robust evidence of differences among primates and murids.

6.3 Differences Among Humans and Chimpanzees

Overall, the rates of chimp are slightly generally higher than those of human for most of the categories examined (Figure 6.2) and across all databases (Table 6.3). However, the differences are largely non significant when comparing SEWM ω with a few exceptions. The only categories showing differences among human and chimp are: *brain* and partially, *other* within the GNF set, persisting even after the correction for lineage effects; *Male reproductive gland* (MRG) in the TDDB set; and both *muscle* and *lung* in the GeneNote set. Notably, these few differences are due to a deviation in the rates of chimp with respect to the general estimates for the three other lineages. In the *other* category chimp shows rates that are lower than those of human, rat, and mouse which are largely similar (Figure 6.2). In all other cases, a similar but opposite tendency is observed. Markedly, the differences within primates are driven by elevated rates in chimpanzee that deviate the most in comparison to the other three lineages (Table 6.3). In the GNF set, the difference among primates and murids after correction for lineage effects is clearly driven by an elevation in the rates of chimp.

6.4 Differences Among Tissue Specific Gene Categories

As observed in previous studies (e.g. Khaitovich *et al.* , 2005), a simple inspection shows that the overall rates of brain specific and housekeeping genes across lineages tend to be more conserved than other genes (Figure 6.2). Genes ex-

pressed in *testis*, *pancreas*, and *liver* show higher evolutionary rates than other tissues when ω are reported as the standard error weighted mean values derived from a free branch ML model. These results are supported Table 6.3 both in the TDDB set, for HK, *brain*, and *male reproductive gland* categories where *liver* and *pancreas* are not found, and in the GeneNote set for *liver* and *prostate*, where HK and and *pancreas* are not found. Additionally, *lung* (GeneNote) and *pleura* (TDDB) also show rates among the highest in tissue specific categories, where as in GNF *respiratory* TSGs show high but not particularly elevated rates of SEWM ω. These observations remain after corrections for lineage effects. Across all databases, human brain specific genes have the lowest rate among all other tissues, including HK genes, however in both GNF and TDDB, and marginally, although in the same direction in GeneNote, chimp brain specific genes are those with the highest rates among lineages.

In order to assess these observations, specific permutation tests on the GNF set of all against all tissue categories, individual tissues against the genomic set, and individual tissues against random for human rates, yield a common pattern. With the exception of one or two significantly different comparisons among tissues showing higher rates in *other*, *brain* is the only class showing significantly different rates for more than two tissues when human SEWM ω estimates are considered (threshold, $p < 0.05$). None of the human categories show significant differences when compared to the genome SEWM ω or against random. In chimp, *pancreas* is the only category that shows significant differences in these tests, where it is lower than *testis*, *liver*, and *respiratory* TSG categories. When mouse and rat are considered, the pattern observed for in human *brain* is reinforced together with the general trends mentioned at the beginning of this section. As such, brain specific genes have significantly lower rates than all other TSG categories, including housekeeping genes, and are significantly lower than genomic and random estimates ($p \leq 0.0004$). *Brain*, by far shows the most highly significant differences among all differences observed. In each of the murid lineages as well, categories such as *testis*, *other*, and *liver* show significantly higher estimates than most other categories including both, random, and genomic estimates. HK also shows values that are lower than most TSG categories. However, *brain* is the only TSG category that is significantly lower than the genomic and random estimates ($p << 0.001$ for both). As such, these results are in direct opposition of studies where a general elevation of rates in brain are observed within primate lineages and among those of primates and murids. Certainly, the most marked pattern that emerges is the conservation of brain specific genes which seem to show some evidence of elevation in chimp.

It is also important to highlight that while HK genes are included among the categories used for comparison since they have become a reference in various types of analyses, it must be noted that it is not a TSG category. Certainly, HK genes are by definition constitutively expressed in most cells as they are required for basic housekeeping functions. As such, they are possibly the category of genes that is the furthest from being a suitable point of comparison for drawing conclusions about the behavior of TSGs. This becomes evident In Figure 6.3

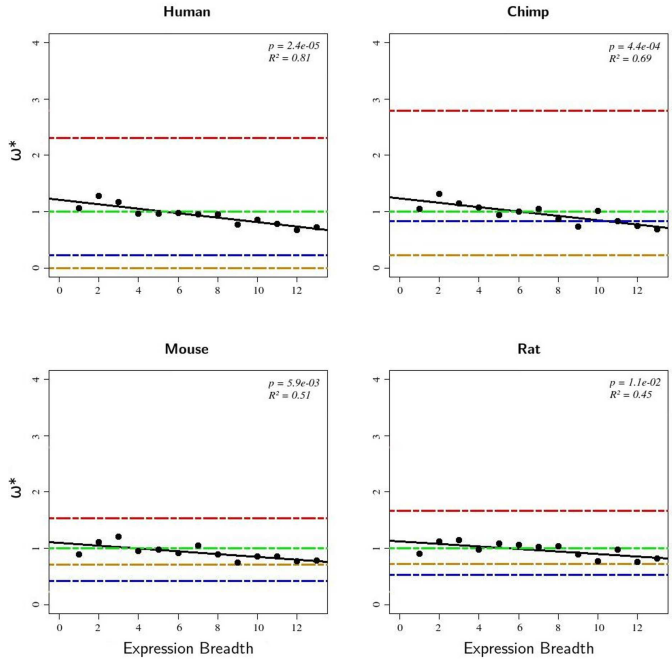


Figure 6.3: Evolutionary Rates and Expression Breadth The median ω for genes, determined to be expressed in a given number of tissues (Expression Breadth) is shown divided by the median ω for the genome (ω*) in each graph for human, chimp, mouse, and rat. The p-value and R^2 values associated to the linear regression (black line) are shown on the top right corner. The median ω* value for *brain* (blue), *testis* (red), HK (ocre), and the genome (green), are represented by horizontal dotted lines. The regressions in each of the lineages are highly significant showing a strong correlation between ω* and expression breadth.

where all genes that are expressed in one or more tissues from the GNF set show strong ($0.45<R^2<0.81$) and highly significant negative correlations for ω^* (median ω for all genes in each tissue bin / median ω for the genome) and the number of tissues they are expressed in (expression breadth) in all of the species assayed. The green, dotted, horizontal line is set at $\omega^*=1$ which is the value of the genomic estimate. ω estimates for genes expressed in 6 tissues or less tend to rise above the genomic estimate, while those expressed above this value, tend to be found below. The median ω^* for *testis* (dotted red line), which is one of the fastest evolving categories, as mentioned earlier, is always found above the highest estimate for any of the expression breadth bins shown. Indeed, with the exception of *pancreas* TSGs which show particularly low values in chimp, only *brain* and HK categories show values that are lower than the genomic estimate. While HK genes have an average appropriately found well below the genomic estimate. The mean for *Brain* genes, which have an expression breadth value of one, with the exception of in chimp, is also found not only well below the genomic estimate, but also further below any other estimate for any of the expression breadth bins.

The fact that ω is negatively correlated with expression breadth shows that selective constraints are additive with the number of tissues a gene is expressed. A relationship that has indeed been reported before (e.g. Khaitovich *et al.* , 2006). However, none of the evidence found here points towards an elevation of human brain specific genes as seen in Dorus *et al.* 2004. In fact, what seems most remarkable when comparing TSG categories amongst themselves, is that brain TSGs have such markedly constrained rates. More so than any other TSG category, and even above those of the HK category which are found at the other end of the scale of additive constraints exerted by an elevated expression breadth.

Finally, the analysis was repeated considering TSG categories where only genes that were expressed in only one of the source tissues, excluding whole tissue samples: e.g. *whole brain*, that make up each TSG category were considered. Given that expression breadth affects the constraints of genes additively, the possibility that results are influenced by the number of actual source tissues in the TSG categories is thus addressed. More precisely the TSG categories used here and in several other studies (Dorus *et al.* , 2004; Nielsen *et al.* , 2005; Khaitovich *et al.* , 2006; Wang *et al.* , 2006; Zhang & Li, 2004) are actually mostly situated at a system's organ level where, to cite an example, tissues with different cell types, such as *cortex*, *substantia nigra*, among others are included under the TSG category of *brain*. Linear regressions considering source tissues for all of the genes when comparing ω and expression breadth also show a negative correlation ($2.7e\text{-}08\leq p\leq 8.4e\text{-}09$ and $0.33<R^2<0.35$ for all lineages except chimp where $p = 2.2e\text{-}03$ and $R^2 = 0.12$). Standard error weighted means of these TSG categories show the exact same patterns as described throughout previous section of this analysis, with the exception that no significant difference is observed within or among primate and murid lineages in the *other* and *brain* TSG categories. While in this case the categories have less than four times

as many genes and their SEWM ω show higher errors, brain is still the category which is most constrained, still showing estimates even lower than those of HK genes. Also, other classes with high estimates of ω are again confirmed (Data not shown). These results support that tissue heterogeneity, in light of both, the additive nature of expression breadth, or with regard to the effect that higher heterogeneity might produce lists of genes that are biased towards those expressed among all cells of a certain tissue type, is not responsible for the evolutionary patterns observed among tissues here.

6.5 Comparison Between Statistical Methods

To examine how robust conclusions are to the methods used to infer values of ω, estimates were also obtained by using pairwise comparisons based on Li (1993) and ML methods. The use of both the simple arithmetic mean and the standard error weighted means in order to estimate and compare values of ω for classes was also explored. Table 6.1 shows the results obtained for both pairwise and lineage estimates together with different methods for combining ω. Here the numerical results in Figure 6.2 are shown under SEWM free branch ML for comparison. The results based on lineage specific estimates and the pairwise estimates and different methods for combining estimates are virtually identical with respect to the comparison of rates among lineages. This is a good indication that the patterns observed here are not an effect of the methodology employed for the study. However, two critical points merit mention. Comparison among tissues classes within particular lineages are largely similar for the murid lineages independent of the method used for estimating means. For both primate lineages however, the use of a simple harmonic mean instead of SEWM produces largely significant differences when comparing among tissue specific cases within lineages. Since divergence times are relatively small in the primate lineages analyzed, the presence of few mutations (i.e. 1 or 2) being predominantly nonsynonymous may result in an elevation of ω associated to a large variance. While the general lack of differences when contrasting both types of means among lineages suggests the possibility that the apparent difference between murids and primates is not an artifact caused by the higher variance associated with the estimates in the primate lineages, the difference among comparisons within lineages suggest that an effect may exist and should be addressed when looking at sequences with low divergence.

6.6 PS Tests on Tissue Specific Genes

While evidence for an elevated ω for brain specific genes within in the human lineage is weak, this does not exclude the possibility that there are more of these genes that have been targeted by positive selection in human than in chimpanzee. An elevated dN/dS ratio can be caused both by a relaxation of constraints and by positive selection. Only ω values significantly > 1 provide

	SM Li93			SM pairwise ML			SEWM pairwise ML			SEWM free branch ML						
	P	M	***p*-val P-M**	**P**	M	***p*-val P-M**	**P**	M	***p*-val P-M**	**human**	**chimp**	**mouse**	**rat**	***p*-val P**	***p*-val M**	***p*-val P-M**
HK	0.254	0.142	0.000	0.203	0.117	0.000	0.095	0.063	0.000	0.101	0.112	0.060	0.068	0.706	0.049	0.000
n	831	851	-	828	859	-	514	780	-	329	400	711	720	-	-	-
Nrm	0.794	0.766	0.600	0.769	0.774	0.920	0.672	0.811	0.044	0.774	0.768	0.856	0.924	0.971	0.243	0.080
Pancreas	0.304	0.224	0.104	0.279	0.183	0.097	0.221	0.123	0.068	0.213	0.106	0.149	0.119	0.107	0.358	0.855
n	27	28	-	27	28	-	21	27	-	13	16	23	26	-	-	-
Nrm	0.950	1.206	0.183	1.054	1.205	0.543	1.568	1.585	0.963	1.634	0.725	2.129	1.619	0.071	0.270	0.054
Testis	0.484	0.219	0.000	0.405	0.194	0.000	0.205	0.095	0.000	0.211	0.253	0.097	0.088	0.564	0.579	0.000
n	99	105	-	99	105	-	88	103	-	68	60	100	99	-	-	-
Nrm	1.511	1.179	0.052	1.531	1.279	0.202	1.454	1.227	0.382	1.614	1.736	1.384	1.196	0.810	0.392	0.099
Other	0.348	0.235	0.000	0.300	0.203	0.000	0.147	0.103	0.000	0.174	0.132	0.091	0.098	0.172	0.393	0.000
n	318	332	-	317	332	-	264	327	-	172	194	315	320	-	-	-
Nrm	1.086	1.266	0.038	1.135	1.340	0.057	1.046	1.330	0.032	1.336	0.907	1.296	1.337	0.025	0.728	0.047
Resp.	0.357	0.192	0.001	0.287	0.164	0.003	0.177	0.118	0.069	0.142	0.259	0.111	0.105	0.127	0.792	0.002
n	45	46	-	44	46	-	39	45	-	33	29	43	43	-	-	-
Nrm	1.115	1.036	0.680	1.087	1.080	0.976	1.259	1.520	0.453	1.090	1.783	1.578	1.432	0.223	0.658	0.606
Brain	0.237	0.111	0.000	0.177	0.085	0.000	0.125	0.046	0.000	0.099	0.195	0.040	0.052	0.000	0.040	0.000
n	200	200	-	199	200	-	141	191	-	87	97	171	175	-	-	-
Nrm	0.738	0.599	0.072	0.670	0.563	0.176	0.890	0.593	0.002	0.755	1.339	0.565	0.706	0.000	0.091	0.000
B-lymph	0.337	0.165	0.001	0.310	0.140	0.000	0.173	0.097	0.002	0.146	0.186	0.088	0.092	0.440	0.865	0.000
n	70	71	-	70	71	-	55	67	-	36	39	64	63	-	-	-
Nrm	1.053	0.890	0.381	1.172	0.921	0.244	1.225	1.245	0.949	1.116	1.276	1.257	1.251	0.663	0.984	0.830
Liver	0.407	0.251	0.003	0.331	0.220	0.016	0.211	0.167	0.149	0.188	0.269	0.140	0.146	0.188	0.789	0.002
n	51	51	-	50	51	-	46	49	-	31	31	49	49	-	-	-
Nrm	1.270	1.354	0.662	1.253	1.450	0.339	1.502	2.155	0.061	1.444	1.852	1.991	1.986	0.437	0.990	0.297
Genome	0.321	0.186	0.000	0.264	0.152	0.000	0.141	0.078	0.000	0.130	0.145	0.070	0.074	0.089	0.043	0.000
n	8556	8754	-	8518	8807	-	6533	8455	-	4442	4645	7967	8074	-	-	-

Table 6.1: Comparison Between Models and Methods Results for estimates of ω using both Li and ML methods are shown along the columns for each of the TSG categories in GNF (rows). Pairwise and free branch results are combined for each category using the simple harmonic mean (SM) or the standard error weighted mean (SEWM). For each TSG category, the corresponding estimates for each lineage are presented, followed by the number of genes that could be analyzed by that method (n), and the value of the estimate of ω divided by the genomic estimate of ω for the lineage using the same method. The last rows show the genomic value of ω (Genome) and number of genes used in its calculation (n) for all genes that had orthologs in human, chimp, mouse, rat, and dog.

Tissue	H	Ch	Total	% H	% Ch
Other	5	10	182	2.7	5.5
BM	1	2	38	2.6	5.3
Pancreas	1	2	40	2.5	5.0
Not.TSG	134	215	6122	2.2	3.5
Muscle	1	1	48	2.1	2.1
Brain	4	9	232	1.7	3.9
MRG	2	3	190	1.1	1.6
Germ	1	2	163	0.6	1.2
HK	5	18	798	0.6	2.25
Eye	0	1	51	0.0	2.0
GI	0	1	43	0.0	2.3
Kidney	0	3	58	0.0	5.2
Liver	0	1	70	0.0	1.4
Lung	0	1	35	0.0	2.9
Respiratory	0	1	29	0.0	3.4
SC	0	0	28	0.0	0.0
Blymphoblasts	0	0	38	0.0	0.0
Total	154	270	8165	1.9	3.3

Table 6.2: Genes Under Positive Selection for Each of the TSG Categories from All Databases The number of PS genes deduced from branch-site test II after correction for multiple testing in human (H) and chimp (Ch) is shown for each of the TSG categories (Tissue) analyzed from the GNF, TDDB, and GeneNote databases. The total number of genes under each TSG category is shown in the middle, followed by the corresponding percentage of PS genes found in human (% H) and chimp (% Ch). All categories found in more than one DB were merged (see Table 6.3) BM: *bone marrow*; Not TSG: all genes that are not tissue specific within any of the databases; Muscle: *skeletal muscle*; MRG: *male reproductive gland* in TDDB, *prostate* in GeneNote, and *testis* in GNF; GI: *gastrointestinal tract*; SC: *spinal cord*;

evidence for positive selection. Likewise, absence of an elevated ω could be caused by a cancellation of effects of increased negative and positive selection. To examine if there is more evidence for positive selection in the brain specific genes in the human lineage, the statistical test described in Chapter 4 for detecting lineage specific selection was used. Table 6.2 shows in agreement with previous results (Chapter 4), that the numbers of genes showing evidence for positive selection in human are considerably smaller than those found in chimp. Moreover a general lack of correlation can be found when comparing TSG categories mentioned earlier to have particularly high or low values with the number, or percent representation, of PS genes within each. As such, categories like *respiratory*, *liver*, and *germ* show less positive selection than categories with particularly low rates such as HK and *pancreas*. Similarly and vice versa, categories with high and low rates respectively, such as *other* and *muscle*, show high and low occurrence of PS genes. *Brain* actually shows 4 and 9 PS genes in human and chimp respectively. However when compared to other tissues, it is not particularly striking as not only do other tissues show higher percent representations and counts, but also in that the actual signal found in any of the TSG categories is not particularly outstanding. In fact, brain specific genes show a pattern similar to that observed for housekeeping genes. All in all, the results of the lineage specific branch-site tests reinforce the general conclusion that brain specific genes do not show more evidence for positive selection in the

DB	Tissue	Human	Chimp	Mouse	Rat	*p*-val P	*p*-val M	*p*-val P-M
GNF	HK	0.101	0.112	0.060	0.068	0.706	0.049	0.000
	Brain	0.099	0.195	0.040	0.052	0.000	0.040	0.000
	Testis	0.211	0.253	0.097	0.088	0.564	0.579	0.000
	Pancreas	0.213	0.106	0.149	0.119	0.107	0.358	0.855
	Other	0.174	0.132	0.091	0.098	0.172	0.393	0.000
	Respiratory	0.142	0.259	0.111	0.105	0.127	0.792	0.002
	B-lymph.	0.146	0.186	0.088	0.092	0.440	0.865	0.000
	Liver	0.188	0.269	0.140	0.146	0.271	0.789	0.002
TDDB	HK	0.174	0.145	0.056	0.064	0.362	0.204	0.000
	Brain	0.104	0.180	0.063	0.062	0.052	0.958	0.000
	MRG	0.206	0.278	0.117	0.129	0.047	0.575	0.000
	Germ	0.142	0.240	0.109	0.082	0.146	0.441	0.030
	Kidney	0.164	0.184	0.090	0.112	0.730	0.390	0.005
	Eye	0.140	0.158	0.066	0.059	0.694	0.692	0.000
	GI	0.192	0.279	0.115	0.108	0.123	0.904	0.007
GeneNote	Brain	0.092	0.123	0.052	0.060	0.585	0.371	0.000
	Prostate	0.177	0.163	0.082	0.091	0.758	0.610	0.000
	SC	0.124	0.182	0.069	0.074	0.195	0.741	0.000
	Liver	0.152	0.206	0.220	0.104	0.263	0.709	0.000
	Pancreas	0.166	0.180	0.074	0.096	0.743	0.198	0.000
	BM	0.174	0.237	0.106	0.112	0.383	0.778	0.000
	Muscle	0.103	0.204	0.087	0.103	0.046	0.554	0.193
	Kidney	0.156	0.238	0.077	0.092	0.211	0.515	0.000
	Lung	0.157	0.248	0.114	0.099	0.039	0.470	0.000

Table 6.3: Evolutionary Rate Estimates for the TSG Categories Analyzed in All of the Databases SEWM ω is shown for the TSG categories from GNF, TDDB, and GeneNote for each of the lineages under consideration. *p*-values for the permutation test within and among primate and murid lineages are shown labeled as in Table 6.1 and Figure 6.2. Tissue categories as in Table 6.2.

human than in the chimpanzee lineage.

Chapter 7

Natural Selection and Disease

7.1 Distribution of Disease-Associated Mutations in p53

More than half of human cancers are associated with one or more alterations in the tumor suppressor gene p53 (Brachmann *et al.* , 1996). The main role of p53 in normal cells is the induction of cell-cycle arrest or apoptosis in response to cellular stress, particularly DNA damage (Kuerbitz *et al.* , 1992). The p53 tumor suppressor is a 393 amino acid residue transcription factor that activates the transcription of a number of genes. Structurally and functionally, it can be divided into five regions (Ayed *et al.* , 2001): an acidic N-terminal transactivation domain (p53TA, residues 1–60), a proline-rich domain (p53PR, residues 61–97), a hydrophobic DNA-binding domain (p53DB, 100–300), a tetramerization domain (p53TR, 320–360), and a basic C-terminal domain (p53CO 361–393). Figure 7.1 shows the codon mutation frequencies, derived from the IARC TP53 mutation database (Olivier *et al.* , 2002), associated with each of the protein domains. p53 missense mutations are scattered throughout the coding sequence, although 96% of them (17,389/18,135) cluster within the p53DB domain. Six different "mutational hotspots" (defined by a mutation frequency higher than 2% of all mutations (Olivier *et al.* , 2002)) have been identified at residues Arg175, Gly245, Arg248, Arg249, Arg273 and Arg282. According to this description, these mutational hotspots fall within the p53DB domain, and since they are structurally relevant to protein function (Cho *et al.* , 1994) they would be expected to be protected against mutations by strong purifying selection.

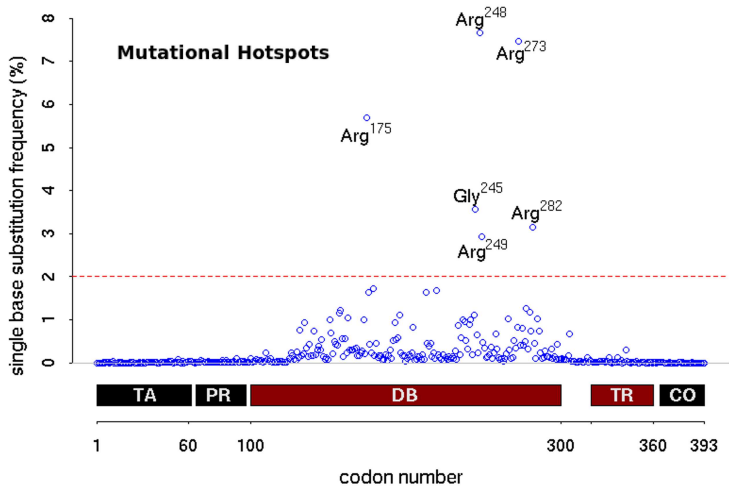


Figure 7.1: Distribution of Disease Causing Mutations in p53 Mutation frequencies collected in the IARC TP53 R10 database (18,145 nonsynonymous mutations) are plotted against the protein domains. The DNA-binding (p53DB) domain contains six residues considered mutational hotspots in cancer.

7.2 Characterization of Selective Constraints on p53 Codon Sites

There is no *a priori* ω distribution that appropriately describes how selective pressures vary along the sequences of specific genes. In order to know which is the best estimate of ω for a specific residue in a protein, two different maximum likelihood site based tests of positive selection (the nearly neutral M1a and M7(β) models versus the M2a and M8(β+ω) selection models, respectively) were evaluated using likelihood-ratio tests (LRT). The log-likelihood value of the M1a model was the same as that for the M2a model, providing no evidence of positive selection in p53. ML adjustment of parameters for the M2a model suggested that about 80% (p_0=79.07%) of the sites have evolved under the action of purifying selection (ω_0=0.077), (see Materials and Methods for a full description of parameter notations). Model M2a estimated that 16% of the codons in p53 changed according to a neutral model of evolution (ω_2=1). Posterior probabilities obtained from the empirical Bayes approach were not significant (p<95%) for any of the protein residues, suggesting again the absence of positive selection on p53 under this model.

ML adjustment estimates were different once the more complex codon-based likelihood models were evaluated for p53. The LRT statistic (Equation 3.4) for the M7 and M8 comparison was 2Δ=0.0006, which is not greater than the

critical values from a $\chi 2$ distribution with d.f.=2. Thus M7(β) is not rejected in favor of M8($\beta+\omega$), providing additional evidence for the absence of positive selection under a richer parameter model. Empirical Bayes analysis did not identify positively selected sites with a posterior probability greater than 95%.

Model M8($\beta+\omega$) (hereinafter M8) fits the data better than the M2a model (selection). ML estimations of the M8 parameters suggest that almost 100% of p53 codon-sites are constrained under the influence of purifying selection, and only a minimum proportion of sites ($p_1<1\times10-06$) evolved with $\omega>1$ ($\omega_1=21.21$). The parameters of the β-distribution (p=0.385, q=2.076) suggest that the distribution of ω over the sites is L-shaped, with most sites in p53 being highly conserved.

The site-wise likelihood-ratio method (SLR) found 228 codon sites under the influence of strong purifying selection after correcting for multiple testing (202 at $p<0.01$ and 26 at $p<0.05$). It is interesting to notice that this number is higher than the 109 (47.80%) that are, strictly speaking, phylogenetically conserved and never change in amino acid identity during vertebrate evolution.

Figure 7.2 summarizes the distribution of ω according to each of the different evolutionary models employed ((a)–(c)) and within the different domains in p53 ((d)–(f)). Figure 7.2(a) shows that model M2a estimates higher ω values than the M8 model. The same occurs when comparing the M2a and SLR estimations (Figure 7.2(b)). Values of ω estimated by the M8 model (ωM8) and SLR (ωSLR) behave in a similar manner (Figure 7.2(c)). Considering these models, most of the residues showing $\omega<0.3$ contain the set of 228 sites evolving under purifying selection (red squares).

Figure 7.2(d) shows the distribution of ω values along p53 domains according to the three methods used. Independent of the model used (M2a, black triangles; M8, gray crosses; SLR, blue squares), ω values are observed to be relatively higher in domains p53TA and p53PR, suggesting that these regions have evolved under the action of a weaker form of purifying selection. It is interesting to note that, in contrast to the M8 and SLR models, the M2a model predicts many nearly neutral residues ($w\approx1$) along the p53DB, p53TR, and p53CO domains (Figure 7.2(e)). This is probably due to the restriction imposed by the use of only one site class in the M2a model in comparison with the ten site classes used in model M8 for modeling sites under neutral or nearly neutral evolution. Figure 7.2(e) shows a zoomed version of the area where $\omega<0.3$. Finally, Figure 7.2(f) shows that residues with strong evidence of purifying selection (red, SLR $p<0.05$) concentrate in the range of $\omega\leq0.1$ and reduce gradually in the ranges where $0.10<\omega<0.20$ and $0.20<\omega<0.30$. These values are in accordance with previous estimations on mitochondrial protein-coding genes in hominids where highly conserved residues are considered as those where $\omega<0.33$ (Yang *et al.* , 2000).

Table 7.1 summarizes data related to p53 domains, codons, indels, mutations and ω statistics. Independent of the model used to estimate ω, p53DB and p53TR domains showed the highest number of cancer mutations that were associated with the lowest median and mean ω values observed in the analysis.

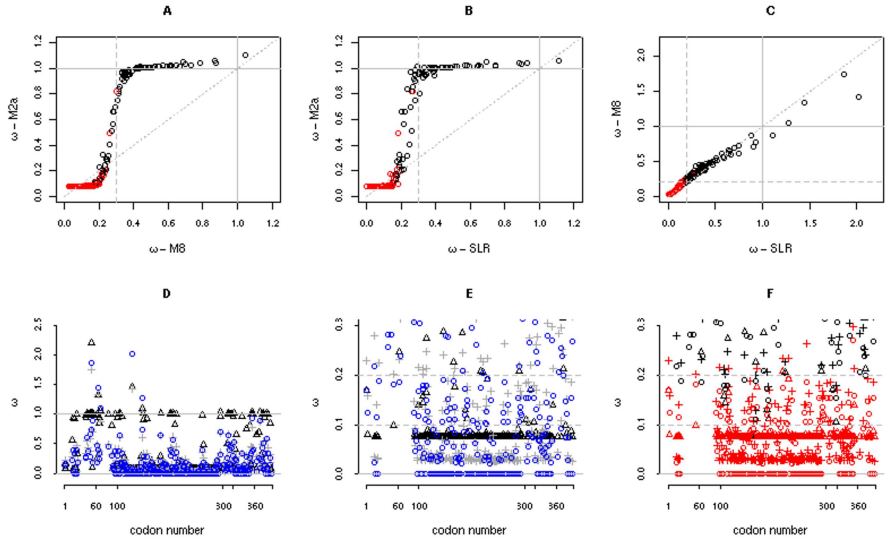


Figure 7.2: Comparisons of Models and ω Distribution in p53 Domains The simplest model, M2a, estimates a sigmoidal distribution of ω in relation to the more parameter-rich M8 and SLR models. The main differences between these models is found at an ω between 0.2–0.3 ((a) and (b)). The ω values between the two parameter-rich models were almost identical (c). Most of the residues under purifying selection detected by the SLR model ($p<0.05$ after correcting for multiple testing) have shown $\omega<0.30$ (red dots to the left of the broken gray line). Using different scales it is evident that, independent of the model used to infer ω, the p53DB and p53TR domains have shown smaller ω values ((d) and (e)). Black triangles, M2a; gray crosses, M8; blue circles, SLR. Codon sites with $\omega<0.1$ concentrate most sites under a statistical definition of purifying selection (f). Red symbols are sites considered to be under purifying selection using the SLR method.

p53 alignment			Mutations		Mdl	ω statistics			
Dom	Codons	Indels	Total	Mps		Min.	Med.	Mean	Max.
TA	1-60	38	96	1.6	M8	0.03	0.334	0.379	1.747
					SLR	0	0.269	0.369	1.865
PR	61-97	22	151	4.2	M8	0.029	0.314	0.376	1.338
					SLR	0	0.307	0.376	1.447
DB	100-300	5	17389	87	M8	0.027	0.039	0.116	1.423
					SLR	0	0.029	0.095	2.018
TR	325-355	0	178	5.1	M8	0.028	0.067	0.126	0.456
					SLR	0	0.068	0.103	0.379
CO	361-393	11	18	1.6	M8	0.027	0.216	0.255	0.878
					SLR	0	0.176	0.226	0.882

Table 7.1: Summary of p53 Domains, Mutations and ω Statistics According to the M8 and SLR Models Mutations were deduced from the IARC TP53 database (Olivier *et al.* , 2002). Mdl is the model used. Med. is the median value of ω. Dom is the domain of p53. Indels are the number of insertions or deletions. Mps is the mean number of mutations per site.

One-tail Kolmogorov–Smirnov (K–S) tests demonstrated that the p53DB and p53TR domains have a significantly low ω value distribution ($p<0.05$) in comparison with the rest of the p53 domains. While the mean ω estimation was close to 0.1 in both domains, the distribution of ω values in the p53DB was lower than in the p53TR domain (data not shown).

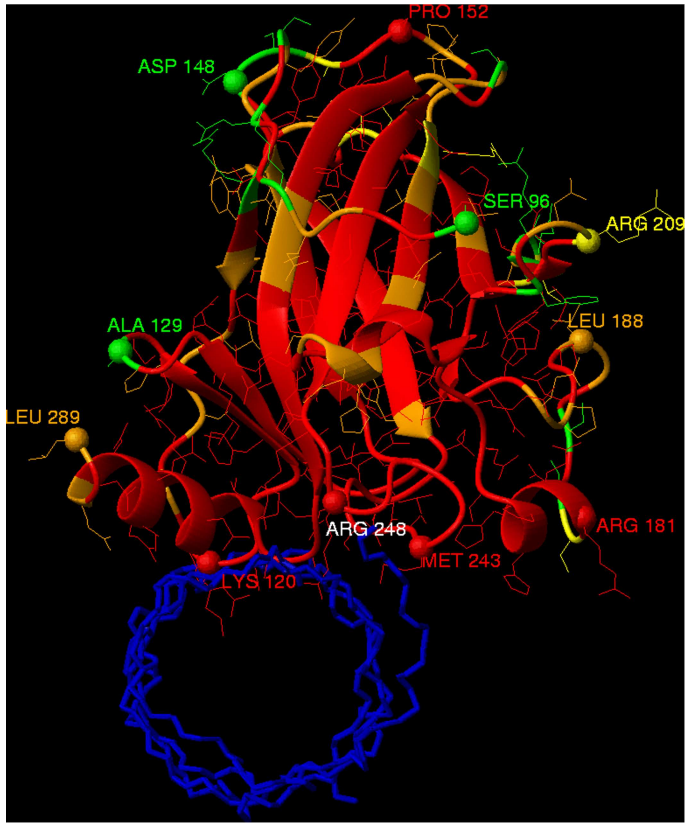
In summary, estimates of natural selection acting on p53 coding sites differentiate, and do so with statistical significance, the relevant functional domains where the prevalence of cancer mutations are the highest in the protein. This pattern is what would be expected if relevant functional structures had been constrained during evolution under strong selective forces avoiding nonsynonymous changes.

7.3 Mapping Selective Pressures in the Structure of p53

In order to study the distribution of ω values within the structure of p53, the core domain in complex with DNA (Cho *et al.* , 1994) and the tetramerization domain (Lee *et al.* , 1994) structure were downloaded from the Protein Data Bank (PDB) (Berman *et al.* , 2000). By defining three different ranges of ω with a gradual distribution of selective constraints (red, $0\leq\omega\leq0.1$; orange, $0.1<\omega\leq0.2$; and yellow, $0.2<\omega\leq0.3$), residues were labeled in the structures of p53DB and p53TR domains in order to test if the deduced ω values are in agreement with the structural and biological knowledge on important residues of these domains. Residues where neutral or nearly neutral evolution was deduced (labeled green, $\omega>0.30$), were expected not to form part of the relevant functional domains of the protein.

Figure 7.3(a) shows the distribution of the ωSLR values, on the core p53DB domain structure. Figure 7.3(b) depicts a schematic representation showing the primary sequence and the secondary structure where the most relevant residues are shown. Residues where denaturing mutants are observed (Cho *et al.* , 1994) (Pro143, Arg175, Gly245, Arg249, Glu258 and Arg282) (red circles), and those involved in Zn2+ coordination (Cys176, His179, Cys238 and Cys242) (white circles) coincided with red labels, suggesting that they are under strong evolutionary constraints imposed by purifying selection. These residues are phylogenetically conserved in the alignment (marked by the asterisk: *), and purifying selection (PFS) was detected on them using SLR at a 99% confidence level (marked by the exclamation symbol: !).

(a)



(b)

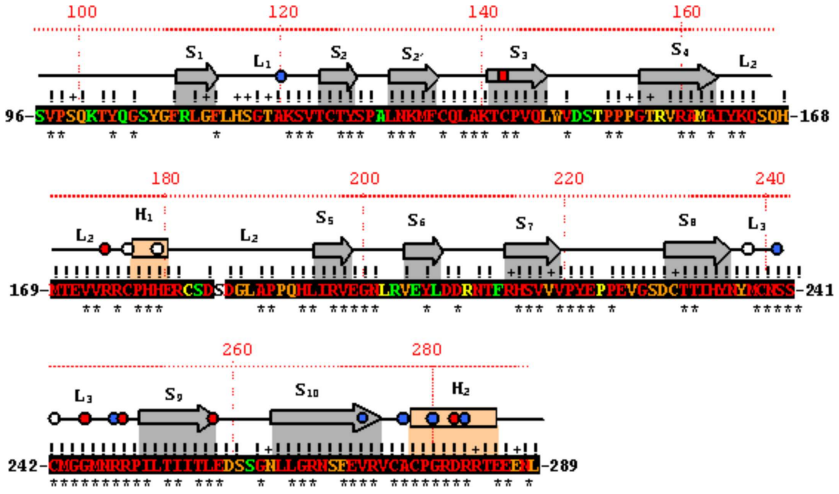


Figure 7.3: Mapping of Selective Constraints in the p53DB Domain (a) The three dimensional structure of the p53DB domain showing residues colored according to different selective pressures. (b) Primary amino acid sequence and secondary structure elements of the p53DB domain. Residues in red, orange, yellow and green show the gradual distribution of the selective constraints represented by ωSLR values. Residues in red ($0 \leq \omega < 0.1$) and orange ($0.1 \leq \omega < 0.2$) are generally associated with DNA contact sites (blue circles), Zn2+ contact (white circles), and sites where mutants are known to be denaturing (red circles) among others. A few of the sites seem to be below the limit considered for selective constraints (yellow, $0.2 \leq \omega < 0.3$). Residues where selective constraints were predicted to be absent (green, $\omega > 0.3$) are distributed along the external regions of the core domain, and most of them are interspersed between β sheets and helices. Arg248 binds in the minor groove of the DNA. Ser185 was conserved in the cluster of primates and rodents, but was discarded in the analysis due to gap insertions in the basal species. * Phylogenetically conserved residue. + SLR detected PFS at 95% confidence after correcting for multiple testing. ! As in +, but at 99% confidence. See the text for a detailed explanation.

The same pattern was observed for most of the residues involved in DNA interaction (Cho *et al.* , 1994) (Lys120, Ser241, Arg248, Arg273, Cys277, Arg280 and Arg283, blue circles). Notably, although Arg283 was not phylogenetically conserved, it was still observed under a red label, with PFS detected at 95% confidence (marked by the plus symbol: +). Asp281 is not directly involved in DNA contact, but it interacts with residues Arg273 and Arg280 and, as could be expected, it is found to be under strong negative selective pressure (ωM8=0.027, ωSLR=0, and SLR at a 99% confidence).

The same occurs for Arg282, a residue of the H2 helix in interaction with other residues under strong selection like Phe134, Thr125 and Ser128, which are involved in the packing of the H2 helix against the β hairpin and the L1 loop (Cho *et al.* , 1994). Other functionally relevant residues are those involved in the hydrophobic core of the β sandwich, which comprises a number of key amino acid residues in maintaining the structural stability of the p53DB domain (Cho *et al.* , 1994). Among them, Cys141, Val143, Val157, Ile195, Val197, Tyr234, Tyr236 and Phe270, were observed to be under the influence of strong PFS effects (SLR at a 99% confidence). Again, some of them were not phylogenetically conserved in all of the vertebrate species. It is interesting to note that at this confidence level, some of these residues are labeled orange, pointing out that residues showing $0.01<\omega<0.02$ are also under strong constraints due to negative selection.

The portion of the hydrophobic core further away from the DNA-binding surface (Phe109, Leu111, Leu145, Pro151, Val157, Val218, Tyr220, Thr230, Ile232, Ile255 and Leu257), although originally thought to be associated with less restrictive functional constraints due to its association with a smaller number of mutations (Cho *et al.* , 1994), was labeled red and orange throughout, suggesting the presence of strong selective constraints. As was previously recognized, L2 and L3 loops have little regular secondary structure and rely extensively on side-chain/side-chain and side-chain/backbone interactions for structural integrity. In agreement with this observation, these long loops were mainly found labeled orange and red, pointing out the prevalence of PFS on these loops. Only one residue (Cys182) showed less restrictive constraints (labeled yellow), although it was very close to the limit of statistical significance for PFS (SLR at a 99% of confidence without multiple testing correction). The only residue observed under nearly neutral evolution in these L2 and L3 loops was Ser183 (labeled green, ωM8=0.646, ωSLR=0.649). This residue is conserved in most mammals but is replaced by Pro in cat and zebra fish, Gly in chicken, Val in the pipid frog, and Asp in catfish. As far as could be determined, there is no evidence of the relative importance of Ser183 in the structure of the p53DB domain. A pattern of variable conservation was observed for the other 13 residues where relaxation of selective constraints was deduced. In agreement with our expectations, most are placed in the external region of the structure (Figure 7.3(a)), outside the β strand or helix regions (Figure 7.3(b)), with the exception of Arg110, Glu204 and Leu206. Interestingly, the first is located at the initial portion of the S1 strand that maps at the outermost external region of core domain. The last

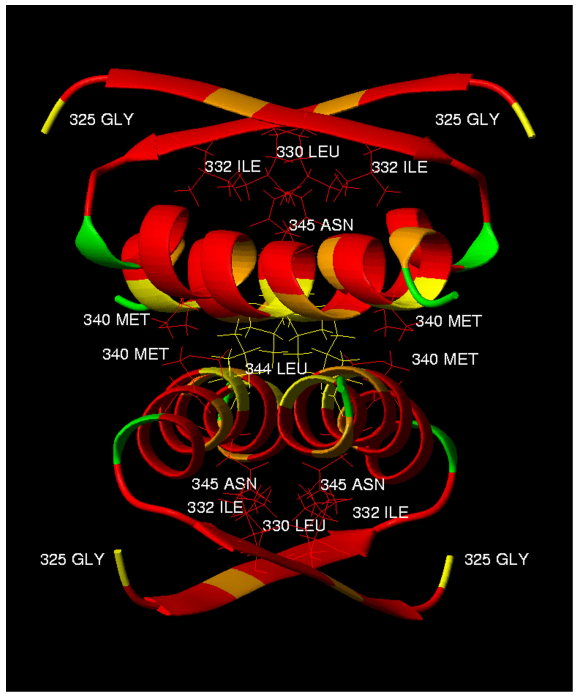
two (Glu204 and Leu206), are located in the S6 strand and their side-chains are pointing out towards the external region of the structure. More interesting is the fact that the side-chains of the neighboring residues, Tyr205 and Asp207 of the S6 strand (labeled in red), point to the internal core of the p53DB domain, and are involved in a number of interactions with neighboring β strands.

Figure 7.4(a) shows the three-dimensional structure of the p53TR domain. The domain forms a symmetric tetrameric structure with a topology made up from a dimer of dimers (Lee *et al.* , 1994). Each of the two primary dimers consists of two antiparallel helices linked by an antiparallel β sheet. The tetramerization interface between the two dimers is formed solely by helix–helix contacts (Lee *et al.* , 1994). Figure 7.4(b) shows a schematic representation of a monomer with its primary sequence and the secondary structure consisting of one β strand and one α helix. Only seven residues were phylogenetically conserved (*), while most of them were deduced as being under strong selective pressures (!). Residues like Leu344 (ωM8=0.304, ωSLR=0.279) and Leu348 (ωM8=0.032, ωSLR=0), labeled yellow and red, respectively, were previously demonstrated to be important for the stability of the tetrameric structure (Mateu & Fersht, 1998). The high ω value of Leu344 (ωM8=0.304, ωSLR=0.279) is due to a non-polar (hydrophobic) change that has occurred during evolution. Specifically, Leu344 changed independently of Ile at least three times in catfish, chicken, and in the common tree shrew. In addition, Lys351 is not phylogenetically conserved, although it was observed as being under the effects of PFS at 95% confidence (+). Lys351 was conserved in most of the species except sheep (Met), chicken (Ala), the pipid frog (Gln), and fish (Ser). In agreement with the low ω value observed (ωM8=0.297, ωSLR=0.269), previous works report that Lys351 is moderately stabilizing but somewhat destabilizing at high temperatures in the structure Mateu & Fersht (1998). The last residue observed at the limit of PFS effects was Gly325 (ωM8=0.278, ωSLR=0.224), which is not involved in the β sheet structure.

All residues in the β strand of p53TR domain were found to be selectively constrained, including Phe328, a site where mutations are strongly destabilizing (Mateu & Fersht, 1998). Only two residues, Glu326 and Leu330, were phylogenetically conserved in the β strand. The residues Arg337, Phe338, Met340 and Phe341 with a demonstrated influence in the p53TR structure (Mateu & Fersht, 1998) were found to be under strong effects of purifying selection (99% confidence), and yet none was phylogenetically conserved. Other residues known to be involved in structural stabilization are Arg333, Asn345, Glu349, Ala347 and Thr329. All of these correspond to red and orange labels, agreeing with the functional role in the p53TR domain. Glu336 is the only core residue within the p53TR domain where neutral or nearly neutral evolution was deduced. As was previously demonstrated by chemical denaturation analysis, this residue shows the highest value of solvent-exposure, and it was recognized as not having substantial effects on the stability of the tetramerization domain (Mateu & Fersht, 1998).

In summary, a large agreement is found between the functional relevance of

(a)



(b)

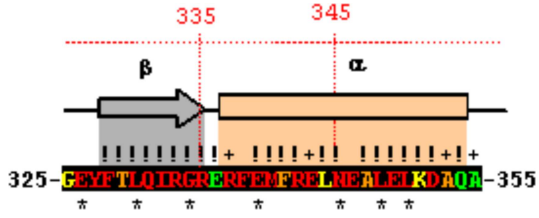


Figure 7.4: Mapping of Selective Constraints in the p53TR Domain (a) The three dimensional structure of the p53 tetramerization domain. (b) Primary amino acid sequence and secondary structure elements of a p53 monomer. Color code and notation as in Figure 3. See the text for a full description.

residues deduced from the estimation of selective constraints using ML models and the functional or structural importance demonstrated experimentally in p53DB and p53TR domains. Moreover, no evidence that residues with neutral or nearly neutral values of ω (green label) play functionally or structurally important roles in p53 was found.

7.4 Selective Pressures and Mutations Associated with Cancer in p53

Evolutionary biologists maintain that natural selection works in proportion to the number of deleterious mutations occurring in the population (Kimura, 1983). Frequent mutations on residues with relevant functional biochemical roles must be targeted by purifying selection and consequently would be expected to show the highest selective constraints in the protein. On the other hand, sites changing under neutral or nearly neutral evolution will not necessarily compromise major functional roles of the protein, and consequently would rarely be expected to be found associated to disease. Therefore, the pattern in the distribution of mutational frequency against ω values should likely approach an L-shaped curve, although this depends ultimately on the number of residues changing by means of purifying selection and neutral or nearly neutral evolution.

Using the high number of mutations collected at the IARC TP53 database and the ω values computed previously, the results demonstrate that p53 residues fit the predicted pattern (Figure 7.5(a)). As expected, disease-associated mutational hotspots have shown the lowest ω values (ωSLR=0, ωM8≤0.033) observed in the study (Figure 7.5(b)). It is interesting to note that residues that are here considered free of the influence of purifying selection show the smallest number of mutants associated with cancer in the database. Moreover, it is important to emphasize that there were no residues showing high ω values (ω>0.3) that also showed a high frequency of mutations associated with human disease (freq>0.5). The absence of such residues not only supports the evolutionary predictions mentioned above, but it also serves to highlight the power of the SLR and M8 models used in computing ω values.

Residues showing a low ω value together with a low or null mutational frequency suggest the existence of lethal residues in the protein and/or the absence of a complete and a representative sample of all the human mutations in the database. While a solution to this problem is not obvious, the conclusion is that the first hypothesis cannot be so easily discarded. Figure 7.5(b) focuses on the distribution of the 228 residues predicted to be under purifying selection according to the SLR model (red circles, p<95%). 108 of such residues have an ωSLR=0 (ωM8≤0.033), and were conserved in all of the species. Particularly, Phe19, Trp23, Asn345, Leu350, Lys382, Lys386, and Asp391, do not have associated mutations in the IARC database (freq=0, ω=0). Notably, previous data confirm the relevance of such residues. Phe19 and Trp23 make extensive contact with the MDM2 hydrophobic binding cleft (Zhong & Carlson, 2005).

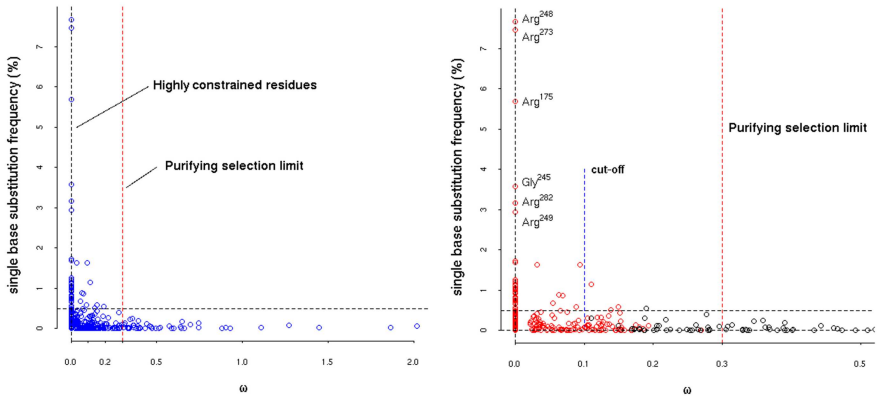


Figure 7.5: Mutation Frequency and ω Distribution in p53 (a) The distribution of p53 residues in the ω-frequency space describe an L-shaped curve where sites under selective constraints (low ω values) are preferentially associated with high mutational frequencies associated to cancer. Conversely, residues above the limit of the effects of purifying selection (high ω) are preferentially associated with low mutational frequencies. (b) Mutational hotspots show high evolutionary constraints imposed by natural selection and the highest mutational frequencies. The cutoff value represents the maximum ω value for which residues were deduced to be under the influence of purifying selection at 95% or 99% confidence using the SLR method. This threshold represents the *a priori* hypothesis used to detect a statistical significance between mutation frequency and ω values using a large set of human disease genes.

Asn345 and Leu350 are two of the four conserved residues in the helix of the TR domain. The acetylation of Lys382 is thought to increase its sequence-specific DNA-binding activity (Gu & Roeder, 1997) as well as protein stability (Ito *et al.* , 2001). Lys386 serves as the major attachment site for SUMO-1, a small ubiquitin-like peptide required for post-translational modifications of p53 (Gostissa *et al.* , 1999), and as far as could be determined no evidence of the functional importance of Asp391 has been published.

If selection has modeled ω values for generations by rejecting deleterious mutations associated with the more frequent disease mutations in the population, one would expect a gradual increase in the selective constraints in p53 associated with the more common cancer mutations. Table 7.2 shows the total number of mutations, residues (in parentheses), and the mean number of mutations per residue (numbers in bold) for the p53DB and p53TR domains, computed according to M8 and SLR models. The values were arranged according to eight different ω categories (columns): the four previously defined in the coloring scheme (green, yellow, orange and red, ranging from neutral or nearly neutral evolution to highly constrained values) and another four categories. These last show residues where purifying selection was statistically deduced using the SLR method (SLR*), ω values observed in mutational hotspots in both the M8 model ($\omega M8 \leq 0.033$) and the SLR method ($\omega SLR = 0$), and finally, a category collecting phylogenetically conserved residues throughout the alignment (PC). In agreement with the above-mentioned expectation, the mean number of mutations per site show a gradual increase according to the strength of natural selection: from 9.5 to 87.7 depending on whether residues have evolved under neutral or nearly neutral conditions ($\omega > 0.3$), or if they have been under the effects of strong purifying selection ($\omega < 0.1$). The pattern is consistent throughout the whole protein, and is independent of the method used to estimate the ω values. The only exception occurred when considering the category where $0.1 \leq \omega < 0.2$ under the SLR method, but it seems justifiable given the low number of mutations observed in the p53TR domain. As expected, the category where residues are phylogenetically conserved (PC) showed the highest rate of mutations per site estimated for the full protein (120.6), which was very close to that estimated by SLR* (120.3) and $\omega M8 < 0.033$ (119.5). A detailed analysis of these three categories demonstrates that, with the exception of Ser241 ($\omega SLR = 0.063$, $\omega M8 = 0.078$) coded by the TCC codon in most of the species and by AGC in zebra fish and catfish, residues with an $SLR = 0$ and an $\omega M8 \leq 0.033$ are those that are phylogenetically conserved. Conversely, the only residue with $\omega M8 \leq 0.033$ not conserved in the alignment was Asn235, which changes to Lys in rodents.

Finally, SLR* seems to be the more informative category when deducing the selective constraints imposed on the protein, since it contains all of the PC residues and shows that, in total, twice as many as those that are phylogenetically conserved are selectively constrained. In addition, SLR* seems to possess the ability to detect a greater number of mutations associated with disease per residue (87.0) for a greater proportion of residues (228).

p53	Mdl	ω Rng3	ω Rng2	ω Rng1	ω<0.1	SLR*	ωM8	ωSLR	PC
Full protein	M8	570	382	1714	15165	16883	13028	12992	13152
		9.5	13.2	35	87.7				
	SLR	430	250	1495	15656	87	119.5	120.3	120.6
		8.6	11.4	25.3	87				
DB	M8	437	337	1669	14814	16471	12998	12952	13112
		23	25.9	50.6	113.1				
	SLR	306	223	1436	15292	99.2	139.6	140.8	141
		20.4	31.9	36.8	113.3				
TD	M8	8	7	12	152	164	30	30	30
		2	2.3	2.4	7.6				
	SLR	6	8	6	158	6.3	5	4.3	4.3
		2	2.5	1.5	7.5				

Table 7.2: Summary of the Number of Mutations, and Mutations per Residue (bold), in p53 and 2 Domains Evaluated Under Different Categories of Selective Constraints Note that phylogenetically conserved (PC) residues show higher values of mutations per residue (120.6 and 141.0 for p53 and p53DB) although SLR* contains a higher number of sites with statistical evidence of strong purifying selection (228 and 166). The increasing number of mutations per residue observed in ranges of ω with higher selective constraints demonstrates that natural selection works in proportion to the number of mutations in the population (see the text). SLR: Residues under the constraints of purifying selection evaluated by the SLR method at 95% and 99% statistical confidence; ωM8: Residues with ωM8≤0.033; ωSLR: Residues with ωSLR=0; PC: Residues phylogenetically conserved throughout the p53 alignment. ω Rng1: ω>0.3; ω Rng2: 0.2≤ω<0.3; ω Rng3: 0.1≤ω<0.2.

7.5 Testing Associations Between ω and Disease in Human Genes

Genes from the immunodeficiency resource (IDR) (Valiaho *et al.* , 2002), the MeCP2 database (RettBASE), and the catalog of somatic mutations in cancer (COSMIC) (Bamford *et al.* , 2004) were analyzed in order to test for the statistical association between ω values and the mutational frequency of residues in different human disease genes. A total of 264 genes were collected in which codons have a variable number of mutations observed in patients. After the alignment of orthologous sequences, two datasets consisting of mammals and vertebrates were obtained. Once genes with a very low or high number of mutations were discarded from the statistical analysis (see Part III: Materials and Methods), 43 genes adding up to a total of 24,375 residues and 8970 mutations remained in the mammalian dataset. The vertebrate dataset included the same number of genes, although it had a smaller number of total residues (˜17,400) and mutations (˜8080).

If selective constraints imposed by PFS are defined independently of codon mutational frequency, a nonsignificant association would be expected between mutation frequency and ω values. This represents the null hypothesis against which to test the prediction, which assumes that residues below a particular ω cutoff are preferentially associated with the highest values of mutational fre-

quency in human disease genes. The threshold that was used to test the hypothesis was deduced from p53 and represents the highest ω value under which all the residues were statistically found to be constrained by purifying selection (ω=0.1, Figure 7.5(b)). Hence, this cutoff can be used as an *a priori* value to test the hypothesis using a two sample K–S test.

Table 7.3 shows the p-values obtained after one-tail K–S tests. Depending on the method used to compute ω values (SLR or PAML: M2a or M8, depending on which model fit the data better on alternative genes) and the ω cutoff used to test for statistical differences in frequencies, the null hypothesis: H_0:f[ω≤cutoff]=f[ω>cutoff], H_1:f[ω≤cutoff]>f[ω>cutoff] is accepted or rejected at different levels of significance. Specifically, using the *a priori* cutoff and PAML estimations in mammals, the K–S test detected highly significant differences in frequencies above and below the selected cutoff (p=3.027×10−05). Alternative cutoffs did not reject the null hypothesis with a higher confidence (lower p-values) in mammals. Within the vertebrate dataset, the highest differences in frequency (p=0.0010) were found at an ω cutoff=0.12. In sum, using PAML ω estimations, a significant statistical differentiation between residues with frequencies above and below a specific ω value was detected using the K–S test. Therefore, an evolutionary parameter that statistically differentiates amino acid residues for which mutations are highly associated to human diseases has been found.

When considering the results of the SLR method on both the mammalian and vertebrate datasets, a statistical differentiation was observed for a wide range of values for the ω cutoff using the K–S test. This seems to be due to the anomalous behavior of the SLR method, which collapses residues showing very low ω values to 0 (associated to high mutational frequencies). This unfortunate property of the SLR method produces a gap of ω values from 0 to 0.02 in mammals and from 0 to 0.003 in vertebrates, leading to a rejection of the null hypothesis in the K–S test. While SLR produces a discontinuous distribution of ω, PAML produces a continuous ω distribution that never reaches 0 (ω minimum=0.025 in both datasets). Although both methods fail at the moment of describing a continuous distribution of ω values ranging from 0 to positive values, only the ωSLR distribution produces a fatal condition to test for our hypothesis. Figure 7.6 shows the ω frequency distribution of residues using PAML and the SLR method in mammals and vertebrates. The zoomed graphs (internal graphs) show the main differences in the ω frequency distributions resulting from PAML and SLR estimations. While PAML retained a high number of residues between ω=0.025 and ω=0.1, the SLR method vanished ω differences for residues in the same area. Such differences in the distribution of residues within the ω frequency space produce a dissimilar behavior at the moment of rejecting the null hypothesis.

Therefore, PAML ω estimations allow distinguishing the best ω cutoff under which mutations are more frequently associated with disease in humans. This value is exactly the same as that deduced in an *a priori* fashion with p53 for mammals, and very close to the value observed in vertebrates. Although SLR

	Mammals		Vertebrates	
ω cutoff	PAML	SLR	PAML	SLR
0.03	0.9748	0.0095	0.0504	0.0061
0.05	0.0114	0.0075	0.0026	0.0008
0.1	3.0×10^{-05}	0.0076	0.0016	0.0009
0.12	0.0007	0.0077	0.001	0.0023
0.15	0.0025	0.0078	0.0012	0.0018
0.2	0.0715	0.0074	0.0019	0.0019
0.25	0.1938	0.0074	0.0044	0.0043
0.3	0.0188	0.0076	0.0035	0.0065
0.4	0.0486	0.0101	0.0176	0.0254
0.5	0.1849	0.0223	0.0534	0.101
G	43	43	43	43
R	24375	24375	17424	17435
M	8970	8970	8081	

Table 7.3: Evaluation of Alternative ω Cutoff Values and Mutational Frequencies in Disease One-tail K–S tests reject the null hypothesis, which considers that the frequency of mutations are not differentially distributed above and bellow the given ω cutoff. The alternative hypothesis, which considers that disease associated mutations are preferentially associated with values below the ω cutoff, is accepted with the highest confidence using ωPAML estimations on mammal (ω cutoff=0.10) and vertebrate (ω cutoff=0.12) datasets. The K–S test on SLR estimates reject the null hypothesis for all values of ω cutoff evaluated. This is the consequence of the undesirable behavior of the SLR method, which drops low values of ω to 0 (see the text and Figure 7.6 for explanation). G: Number of genes evaluated; R: Number of residues evaluated; M: Number of mutations evaluated.

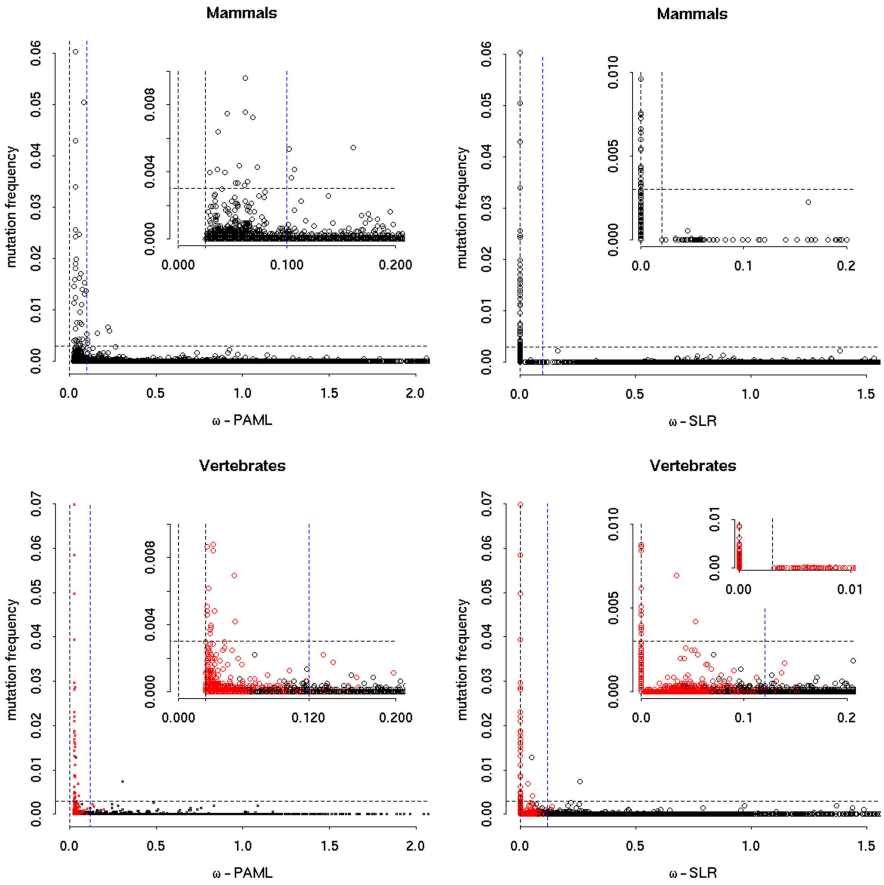


Figure 7.6: Mutation Frequency and ω Distribution in Human Disease Genes Mutation frequency against ω for 43 human disease genes follows an L-shaped distribution. Independent of the number of species used (four mammals, 12 vertebrates), PAML and SLR estimations differ when ω is close to zero. The internal graphics for mammalian and vertebrate datasets show that, while PAML produces a continuous distribution of ω starting from a minimum value of 0.025, the SLR method collapses ω to zero in many of the residues with ωSLR<0.02. This unfortunate behavior of the SLR method removes residues in the area of the distribution where the null hypothesis is rejected using PAML estimations (see Table 7.3). The SLR method detects statistically significant signals of purifying selection in the vertebrate dataset only (red dots).

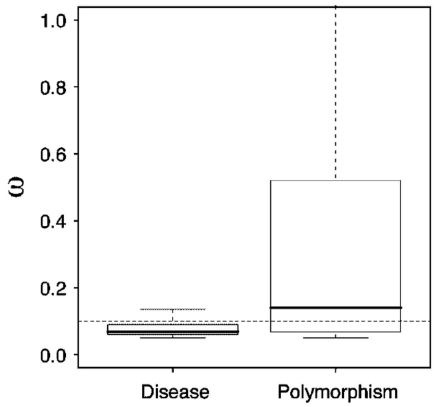


Figure 7.7: ω Distribution for Disease-Related and Simple Polymorphic SNPs at a Large Scale The boxplots represent the distribution of ω values for protein variants with disease related and neutral polymorphisms (box) from the SwissProt database annotation. Their respective means, 0.068 and 0.14, are shown as by horizontal bold lines. The dotted horizontal line is set at the ω=0.1 cutoff as determined earlier. Whiskers delimit 1.5 times the interquartile distance.

ω estimations did not allow for the determination of a clear cutoff value, the method is definitively useful for detecting which residues are under the strict control of negative selection. As can be seen in Figure 7.6, the test has only enough statistical power to detect significant deviations from neutrality when using a high number of species is used (red dots).

Finally, in order to obtain a broader perspective of the association of selective pressures at a codon level and disease, an analysis was performed on annotations of disease related and neutral polymorphic variants available from the SwissProt database (Wu *et al.* , 2006). The annotation available in SwissProt represents one the largest publicly available resources on the annotation of of the phenotypic effects of polymorphisms. Since the previous databases were mostly associated with cancer and immune related diseases, the purpose was to test the broad applicability of the ω cutoff based predictor, discussed above, at a large scale. The hypothesis was that coding nonsynonymous single nucleotide polymorphisms (SNPs) that affect human health in general would again be observed more frequently under strong selective pressures (i.e. below the ω<0.1 cutoff). Figure 7.7 shows that the distribution for values of ω associated to both disease and neutral polymorphic variants which have considerably different distributions with corresponding means (bold lines in boxplots) of 0.068 and 0.14, respectively. The disease related median value is 0.072 lower than that for neutral polymorphisms. This difference, although small, is very significant given a much larger distribution of ω values for neutral polymorphisms (p-value of 2.2e–16). The result indicates that ω values smaller than 0.1 is a general pattern of disease related variants which is indeed broadly applicable.

Chapter 8

Contributed Resources

8.1 Predicting Deleterious Mutations within the Pupas Suite Server

The combined effect of all selective pressures causes the preservation of the functionally relevant parts of genes. Under this perspective, comparative and evolutionary studies have been used to predict the putative functional effect of single nucleotide polymorphisms (SNPs) (Ferrer-Costa *et al.* , 2002; Miller & Kumar, 2001) although these have mainly ignored the underlying phylogeny. This method, presented in Chapter 7, is a more direct estimator of selective pressures that additionaly takes into account phylogenetic information serving as an estimator of functional effect. While these estimates can be reproduced by running a pipeline of programs as described in Section Materials and Methods, a precomputed database of all of the SNPs in human, mouse, and rat was pregenerated and integrated into the PupaSuite web server (Conde *et al.* , 2006).

PupaSuite was developed for the selection of good candidate SNPs for large scale genotyping purposes. The clever design and flexibility of its interface (Figure 8.1) permits obtaining data for a few or single genes, and for all of the genes in selected chromosomal regions. As such it serves as a great tool for querying and accessing the precomputed data produced here. It also has advanced visualization capabilities that allow users to place features at a sequence level and easily compare the information with the results of other methods available in PupaSuite (Figure 8.2). Briefly, the methods in PupaSuite can best be understood in the light of the two main types of studies for which it is tailored: genes probably related to a disease because they are functionally related (e.g. they belong to a pathway affected in the disease), or genes present in a chromosomal region linked to a disease. With this in mind, PupaSuite provides minor allele frequency (MAF) data in different populations from dbSNP data in Ensembl (Wheeler *et al.* , 2006), as well as linkage disequilibrium (LD) parameters and haplotype blocks associated to genes and chromosomal regions. Together these

estimates provide a measure of the recent history of polymorphic variation, and the longitudinal genomic partitions under which patterns of variation are associated. The information provided by the ω parameter is instead derived from the ancestral history for all sites and is thus specifically complimentary. In particular, it sheds light on the magnitude of current variation as compared with the history of variation of sites at a much larger time scale and, in this manner, it also provides information about which polymorphisms in a block where recent variations are associated, are particularly striking, and may be causative. Importantly, PupaSuite also provides estimates of the possible functional effect of SNPs through a combination of methods that include other functional estimators based sequence and structural information (Pmut Ferrer-Costa *et al.* , 2002, 2004; SNPeffect Reumers *et al.* , 2005; Fernandez-Escamilla *et al.* , 2004; Schymkowitz *et al.* , 2005). Since strong selective constraints are ultimately a good indicator of functionality, it is among these methods, that the ω estimator has been included (red arrow in Figure 8.1).

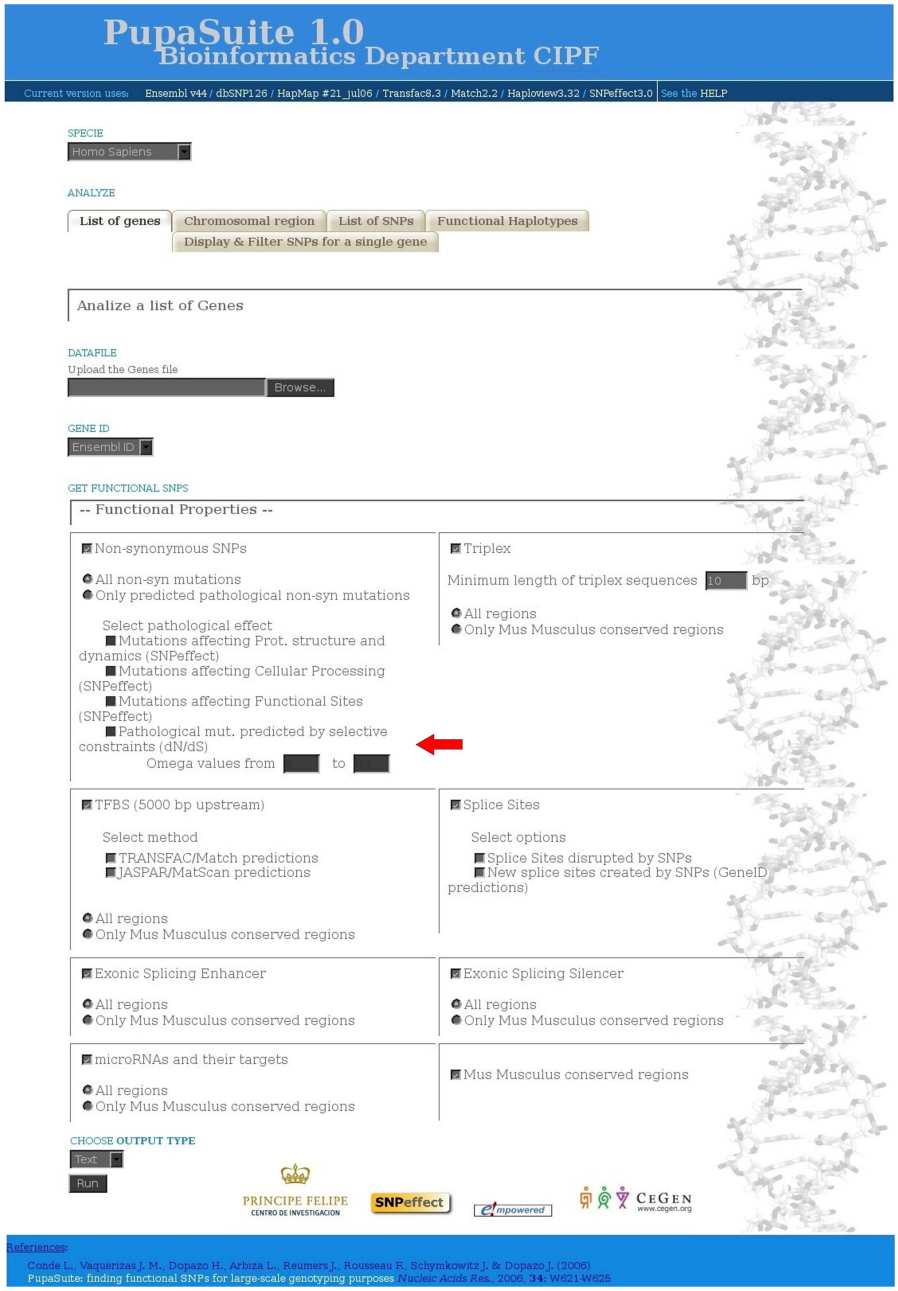


Figure 8.1: The PupaSuite Webserver Interface takes a list of SNPs, chromosomal regions, genes, or genotyping data permitting the discovery of functional haplotype blocks or analysis of all SNPs in a region through a wide range of functional properties that can be selected by the user. The check box labeled "Pathological mut. predicted by selective constraints" (red arrow) together with the ability to select a range in the value of ω to be considered, are the only fields required to obtain the estimates obtained in this work from the database (see Chapter 7). The server is available at http://pupasuite.bioinfo.cipf.es/

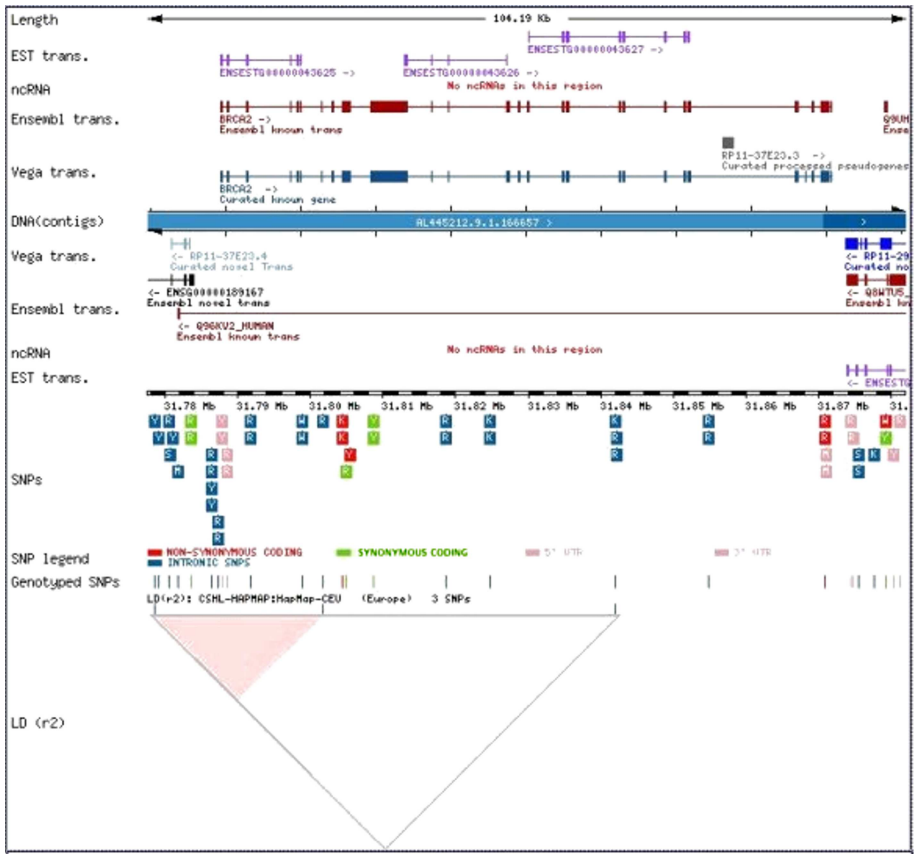


Figure 8.2: PupaSuite Results Different types of SNPs are shown within a genomic context after filtering by the selected constraints, together with the output provided by other programs and genomic estimates. Mousing over ncSNPs or asking for the output in text format reveals the estimates of $\omega$from PAML and SLR methods for each SNP.

As such, PupaSuite is an ideal framework to obtain, apply, and benefit, from the use of selective pressures as an estimator of the putative phenotypic effects of mutations. First and foremost, it provides a simple yet powerful interface. This makes the precomputed data readily available for its application in single gene or genomic scale analysis allowing the retrieval of estimates for any number of genes. The powerful visualization capabilities shown in Figure 8.2, also allow more detailed analysis of results by placing them in a genomic context. In this genomic context, results provided by the combination of methods for functional prediction and other genetic metrics are conveniently laid out for the user facilitating interpretation and SNP selection.

Second, although PupaSuite is a relatively new, it originates and is the latest improvement of two previous servers, PupaSNP (Conde *et al.* , 2004) and

PupasView (Conde *et al.* , 2005), that have been around for more than 3 years, and has had, for example, an approximate average of 60 daily experimental designs throughout 2007 (Conde *et al.* , 2006). Consequently, the integration of the method developed here into this platform allows users to access the information in a familiar environment without the need to learn the details of novel, single purpose frontends.

Finally, since the method itself is geared towards predicting the functional effect of mutations in disease, its integration into PupaSuite is extremely well suited. Particularly, it allows both, for the evaluation of these estimates in the light of complementary methods that are available in PupaSuite, and becomes in itself an additional and novel resource for selecting optimal SNPs. It is particularly relevant when considering that SNPs, which are the simplest and one of the most frequent type of DNA sequence variation among individuals, constitute one of the most powerful tools in the search for disease susceptibility genes, drug response-determining genes, and the like (Collins *et al.* , 2003; Risch, 2000). Moreover, in the selection of an optimal set of SNPs for genotyping experiments (among several thousands of candidates in some cases), algorithms such as those implemented in PupaSuite, which use information to facilitate the posterior analysis of the results, are expected to have a major impact on the efficiency of a large-scale genotyping study (Conde *et al.* , 2006). Indeed, optimal SNPs must be the best possible markers for traits, which often are multigenic, usually reflecting disruptions in proteins that participate in a protein complex or a in a pathway (Badano & Katsanis, 2002), and the consideration of the likelihood of being the causative agent of any type of damage including a consideration of the evolutionary constraints that have marked that position, are among the most relevant guidelines available to date for this task. Indeed, the predicted functional effect of SNPs have been gaining importance as a selection criteria given that this constitutes a potentially important factor for increasing the sensitivity of association tests significantly (Badano & Katsanis, 2002; Botstein & Risch, 2003; Conde *et al.* , 2004, 2005). Importantly, this method was chosen and used by the STAR consortium to analyze the genetic variation in rat based on almost 3 million newly identified SNPs and published in Nature Genetics (Consortium, 2008).

8.2 Testing for Molecular Adaptation and Rate Estimation within the Phylemon Server

While the previous section has dealt with fulfilling the purpose of making the results of the pipeline for computing selective pressures at a codon level available to the scientific community, it also important to consider that slight variations of the methods, or the reproduction of other various methods implemented in other sections of this work or related methodologies, can be extremely useful to users in various fields. In particular since some of the software used here is only available from the authors, the integration of particular programs requires

manual or automated data manipulation of formats to extract relevant results for use among programs. Many of these programs have complex functions that require a compilation of several publications, manuals, and trial and error in order to assure their proper use. As such compiling the methods to build any of the analyses presented throughout this work requires the installation of several programs, dealing with specific platform requirements, the compilation of software from source code, moderately powerful CPUs and memory availability, and a series of other technical details that are likely to become overwhelming for non expert users. This becomes even more daunting when considering that the analyses of data requires previous preparation through an additional pipeline of tools required to generate sequence alignments and phylogenetic trees, format their output for use in each of the different analyses, and being able to obtain other evolutionary metrics that may be important in order to asses the applicability of particular methods.

The Phylemon server (Tárraga *et al.* , 2007) is an online platform for phylogenetic and evolutionary analyses of molecular sequence data. It has been developed as a frontend that integrates a suite of different tools selected among the most popular standalone programs in phylogenetic and evolutionary analysis. It has been conceived as a natural response to the increasing demand of data analysis of many experimental scientists wishing to add a molecular evolution and phylogenetics insight into their research. Tools included in Phylemon cover a wide yet selected range of programs: from the most basic for multiple sequence alignment to elaborate statistical methods of phylogenetic reconstruction including methods for evolutionary rates analyses and molecular adaptation. Figure 8.3, shows some of the important features in Phylemon. It has several that differentiate it from other resources: (i) It offers an integrated environment that enables the direct concatenation of evolutionary analyses, the storage of results and handles required data format conversions transparently, (ii) Once an outfile is produced, Phylemon suggests the next possible analyses, thus guiding the user and facilitating the integration of multi-step analyses, and (iii) users can define and save complete pipelines for specific phylogenetic analysis to be automatically used on many genes in subsequent sessions or multiple genes in a single session (phylogenomics).

With this in mind, all of the computational resources have been made available that allow both, preparing the data, and running all of the different parts of the analyses mentioned in this work, through the incorporation of the required software into the Phylemon webserver platform designed by Tárraga *et al.* 2007. The RRTree program (Robinson-Rechavi & Huchon, 2000), has been made available for running relative rates tests as part of the methods employed in Chapter 4. The Codeml program from the PAML package (Yang, 1997), for the estimation of rates through the ML methods, and running branch, site, and branch-site models for the detection of positive selection, relaxation, and other evolutionary hypothesis on coding nucleotide or amino acid sequences, has also been made available. Many of the tests for positive selection in the Hyphy Package (Pond *et al.* , 2005) have also been made available. While in some case the analyses

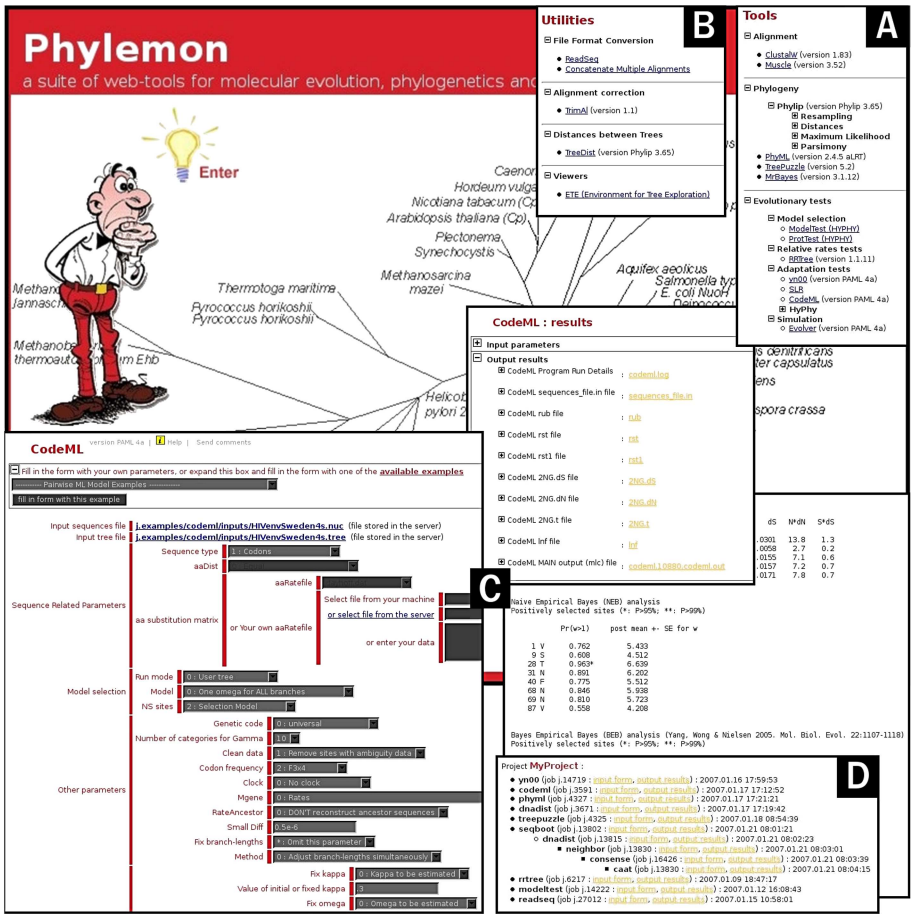


Figure 8.3: The Phylemon Web Server has been used as the integrated framework of choice for making all of the tools used in this work available publicly, requiring only a java enabled web brower. A) shows the Tools page of Phylemon where a list of available software for phylogenetic and evolutionary analyses is shown. B) The Utilities section, shows some of the additional software provided for data manipulation and other purposes. C) From left to right shows: the main interface for the Codeml program from the PAML package, the results page with all of the output files, and part of the Codeml output. D) The project view where users can access all of the analyses that they have run previously and can retrieve or select previous results as input to other analyses.

available here, have some overlap with those in Codeml and SLR, mentioned earlier, they do not only serve as verification, but provide many novel capabilities that are not available elsewhere. To mention just a few of these, variation in synonymous rates, automatic checks of model convergence, and the use of alternative markov models in combination with different models of nucleotide substitution, are available in Hyphy, among many others. The site-wise likelihood ratio test (Massingham & Goldman, 2005) for running site specific tests of neutrality, positive, and negative selection, used in Chapter 7 of this work has also been included. The program yn00 of the PAML package which implements the *ad hoc* method of Yang and Nielsen (Yang & Nielsen, 2002) for estimating synonymous and nonsynonymous substitution rates in a pairwise fashion was added as well.

The Evolver program from the PAML package (Yang, 1997), has also been included. Its main purpose is the simulation of datasets under specific evolutionary models and serves as a necessary tool in being able to exploit the full potential of the flexibility of the models included in Codeml and Hyphy where tests can be conceived and run, whose appropriate statistical behavior has not been tested and described elsewhere. More precisely, as stated in the chapter on the estimation of selection pressures on alignments of codon sequences of the phylogenetic handbook (Pond *et al.* , 2009), the use of any test requires a prior evaluation of the statistical properties and parameter estimates for particular testing procedures. For this reason it is crucial, to be able to generate sequence data that has evolved according to known parameter values so that at the very least, the procedure should be consistent, in being able to recover correct parameter estimates, and efficient, in providing these with an acceptable accuracy. Additionally, when testing the theoretical cutoffs employed when comparing the fit of models in a test to compare specific hypothesis, these should be evaluated to ensure that specific type I and type II error rates observed fall within those expected by theory.

Additionally, accessory scripts for the generation of protein based codon alignments, which can optionally be gapless as required for some of the analyses in Codeml, and for the concatenation of several alignments into one larger dataset for analyses, have also been included within the utilities section of the phylemon webserver. Together with the tools that were previously part of phylemon and that allow multiple sequence alignments (Muscle (Edgar, 2004), and ClustalW (Thompson *et al.* , 1994)), the not yet available but scheduled incorporation of phylomeDB for the retrieval of various kind of homologous sequences from more than 35 eukariotic species, together with tools for format conversion, pipelining mutli-step analyses, and other phylogenetic and molecular evolutionary analyses software, the methods used in this work are made available in one of the most complete and integrated environments publicly accessible requiring only a java enabled web browser. Furthermore, all of the programs incorporated have been provided with examples of the most common analysis, permitting users to have a working example of many times complex configurations that can act as barriers when first seeking to understand their functionality. Finally,

in order to provide some idea of the use of this software, Codeml is the third most used program, just under MrBayes (Huelsenbeck & Ronquist, 2001) and Phyml (Guindon & Gascuel, 2003), in Phylemon which in itself has had an average of 30 analyses run per day (May-July, 2008). The programs included as part of this effort to bring these tools closer in availability and facilitate their use by the community also constitute and important portion of all the software available in Phylemon.

Part V

Discussion

Chapter 9

Summarizing Discussion

9.1 Test Based Inferences of Natural Selection

For years evolutionary biologists have been interested in knowing to what extent natural selection and genetic drift have shaped the genetic variation of populations and species (Gillespie, 1991; Kimura, 1968, 1983; King & Jukes, 1969; Li, 1997). Neutrality tests have provided powerful tools for developing hypotheses regarding this issue. The first objective of related studies had been to make general inferences about the causes of molecular evolution, and many efforts have been made to search for deviations from the molecular clock hypothesis. However, in the past ten years the focus has changed towards finding molecular events showing positive selection (PS) (Nielsen, 2001).

Since the sequencing of the human and the chimp genomes, the grounds have been set for addressing one of the major challenges in evolutionary biology: namely, the search for positively selected genes that have shaped the differences among us as humans and our closest living evolutionary relative, the chimpanzee. With the growing framework available for comparative genomic studies, it has been possible to test for neutrality against positive (or negative) selection at a genomic level. Indeed, efforts at a large or a complete genomic scale have been conducted to search for positively selected genes in human and in chimp. However, recently developed methods allowing for a more sensitive and controlled approach in the detection of positive selection can be employed to answer many important questions that have had only partial answers. Chapter 4, presents a complete genomic evolutionary analysis of molecular clock, relaxation of constraints (RSC), and positive selection (PS) of protein coding genes in human and chimp. It is the first study where the differentiation of likely events of RSC has been addressed when using branch-site tests for precision and comparison to detect events of PS. As such, a characterization of the frequency of these events together with how these phenomena are related at a genomic scale, how these processes differ within species, and how they vary among them, is considered. Additionally, a comparison with the ancestral lineage of hominids

is addressed in order to differentiate adaptive trends in evolution after the speciation process differentiating human and chimpanzee. Methodologically, specific tests for acceleration, as well as different branch site tests of positive selection are employed in order to determine which sets of genes can be deduced under each of these processes.

Based on testing deviations from neutrality in a gene-by-gene approach, a total of 1,182 (9.0%) human and 1,948 (14.8%) chimp genes with statistically significant deviations were observed in at least one of the mentioned processes. However, after correcting for multiple testing only 665 (5.0%) human and 1,341 (10.2%) chimp were considered as a better estimate of the minimal sets under non-neutral evolution in these species. From this perspective, one could thus conclude that these non-neutral evolutionary processes do not show signs of being frequent events shaping the pattern of divergence between human and chimp genomes. Whole-genome analyses of evolutionary properties were made without any *a priori* hypothesis about the resulting genes, and consequently the analyses presented here are exhaustive and at the same time conservative regarding individual results. However, the necessity of keeping the type I error rate at an acceptable level leads to an unavoidable increase in the rejection of true positive results (Ge *et al.* , 2003). Therefore, the complete sets of accelerated and PSGs reported here can only be considered their respective most significant parts.

Differences in evolutionary rates exist between the species although there were no net significant differences. The number of genes showing a significant acceleration in nonsynonymous rates exceeds those evolving by synonymous changes, and is greater for chimp than for humans. This excess of nonsynonymous changes favoring chimp correlates with the greater number of PS events observed in this species, and could be due in part to the comparatively smaller population size that has shaped human evolution (Chen & Li, 2001). Nevertheless, the difference is important and suggests that chimp has shown a frequency of PS and RSC larger than that in human. The results obtained in Chapter 4 of this work, can be contrasted with those in Chapter 6. In the earlier case, the analysis was made with the first assembly of the chimp genome, which had a considerably lower coverage and higher contig fragmentation than those of human and posterior chimp assemblies. Although in the later chapter a considerably lower number of genes is used due to the required availability of expression data, two important improvements are included: the alignments were manually inspected to discard obvious cases with errors, and the second chimp assembly with increased coverage and less fragmentation is used. In this more controlled approach, the results again suggest a similar although less elevated difference, maintaining that indeed chimp has undergone more PS in comparison with human. Short after the work in the first chapter was published and the second chimp assembly was made available, the study by Bakewell et al. (Bakewell *et al.* , 2007) citing the results here, employed an ingenious strategy to confirm this assertion. They found that PS events in chimp where 51% more frequent than in human which is quite similar to that obtained in Chapter 6

using the same chimp assembly (˜43%, see Table 6.2). Using sequence quality scores for individual bases to partition the analysis into three increasing quality sets, and by downgrading the human sequence quality through the introduction of random errors to one that was comparable with that of the available chimp assembly, they provided further support for this conclusion.

The comparison of the different sets of genes under different evolutionary processes is also illustrative. For years, evolutionary biologists have known that deviations from the molecular clock, or rate acceleration in general, are not necessary, nor sufficient, to infer adaptive processes occurring during evolution of species. The comparisons shown here highlight that a consideration of genes with a Ka/Ks > 1 yields a set where only 7%–20 % of genes show evidences of PS. Similarly, using a relative rates test (RRT) approach on nonsynonymous mutations, those showing significant deviations are enriched for PS events from 10%–30%. With the addition of a minimal divergence value in nonsynonymous rates (dKa > 0.0006) the number of genes is reduced considerably, but PS events reach a concentration of 80%–95%. However, in all of these cases a high proportion of positively selected genes (PSG) are discarded in comparison with the number of PS events found by using the maximum likelihood (ML) branch-site models of Test II used in this study. These results serve to highlight a couple of important considerations. First, they emphasize one of the downfalls of using elevated normalized Ka rates as a means of concentrating likely cases of PS in an *a priori* fashion. Second, they serve to show the importance and contribution of the approaches that have been recently developed, and are employed here, which are more sensitive as they make more realistic assumptions about the nature of events of positive selection. While many advances have been made in the past decades to build probabilistic models that take into account a phylogenetic framework, and permit testing for events of adaptation that can take place at varying numbers of sites within coding sequences, many studies still employ simple pairwise comparisons averaging rates along the whole sequence in their analysis. While these earlier methods are useful for obtaining a quick and approximate glimpse of extensive patterns affecting a high number of residues of sequences, these results show, that in order to obtain a better notion of the frequencies of the neutral and non-neutral evolutionary processes, the choice of methodology can have an important effect on conclusions. This is not only evident from the lack of overlap among the different sets in numbers, but is also evident when comparing their functional representation as noted in the results and in the next section.

9.2 Natural Selection and Biological Function

Darwin's concepts of natural and sexual selection, are still to date the best fitting and among the best available explanations for the existence of adaptation. Understanding the degree to which non neutral evolutionary processes have

played a role at the genome level in contrast to neutral evolutionary processes, can provide a deeper understanding of the forces and events that have shaped human evolutionary history. However, and quite importantly, the signatures of rate acceleration, relaxation of selective constraints (RSC), and more importantly positive selection (PS), are also important beacons that signal regions within the genome that are, have been, or are no longer functionally important. Having identified the most robust set of genes under these processes, a natural question is to ask which are the organismal functions they have acted on.

For this purpose, the annotation of the biological functions of genes through Gene Ontology (GO) (Ashburner *et al.* , 2000) are employed. This controlled vocabulary, has many properties that make it flexible and one of the best available means in order to analyze the functional representation of sets of genes. It describes the biological function of genes permitting the realistic ability of having multiple annotations of the various functions of genes all related to each other by levels in a directed hierarchy across organisms. Indeed, GO has had a wide acceptance and has been used in many previous studies in order to assert that certain biological properties have been enriched for cases of positive selection (Clark *et al.* , 2003a; Nielsen *et al.* , 2005; Sequencing & Consortium, 2005). However, as we can observe from the results, there are important considerations regarding both, the methods used to infer positive selection, and statistical approaches used to compare functional representation, that are critical and must be treated with care before hoping to draw robust conclusions in this regard.

A previous genomic study focusing on PS in human and in chimp has found that many functional categories were over- and under-represented in both species (Clark *et al.* , 2003a). This was in disagreement with the results obtained in a posterior study (Sequencing & Consortium, 2005) where only one GO category (*developmental regulators*) showed a possible over-representation in human in relation to chimp. From the results obtained here (Chapter 5), the possibility that the results of Clark *et al.* , 2003a were either likely to contain false positives involved in RSC, or had RSC and PS correlated, is proposed. The results here tend to agree more with this last study, providing evidence for the lack of differentiation in functional classes of PSG in human and in chimp when using this kind of approach for functional inference. They also support the notion that Clark's results may have included cases of RSC given that the model 2 based test used in that study, is very similar to Test I used here, and that many of the deduced classes are here observed with a marked presence under RSC (i.e. *G-protein coupled receptor* and *sensory perception*). However, a probable correlation between PS and RSC could not be discarded since highly represented functional classes under one of the processes are also highly represented in the other.

The sets of genes deduced without correction for multiple testing in molecular clock and PS analyses produced similar results for most of the GO representation comparisons observed after corrections. The only exception was the term *G-protein coupled receptor protein signaling pathway* found to be additionally over-represented in human in relation to chimp under PS (Test II, $p = 0.005$).

As previously mentioned, after correction for multiple testing no GO terms were found as being over- or under-represented through this approach between both species. However, if differences between human and chimp are considered as independent trends evolving from the ancestral condition, a certain pattern seems apparent (although ancestral and descendant differences were not statistically significant). That is, the results show that a relative increase of PSG occurred in human for 41 out of the 59 GO categories common to all of the lineages, while only 11 showed a relative increase in chimp even though PSG in human are approximately six times less frequent than those in chimp. Although further studies would be required, this might suggest that in at least common functional GO classes, human has grown further apart from the ancestral lineage than chimp has through adaptive evolution. Finally, since most of the PSG are different between both species, the individual roles of the alternative PSG found associated under the same functional categories may be an important factor underlying biological differences between human and chimp.

Given that patterns of genetic variation are marked by the interplay of natural selection and genetic drift in finite populations, tests of neutrality have become one of the most appropriate methods for distinguishing cases of positive selection. All tests, make assumptions and may or may not be free of the evidence of confounding factors that come to play in the models that are used to contrast these two alternative evolutionary scenarios. As mentioned in the introduction, the ML tests used here circumvent certain important problems associated to population demography and are among the best alternatives available for obtaining convincing evidence of positive selection. However, as is evident from the results, the consideration of alternative approaches to stud functional implications cannot only circumvent problems associated to threshold-based approaches requiring multiple corrections for multiple testing (two step approaches), as those used in the first sections of Chapter 5, but can also produce important differences in results.

More precisely, the rank based partition test implemented in the FatiScan program (Al-Shahrour *et al.* , 2005a, 2007) can be used directly to rank all of the genes that can be tested at a genomic scale to perform an analysis of asymmetries in the distributions of the functional classes themselves, according to evidence of PS, and in a threshold free manner. For this purpose, the 2Δ statistic, which had been used originally in test II to determine cases of PS by comparison to the appropriate cutoff and then correcting for multiple testing, is now used to rank the genes. The importance of this change in methodology becomes evident when considering that different authors (Clark *et al.* , 2003a; Nielsen *et al.* , 2005) using the classical threshold-based two step approach, claimed to have found GO categories significantly over-represented among positively selected genes in humans. Nevertheless, if the associated p-values were adjusted for multiple testing, as can be observed from the earlier results in Chapter 5, the GO categories previously found do not show significant differences. However, when switching to the more efficient methodology of using 2Δ to rank a genomic set of genes according to evidence of PS and testing with

the FatiScan threshold-free test, many functional classes are seen to have distributions biased towards higher evidences of PS. The classes shown for human and chimp, show significant biases in categories that could not be appropriately determined in previous studies, and in others that have previously not been reported. The patterns observed in these results, also suggest a different perspective than the earlier results. Approximately 30% and 60% of the biological classes showing significant skews towards the extremes of values representing evidence of PS under human and chimp, respectively, are not common to both species.

Finally, it is also important to mention that this change in methodology also represents an important change in perspective. The hypothesis tested is now not about individual genes, but about functional classes. Mutations occur on single genes but natural selection acts on phenotypes by operating on whole sub-cellular systems. Mutations in genes either remain finally fixed or disappear because of their beneficial or disadvantageous effect, respectively. This effect on the function of individual proteins can only be understood in the context of the system (e.g. a pathway, GO functionality, etc.) in which the proteins are involved. If a list of genes arranged by some parameter that accounts for the evidence of PS is examined, it is to be expected if genes under functional classes were favored or disfavored by selection, they will tend to appear towards the extremes. Being able to test natural selection at the scope of functional classes, from a system's level perspective, can thus be a crucial and enlightening methodological improvement.

9.3 Natural Selection at the Organ System Level

Thus far we have examined the interplay of natural selection and function both, by asking questions about the functional roles of genes showing signals of natural selection, and also, by extending the perspective to test evolutionary hypothesis at the scope of functional classes themselves. In so doing, we have discussed that natural selection, although ultimately acting by fixing, maintaining, or eliminating mutations at the genetic level, its scope of action can be perceived from the more global point of view of a biological system perspective. However, the precise definition of functionality as such, is elusive. What is clear is that it expands the range of levels in the scale of organismal functions: from the genome itself to the plethora of products and their emerging interactions, and from the molecular scale to higher organismal system's levels. Accordingly, one of the most intriguing questions to date, which has received much attention in the past years, deals with possibly one of the furthest extensions of the relation between the role of forces acting on higher levels of organismal system functions and their role on molecular adaptation: the phenomenon of humanness. A characteristic that undoubtedly finds its strongest basis in one of the higher system levels. Namely, at the system's organ level, and in particular, in the human brain.

Indeed, this question has been a subject of much interest and debate in the past few years where several studies have focused on the rates of evolution of human brain specific genes. Some concrete example cases of positive selection have been found in genes that are known to affect interesting properties in the development of this organ (Mekel-Bobrov *et al.* , 2005). However, even though some studies have claimed observing wide spread traces of the evidence of molecular adaptation in genes belonging to this system level structure, the conclusions from these studies have varied, and have at times been contradictory and controversial. At the same time, as noble as the pursuit might be, it seems particularly inappropriate to focus just on brain, undermining the other relevant implications that changes due to the action of natural selection at other organ system level components could provide if they existed. As such, the original question had still remained: Has adaptive molecular evolution operated on the set of genes coding for function at the organ's system level, or is the phenotype we observe due to other targets of action such as the adaptation of a few genes, or perhaps at other lower levels in the orders of biological organization? Undoubtedly, in order to address this question, the proper approach cannot elude an understanding of how the different components at the organ's system level behave, before hoping to draw conclusions about the possible evolutionary pattern of any single one of the components on its own. What happens in other tissues, together with an understanding of the relationship between evolutionary patterns of genes that are more or less tissue specific, are themselves the most important factors to act as controls and points of comparison.

The results in Chapter 6, show that no evidence for a higher dN/dS ratio in primate than in murid in brain specific genes can be observed. In contrast, using the genes and methodology used by Dorus *et al.* 2004, there appears to be a difference. An important difference between this study, and some of the previous ones, is that the estimates of the average value of ω here are based on the standard error weighted average. The motivation for doing so is that the variance may be very large for some genes. Since ω essentially is a ratio between divergence in nonsynonymous and synonymous sites, the arithmetic mean will be biased towards larger values for species with short divergence times. Weighting by the estimated standard error will help to alleviate this problem, and will result in an estimator of the average value with lower standard error. An alternative is to use the average dN divided by the average dS as an estimator like in Dorus *et al.* 2004, although it should be noted that this estimator is not an estimator of the average value of dN/dS = w, but an estimator of the ratio of averages. As with the simple means estimator, it will also be biased, especially if the variance of dS is high (see e.g. Rice, 1994, pp. 208). However, it is clear that both the method used to combine dN/dS estimates, the estimator of dN/dS, and the particularities of the genes included in the analysis influences conclusions regarding comparisons of rates of evolution among different categories of genes.

In the study of Dorus et al., housekeeping (HK) genes were used as a point of comparison. A simple inspection of the remarkably low rates of HK genes as compared to other tissue categories suggests that this group of genes is not a

suitable point of comparison when hoping to draw conclusions on the evolutionary patterns of TSGs in general. Moreover, HK genes are quite the opposite of TSGs as they are expressed ubiquitously and their rates are influenced by the additive effects of high expression breadth. Even when excluding HK genes as the reference for comparison, it should be noted that the difference observed by Dorus et al., is not larger than what is observed when comparing against many other tissues, such as in *Testis*, *Liver, B-lymphoblasts*, and *Other*. When considering a correction for lineage effects, none of the classes show significant differences within or among primate and murid lineages (except *Other* and *Brain* which show differences associated to an elevation of ωin chimp). No method or set of genes seems to indicate an acceleration of evolution specific to the human lineage, and what is more, brain specific genes actually represent one of the most notable patterns among all TSG categories by having the absolutely and comparatively most significantly low rates. As such, the evidence for accelerated evolution in brain specific genes during human evolution is absent at the moment, and the question to address is why brain specific genes have such remarkably low rates. Certainly all of this evidence serves to illustrate that the remarkable functional changes in the human brain have not been followed by a general acceleration of brain specific genes. However, this does not preclude positive selection acting on these genes as it does not preclude that a smaller subset of coding or non-coding loci have experienced accelerated evolution. Indeed, 4 genes are seen as showing evidence for positive selection in human brain TSGs (as well as others with brain associated functions in Chapter 5), but these are less than half as many as in chimp (9 brain TSGs), and not particularly elevated when compared to other TSG categories. *Other, Bone Marrow, Pancreas,* and *Muscle* show higher percentages of genes under PS.

9.4 Natural Selection and Disease

Selection against deleterious mutations (purifying selection) is accepted by most evolutionary biologists as the predominant form of selection at a molecular level. Earlier attempts at predicting functional consequences of nonsynonymous mutations (Sunyaev *et al.* , 2000; Guerois *et al.* , 2002; Ng & Henikoff, 2001; Miller & Kumar, 2001; Mooney & Altman, 2003; Chasman & Adams, 2001) represent indirect approaches for evaluating the strength of natural selection acting on polymorphic variation. Structural information alone, with or without the assistance of protein sequence alignments, was used in these methods as a multivariate proxy to deduce if a particular coding single nucleotide polymorphism (cSNP) produces a deleterious change in phenotype. The methods used here differ from previous approaches through the explicit definition of the selective strengths occurring at a codon level, without the assistance of other physicochemical parameters, and based exclusively in methods of comparative genomics.

Previous attempts using evolutionary approaches make use of straightforward and phylogenetic tree-independent parameters, do not distinguish among orthologous and paralogous relationships, or cannot be easily applied to more than a few genes. Some of the parameters used by these methods include conservation of residues in an alignment or the conservative/non-conservative nature of amino acid substitutions. Although both properties are indeed considered in the estimations of the ω values when using ML methods, many other parameters are taken into account when using codon-based maximum likelihood (ML) methods. In particular, ML methods consider a transition probability matrix specified by a relative substitution rate between codons (which are differentially affected by the transition/transversion ratio and the frequency of codons in the alignment, depending on whether DNA codon changes are synonymous or nonsynonymous), and the relations of sequences in a specific phylogenetic framework. Such codon-based models seem to be sufficient for defining which of the sites are functionally relevant in the p53 protein. This evolutionary approximation accurately detects which residues the action of natural selection, working throughout millions of years, seems to have acted upon with more or less strength. The method detected a high number of residues under the influence of purifying selection, which cannot be detected by previous methods evaluating conservation, or lack thereof, solely through inspection of sequence alignments.

Since the methods used here rely on different assumptions made within the models employed, they have a number of limitations that are worth considering. One of the most important is the assumption that protein function has been maintained during the evolution of the species compared. Indeed it is important, at the moment of performing the tests, to maintain the correct comparative framework in order to avoid including orthologous sequences where changes in gene function have occurred. The inclusion of orthologous proteins with different biological functions in the alignment could give rise to incorrect estimations of ω, since a change of function frequently compromises the change of constraints by means of amino acid changes from the ancestral to the derived function in descendants.

Codon-based ML models employed here make use of a number of parameters representing the more frequent changes occurring in the sequences during evolution. Recent evidence pointing out that synonymous substitution can be the object of selection represents a challenge for the development of more accurate methods seeking evidence for selection (Chamary *et al.* , 2006; Hellmann *et al.* , 2003; **?**). Until an improved version of these models is completely defined, the approximation used here represents one of the most accurate definitions we have to detect natural selection at a codon level.

The use of disease-related databases requires the consideration of a trade-off. A smaller number of genes with high confidence in the values of mutation frequencies were used, rather than using a high number of genes representing human diseases with low confidence in codon-mutation representation. While the first decision would produce a bias in the scope of our conclusions for all

human diseases, the second would produce an increase in type II error at the moment of testing the hypothesis. That is, without confidence in the quality of the frequency data, it is more probable to accept the hypothesis when it is false. COSMIC and IDR databases contain the adequate data to evaluate the hypothesis posed here because mutations have been scanned for in different patients for each of the genes. The study was additionally extended through the use of annotations of disease-related and non-disease-causing polymorphisms available in swissprot. This shows that the original hypothesis tested is not limited to cancer and immunodeficiency genes, but is instead a general pattern of human diseases.

In summary, an evolutionary parameter that allows the differentiation of codon sites where human disease is frequent is described. The underlying hypothesis being that nonsynonymous changes of amino acid residues showing $\omega<0.1$ probably affect the normal function of proteins. In so doing, the relation between patterns of natural selection and phenotypic outcome has been addressed, which is importantly one of the most relevant associations sought after when seeking to understand the effects of forces that have marked evolutionary history. Unfortunately, the results apply mostly to Mendelian diseases. Some complex diseases are included within the SwissProt dataset analyzed. However, more extensive data is needed in order to explore the relationship between selective pressures at a codon level and complex functions or disease. Speculatively, selective pressures at a codon level provide a finer-grained evolutionary perspective on the internal structure of molecular actors from the point of view of their most basic building blocks. As such, it also likely to be a key perspective for the visualization of the projection of the action of natural selection on aspects of the phenotype that are governed by system level functions within its ultimate obligate plane of action: the genotype. A projection, that tracing back down through the levels of biological organization and their interactions, is most reasonably spread beyond the boundaries of genes, to the different sites or collections of sites, that conform the "metagenes[1]" of these complex functions.

9.5 Contributed Resources

Throughout the development of the work discussed herein, a collection of methods has been employed for exploring and contrasting different evolutionary hypotheses. In most cases the analyses have involved multi-step approaches drawing from the use of several programs running on computational clusters. They have required carefully developed bioinformatic pipelines able to handle input output format conversions, data manipulation, and a series of technical skills together with a certain degree of familiarity with informatic platforms. In

[1]The term metagenes is classically defined in as collections of genes that show similar patterns of expression. Here it is used to describe the notion that collections of sites or groups of sites spreading beyond individual genes compromise the genomic locations coding for higher system level / complex function.

many cases these requirements may be limiting and possibly overwhelming for non-expert users or those wishing to perform several computationally intense analyses.

Indeed, this is a general trend with many bioinformatic analyses, and it is thus not surprising that web servers providing functional and evolutionary tools are among the most often used methodologies in laboratories working in functional, comparative, and structural genomics. As a whole, they provide a direct means for addressing several evolutionary questions, ranging from the computation of multiple alignments and neighbor-joining trees, to the latest analyses for detection of positive selection from molecular data. A great variety of these resources exist (an example of some of these can be found in Felsenstein's Page at http://evolution.genetics.washington.edu/phylip/software.html), but few have been developed as integrated suites where tools for running a wide array of analyses, as well as confirming and testing the robustness of results with the most relevant alternative approaches, are available.

As such, in lieu of providing separate online frontends for the methods that have been developed in this work, or only for those developed by other authors but that were previously unavailable online, they have been integrated instead, together with additional complementary programs, into two online software suite platforms : the PupaSuite (Conde *et al.* , 2006) and Phylemon webservers (Tárraga *et al.* , 2007) . In this manner, all of the methods used here can now be run, or retrieved directly, from single frontends providing a seamless integration of programs. This also provides the additional resources that are necessary for both, obtaining and preparing the data necessary to run them, as well as testing a wide range of related hypotheses, or verifying results with a wide array of alternative and/or complementary methods.

9.6 Further Considerations

For years it has been thought that the availability of the chimpanzee genome sequence and its comparison to that of human would reveal some of the molecular bases underlying the observable differences and possibly provide clues to that which makes us human. Certainly, estimable and tentative as this pursuit might be, it is evident that, as might have been expected, the patterns of genetic variation within coding sequences, have not provided a conclusive answer. While some functional classes have been observed here as showing evidence for molecular adaptation, part of the reason for these results may be that neither the methodologies existing nor the detail and quality of the available sequence data and functional annotation have allowed for a conclusive answer. An even more important consideration is that functionally important genetic variability in non-coding regions or other variations not perceivable from this perspective are likely to be at play. Many different types of variation are starting to emerge as possible players underlying the differences between humans and chimpanzee. These changes may be found in: non coding conserved elements of known and

unknown function –i.e. promoters, enhancers, miRNAs, genes showing differential deletions and duplications, copy number variations, differences in mRNA and protein expression patterns, epigenetic mechanisms, repetitive sequences, and genomic organization, among many others (see Kehrer-Sawatzki & Cooper, 2007; Varki *et al.* , 2008 for a more extensive review). This being the case, the growing number of data available situates us at an exciting point in time where we are better suited than ever to address this age old question from the required perspective of a multidisciplinary approach.

However, even though a multidisciplinary approach is likely to be the key for delving deeper into an understanding of humanness, there is still much ground to cover with respect to improvements in methodology, approaches, and availability of sequence data, that far from rendering the search for an understanding how patterns of genetic variation in coding sequences and events of molecular adaptation as demotivating, indicate on the contrary, that there is much to be gained from these types of analysis. It is very important to consider that several aspects still limit the power associated to tests of neutrality as utilized here. First, the power they have in the detection of positive selection grows strongly with the amount of sequences used in the analysis (Zhang *et al.* , 2005), and we have only been able to look at the tip of the iceberg in this respect. Second and similarly, the phylogenetic framework used for comparison is also important. The comparison of sequences with low divergence is problematic, and the species included in the analysis, which act as a references for comparison, can also influence results. With the growing amount genomic sequences available, specially for primates, it is likely that many more genes under PS will be deduced from both, the increase in the number of sequences, and the additional power these provide for comparisons among more primates and closely related species. Third, we have already seen how a slight change in the statistical methodology and perspective used for understanding the possible functional implications of patterns of molecular adaptation can have important effects on the results. Both the increase in amount of functional annotation, together with approaches that incorporate a post-genomic perspective instead of a single gene centered view are likely to provide further insights than those that have already been obtained to this regard.

Also, important improvements in the models and comparisons used to perform tests of neutrality are being introduced that may be able to yield further insights. The models used in this work, although representing the sate-of-the-art developments at the time they were employed, have been surpassed by newer models incorporating less unrealistic assumptions. Models considering heterogeneous distributions for synonymous rates have been implemented (Pond *et al.* , 2005) which represent an important adjustment given that some synonymous sites are known to be under selective constraints – affecting mRNA stability, coding regulatory regions with in genes such as enhancers and promoters, and alternative splicing sites, among many others. Importantly the test II employed here also makes one particularly unrealistic assumption in the null model. It considers that all lineages in the background, evolve under similar rates. While

no available method has addressed this problem, false positives may result from the alternative model having a better fit to the data due to a relaxation of this unrealistic constraint. Considering that ample room is left for improvement in the robustness and power of these methods, that improvements are being implemented at a faster pace, and that the amount of data from which these tests can draw power is growing, we are left at a very exciting point in time where future ventures may still provide quite interesting discoveries with respect to the nature and frequency of natural selection as a force shaping the evolution of our genomes and its functional implications.

Part VI

Conclusions

Conclusions

1. The application of recently developed, more sensitive, and robust methodologies for the detection of positive selection, together with the appropriate statistical considerations, results in important differences in the results obtained regarding the frequency of occurrence of these events and their possible functional implications. Remarkably, no differences are found are found among functional classes represented under positive selection when necessary corrections to the effects of multiple testing are taken into account using the classical two step approach employed in previous studies.

2. The distinction of likely cases of relaxation of selective constraints, from events of positive selection, provides important insights with respect to the identity, frequency, an the functional implications, of non neutral evolutionary processes shaping the evolution of human and chimp genomes. Relaxation of constraints occurs more frequently than positive selection in human and chimp, and may be responsible for some of the previous controversial results claiming an over representation of the functional classes associated to positive selection. Positive selection has been more frequent in chimp protein coding genes than those of human.

3. Approaching the exploration of the possible functional implications of positive selection with alternative methodologies, and through a comparison with the ancestral lineage, shows that many classes are asymmetrically distributed towards higher evidence of positive selection and yields some observable differences in the functional changes common and different to both lineages, respectively.

4. The analysis of selective pressures among different tissue specific gene classes shows that although few differences may be observed when comparing among lineages, the effect of natural selection at an organ system level is not uniform, and some tissues seem to have evolved under more elevated or constrained selective pressures. One of the most striking patterns observed is the significantly strong constraints on brain specific genes, which show no evidence of elevated cases of positive selection.

5. Selective pressures at a codon level map well with functional and structural residues of proteins, and can be used to predict mutations that are likely to have a phenotypic effect resulting in disease.

Bibliography

Adams, J, Crosbie, J, Wigg, K, Ickowicz, A, Pathare, T, Roberts, W, Malone, M, Schachar, R, Tannock, R, Kennedy, J L, & Barr, C L. 2004. Glutamate receptor, ionotropic, n-methyl d-aspartate 2a (grin2a) gene as a positional candidate for attention-deficit/hyperactivity disorder in the 16p13 region. *Molecular psychiatry*, **9**(5), 494–9.

Adato, A., Lefèvre, G., Delprat, B., Michel, V., Michalski, N., Chardenoux, S., Weil, D., El-Amraoui, A., & Petit, C. 2005. Usherin, the defective protein in usher syndrome type iia, is likely to be a component of interstereocilia ankle links in the inner ear sensory cells. *Human molecular genetics*, **14**(24), 3921–32.

Al-Shahrour, F., Díaz-Uriarte, R., & Dopazo, J. 2004. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics (oxford, england)*, **20**(4), 578–80.

Al-Shahrour, F., Minguez, P., Vaquerizas, J. M., Conde, L., & Dopazo, J. 2005a. Babelomics: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic acids research*, **33**(Web Server issue), W460–4.

Al-Shahrour, F., DÃaz-Uriarte, R., & Dopazo, J. 2005b. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics (oxford, england)*, **21**(13), 2988–93. PMID: 15840702.

Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., Mínguez, P., Montaner, D., & Dopazo, J. 2007. From genes to functional classes in the study of biological systems. *Bmc bioinformatics*, **8**, 114. PMID: 17407596.

Amato, F., Warnes, G. M., Kirby, C. A., & Norman, R. J. 2002. Infertility caused by hcg autoantibody. *The journal of clinical endocrinology and metabolism*, **87**(3), 993–7.

Anisimova, M, Bielawski, J P, & Yang, Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular biology and evolution*, **18**(8), 1585–92.

Anisimova, M., Bielawski, J. P., & Yang, Z. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. *Molecular biology and evolution*, **19**(6), 950–8.

Aoki, H., Motohashi, T., Yoshimura, N., Yamazaki, H., Yamane, T., Panthier, J. J., & Kunisada, T. 2005. Cooperative and indispensable roles of endothelin 3 and kit signalings in melanocyte development. *Developmental dynamics: An official publication of the american association of anatomists*, **233**(2), 407–17.

Arbiza, L., Dopazo, J., & Dopazo, H. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *Plos comput biol*, **2**(4), e38.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature genetics*, **25**(1), 25–9.

Ayed, A., Mulder, F. A., Yi, G. S., Lu, Y., Kay, L. E., & Arrowsmith, C. H. 2001. Latent and active p53 are identical in conformation. *Nat struct biol*, **8**(9), 756–60.

Badano, J. L., & Katsanis, N. 2002. Beyond mendel: an evolving view of human genetic disease transmission. *Nature reviews. genetics*, **3**(10), 779–89. PMID: 12360236.

Bakewell, Margaret A, Shi, Peng, & Zhang, Jianzhi. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the national academy of sciences of the united states of america*, **104**(18), 7489–94. PMID: 17449636.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P. A., Stratton, M. R., & Wooster, R. 2004. The cosmic (catalogue of somatic mutations in cancer) database and website. *Br j cancer*, **91**(2), 355–8.

Bearer, E. L., Chen, A. F., Chen, A. H., Li, Z., Mark, H. F., Smith, R. J., & Jackson, C. L. 2000. 2e4/kaptin (kptn)–a candidate gene for the hearing loss locus, dfna4. *Annals of human genetics*, **64**(Pt 3), 189–96.

Benjamini, Y., & Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J roy stat soc*, **B**(57), 289–300.

Benjamini, Y., & Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann stat*, **29**, 1165–1188.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. 2008. Genbank. *Nucleic acids research*, **36**(Database issue), D25–30. PMID: 18073190.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. 2000. The protein data bank. *Nucleic acids res*, **28**(1), 235–42.

Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Graf, S., Hammond, M., Herrero, J., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Kokocinski, F., Kulesha, E., London, D., Longden, I., Melsopp, C., Meidl, P., Overduin, B., Parker, A., Proctor, G., Prlic, A., Rae, M., Rios, D., Redmond, S., Schuster, M., Sealy, I., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Stabenau, A., Stalker, J., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., & Hubbard, T. J. 2006. Ensembl 2006. *Nucleic acids res*, **34**(Database issue), D556–61.

Biswas, S., & Akey, J. M. 2006. Genomic insights into positive selection. *Trends in genetics: Tig*, **22**(8), 437–46. PMID: 16808986.

Botstein, D., & Risch, N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics*, **33 Suppl**(Mar.), 228–37. PMID: 12610532.

Brachmann, R. K., Vidal, M., & Boeke, J. D. 1996. Dominant-negative p53 mutations selected in yeast hit cancer hot spots. *Proc natl acad sci u s a*, **93**(9), 4091–5.

Cassidy, A. J., van Steensel, M. A., Steijlen, P. M., van Geel, M., van der Velden, J., Morley, S. M., Terrinoni, A., Melino, G., Candi, E., & McLean, W. H. 2005. A homozygous missense mutation in tgm5 abolishes epidermal transglutaminase 5 activity and causes acral peeling skin syndrome. *American journal of human genetics*, **77**(6), 909–17.

Chamary, J. V., Parmley, J. L., & Hurst, L. D. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat rev genet*, **7**(2), 98–108.

Chasman, D., & Adams, R. M. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J mol biol*, **307**(2), 683–706.

Chen, F. C., & Li, W. H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *American journal of human genetics*, **68**(2), 444–56.

Chen, M., Kasahara, N., Keene, D. R., Chan, L., Hoeffler, W. K., Finlay, D., Barcova, M., Cannon, P. M., Mazurek, C., & Woodley, D. T. 2002. Restoration of type vii collagen expression and function in dystrophic epidermolysis bullosa. *Nature genetics*, **32**(4), 670–5.

Chen, W. J., Orti, G., & Meyer, A. 2004. Novel evolutionary relationship among four fish model systems. *Trends genet*, **20**(9), 424–31.

Cho, Y., Gorina, S., Jeffrey, P. D., & Pavletich, N. P. 1994. Crystal structure of a p53 tumor suppressor-dna complex: understanding tumorigenic mutations. *Science*, **265**(5170), 346–55.

Ciaccio, P. J., Jaiswal, A. K., & Tew, K. D. 1994. Regulation of human dihydrodiol dehydrogenase by michael acceptor xenobiotics. *The journal of biological chemistry*, **269**(22), 15558–62.

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., C., D., Lu, F., Murphy, B., Ferriera, S., Wang, G., Zheng, X., White, T. J., Sninsky, J. J., Adams, M. D., & Cargill, M. 2003a. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science (new york, n.y.)*, **302**(5652), 1960–3.

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P., Kejariwal, A., Todd, M. J., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., Ferriera, S., Wang, G., Zheng, X., White, T. J., Sninsky, J. J., Adams, M. D., & Cargill, M. 2003b. Positive selection in the human genome inferred from human-chimp-mouse orthologous gene alignments. *Cold spring harb symp quant biol*, **68**, 471–7.

Collins, F. S., Green, E. D., Guttmacher, A. E., & Guyer, M. S. 2003. A vision for the future of genomics research. *Nature*, **422**(6934), 835–47. PMID: 12695777.

Conde, L., Vaquerizas, J. M., Santoyo, J., Al-Shahrour, F., Ruiz-Llorente, S., Robledo, M., & Dopazo, J. 2004. Pupasnp finder: a web tool for finding snps with putative effect at transcriptional level. *Nucleic acids res*, **32**(Web Server issue), W242–8.

Conde, L., Vaquerizas, J. M., Ferrer-Costa, C., de la Cruz, X., Orozco, M., & Dopazo, J. 2005. Pupasview: a visual tool for selecting suitable snps, with putative pathological effect in genes, for genotyping purposes. *Nucleic acids res*, **33**(Web Server issue), W501–5.

Conde, L., Vaquerizas, J. M., Dopazo, H., Arbiza, L., Reumers, J., Rousseau, F., Schymkowitz, J., & Dopazo, J. 2006. Pupasuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic acids research*, **34**(Web Server issue), W621–5. PMID: 16845085.

Consortium, The STAR. 2008. Snp and haplotype mapping for genetic analysis in the rat. *Nature genetics*, **40**(5), 560–6. PMID: 18443594.

Dard, P., Lefranc, M. P., Osipova, L., & Sanchez-Mazas, A. 2001. Dna sequence variability of ighg3 alleles associated to the main g3m haplotypes in human populations. *European journal of human genetics: Ejhg*, **9**(10), 765–72.

Del Cornò, M, Gauzzi, M. C., Penna, G., Belardelli, F., Adorini, L., & Gessani, S. 2005. Human immunodeficiency virus type 1 gp120 and other activation stimuli are highly effective in triggering alpha interferon and cc chemokine production in circulating plasmacytoid but not myeloid dendritic cells. *Journal of virology*, **79**(19), 12597–601.

Dorus, S., Vallender, E. J., Evans, P. D., Anderson, J. R., Gilbert, S. L., Mahowald, M., Wyckoff, G. J., Malcom, C. M., & Lahn, B. T. 2004. Accelerated evolution of nervous system genes in the origin of homo sapiens. *Cell*, **119**(7), 1027–40.

Edgar, R. C. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids res*, **32**(5), 1792–7.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. 2005. The sequence ontology: a tool for the unification of genome annotations. *Genome biology*, **6**(5), R44.

Eisenberg, E., & Levanon, E. Y. 2003. Human housekeeping genes are compact. *Trends genet*, **19**(7), 362–5.

Ernst, C., & Christie, B. R. 2006. The putative neural stem cell marker, nestin, is expressed in heterogeneous cell types in the adult rat neocortex. *Neuroscience*, **138**(1), 183–8.

Fernandez-Escamilla, A., Rousseau, F., Schymkowitz, J., & Serrano, L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature biotechnology*, **22**(10), 1302–6. PMID: 15361882.

Ferrer-Costa, C., Orozco, M., & de la Cruz, X. 2002. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J mol biol*, **315**(4), 771–86.

Ferrer-Costa, C., Orozco, M., & de la Cruz, X. 2004. Sequence-based prediction of pathological mutations. *Proteins*, **57**(4), 811–9. PMID: 15390262.

Ge, H., Walhout, Albertha J. M., & Vidal, M. 2003. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends in genetics: Tig*, **19**(10), 551–60.

Gillespie, J. H. 1991. *The causes of molecular evolution*. Oxford University Press.

Goldberg, M., Septier, D., Rapoport, O., Iozzo, R. V., Young, M. F., & Ameye, L. G. 2005. Targeted disruption of two small leucine-rich proteoglycans, biglycan and decorin, excerpts divergent effects on enamel and dentin formation. *Calcified tissue international*, **77**(5), 297–310.

Goldman, N., & Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution*, **11**(5), 725–36.

Gostissa, M., Hengstermann, A., Fogal, V., Sandy, P., Schwarz, S. E., Scheffner, M., & Del Sal, G. 1999. Activation of p53 by conjugation to the ubiquitin-like protein sumo-1. *Embo j*, **18**(22), 6462–71.

Gu, W., & Roeder, R. G. 1997. Activation of p53 sequence-specific dna binding by acetylation of the p53 c-terminal domain. *Cell*, **90**(4), 595–606.

Guerois, R., Nielsen, J. E., & Serrano, L. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J mol biol*, **320**(2), 369–87.

Guindon, S., & Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, **52**(5), 696–704. PMID: 14530136.

Hedges, S. Blair, & Kumar, S. 2003. Genomic clocks and evolutionary timescales. *Trends in genetics: Tig*, **19**(4), 200–6.

Hellmann, I., Zollner, S., Enard, W., Ebersberger, I., Nickel, B., & Paabo, S. 2003. Selection on human genes as revealed by comparisons to chimpanzee cdna. *Genome res*, **13**(5), 831–7.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., & Birney, E. 2005. Ensembl 2005. *Nucleic acids research*, **33**(Database issue), D447–53.

Hubby, J. L., & Lewontin, R. C. 1966. A molecular approach to the study of genic heterozygosity in natural populations. i. the number of alleles at different loci in drosophila pseudoobscura. *Genetics*, **54**(2), 577–94. PMID: 5968642.

Huelsenbeck, J. P., & Ronquist, F. 2001. Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics (oxford, england)*, **17**(8), 754–5. PMID: 11524383.

Hughes, A. L. 2007. Looking for darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity*, **99**(4), 364–73. PMID: 17622265.

Ihaka, R., & Gentleman, R. 1996. R: A language for data analysis and graphics. *J comput graph stat*, **5**, 229–314.

Ito, A., Lai, C. H., Zhao, X., Saito, S., Hamilton, M. H., Appella, E., & Yao, T. P. 2001. p300/cbp-mediated p53 acetylation is commonly induced by p53-activating agents and inhibited by mdm2. *Embo j*, **20**(6), 1331–40.

Jonnakuty, S., Hotz-Wagenblatt, A., Glatting, K-H., & Suhai, S. 2006. *Tissuedistributiondbs: a repository of organism-specific tissue distribution profiles.* Personal communication.

Kabashima, K., Sakata, D., Nagamachi, M., Miyachi, Y., Inaba, K., & Narumiya, S. 2003. Prostaglandin e2-ep4 signaling initiates skin immune responses by promoting migration and maturation of langerhans cells. *Nature medicine*, **9**(6), 744–9.

Karkkainen, M. J., Haiko, P., Sainio, K., Partanen, J., Taipale, J., Petrova, T. V., Jeltsch, M., Jackson, D. G., Talikka, M., Rauvala, H., Betsholtz, C., & Alitalo, K. 2004. Vascular endothelial growth factor c is required for sprouting of the first lymphatic vessels from embryonic veins. *Nature immunology*, **5**(1), 74–80.

Kearns, A. E., Donohue, M. M., Sanyal, B., & Demay, M. B. 2001. Cloning and characterization of a novel protein kinase that impairs osteoblast differentiation in vitro. *The journal of biological chemistry*, **276**(45), 42213–8.

Kehrer-Sawatzki, H., & Cooper, D. N. 2007. Understanding the recent evolution of the human genome: insights from human-chimpanzee genome comparisons. *Human mutation*, **28**(2), 99–130. PMID: 17024666.

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

Khaitovich, P., Enard, W., Lachmann, M., & PÃ€Ã€bo, S. 2006. Evolution of primate gene expression. *Nature reviews. genetics*, **7**(9), 693–702. PMID: 16921347.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature*, **217**(5129), 624–6.

Kimura, M. 1983. *The neutral theory of molecular evolution*.

King, J. L., & Jukes, T. H. 1969. Non-darwinian evolution. *Science (new york, n.y.)*, **164**(881), 788–98.

Koradi, R., Billeter, M., & Wuthrich, K. 1996. Molmol: a program for display and analysis of macromolecular structures. *J mol graph*, **14**(1), 51–5, 29–32.

Koref, M. F. Santibanez, Gangeswaran, R., Shanahan, N., & Hancock, J. M. 2003. A phylogenetic approach to assessing the significance of missense mutations in disease genes. *Hum mutat*, **22**(1), 51–8.

Kuerbitz, S. J., Plunkett, B. S., Walsh, W. V., & Kastan, M. B. 1992. Wild-type p53 is a cell cycle checkpoint determinant following irradiation. *Proc natl acad sci u s a*, **89**(16), 7491–5.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H.,

Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., & Szustakowki, J. 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Lee, J. W., Choi, H. S., Gyuris, J., Brent, R., & Moore, D. D. 1995. Two classes of proteins dependent on either the presence or absence of thyroid hormone for interaction with the thyroid hormone receptor. *Molecular endocrinology (baltimore, md.)*, **9**(2), 243–54.

Lee, W., Harvey, T. S., Yin, Y., Yau, P., Litchfield, D., & Arrowsmith, C. H. 1994. Solution structure of the tetrameric minimum transforming domain of p53. *Nat struct biol*, **1**(12), 877–90.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J mol evol*, **36**(1), 96–9.

Li, W. S. 1997. *Molecular evolution*. Sinauer Associates, Inc.

Lu, Q., Kaplan, M., Ray, D., Ray, D., Zacharek, S., Gutsch, D., & Richardson, B. 2002. Demethylation of itgal (cd11a) regulatory sequences in systemic lupus erythematosus. *Arthritis and rheumatism*, **46**(5), 1282–91.

Mahler, B., Gocken, T., Brojan, M., Childress, S., Spandau, D. F., & Foley, J. 2004. Keratin 2e: a marker for murine nipple epidermis. *Cells, tissues, organs*, **176**(4), 169–77.

Massingham, T., & Goldman, N. 2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**(3), 1753–62.

Mateu, M. G., & Fersht, A. R. 1998. Nine hydrophobic side chains are key determinants of the thermodynamic stability and oligomerization status of tumour suppressor p53 tetramerization domain. *Embo j*, **17**(10), 2748–58.

Mekel-Bobrov, N., Gilbert, S. L., Evans, P. D., Vallender, E. J., Anderson, J. R., Hudson, R. R., Tishkoff, S. A., & Lahn, B. T. 2005. Ongoing adaptive evolution of aspm, a brain size determinant in homo sapiens. *Science*, **309**(5741), 1720–2.

Miller, M. P., & Kumar, S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. *Hum mol genet*, **10**(21), 2319–28.

Mooney, S. D., & Altman, R. B. 2003. Mutdb: annotating human variation with functionally relevant data. *Bioinformatics*, **19**(14), 1858–60.

Muse, S. V., & Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular biology and evolution*, **11**(5), 715–24.

Möröy, T., & Geisen, C. 2004. Cyclin e. *The international journal of biochemistry & cell biology*, **36**(8), 1424–39.

Nagata, K., Asano, T., Nozawa, Y., & Inagaki, M. 2004. Biochemical and cell biological analyses of a mammalian septin complex, sept7/9b/11. *The journal of biological chemistry*, **279**(53), 55895–904.

Naruse, T. K., Kawata, H., Inoko, H., Isshiki, G., Yamano, K., Hino, M., & Tatsumi, N. 2002. The hla-dob gene displays limited polymorphism with only one amino acid substitution. *Tissue antigens*, **59**(6), 512–9.

Ng, P. C., & Henikoff, S. 2001. Predicting deleterious amino acid substitutions. *Genome res*, **11**(5), 863–74.

Nicklin, M. J., Barton, J. L., Nguyen, M., FitzGerald, M. G., Duff, G. W., & Kornman, K. 2002. A sequence-based map of the nine genes of the human interleukin-1 cluster. *Genomics*, **79**(5), 718–25.

Nielsen, R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity*, **86**(Pt 6), 641–7.

Nielsen, R., & Yang, Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the hiv-1 envelope gene. *Genetics*, **148**(3), 929–36.

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., J, J. Sninsky, Adams, M. D., & Cargill, M. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *Plos biol*, **3**(6), e170.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature*, **246**(5428), 96–98.

Ohta, T. 1993. An examination of the generation-time effect on molecular evolution. *Proc natl acad sci u s a*, **90**(22), 10676–80.

Ohta, T., & Ina, Y. 1995. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. *Journal of molecular evolution*, **41**(6), 717–20.

Olivier, M., Eeles, R., Hollstein, M., Khan, M. A., Harris, C. C., & Hainaut, P. 2002. The iarc tp53 database: new online mutation analysis and recommendations to users. *Hum mutat*, **19**(6), 607–14.

Onochie, C. I., Korngut, L. M., Vanhorne, J. B., Myers, S. M., Michaud, D., & Mulligan, L. M. 2000. Characterisation of the human gfralpha-3 locus and investigation of the gene in hirschsprung disease. *Journal of medical genetics*, **37**(9), 674–9.

Pedrini, E., De Luca, A., Valente, E. M., Maini, V., Capponcelli, S., Mordenti, M., Mingarelli, R., Sangiorgi, L., & Dallapiccola, B. 2005. Novel ext1 and ext2 mutations identified by dhplc in italian patients with multiple osteochondromas. *Human mutation*, **26**(3), 280.

Perez, C., Auriol, J., Gerst, C., Bernard, B. A., & Egly, J. M. 1999. Genomic organization and promoter characterization of two human uhs keratin genes. *Gene*, **227**(2), 137–48.

Pfankuch, T., Rizk, A., Olsen, R., Poage, C., & Raber, J. 2005. Role of circulating androgen levels in effects of apoe4 on cognitive function. *Brain research*, **1053**(1-2), 88–96.

Pochampally, R. R., Horwitz, E. M., DiGirolamo, C. M., Stokes, D. S., & Prockop, D. J. 2005. Correction of a mineralization defect by overexpression of a wild-type cdna for col1a1 in marrow stromal cells (mscs) from a patient with osteogenesis imperfecta: a strategy for rescuing mutations

that produce dominant-negative protein defects. *Gene therapy*, **12**(14), 1119–25.

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. 2005. Hyphy: hypothesis testing using phylogenies. *Bioinformatics (oxford, england)*, **21**(5), 676–9. PMID: 15509596.

Pond, S. L. K., Poon, A. F. Y., & Frost, S. D. W. 2009. *The phylogenetic handbbok (2nd edition)*. Cambridge University Press. Chap. Estimating selection pressures on alignments of coding sequences.

Raymond, M. H., Schutte, B. C., Torner, J. C., Burns, T. L., & Willing, M. C. 1999. Osteocalcin: genetic and physical mapping of the human gene bglap and its potential role in postmenopausal osteoporosis. *Genomics*, **60**(2), 210–7.

Reumers, J., Schymkowitz, J., Ferkinghoff-Borg, J., Stricher, F., Serrano, L., & Rousseau, F. 2005. Snpeffect: a database mapping molecular phenotypic effects of human non-synonymous coding snps. *Nucleic acids research*, **33**(Database issue), D527–32. PMID: 15608254.

Rice, J. A. 1994. *Mathematical statistics and data analysis. 2nd edition*. 2nd edition edn. Duxbury advanced series. Duxbury Press.

Risch, N. J. 2000. Searching for genetic determinants in the new millennium. *Nature*, **405**(6788), 847–56. PMID: 10866211.

Rittenberg, B., Partridge, E., Baker, G., Clokie, C., Zohar, R., Dennis, J. W., & Tenenbaum, H. C. 2005. Regulation of bmp-induced ectopic bone formation by ahsg. *Journal of orthopaedic research: Official publication of the orthopaedic research society*, **23**(3), 653–62.

Ritter, M., Ali, M. Y., Grimm, C. F., Weth, R., Mohr, L., Bocher, W. O., Endrulat, K., Wedemeyer, H., Blum, H. E., & Geissler, M. 2004. Immunoregulation of dendritic and t cells by alpha-fetoprotein in patients with hepatocellular carcinoma. *Journal of hepatology*, **41**(6), 999–1007.

Robinson-Rechavi, M., & Huchon, D. 2000. Rrtree: relative-rate tests between groups of sequences on a phylogenetic tree. *Bioinformatics (oxford, england)*, **16**(3), 296–7.

Sargent, I. L. 2005. Does 'soluble' hla-g really exist? another twist to the tale. *Molecular human reproduction*, **11**(10), 695–8.

Saunders, C. T., & Baker, D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J mol biol*, **322**(4), 891–901.

Schipani, E., Karga, H., Karaplis, A. C., Potts, J. T., Kronenberg, H. M., Segre, G. V., Abou-Samra, A. B., & Jüppner, H. 1993. Identical complementary deoxyribonucleic acids encode a human renal and bone parathyroid hormone (pth)/pth-related peptide receptor. *Endocrinology*, **132**(5), 2157–65.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., & Schomburg, D. 2004. Brenda, the enzyme database: updates and major new developments. *Nucleic acids res*, **32**(Database issue), D431–3.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. 2005. The foldx web server: an online force field. *Nucleic acids research*, **33**(Web Server issue), W382–8. PMID: 15980494.

Sebastiani, S., Allavena, P., Albanesi, C., Nasorri, F., Bianchi, G., Traidl, C., Sozzani, S., Girolomoni, G., & Cavani, A. 2001. Chemokine receptor expression and function in cd4+ t lymphocytes with regulatory activity. *Journal of immunology (baltimore, md.: 1950)*, **166**(2), 996–1002.

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

Shmueli, O., Horn-Saban, S., Chalifa-Caspi, V., Shmoish, M., Ophir, R., Benjamin-Rodrig, H., Safran, M., Domany, E., & Lancet, D. 2003. Genenote: whole genome expression profiles in normal human tissues. *Comptes rendus biologies*, **326**, 1067–72. PMID: 14744114.

Smanik, P. A., Ryu, K. Y., Theil, K. S., Mazzaferri, E. L., & Jhiang, S. M. 1997. Expression, exon-intron organization, and chromosome mapping of the human sodium iodide symporter. *Endocrinology*, **138**(8), 3555–8.

Smith, E. L., & Margoliash, E. 1964. Evolution of cytochrome c. *Federation proceedings*, **23**, 1243–7. PMID: 14236132.

Sokal, R. 1981. *Biometry. the principles and practice of statistics in biological research*. 2nd edn. W. H. Freeman and Company.

Springer, M. S., J., Stanhope M., Ole, Madsen, & de Jong W.W. 2004. Molecules consolidate the placental mammal tree. *Trends in ecology and evolution*, **19**, 430–438.

Stolze, I., Berchner-Pfannschmidt, U., Freitag, P., Wotzlaw, C., Rössler, J., Frede, S., Acker, H., & Fandrey, J. 2002. Hypoxia-inducible erythropoietin gene expression in human neuroblastoma cells. *Blood*, **100**(7), 2623–8.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., & Hogenesch, J. B. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the national academy of sciences of the united states of america*, **101**(16), 6062–7. PMID: 15075390.

Sunyaev, S., Ramensky, V., & Bork, P. 2000. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends genet*, **16**(5), 198–200.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., Kondrashov, A. S., & Bork, P. 2001. Prediction of deleterious human alleles. *Hum mol genet*, **10**(6), 591–7.

Suzuki, Y., & Gojobori, T. 1999. A method for detecting positive selection at single amino acid sites. *Mol biol evol*, **16**(10), 1315–28.

Thompson, J. D., Higgins, D.. G., & Gibson, T. J. 1994. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, **22**(22), 4673–80.

Tseng, J., Do, J., Widdicombe, J. H., & Machen, T. E. 2006. Innate immune responses of human tracheal epithelium to pseudomonas aeruginosa flagellin, tnf-alpha, and il-1beta. *American journal of physiology. cell physiology*, **290**(3), C678–90.

Tárraga, J., Medina, I., Arbiza, L., Huerta-Cepas, J., Gabaldón, T., Dopazo, J., & Dopazo, H. 2007. Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics. *Nucleic acids research*, **35**(jul), W38–42. PMID: 17452346.

Valiaho, J., Pusa, M., Ylinen, T., & Vihinen, M. 2002. Idr: the immunodeficiency resource. *Nucleic acids res*, **30**(1), 232–4.

Vamvakopoulos, N. C., Monahan, J. J., & Kourides, I. A. 1980. Synthesis, cloning, and identification of dna sequences complementary to mrnas for alpha and beta subunits of thyrotropin. *Proceedings of the national academy of sciences of the united states of america*, **77**(6), 3149–53.

Van der Bliek, A. M., Baas, F., Ten Houte de Lange, T., Kooiman, P. M., Van der Velde-Koerts, T., & Borst, P. 1987. The human mdr3 gene encodes a novel p-glycoprotein homologue and gives rise to alternatively spliced mrnas in liver. *The embo journal*, **6**(11), 3325–31.

Varki, A., Geschwind, D. H, & Eichler, E. E. 2008. Explaining human uniqueness: genome interactions with environment, behaviour and culture. *Nature reviews. genetics*, **9**(10), 749–63. PMID: 18802414.

Velculescu, V. E., Madden, S. L., Zhang, L., Lash, A. E., Yu, J., Rago, C., Lal, A., Wang, C. J., Beaudry, G. A., Ciriello, K. M., Cook, B. P., Dufault, M. R., Ferguson, A. T., Gao, Y., He, T. C., Hermeking, H., Hiraldo, S. K., Hwang, P. M., Lopez, M. A., Luderer, H. F., Mathews, B., Petroziello, J. M., Polyak, K., Zawel, L., Kinzler, K. W., & et al. 1999. Analysis of human transcriptomes. *Nat genet*, **23**(4), 387–8.

Wang, D., Oakley, T., Mower, J., Shimmin, L. C., Yim, S., Honeycutt, R. L., Tsao, H., & Li, Wen-Hsiung. 2004. Molecular evolution of bat color vision genes. *Molecular biology and evolution*, **21**(2), 295–302.

Wang, H. Y., Chien, H. C., Osada, N., Hashimoto, K., Sugano, S., Gojobori, T., Chou, C. K., Tsai, S. F., Wu, C. I., & Shen, C. K. 2006. Rate of evolution in brain-expressed genes in humans and other primates. *Plos biol*, **5**(2), e13.

Wang, H. Y., Chien, H. C., Osada, N., Hashimoto, K., Sugano, S., Gojobori, T., Chou, C. K., Tsai, S. F., Wu, C. I., & Shen, C. K. 2007. Rate of evolution in brain-expressed genes in humans and other primates. *Plos biol*, **5**(2), e13.

Wang, Z., & Moult, J. 2001. Snps, protein structure, and disease. *Hum mutat*, **17**(4), 263–70.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Helmberg, W., Kapustin, Y., Kenton, D. L., Khovayko, O., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Pruitt, K. D., Schuler, G. D., Schriml, L. M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Suzek, T. O., Tatusov, R., Tatusova, T. A., Wagner, L., & Yaschenko, E. 2006. Database resources of the national center for biotechnology information. *Nucleic acids res*, **34**(Database issue), D173–80.

Wong, W., Yang, Z., Goldman, N., & Nielsen, R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, **168**(2), 1041–51.

Wrabetz, L, & Feltri, M L. 2001. Do schwann cells stop, dr(o)p2, and roll? *Neuron*, **30**(3), 642–4.

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N., & Suzek, B. 2006. The universal protein resource

(uniprot): an expanding universe of protein information. *Nucleic acids research*, **34**(Database issue), D187–91. PMID: 16381842.

Xu, C., Li, C. Yong-Tao, & Kong, Ah-Ng T. 2005. Induction of phase i, ii and iii drug metabolism/transport by xenobiotics. *Archives of pharmacal research*, **28**(3), 249–68.

Yamada, K., & Nabeshima, T. 2004. Interaction of bdnf/trkb signaling with nmda receptor in learning and memory. *Drug news & perspectives*, **17**(7), 435–8.

Yang, Z. 1997. Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput appl biosci*, **13**(5), 555–6.

Yang, Z., & Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol biol evol*, **17**(1), 32–43.

Yang, Z., & Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol biol evol*, **19**(6), 908–17.

Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A. M. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**(1), 431–49.

Yang, Z., Wong, W. S., & Nielsen, R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol biol evol*, **22**(4), 1107–18.

Zhang, J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. *Molecular biology and evolution*, **21**(7), 1332–9.

Zhang, J., Nielsen, R., & Yang, Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol biol evol*, Aug 17.

Zhang, L., & Li, W. H. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol biol evol*, **21**(2), 236–9.

Zhong, H., & Carlson, H. A. 2005. Computational studies and peptidomimetic design for the human p53-mdm2 complex. *Proteins*, **58**(1), 222–34.

Zimek, A., & Weber, K. 2005. Terrestrial vertebrates have two keratin gene clusters; striking differences in teleost fish. *European journal of cell biology*, **84**(6), 623–35.

Zuckerkandl, E., & Pauling, L. 1965. Molecules as documents of evolutionary history. *Journal of theoretical biology*, **8**(2), 357–66.