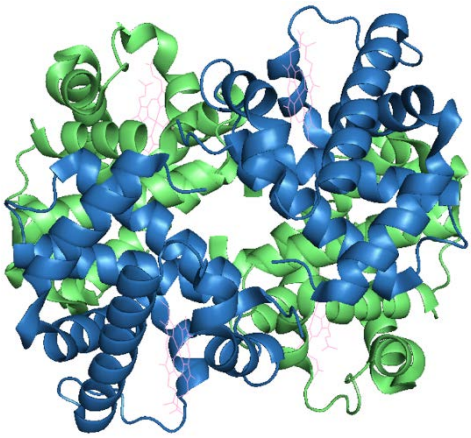
U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Study of Hemoglobin A_2: the paradox of *δ-globin* gene conservation and its supposed physiological irrelevance

Ana Sofia dos Santos Moleirinho

Tese de Doutoramento apresentada à Faculdade de Ciências da Universidade do Porto
Biologia
2015



Study of Hemoglobin A_2: the paradox of *δ-globin* gene conservation and its supposed physiological irrelevance

Ana Sofia dos Santos Moleirinho
Programa Doutoral em Biologia
Biologia
2015

Orientador
Professor Doutor António Amorim dos Santos, Professor Catedrático, Faculdade de Ciências da Universidade do Porto

Coorientador
Professora Doutora Maria João Prata Martins Ribeiro, Professora Associada com Agregação, Faculdade de Ciências da Universidade do Porto

Acknowledgments/Agradecimentos

São muitas as pessoas às quais não posso deixar de agradecer por todo o apoio ao longo destes últimos quarto anos.

As minhas primeiras palavras vão para o meu orientador, Professor António Amorim, por todo o apoio constante ao longo deste trabalho. Esteve sempre presente e disponível para responder a todas as minhas dúvidas, por vezes intermináveis. Agradeço pelo tempo dispensado a discutir e a partilhar ideias, o espirito crítico e a confiança depositada em mim, que foram cruciais para o meu crescimento a nível científico.

À Professora Maria João por ter aceitado ser minha coorientadora e pela sua disponibilidade e prontidão em ajudar durante todo o desenvolvimento do trabalho.

Porque nem tudo ao longo destes quatro anos foi fácil, quero agradecer à Alexandra Lopes e à Susana Seixas por me “aturarem” em todas as fases boas e menos boas, pelo encorajamento quando me sentia perdida, e por toda a ajuda ao longo deste percurso.

A todos os membros do grupo de genética populacional, em especial aos que partilharam comigo o dia a dia (vocês sabem quem são), por me ouvirem (tarefa que nem sempre é fácil), por me apoiarem e por permitirem que estes anos fossem mais leves e fáceis de levar sempre com uma boa “risada”.

Ao IPATIMUP por me ter dado a oportunidade e todas as condições necessárias para desenvolver o meu trabalho de investigação.

A todos os co-autores e a todos os que colaboraram de alguma forma para a realização deste trabalho o meu obrigado.

À Fundação para a Ciência e Tecnologia (FCT) por me ter concedido a bolsa de estudo (SFRH/BD/73508/2010) sem a qual não poderia ter realizado este trabalho.

Aos amigos que tenho quase desde que me conheço, Ana Daniela e Inês Carlos, obrigado por estarem sempre lá quando preciso e por todas as aventuras que vivemos juntas, e esperemos que venham muitas mais.

À Rita e ao Tó, amigos de faculdade que ficaram para a vida, obrigado por todas as alegrias e gargalhadas que me fazem dar.

E por último o agradecimento maior vai para a minha família. Obrigada ao meu avô Zé e ao meu avô Joaquim, que não assistem ao terminar desta etapa mas que sei me apoiam e que contribuíram de forma carinhosa para a pessoa que hoje sou. Às minhas duas avós, Vitória e Carminha, pelo exemplo de força e inspiração de vida. À Joana, a minha irmã “canita” que entretanto ficou grande, obrigada por me fazer crescer, pelo seu otimismo e alegria. Ao Pedro, por estar lá todos os dias e tornar a vida mais alegre. Aos meus pais, Helena e Júlio, por me permitirem ser o que sou hoje, pelo orgulho que têm em mim e por todo o apoio, muito obrigada por tudo.

Abstract

In most vertebrates, hemoglobin (Hb) is a heterotetramer composed of two dissimilar globin chains that change during development according to the patterns of expression of the α- and β-globin gene families. These two gene clusters exhibit substantial differences in gene content across vertebrate taxa, which are especially pronounced in the β-globin gene cluster. In placental mammals (eutherians), the β-globin gene cluster includes three early-expressed genes, located at the 5' end of the cluster in the order ε-(*HBE*)-γ(*HBG*)-ψβ(*HBBP1*), and a pair of late expressed genes, δ(*HBD*) and β (*HBB*), at the 3' end. *HBB* codes for the major adult β-globin chain, while *HBD* is either marginally or not transcribed at all. It is generally considered that *HBD* is under a process of pseudogenization due the physiological irrelevance of the encoded polypeptide chain. Paradoxically, reduced diversity levels have been reported for this gene and, in humans and in some primate species, the region encompassing *HBD* reveals a degree of conservation not reconcilable with any process of evolutionary inactivation. A regulatory role for *HBD* in the fetal to adult Hb switch, which is only seen in Anthropoid primates, has been proposed decades ago, but the conjecture has received little attention. Therefore, in this thesis we sought to clarify the biological relevance of *δ-globin* gene conservation in a subgroup of primates, using population genetics and comparative genomics tools to gain insight into the evolution and functional divergence of *HBD* in placental mammals. The results presented here reveal that strong purifying selection has shaped the evolutionary history of *HBD* in humans and chimpanzees, and surprisingly, the same happened to the pseudogene *HBBP1*. Moreover, the study of *HBD* evolutionary trajectory across placental mammals showed that δ*-globin* gene conservation is also observed in Anthropoids, suggesting a long-term effect of purifying selection, with similar strong functional constrains acting over 65 Myr of primate evolution. We further documented that not only the level of sequence conservation but also the mode of evolution of the *HBD* in higher primates are strictly associated with the fetal/adult β-cluster developmental switch. Altogether the studies presented in this thesis disclose that in some primate lineages *HBD* and *HBBP1* play an essential and nonredundant biological role, which in contrast with *HBB*, is not – at least directly – associated with oxygen transport. In the light of recent advances in understanding the mechanism coordinating β*-globin* gene expression, we present evidence that the strong functional constraints underlying the decreased contemporary diversity at these two

genomic regions were not driven by protein function but instead are compatible with a regulatory role in the ontogenic switches of gene expression at the *β-globin* cluster.

Key words: *β-globin* cluster, hemoglobin switch, gene diversity, gene evolution, chromatin interactions

Resumo

Na maioria dos vertebrados, a hemoglobina (Hb) é um heterotetrâmero composto por dímeros de cadeias globínicas que variam durante o desenvolvimento conforme os padrões de expressão das famílias de genes da α- e a β-globina. Dentro dos vertebrados, estes dois conjuntos de genes exibem diferenças substanciais quanto ao seu conteúdo genético, sendo especialmente acentuadas no da *β-globina*. Nos mamíferos placentários (Eutheria), o agrupamento génico *β-globina* é constituído por 3 genes expressos nas fases embrionária e fetal, que se encontram na extremidade 5' do agrupamento ε-(*HBE*)-γ(*HBG*)-ψβ(*HBBP1*), e por um par de genes expressos na fase adulta, δ(*HBD*) and β (*HBB*), localizados na extremidade 3'. *HBB* codifica a cadeia β da Hb adulta mais abundante enquanto *HBD* é expresso em níveis comparativamente muito baixos. Considera-se geralmente que *HBD* está em via de pseudogenização dada a irrelevância fisiológica da cadeia polipeptídica que codifica. Paradoxalmente, foram descritos baixos níveis de diversidade para este gene e aparentemente, em humanos e em alguns outros primatas, a região que engloba *HBD* revela um grau de conservação incompatível com um processo evolutivo de pseugogenização. Há várias décadas postulou-se para *HBD* um papel na regulação do *switch* da Hb fetal para a adulta, que ocorre apenas em Antropóides, mas esta conjetura não despertou interesse e passou praticamente despercebida. Assim, nesta tese, procurámos clarificar a relevância funcional da conservação do gene da δ-globina num subgrupo de primatas, usando ferramentas de genética populacional e genómica comparativa de forma a obter uma melhor compreensão acerca da evolução e divergência funcional de *HBD* em mamíferos placentários. Os resultados apresentados nesta tese revelam que a história evolutiva de *HBD* em humanos e chimpanzés tem sido moldada por seleção purificadora, e surpreendentemente, o mesmo parece aplicar-se ao pseudogene *HBBP1*. Adicionalmente, o estudo da trajetória evolutiva do *HBD* em mamíferos placentários, mostra que a conservação do gene da δ-globina se estende aos Antropóides, o que sugere um efeito de seleção purificadora ao longo da evolução dos primatas, em que as mesmas constrições funcionais atuam há cerca de 65 Myr. Demonstrámos ainda que não só a conservação da sequência, mas também o modo de evolução do *HBD* em primatas superiores estão intimamente relacionados com o *switch* da Hb fetal para a adulta. Os estudos apresentados nesta tese mostram que, em algumas linhagens de primatas, *HBD*

e *HBBP1* têm um papel essencial e não redundante, que, ao contrário do que acontece com HBB não está, pelo menos diretamente, relacionado com o transporte de oxigénio. À luz dos recentes avanços na compreensão do mecanismo que coordena a expressão dos genes β-globínicos, nós apresentamos provas que apoiam a hipótese de que os fortes constrangimentos funcionais que levam á atual baixa diversidade encontrada nestas duas regiões genómicas, não são devidos a uma função proteica mas, provavelmente, a um papel regulador na expressão diferencial dos genes β-globínicos ao longo do desenvolvimento.

Palavras chave: *β-globin* cluster, switch da hemoglobina, diversidade genética, evolução, interações de cromatina

Index

Figure Index

Abbreviations

ACH	Active chromatin hub
BCL11A	B-cell lymphoma/leukemia 11A
CRMs	CRMs
DNA	Deoxyribonucleic Acid
EKLF	Erythroid kruppel like factor
Hb	Hemoglobin
HBB	Beta globin gene
HBBP1	Beta globin pseudogene 1
HBD	Delta globin gene
HBE	Epsilon globin gene
HbF	Fetal hemoglobin
HBG	Gamma globin gene
HbS	Sickle hemoglobin
HbVar	Database of human hemoglobin variants and thalassemias
HPFH	Hereditary persistence of fetal hemoglobin
HS	Hypersensitive site
Kb	kilobases
KLF	Kruppel-like zinc finger
LCR	Locus control region
LD	Linkage disequilibrium
Myr, Mya	Million years, million years ago
NWM	New World Monkeys
NuRD	Nucleosome-remodeling and histone deacetylation
OWM	Old World Monkeys
RNA TRAP	Ribonucleic Acid Tagging and recovery of associated proteins
SCD	Sickle cell disease
SNP	Single nucleotide polymorphism
3C, 5C	Chromosome conformation capture, Chromosome Conformation Capture Carbon Copy

CHAPTER 1. INTRODUCTION

1.1 Hemoglobin

Hemoglobins (Hbs) were originally discovered as abundant proteins in red blood cells of mammals and other jawed vertebrates (gnathostomes), which carry oxygen from the lungs, gills or other respiratory organs to peripheral tissues. Almost all jawed vertebrates express different forms of hemoglobin at progressive developmental stages. All species that have been examined to date make embryonic-specific hemoglobins in primitive erythroid cells derived from the yolk sac, some species make a fetal-specific form in the liver, and all species produce an adult hemoglobin in erythroid cells produced in the bone marrow. Each of these hemoglobin molecules are heterotetramers composed of two α-like and two β-like globin chains, each with its associated heme group. These globin chains are encoded by members of the α- and β-globin gene families. Expression of the two gene families must be strictly coordinated for balanced expression of *α*- and *β-globin* genes. The hemoglobin system represents a paradigm for tissue-specific and developmental gene expression (for a review see (Hardison 2012b)).

1.2 Evolution of Hemoglobin Gene Clusters

The *α*- and *β-globin* genes arose via tandem duplication of an ancestral single-copy globin gene approximately 450–500 Mya, in the common ancestor of jawed vertebrates (Czelusniak, et al. 1982; Goodman, et al. 1987; Goodman, et al. 1975). The history of the two gene clusters during vertebrate evolution is dynamic and complex and the *α*- and *β-globin* gene clusters have diverged considerably since their duplication. In amphibians and teleost fish the *α*- and *β-globin* genes retain the ancestral arrangement and are found closely linked in the same locus (Chan, et al. 1997; Fuchs, et al. 2006; Hosbach, et al. 1983; Jeffreys, et al. 1980; Kay, et al. 1980; McMorrow, et al. 2003; Pisano, et al. 2003). In amniotes (birds and mammals), by contrast, they are located on different chromosomes (Hardison 2008; Patel, et al. 2008; Patel, et al. 2010) and have thus evolved independently to generate the contemporary gene clusters. Multiple rounds of duplication and divergence have produced diverse repertoires of *α*- and *β-like* globin genes that are ontogenetically regulated (Hoffmann, et al. 2010). The genomic location of *α*- and *β-globin* genes in

amniotes requires coordination of expression between different chromosomes for balanced production of α- and β-globin chains and efficient formation of functionally distinct Hb isoforms during different stages of prenatal development and postnatal life. The composition and expression patterns of the *α-globin* gene cluster are remarkably stable among jawed vertebrate taxa. By contrast, the *β-globin* cluster exhibits a number of significant differences in gene content. Distinct repertoires of mammalian *β-like globin* genes with different developmental regulation have evolved multiple times by independent lineage-specific duplications followed by functional divergence (Hoffmann, et al. 2008; Hoffmann and Storz 2007; Hoffmann, et al. 2010; Opazo, et al. 2008a, b; Storz, et al. 2013, 2011).

1.2.1 Mammalian *β-globin* Gene Clusters

Within the three major subclasses of mammals, the *β-like globin* genes have been duplicated and lost independently in specific lineages (Figure 1). In both monotremes and marsupials, the *β-globin* gene cluster contains a single pair of genes, the early expressed *ε-globin* and the late expressed *β-globin* (Opazo, et al. 2008b). However, the *ε-globin* gene in monotremes is not orthologous to the *ε-globin* gene in marsupials, since they were independently derived from lineage-specific duplications of a proto *β-globin* gene (Opazo, et al. 2008b). The duplication event, which occurred before the divergence of marsupials and eutherians, approximately 160 Mya, originated the early- and late-expressed globin genes in therian mammals (the embryonically expressed *ε-globin* genes and the adult expressed *β-globin* genes). In the eutherian stem, further tandem duplications gave rise to a cluster of five paralogous *β-like globin* genes and one pseudogene, arranged 5'-ε-(*HBE*)-γ(*HBG*)-ψβ(*HBBP1*)-δ(*HBD*)-β(*HBB*)-3' consistent with the orientation in contemporary species (Hardies, et al. 1984; Hardison 1984; Opazo, et al. 2008b). The *β-* and *δ-globin* genes, located at the 3' end of the gene clusters, descended from the ancestral *HBB* gene, and, if functional, are expressed in fetal and/ or adult erythroid cells. The *ε-*, *γ-*, and *ψβ-globin* genes, located at the 5' end of the gene clusters, descended from the ancestral *HBE* gene and are expressed in embryonic erythroid cells, except for the *γ-globin* genes in anthropoid primates, which after duplication in the anthropoid branch were recruited for fetal-specific expression. The full cluster of five *β-globin* genes is not found in any extant

mammalian species. Gene duplication, deletion and inactivation have occurred in different lineages generating distinct repertoires of mammalian *β-like globin* genes, as shown in figure 1 (Hoffmann, et al. 2008; Hoffmann and Storz 2007; Hoffmann, et al. 2010; Opazo, et al. 2008a, b; Storz, et al. 2013, 2011). The *ε-globin* gene has the most conserved features across eutherian species, being always located in the 5' end of the gene cluster and embryonically expressed. In contrast, the *γ-*, *ψβ-* and *δ-globin* genes have been gained and lost frequently, sometimes in entire orders of mammals (Opazo, et al. 2008a; Song, et al. 2012). Furthermore, the developmental expression specificity of the different genes in the cluster varies drastically between clades. A remarkable example is the delay in the *γ-globin* (fetal) and *β-globin* (adult) gene expression in anthropoid primates (Johnson, et al. 2000; Johnson, et al. 2002b).

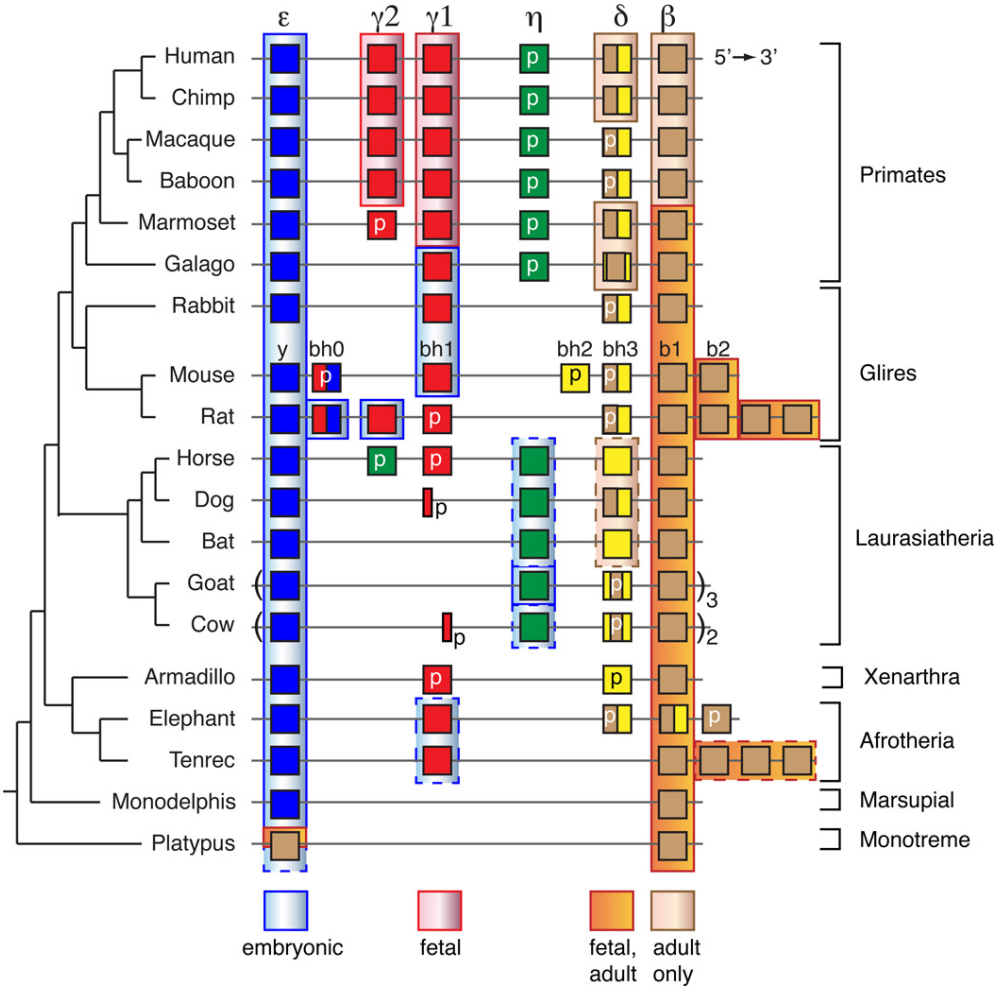


Figure 1 - Maps of orthologous β-like globin genes and expression timing in mammals.

(Legend in the next page)

Figure 1 - Maps of orthologous β-like globin genes and expression timing in mammals.

When the timing of expression is predicted rather than experimentally determined (or highly likely, as in the case of embryonic expression of *ε-globin* gene orthologs), the background shading is outlined with a dashed line. The gene clusters are triplicated in goats and duplicated in cows, indicated by the parentheses and subscripts. The orthology and expression assignments for these are based largely on the first copy of the segmental duplication; the three *β-globin* orthologs are expressed in juvenile, adult, and fetal goats (Townes, et al. 1984). The assignments of orthologous relationships are based on grouping within phylogenetic comparisons of coding sequences (Koop and Goodman 1988; Opazo, et al. 2008a, b; Opazo, et al. 2009; Patel, et al. 2008) and automated determination of orthologs using a method that recognizes gene conversions (Song, et al. 2012). The timing of expression is based on multiple reports in the literature (Efstratiadis, et al. 1980; Johnson, et al. 1996; Johnson, et al. 2000; Koop and Goodman 1988; Lecrone 1970; Patel, et al. 2008; Rohrbaugh and Hardison 1983; Satoh, et al. 1999; Schimenti and Duncan 1985; Shapiro, et al. 1983; Stockell, et al. 1961; Tagle, et al. 1988; Townes, et al. 1984; Whitelaw, et al. 1990). From (Hardison 2012a).

1.2.2 Regulation of *β-like Globin* Genes

To ensure high-level, tissue- and stage-specific activation and repression of the individual genes during development, expression of *β-like globin* genes must be tightly regulated. The *β-globin* cluster has been regarded as a complex genetic system and a paradigm of gene expression regulation. Over the past three decades, a boost of studies on the *β-globin* cluster have contributed to a better understanding of the several mechanisms known to be involved in the molecular control of *β-globin* gene switching (for a review see (Chakalova, et al. 2005; Harju, et al. 2002). This intricate regulation is exerted, at least partially, by the binding of specific transcription factors to DNA sequences that serve as cis-regulatory modules (CRMs). Some are found proximal to the genes, such as promoters, and others are located distal to the genes such as the major regulatory element of the *β-globin* cluster, known as the locus control region (LCR) (for a review see (Noordermeer and de Laat 2008)). Long-range interactions between the distal LCR and the promoter of the targeted genes, apparently by chromatin looping, also has a central role in the control of *β-like globin* gene expression (Bulger and Groudine 1999; Tolhuis, et al. 2002). Though the mammalian *β-globin* gene cluster has long served as a paradigm for tissue-specific and developmentally regulated expression, its complex structure and developmental gene activation pattern are not fully understood.

1.2.2.1 Promoters and Transcription Factors

Genetic information governing the stage specificity for all *β-like globin* genes is located in gene proximal regions known as promoters. These motifs represent transcription factor binding sites that recruit proteins or protein complexes in a stage-specific manner. Evidence exist for the presence of both positive and negative acting factors that activate or repress genes at a specific developmental stage (Stamatoyannopoulos 2005). Studies focused on the identification of conserved sequences in orthologous genes in different eutherian mammals have revealed protein binding sites essential for the regulation of differentially expressed *β-like globin* genes (Gumucio, et al. 1996; King, et al. 2005). Among the conserved mammalian binding sites, only two, the TATA box and CCAAT box are found in all highly expressed *β-like globin* genes (Efstratiadis, et al. 1980). There is a set of binding sites which are distinctive for each gene. An example is the CACCC box which is bound by members of the family of kruppel-like zinc finger (KLF) proteins (Pearson, et al. 2008). The most extensively studied erythroid KLF is the EKLF (erythroid kruppel like factor), which has been specifically implicated in the regulation of the *β-globin* gene (Miller and Bieker 1993; Wijgerde, et al. 1996). Other KLFs bind similar but distinctive CACCC motifs in the *ε-* and *γ-globin* gene promoters (Asano, et al. 1999). An additional sequence motif commonly found in upstream regulatory regions is the binding site for the transcription factor GATA-1 (globin transcription factor 1), which plays a critical role in erythroid-specific gene activation and repression (Johnson, et al. 2002a; Welch, et al. 2004). However, across mammals, the GATA-1 binding site is not conserved in all *β-like globin* genes. Whereas in most mammals the GATA-1 binding site is found at about the same position in both *ε-* and *γ-globin* genes, the homologous region in the *β-globin* genes does not have a conserved GATA-1 binding motif (Hardison 2001). Despite the high homology between *β-like* globin gene promoters, they show unique features that may account for the developmental stage specificity of gene expression.

1.2.2.2 LCR and Chromatin Structure/Looping

The major regulatory element of the mammalian *β-globin* gene cluster, the LCR, is located at the 5' end of the cluster, far distal to the genes and consists of multiple DNase I Hypersensitive sites (HSs). It was shown that it is required for high-level expression of the

β-like globin genes (Bender, et al. 2000; Hardison, et al. 1997) but the mechanism by which the LCR controls gene expression over a large distance has been the subject of intense study for a long time. Mostly due to the appearance of 3C (chromosome conformation capture)-based technologies (Dekker, et al. 2002), which were extremely useful for investigations on the topology of the *β-globin* gene cluster, over the past twelve years we have witnessed major advances in the understanding of the mechanism underlying *β-globin* gene activation (Dostie, et al. 2006; Tolhuis, et al. 2002). Although still not entirely understood it is now clear that it involves physical contacts between the distal LCR HSs and the promoters of the activated genes through a chromatin looping mechanism, to form what is called an active chromatin hub (ACH) (Figure 2) (Carter, et al. 2002; Dostie, et al. 2006; Patrinos, et al. 2004; Tolhuis, et al. 2002; Vakoc, et al. 2005). It was shown that the 3D configuration of the *β-globin* cluster changes in a developmentally dynamic manner and during erythroid differentiation. In cells that do not express the *β-globin* genes, long-range interactions between the LCR and the genes are not observed. During development, the LCR switches its interactions progressively from early- to late-expressed *β-globin* genes to ensure their activation at the proper developmental stage (Palstra, et al. 2003). Many transcription factors were shown to be involved in chromatin looping formation and stabilization, such as EKLF, GATA-1, and Ldb1 that bind to both the LCR and gene promoter regions and in turn recruit other chromatin remodeling complexes resulting in conformational changes in the *β-globin* cluster (Drissen, et al. 2004; Song, et al. 2007; Vakoc, et al. 2005). Although knowledge of CRMs in the *β-globin* cluster is still incomplete, over the last decades comparative genomic approaches have revealed a huge number of other transcription factors involved in *β-globin* gene regulation. Some of these are preserved throughout mammalian evolution but others are species specific and possibly have driven interspecies differences in critical aspects of the expression patterns or mechanisms of regulation in globin gene clusters. The best example is the newly discovered BCL11A (B-cell lymphoma/leukemia 11A), which plays a central role in the most noticeable gene expression change that has occurred in eutherian evolution, the emergence of the fetal-to-adult switch in humans and other primates (Sankaran, et al. 2009) (see below). Besides the multi-protein complexes involved in chromatin remodeling, also intergenic transcription was shown to set up transcriptionally active chromatin subdomains in the human *β-globin* cluster that are developmentally regulated (Gribnau, et al. 2000).

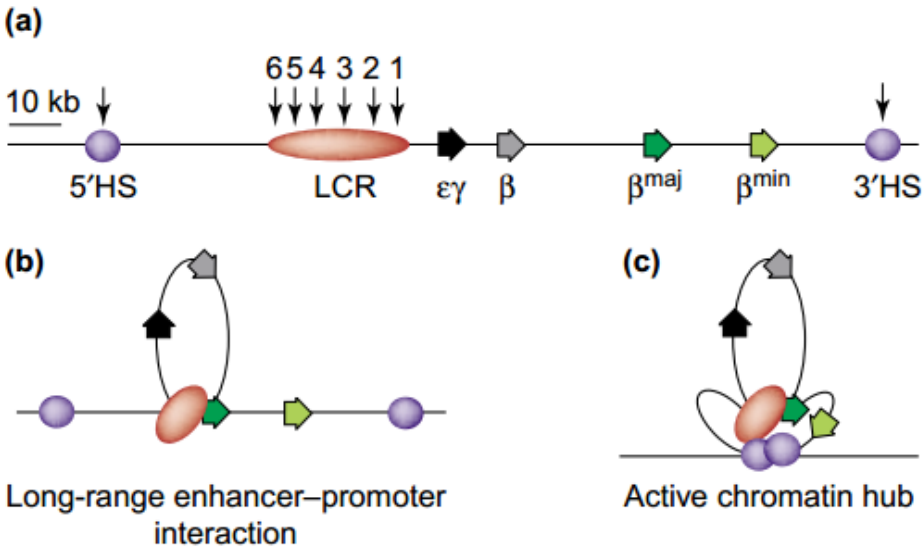


FIGURE 2 - Looping interactions in the β-globin cluster.

(a)Schematic representation of the murine *β-globin* cluster. Vertical arrows indicate *DNaseI* hypersensitive sites (HS) present in the locus control region (LCR) and at the 5' and 3' ends of the ~130-kb *β-globin* gene cluster (purple). Horizontal arrows indicate the four globin genes. (b) RNA TRAP (tagging and recovery of associated proteins) analysis showed interactions between the LCR and active globin genes (dark and light green arrows) whereas inactive genes (black and gray arrows) are looping out. (c) Chromosome conformation capture (3C) analysis demonstrated interactions between 5' HS, LCR, active globin genes and 3' HS, leading Tolhuis et al. (Tolhuis, et al. 2002) to propose that all those sites cluster to form an active chromatin hub. From (Dekker 2003).

1.2.2.3 The Fetal-to-Adult Switch

In mice and in most other mammals that have been well studied, the switch from embryonic hemoglobin to adult hemoglobin, which is known as the primitive to definitive hemoglobin switch, appears to be the major event at the *β-globin* gene cluster. In the course of evolution, Simian primates acquired a unique stage of hemoglobin expression, characterized by a subunit expressed predominantly in the early fetal definitive erythrocytes that remains for much of gestation (Sankaran, et al. 2010a). This molecule is encoded by two duplicated *γ-globin* genes copies, which differ by only a single amino acid, originated by a tandem duplication event in the common ancestor of Simian primates (Fitch, et al. 1991). Thus, unlike what happens in most mammals, in Simian primates, two developmental switches take place at the *β-globin* cluster. In catarrhines, this switch occurs shortly after the time of birth, which is reflected by a transcriptional switch from *γ-* to

β-globin genes (Johnson, et al. 2000). However, platyrrhines exhibit a tendency to inactivate one of the *γ-globin* gene copies, resulting in the expression of only one γ-globin chain and its replacement by the β-globin chain of the adult Hb occurs well before birth (Johnson, et al. 1996; Johnson, et al. 2002b). The regulation of the switch from fetal to adult hemoglobin that occurs after birth in humans and Old World Monkeys (OWM), has been of long-standing interest, given that an improved understanding of this process would open new perspectives leading toward novel therapeutic strategies for fetal hemoglobin induction in β-hemoglobinopathies (Bauer and Orkin 2011; Sankaran and Nathan 2010; Wilber, et al. 2011) (see below). Despite years of study, the molecular mechanisms mediating this switch remained obscure until recently. New insights into the regulation of this switch have come from studies of the genetic basis for inherited conditions associated with increased levels of HbF (fetal hemoglobin) in adulthood (Galarneau, et al. 2010; Lettre, et al. 2008; Menzel, et al. 2007; Thein, et al. 2009; Uda, et al. 2008). The newly discovered BCL11A represents one of the major factors regulating the fetal-to-adult switch, that plays a key role in the silencing of *γ-globin* genes (figure 3) (Sankaran, et al. 2010b). It has been shown that BCL11A is a component of a protein complex that contains the transcription factor GATA-1 and the NuRD (Nucleosome-remodeling and histone deacetylation) chromatin remodeling complex (Sankaran, et al. 2008). Additionally, it has been demonstrated that BCL11A may cooperate with the transcription factor Sox6 to silence fetal globin gene expression, through a reconfiguration of the *β-globin* cluster, which in turn mediates long-range physical interactions between distal regulatory elements located in the LCR (Xu, et al. 2010). BCL11A occupies the *β-globin* gene cluster in several discrete sites, including the upstream locus control region and the *HBG1-HBD* intergenic region (Figure 3). For decades, the intergenic region between the *γ-* and *δ-globin* genes has been a focus of attention because of its possible role in the hemoglobin switch, given the different effects of deletions involving different portions of this region on HbF expression (Bank, et al. 2005; Calzolari, et al. 1999). Deletions within the *β-globin* gene cluster result either in δβ-thalassemia or in hereditary persistence of fetal hemoglobin (HPFH), resulting in a modest increase in HbF and a mild thalassemia phenotype (see below) or higher HbF levels with no associated phenotype, respectively (Forget 1998). Recently, through the characterization and comparison of the breakpoints of these rare deletions, a 3.5-kb intergenic region near the 5' end of *HBD* was identified as being crucial for γ-globin silencing (Figure 3) (Sankaran, et al. 2011a). This region overlaps with the

binding region of BCL11A and contains other structural elements which may also be essential for the temporal coordination of gene expression. This region of DNA may act as a boundary between fetal and adult domains of the *β-globin* gene cluster (Forget 2011).

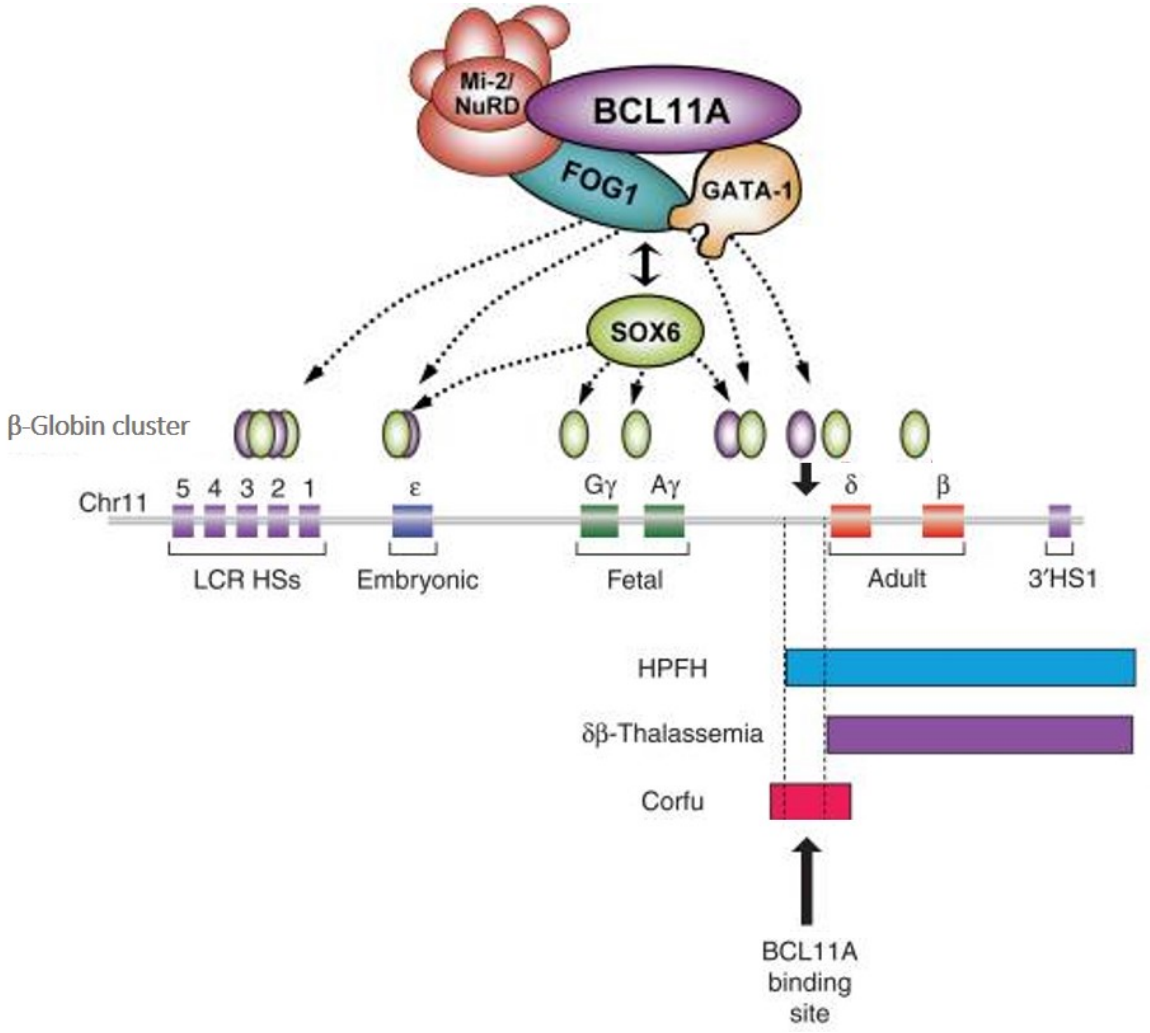


Figure 3 - A model for regulation of γ-globin silencing in the human β-globin gene cluster

The diagram illustrates the physical interactions between BCL11A and the Mi-2/NuRD complexes, erythroid transcription factors GATA1 and FOG1, and the SOX6 protein. Rather than binding to the promoters of the *γ-globin* or *β-globin* genes as these latter factors do, BCL11A occupies the upstream LCR and γδ-intergenic regions of the *β-globin* cluster in adult human erythroid progenitors. This illustration depicts the ~3-kb region upstream of the *δ-globin* gene critical for the switching mechanism, which was found by comparing the regions deleted in HPFH with those removed in δβ-thalassemia deletions (Sankaran, et al. 2011b). Typical deletions are illustrated in the model below the cluster. In addition, the Corfu thalassemia deletion is also known to remove this region, as shown by the model below. BCL11A has been shown to bind to chromatin within this 3-kb region, along with its partners (Sankaran, et al. 2011b). Adapted from (Xu, et al. 2010) and (Sankaran and Orkin 2013).

1.3. Human Hemoglobins

The normal human Hbs are composed of two dimers of α- and *β-like* globin chains, consisting of 141 and 146 amino acids respectively. These heterotetramers of globin chains change their composition during ontogeny, corresponding to the patterns of expression of the α- and *β-globin* clusters on chromosomes 16 and 11, respectively (figure 4) (Patrinos and Antonarakis 2010). From shortly after early embryonic development up to adulthood, human hemoglobins maintain identical α-globin chains, while β-like chains are replaced, as result of two critical switches in gene expression at embryonic-to-fetal and at fetal-to-adult transitions (Johnson, et al. 2002b; Sankaran, et al. 2010a; Schechter 2008).

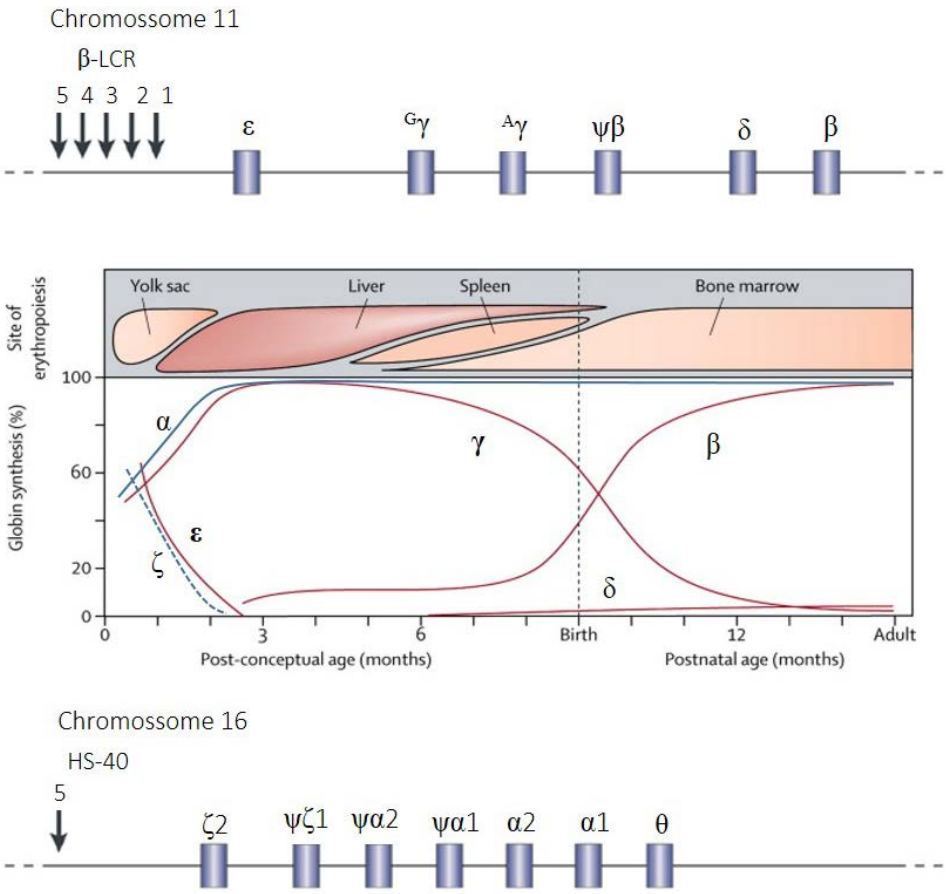


Figure 4 - Normal developmental switches in human globin gene expression

The structure of the *α*- and *β*-like globin gene clusters are shown together with the sites of hematopoiesis at different stages of development and the levels of expression of the embryonic, fetal, and adult globin chains at various gestational ages. The important control elements, HS-40 and the LCR, are also shown at their approximate locations and the vertical arrows indicate the location of *DNaseI* hypersensitive sites. Adapted from (Higgs, et al.) and (Weatherall 2001).

In the embryo, ζ-chains combine with γ (Hb Portland, $\zeta_2\gamma_2$) or ε-chains (Hb Gower 1, $\zeta_2\varepsilon_2$), and α- and ε-chains form Hb Gower 2 ($\alpha_2\varepsilon_2$). Embryonic hemoglobins are replaced in early embryonic life by HbF ($\alpha_2\gamma_2$), the predominant hemoglobin expressed in the fetus. After birth, HbF is replaced by the major and minor adult forms, HbA ($\alpha_2\beta_2$) and HbA_2 ($\alpha_2\delta_2$) respectively, although in normal adults small amounts of HbF, reaching approximately 1% of the total hemoglobin, continue to be produced (figure 5) (Patrinos and Antonarakis 2010).

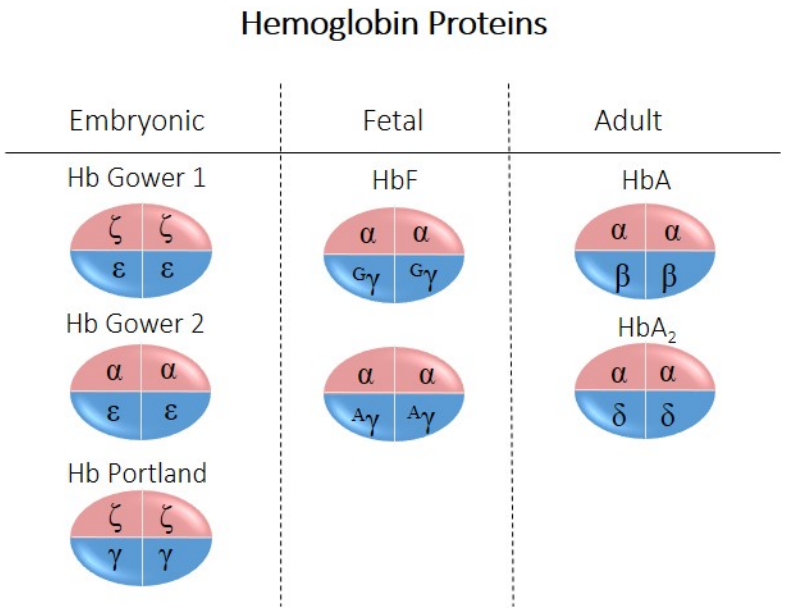


Figure 5 - Types of hemoglobin produced at each developmental stage.
The α-like and β-like chains are shown in pink and blue, respectively. Adapted from (Schechter 2008).

1.3.1 Inherited Disorders of Hemoglobin

It has been estimated that approximately 7% of the world's population are carriers of different inherited disorders of hemoglobin, which are considered the most common human monogenic disorders (Weatherall and Clegg 2001; Williams and Weatherall 2012). These disorders may be classified into two main groups: structural variants that change the amino acid sequence and produce an abnormal hemoglobin, and thalassemias that result from defective synthesis of the globin chains. There is a third group of conditions in which there is a defect in the normal switch from fetal to adult hemoglobin that is called HPFH, characterized by a persistence of HbF production through adult life and the

absence of any hematological disorder (Steinberg 2009). According to the HbVar (database of human hemoglobin variants and thalassemias) (Giardine, et al. 2014) over 1000 structural variants have been identified, but only four, sickle hemoglobin (HbS), HbC, HbE and HbD Punjab occur at a high frequency in different populations (Modell and Darlison 2008; Williams and Weatherall 2012). Their rise in frequency is a reflection of the protection of heterozygotes against severe malaria, the reason why they are most frequent in malaria-exposed populations (Williams and Weatherall 2012). Concerning the thalassemias, the most common types are the α- and β-thalassemia and there is strong evidence from population genetics data that malaria selection also explains their current distribution (Weatherall and Clegg 2008). Each of these forms of hemoglobin disorders present extremely diverse clinical phenotypes, particularly in the case of β-thalassemia in which there is remarkable variability in clinical severity (Weatherall 2001). The mechanisms underlying such phenotypic diversity are numerous, but an important factor in modifying the clinical course of β-thalassemia and sickle cell disease (SCD) is the variation in fetal hemoglobin production. A variety of clinical observations have shown that increased levels of HbF can be highly beneficial in attenuating the severity of these disorders (Akinsheye, et al. 2011; Thein 2004). These observations led to an attempt to develop therapeutic strategies to reactivate HbF production in a targeted manner (Bauer, et al. 2012). Despite the efforts, to date, a strategy that has proven to be effective in achieving the required levels of fetal hemoglobin for therapeutic benefit is still missing (Musallam, et al. 2013; Sankaran 2011). There is a great hope that a better understanding of the molecular mechanism involved in the regulation of the fetal hemoglobin genes and how the fetal to adult globin switch occurs in the course of human ontogeny may have important therapeutic implications (Sankaran and Nathan 2010). Recently, the therapeutic potential of the *δ-globin* gene in β-hemoglobinopathies was demonstrated and therefore increasing *δ-globin* gene expression could represent an alternative approach to the treatment of β-thalassemia and SCD (Manchinu, et al. 2014).

1.3.2. Functional Relevance of Hemoglobin A_2

Of the two adult forms of hemoglobin present in the erythrocytes of normal adults, the predominant is HbA whereas HbA_2 represents only a small fraction of the total adult Hb

(<3%) (Patrinos and Antonarakis 2010). Since HbA and HbA_2 share the same α-globin chain, the imbalance is due to the differential expression of the *δ-* and *β-globin* genes (Steinberg and Adams 1991; Steinberg 2009). Differences in the promoter sequence account for the diminished expression of *HBD* compared to *HBB* (Ristaldi, et al. 1999; Steinberg 2009). Given its low levels of expression, HbA_2 is assumed to be physiologically unimportant without any recognized function (Ranney, et al. 1993; Steinberg and Adams 1991). The δ-globin chain differs in only ten amino acids positions from the β-globin chain and HbA_2 has functional properties that are nearly identical to those of HbA (de Bruin and Janssen 1973). However, even in cases where HbA_2 becomes the predominant oxygen carrier, as occurs in the absence of β-chain production in patients suffering from β-thalassemia major, it never reaches the amount that would be necessary to effectively replace HbA function (Giambona, et al. 2009; Mosca, et al. 2009; Steinberg and Adams 1991). Still, HbA_2 levels are relevant for the diagnosis of β-globin disorders, since its elevated concentration (≥3,5%) is the most significant diagnostic parameter in the thalassemia syndromes, and its quantification plays a key role in β-thalassemia screening programs (Giambona, et al. 2009; Mosca, et al. 2009). Although *HBD* mutations are not pathological, the coinheritance of δ- and β-thalassemia may lead to misdiagnosis of carriers, because HbA_2 levels remain normal or low due to decreased δ-chain production (Bouva, et al. 2006; Morgado, et al. 2007).

1.3.3. Evolution of *δ-globin* Gene

The evolutionary history of eutherian *HBD* is quite complex due to unusually frequent sequence exchanges through extensive gene conversion and unequal recombination with its neighbor, *HBB*. The *HBD* gene has been independently converted by the *HBB* gene in multiple lineages resulting in extensive sequence homogenization and hampering the assignment of orthologous relationships among *HBD* and *HBB* genes. Initially *HBD* was thought to be the result of a recent *HBB* duplication in primate evolution, approximately 40 Mya (Efstratiadis, et al. 1980). Few years later, another study suggested that the *HBD* gene was present in the common ancestor of all mammals (Hardison 1984). Recently, an analysis based on a large number of vertebrate taxa established that the duplication event that generated the *HBD* gene occurred after the marsupial/eutherian

split, and is thus unique to Eutheria (Opazo, et al. 2008b). The *δ-globin* gene, although present in almost all eutherian species examined to date, is frequently pseudogenized (Gaudry, et al. 2014; Goodman, et al. 1984; Hardies, et al. 1984; Hardison 1984; Opazo, et al. 2009). In a few species, namely primates, a transcriptionally active but weakly expressed copy of the δ-globin gene was maintained, encoding the δ-globin chain of the minor fraction of the adult Hb ($\alpha_2\delta_2$), known as HbA_2, which is thus assumed to be physiologically irrelevant. Among primates, expression of the δ-globin chain is absent in Old World Monkeys (OWM) ((Martin, et al. 1983; Martin, et al. 1980) and ranges from 1% concentration in hominoids (Boyer, et al. 1971) to 40% in galago (Tagle, et al. 1991), reaching 6% in New World Monkeys (NWM) (Spritz and Giebel 1988) and 18% in tarsiers (Koop, et al. 1989). These low levels of expression are attributable to the weak *HBD* promoter (Steinberg 2009). Sequence exchanges between *HBB* and *HBD* have most often occurred in the coding regions (Hardies, et al. 1984; Hardison 1984; Hardison and Margot 1984; Opazo, et al. 2008a; Prychitko, et al. 2005), however, higher levels of *HBD* expression are achieved in species in which replacement with *β-globin* gene sequences extends into the promoter region, as in galago (Tagle, et al. 1991). Also, in all extant paenungulate mammals a chimeric β/δ fusion gene, created by unequal crossing-over between *HBD* and *HBB* paralogs, encodes the only β-globin chain produced during adulthood (Gaudry, et al. 2014; Opazo, et al. 2009). These studies reveal that during eutherian evolution *HBD* has not evolved independently but in concert with *HBB*. There is only one reported case of an unequal crossing-over between misaligned *HBD* and *HBBP1* sequences in the ancestry of lemurs, which resulted in a hybrid ψβ/δ pseudogene (Jeffreys, et al. 1982).

1.3.3.1. *δ-globin* Gene Diversity in Humans

Mutations in the β-chain of human HbA are associated with the most common inherited β-globin gene disorders world-wide (Weatherall and Clegg 2001; Williams and Weatherall 2012), as mentioned above, whereas mutations in *HBD* are *per se* clinically silent (Bouva, et al. 2006; Morgado, et al. 2007). Whereas we should expect *HBD* to be subject to a lower functional constraint than *HBB*, according to HbVar 583 *HBB* variants have been identified, while in *HBD* only 67 variants have been described (Giardine, et al. 2014).

Additionally, nucleotide diversity of the *HBD* gene in human populations was found to be lower than in *HBB* (Webster, et al. 2003). A major caveat regarding *HBD* genetic diversity is that its screening has been mainly motivated by diagnostic purposes, implying that estimates of *HBD* diversity are predominantly based on populations where β-globin disorders are prevalent. To clarify the unusual diversity patterns reported so far, a global sampling of populations would be necessary. Although it is consensual that HbA_2 has no important physiological function, *HBD* shows a pattern of sequence conservation typical of genes under strong selective pressures. A regulatory role of *HBD* in the fetal/adult Hb switch has been proposed decades ago (Bank, et al. 1980; Ottolenghi, et al. 1979) but has not been further addressed.

1.3.3.1.1. Tools for Genetic Diversity Analysis

Recent advances in the field of molecular and computational genomics have led to an enormous increase of publicly available genomic data. A major contributor to a deep characterization of human genome sequence variation has been the 1000 Genomes Project, by sequencing a large number of individual genomes from a number of different ethnic groups, which are of great value for population genetics and comparative genomic studies (Altshuler, et al. 2012). However, the currently available tools for genetic diversity analysis cannot handle the data format adopted by massive re-sequencing projects. The VCF format, developed by the 1000 Genomes Project and later adopted by other projects, such as UK10K, dbSNP and the NHLBI Exome Project, is a generic format for storing DNA polymorphism data such as SNPs, insertions, deletions and structural variants (Danecek, et al. 2011). Most of the existing programs for computing a variety of summary statistics of population genetic data, such as the most widely-used software packages DnaSP (Rozas, et al. 2003) and Arlequin (Excoffier, et al. 2005), deal with DNA sequences and not with haplotype sequences of single nucleotide polymorphisms (SNPs) stored in the VCF files. Fast improvements in computational and statistical tools that allow the extraction and analysis of information of large scale data stored in the VCF files became imperative. A program package for working with VCF files (VCFtools) (Danecek, et al. 2011), has been under development providing a set options for processing VCF files such as validating, merging, comparing and calculate some basic population genetic statistics. However, the

options available are still insufficient when the goal concerns the inspection of variation patterns along genomic fragments in which it is required to sequentially compute a variety of summary statistics of population genetic data over a "sliding window". In the course of the work developed in this thesis a new tool was developed to meet our analysis requirements.

CHAPTER 2. OBJECTIVES

In this study we sought a better understanding of the evolutionary forces acting on the *HBD* gene to gain insights into the physiological relevance of *δ-globin* gene conservation in some placental mammals. Our aims were:

1) To perform an unbiased reassessment of the diversity patterns at the *HBD* gene in human populations.

In order to accurately estimate diversity at the *HBD* gene, an unbiased characterization of the genetic diversity was performed using publicly available sequence data from the *β-globin* cluster, from five major population groups: African, European, American, and East and South Asian. The patterns of sequence variation were assessed by means of allele frequency spectrum, linkage disequilibrium (LD) and haplotype analyses.

1.1) Develop a new tool that allows the analysis of information stored in VCF files for computing a variety of summary statistics of population genetic data over a "sliding window", which is the best approach to investigate how patterns of variation change across genomic segments.

2) To gain insight into the evolutionary history of the *HBD* gene in eutherian mammals.

To elucidate the origin and evolution of the *HBD* gene, a comprehensive phylogenetic and comparative analysis of the two adult *β-like globin* genes was performed in a set of diverse mammalian taxa, focusing on the evolution and functional divergence of *HBD* in primates.

CHAPTER 3. RESULTS

3.1. Research Article:

"Evolutionary constraints in the *β-globin* cluster: the signature of purifying selection at the *δ-globin* (*HBD*) locus and its role in developmental gene regulation"

Genome, Biology & Evolution, 2013

(doi: 10.1093/gbe/evt029)

GBE

Evolutionary Constraints in the *β-Globin* Cluster: The Signature of Purifying Selection at the *δ-Globin* (*HBD*) Locus and Its Role in Developmental Gene Regulation

Ana Moleirinho[1,2,*], Susana Seixas[1], Alexandra M. Lopes[1], Celeste Bento[3], Maria J. Prata[1,2], and António Amorim[1,2]

[1]Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Portugal

[2]Department of Biology, Faculty of Sciences, University of Porto, Portugal

[3]Centro Hospitalar e Universitário de Coimbra, Serviço de Hematologia, Portugal

*Corresponding author: E-mail:amoleirinho@ipatimup.pt.

Accepted: February 16, 2013

Abstract

Human hemoglobins, the oxygen carriers in the blood, are composed by two α-like and two β-like globin monomers. The *β-globin* gene cluster located at 11p15.5 comprises one pseudogene and five genes whose expression undergoes two critical switches: the embryonic-to-fetal and fetal-to-adult transition. *HBD* encodes the δ-globin chain of the minor adult hemoglobin (HbA_2), which is assumed to be physiologically irrelevant. Paradoxically, reduced diversity levels have been reported for this gene. In this study, we sought a detailed portrait of the genetic variation within the *β-globin* cluster in a large human population panel from different geographic backgrounds. We resequenced the coding and noncoding regions of the two adult *β-globin* genes (*HBD* and *HBB*) in European and African populations, and analyzed the data from the *β-globin* cluster (*HBE*, *HBG2*, *HBG1*, *HBBP1*, *HBD,* and *HBB*) in 1,092 individuals representing 14 populations sequenced as part of the 1000 Genomes Project. Additionally, we assessed the diversity levels in nonhuman primates using chimpanzee sequence data provided by the PanMap Project. Comprehensive analyses, based on classic neutrality tests, empirical and haplotype-based studies, revealed that *HBD* and its neighbor pseudogene *HBBP1* have mainly evolved under purifying selection, suggesting that their roles are essential and nonredundant. Moreover, in the light of recent studies on the chromatin conformation of the *β-globin* cluster, we present evidence sustaining that the strong functional constraints underlying the decreased contemporary diversity at these two regions were not driven by protein function but instead are likely due to a regulatory role in ontogenic switches of gene expression.

Key words: *β-globin* cluster, hemoglobin switch, gene diversity, chromatin interactions.

Introduction

Hemoglobin (Hb) is the major protein in the circulating human red blood cells and its main function is to transport oxygen (O_2). Human Hb is a tetramer composed of two dimers of α-like and β-like globin chains, which differ according to developmental stage. From shortly after early embryonic development up to adulthood, normal human hemoglobins maintain identical α-globin chains, while β-like chains are replaced, as result of two critical switches in gene expression, the first at the embryonic-to-fetal transition and the second at the fetal-to-adult one (Johnson et al. 2002; Schechter 2008; Sankaran et al. 2010).

The five human *β-globin* paralogs that code for the different β-like chains are clustered at chromosome 11 together with one pseudogene, being arranged as 5′-ε(*HBE*)-^{G}γ(*HBG2*)-^{A}γ(*HBG1*)-ψβ(*HBBP1*)-δ(*HBD*)-β(*HBB*)-3′, in a region extending over approximately 80 kb. The stage-specific expression of each of these genes proceeds sequentially from embryonic (*HBE),* to fetal (*HBG2* and *HBG1*), and finally to adult genes (*HBD* and *HBB*) and relies on the interactions with the locus control region (LCR), located from approximately 6 to 18 kb upstream of *HBE* (fig. 1*A*) (Bulger and Groudine 1999; Tolhuis et al. 2002; Bank 2006). In adulthood, *HBB* expression levels are much higher than those of its

neighbor, *HBD*, resulting in two major Hb tetramers: HbA ($\alpha_2\beta_2$), which accounts for approximately 97% of the total Hb, incorporates β-chains produced by *HBB*, and HbA_2 ($\alpha_2\delta_2$), the minor fraction of adult Hb (generally $< 3\%$), which contains the δ-chains encoded by *HBD* (Schechter 2008).

Mutations in the β-chain of human HbA are associated with the most common inherited *β-globin* gene disorders world-wide (WHO 2011), such as β-thalassemia (Galanello and Origa 2010) and sickle cell disease (Orkin and Higgs 2010). On the contrary, HbA_2, the minor adult Hb, is assumed to be physiologically irrelevant and mutations in *HBD* are per se clinically silent (Steinberg and Adams 1991; Schechter 2008). At the functional level, HbA_2 has features that are nearly identical with those of HbA (de Bruin and Janssen 1973) and in the absence of β-chain production, as occurs in patients suffering from β-thalassemia major, HbA_2 becomes the predominant oxygen carrier. However, HbA_2 never reaches the amount that would be necessary to effectively replace HbA function (Steinberg and Adams 1991; Giambona et al. 2009; Mosca et al. 2009), which implies that its levels are only relevant for the diagnosis of *β-globin* disorders. Indeed, an elevated HbA_2 concentration ($\geq 3.5\%$) is the most significant parameter in the diagnosis of thalassemia syndromes (Cao and Moi 2000; Thein 2005; Galanello and Origa 2010), justifying the key role that HbA_2 measurement plays in β-thalassemia screening programs (Giambona et al. 2009; Mosca et al. 2009).

HBD, encoding the unique δ-globin chain of HbA_2, arose via duplication of the *HBB* gene after the marsupial/eutherian split and is therefore unique to placental mammals (Opazo et al. 2008). *HBD* has been inactivated or deleted in some lineages (Goodman et al. 1984; Hardies et al. 1984), but maintained an intact open reading frame in a few primate species, namely in humans, apes, and New World monkeys (Martin et al. 1980; Spritz and Giebel 1988). Human *HBD* and *HBB* show a high degree of homology (93%), as reflected in the similarity of their encoded proteins that only differ in 10 out of 147 amino acids (Steinberg and Adams 1991). Nonallelic gene conversion is the most commonly accepted explanation for such sequence homogeneity, albeit few gene conversion events have been described in the evolution of *HBD* and *HBB* (Papadakis and Patrinos 1999; Borg et al. 2009).

Given the apparent physiological redundancy of the HbA_2 protein, the maintenance of *HBD* in several primate lineages is intriguing. Even though we might expect *HBD* to be subject to much lower selective pressure than *HBB*, in a previous study analyzing a small number of African individuals *HBD* was found to have lower diversity levels than *HBB* (Webster et al. 2003). In line with this finding, very few *HBD* mutations causing δ-thalassemia have been described (Steinberg and Adams 1991; Patrinos et al. 2004b). This pattern of sequence conservation, typical of genes under strong evolutionary constraints (Zhang 2003), is puzzling given the low expression levels of *HBD* and the presumed negligible functional role of HbA_2 in oxygen transport (Steinberg and Adams 1991). One possible explanation for this apparent inconsistency could be that genetic variation has been more exhaustively assessed for *HBB* than for *HBD*. Even though the *β-globin* cluster is among the most extensively studied regions in the human genome, current genetic diversity estimates for the *HBD* and *HBB* genes across human populations are likely to be biased, because the genetic analysis of *HBD* and *HBB* has been performed mainly for diagnostic purposes and often based on a set of pre-ascertained single nucleotide polymorphisms (SNPs) (Morgado et al. 2007; Lacerra et al. 2008; Liu et al. 2009; Phylipsen et al. 2011).

In this study, we sought a better understanding of the evolutionary forces acting on *HBD* and *HBB* genes as well as of the physiological relevance of δ-globin conservation in placental mammals. To this end, we performed an unbiased characterization of the genetic diversity at the *β-globin* cluster based on Sanger sequencing of both *HBB* and *HBD* in ethnically diverse samples from Europe and Africa and on the sequence data from the 1000 Genomes Project Consortium (Altshuler et al. 2012). Then, we have evaluated the patterns of sequence variation, by means of allele frequency spectrum, linkage disequilibrium (LD) and haplotype structure analyses. Finally, we also assessed the diversity levels in this cluster in nonhuman primates using chimpanzee sequence data provided by the PanMap Project. Our findings indicate that purifying selection has shaped the evolutionary history of *HBD* and surprisingly, the same seems to apply to *HBBP1*. Furthermore, we present evidence that strong functional constraints have contributed to reduce the contemporary diversity of these two regions, probably due to a regulatory role in ontogenic switches of gene expression.

Materials and Methods

Population Samples and Sanger Sequencing

Sequence variation for *HBB* and *HBD*, was surveyed in a total of 71 samples: 25 Portuguese samples (PT) were collected in Coimbra's university hospital (Centro Hospitalar e Universitário de Coimbra), from healthy individuals, under informed consent; 46 samples from the International HapMap Project, 23 CEU (Utah residents with Northern and Western European ancestry from the CEPH collection), and 23 YRI (Yoruba from Ibadan in Nigeria). Two DNA fragments of approximately 2 kb, one spanning the entire *HBD* gene and the other the *HBB* gene, were resequenced (fig. 1*A*). Primers for amplification and sequencing were designed using the latest version of the human genome assembly (GRCh37; http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/index.shtml). DNA fragments were amplified using polymerase chain reaction and sequenced with the BigDye Terminator v.3.1 Cycle Sequencing Kit and run on an Applied

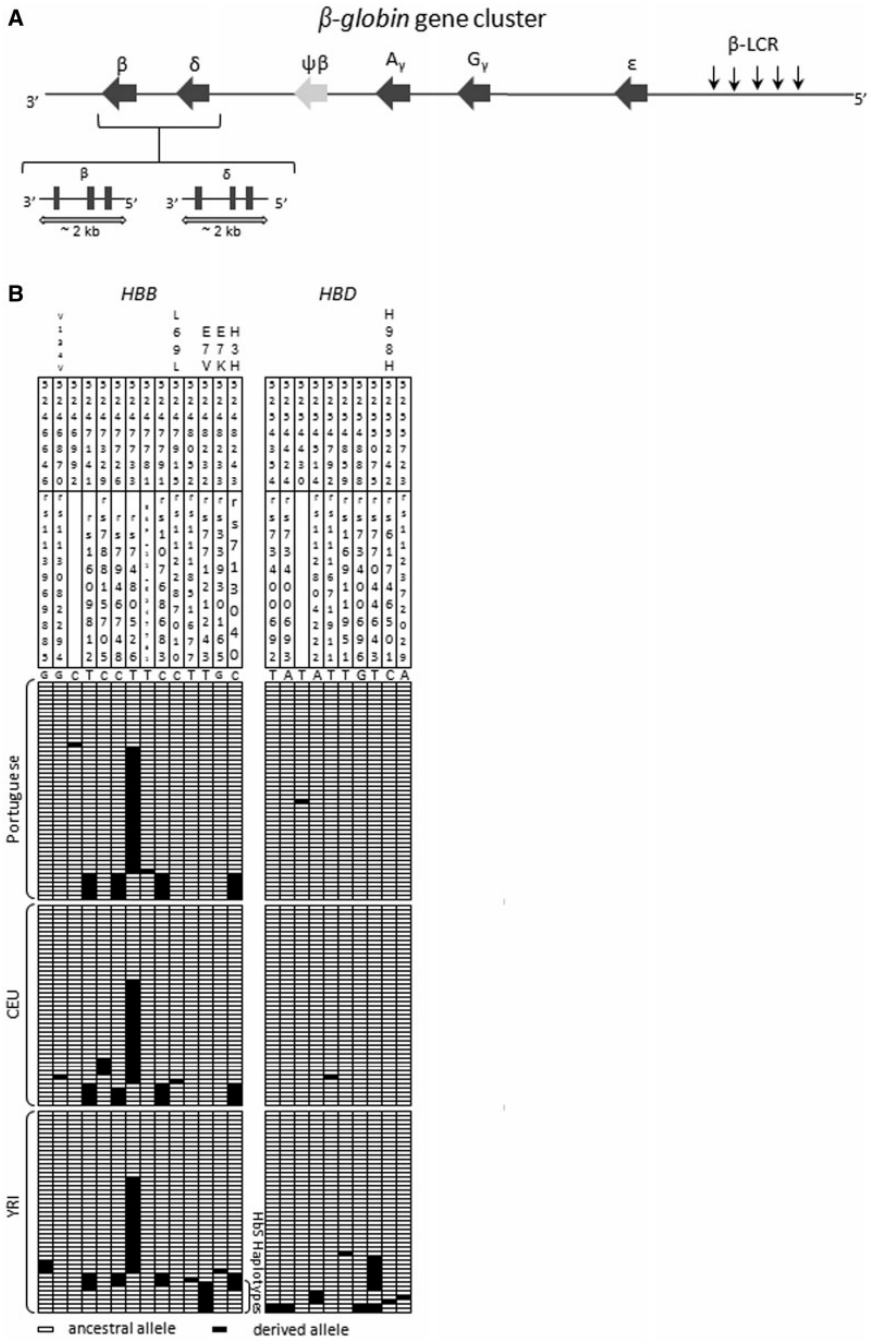


Fig. 1.—(*A*) Schematic representation of the *β-globin* cluster located at chromosome 11p15.5. Top: The relative position of the *β-globin* genes and the LCR in the cluster. Bottom: *δ-globin* (*HBD*) and *β-globin* (*HBB*) gene organization (exons are represented by grey squares). The arrows in the lower diagram indicate the extent of the segments surveyed in the resequencing study of the European and African populations. Small black arrows indicate the five *DNase I* hypersensitive sites (HS) encompassed by the LCR located from approximately 6 to 18 kb upstream of *ε- globin (HBE)*. (*B*) *HBD* and *HBB* haplotypes as inferred

(continued)

Biosystems ABI PRISM 3130xl Genetic Analyzer. All sequences were assembled and analyzed using Geneious version 5.4 created by Biomatters (available from http://www.geneious.com/) and all putative polymorphisms were manually inspected and individually confirmed.

Data Retrieval

Data from the 1000 Genomes Project were retrieved from its website (http://www.1000genomes.org/). Chromosomal locations and genomic segments from *HBB* (5246599–5248441), *HBD* (5253972–5255850), *HBBP1* (5263085–5265019), *HBG1* (5269347–5271272), *HBG2* (5274263–5276195), and *HBE* (5289469–5291330) were obtained using latest version of the human genome assembly GRCh37. Chimpanzee reference sequence was downloaded from the UCSC Genome Browser (UCSC: http://genome.ucsc.edu/) and used as outgroup. Chimpanzee sequence variation was downloaded from PanMap Project website (http://panmap.uchicago.edu/). Additionally, we performed a scan for regulatory elements in *HBD–HBBP1* region using data generated by the Encyclopedia of DNA Elements (ENCODE) Consortium (Myers et al. 2011), available in UCSC Genome Browser (http://genome.ucsc.edu/) in the Human GRCh37/hg19 assembly.

Statistical Analysis

The summary statistics of population genetic variation, number of segregating sites (S), nucleotide diversity (π) (Nei and Li 1979), which is based on average number of pairwise differences between sequences; Watterson's estimator of the population mutation rate parameter (θ_W) (Watterson 1975), which is based on the number of segregating sites and sample size, and Tajima's D (Tajima 1989) and Fay and Wu's H (Fay and Wu 2000), which summarizes information about the spectrum of allele frequencies, were calculated using SLIDER (http://genapps.uchicago.edu/slider/index.html). To assess the statistical significance of Tajima's D, we ran 100,000 coalescent simulations (Hudson 2002) using the previously estimated S statistic. Simulations were produced with the ''ms'' program, assuming distinct demographic models including constant population size and African and European best-fit models (Schaffner et al. 2005; Gutenkunst et al. 2009). To assess the statistical significance of Fay and Wu's H, we computed 10,000 coalescent simulations in DnaSP v.5.10 (Rozas 2009). Haplotypes of *HBB* and *HBD* were inferred using the program PHASE v.2.02 (Stephens et al. 2001; Stephens and Donnelly 2003). Haplotype data were then annotated with additional SNP information and ancestral allele. Ancestral allele state was retrieved from dbSNP (http://www.ncbi.nlm.nih.gov/). LD analyses were performed using Haploview v.4.2 (Barrett et al. 2005) and haplotype blocks were identified through the standard algorithm implemented in the software (Gabriel et al. 2002).

To provide a temporal dimension to the phylogenetic relationships among haplotypes and to estimate the coalescent times and ages of relevant mutations, we used GENETREE v.9.0 (Griffiths and Tavare 1994). Since GENETREE assumes no recombination two incompatible haplotypes were removed from the analysis of *HBB*. Time, scaled in 2Ne generations, was derived from $\theta = 4Ne\mu$. The mutation rate (μ) per gene, per generation, was deduced from the average number of nucleotide substitutions per site between human and chimpanzee reference sequences (Dxy), calculated with DnaSP v.5.10 (Rozas 2009). Time estimates in generations were converted into years using a 25-year generation time. Human/chimpanzee divergence was assumed to have occurred 5.4 Ma (Patterson et al. 2006).

To infer cladistic (network) relationships among the haplotypes, we used Network v.4.6, applying the Median-Joining method (Bandelt et al. 1999).

The evolutionary rates per site, per year, and per generation, were deduced from Jukes and Cantor distance calculated with DnaSP v.5.10 (Rozas 2009) and assuming the same human/chimpanzee divergence time.

Results

HBD and *HBB* Sequence Variation

We characterized the patterns of variation of the *HBD* and *HBB* genes by surveying two DNA fragments, each one spanning approximately 2 kb covering coding, noncoding, and the flanking 5′ and 3′ regions (fig. 1). Both segments were resequenced in a total of 71 samples belonging to populations from different geographic origins: Europeans (PT and CEU) and Africans (YRI). Overall, we identified 25 polymorphic sites, including a 2-bp deletion in *HBD* intron 2, which was excluded from further analyses (fig. 1*B*). In *HBD*, we observed

FIG. 1.— Continued
by PHASE v.2.02. The ancestral state at each site was inferred from ortholog nonhuman primate sequences. From the 25 sites, only 2 lacked a previously associated reference identification code in public databases (dbSNP and Exome Sequencing Project release ESP5400) and were unique to the Portuguese population. Coding variants are labeled. These include the following: one synonymous amino acid replacement in *HBB* (H3H) for the PT population; three synonymous replacements in *HBB* (H3H, L69L, and V134V) for the CEU population; one synonymous replacement in *HBD* (H98H) and one synonymous in *HBB* (H3H) and two nonsynonymous replacements (E7K–HbC allele, E7V–HbS allele) in *HBB* for the YRI population. SNP identifiers as in dbSNP and their chromosomal position based on GRCh37 version are indicated in each column.

10 SNPs: 1 synonymous and 9 noncoding; and in *HBB*, we found 14 SNPs including 2 nonsynonymous, 2 synonymous, and 10 noncoding. The two nonsynonymous replacements, identified only in the YRI population, were the HbS (*HBB*:c.20A > T) and HbC (*HBB*:c.19G > A) alleles, which are known to confer resistance to *Plasmodium falciparum* and to occur at the highest frequencies in Africa, in endemic areas of malaria (Kwiatkowski 2005). Contrary to *HBD* in which the low frequency variants (singletons and doubletons) represented 80% of polymorphic sites, the same category of variants in *HBB* included only 36% of the sites. The frequencies for HbS and HbC alleles were 15% and 2%, respectively. The HbS alleles were linked to divergent haplotypes, which are likely to correspond to 3 out of the 5 "classical" β^S haplotype backgrounds that are named according to their putative geographical origins (Benin, Bantu, Cameron, Senegal, and Arab) (Pagnier et al. 1984).

HBD and *HBB* Polymorphism Levels and Neutrality Tests

Standard population statistics based on the polymorphism levels as summarized by nucleotide diversity (π), and by the estimator of the population mutation rate parameter θ_w (Watterson 1975) are shown in table 1. Tajima's *D* (Tajima 1989) tends to be slightly negative in populations of African descent while it is frequently associated to more positive values in populations of European descent (Wall and Przeworski 2000; Frisse et al. 2001; Akey et al. 2004; Stajich and Hahn 2005; Voight et al. 2005). In our data set, Tajima's *D* statistics at *HBD* differs from the common trend showing significantly negative values in all populations analyzed PT, CEU, and YRI (table 1). This result is mainly due to a skew toward low frequency variants in YRI and to the lack of variation in CEU and PT. On the other hand, the Tajima's *D* values estimated for *HBB* are similar in the three populations, moderately negative but nonsignificant (table 1).

To confirm the significant departure of *HBD* from the expectations under the neutral equilibrium we generated theoretical null distributions for calibrated models of human demography by coalescent simulations (Schaffner et al. 2005; Gutenkunst et al. 2009). The significant results obtained with such tests (table 1) provide arguments for a nondemographic interpretation of the low variation levels at *HBD*. In addition, the diversity patterns observed at the adjacent gene (*HBB*) seem to favor a selective hypothesis for the evolution of *HBD* rather than a population expansion that would have affected both genes equally. Two alternative selective hypotheses could explain the significant departure from neutrality of *HBD*: first, strong purifying selection that would purge any new deleterious mutations and second, a complete selective sweep in which an advantageous variant reached fixation. Considering the functional redundancy of HbA_2, it is difficult to accept a protein modification as the likely target for positive selection. The single fixed difference in *HBD* between humans and great apes replaces two functionally equivalent amino acids, a valine by methionine (V127M) and it is already present at the Denisova sequence (Reich et al. 2011; Meyer et al. 2012). This gives a minimum time frame of 600,000–800,000 years for the origin of M127, which would be incompatible as well with a recent complete selective sweep. The possibility of a noncoding variant under positive selection cannot be excluded; however in that scenario, we would expect other typical features of selection like long and homogeneous haplotypes, high levels of population differentiation and unusual patterns of diversity given the observed divergence, which were not detected (Hudson et al. 1987; Fay and Wu 2000; Sabeti et al. 2002; Zeng et al. 2006; Mathias et al. 2012), as further described below.

Gene Genealogy and Age Estimates of *HBD* and *HBB*

We reconstructed the gene genealogy of *HBD* and *HBB,* and estimated the time to the most recent common ancestor (T_{MRCA}) using a maximum likelihood coalescent analysis (Griffiths and Tavare 1994). The results represented in figure 2, reveal that the tree of *HBD* differs sharply in its

Table 1
Summary Statistics of Population Variation for the Two Adult *β-globin* Genes

Population	*N*[a]	*HBD*					*HBB*				
		L[b]	*S*[c]	π[d]	θ_w[e]	TD[f]	*L*	*S*	π	θ_w	TD
PT	50	1,879	1	0.21	1.18	−1.10*	1,843	7	7.82	8.49	−0.21
CEU	46	1,879	1	0.23	1.21	−1.11*	1,843	8	8.23	9.89	−0.46
YRI	46	1,879	8	4.55	9.66	−1.53**	1,843	9	8.11	11.12	−0.77

[a]Number of chromosomes.
[b]Total number of sites surveyed.
[c]Polymorphic sites.
[d]Nucleotide diversity ($\times 10^4$) (Nei and Li 1979).
[e]Watterson θ per site ($\times 10^4$) (Watterson 1975).
[f]Tajima's *D* statistic (Tajima 1989).
*$P \leq 0.001$ according to the constant size model, to the best fit model from Schaffner et al. (2005) and to the best fit model from Gutenkunst et al. (2009).
**$P \leq 0.05$ according to the constant size model and to the best fit model from Gutenkunst et al. (2009).

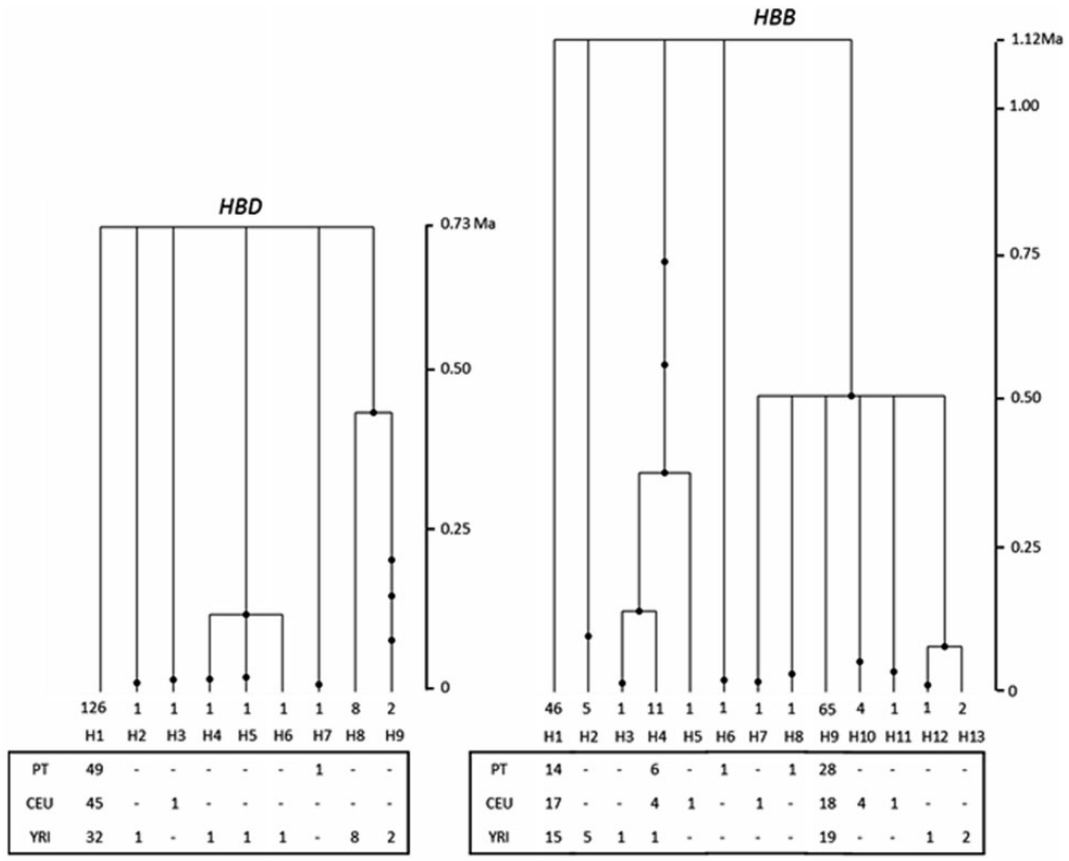


Fig. 2.—*HBD* and *HBB* gene genealogies as estimated by Genetree. Time is scaled in millions of years (Myr). Solid circles represent nucleotide substitutions. The number below each branch of the trees represents the chromosomes observed for each haplotype, and in the lower diagram this information is split by population.

topology from that of *HBB*. Theoretically, *HBD* tree shape could be attributed either to population expansion, to positive or background selection (Harpending et al. 1998; Bamshad and Wooding 2003; Tishkoff and Verrelli 2003). The estimated T_{MRCA} of *HBD*, 0.73 ± 0.33 Myr, is much younger than the estimated for *HBB* (1.12 ± 0.41 Myr), the latter being in full agreement with both observed and expected T_{MRCA} for human autosomal genes (Excoffier 2002; Tishkoff and Verrelli 2003; Garrigan and Hammer 2006; Kim et al. 2010). Given the previous findings and the rejection of a demographic hypothesis, these results provide further support for a nonneutral interpretation of *HBD* evolution. The contrasting tree structure and recent T_{MRCA} of *HBD* relative to *HBB*, can instead be more parsimoniously attributed to strong purifying selection, which is expected to show more recent coalescence times than under neutrality (Bamshad and Wooding 2003).

β-Globin Gene Cluster Diversity Patterns

To obtain a more comprehensive assessment of the patterns of diversity on the entire *β-globin* gene cluster, we analyzed the data available from the 1000 Genomes Project phase 1 release v.3 (Altshuler et al. 2012), generated from the genome sequencing of 1,092 individuals belonging to 14 populations and five major populations groups: African, European, American, and East and South Asian (supplementary table S1, Supplementary Material online). We first analyzed the patterns of LD in the 80 kb region encompassing the entire *β-globin* cluster, in the populations resequenced in our study (CEU and YRI) (supplementary fig. S1, Supplementary Material online). We observed in the full *β-globin* cluster two distinct regions with strong LD, one that contains *HBB* (LD region 1) and the other extending from *HBD* to the LCR (LD region 2). These two regions are separated by a segment that encompasses one of the first recombination hotspots identified in

Table 2
Summary Statistics of Population Variation for the *β-globin* Cluster Genes Using the 1,000 Genomes Project Data

	L^a	CEU					YRI				
		N^b	S^c	π^d	$\theta_w{}^e$	TD[f]	N^a	S^c	π^d	$\theta_w{}^e$	TD[f]
HBB	1,843	170	7	8.28	6.65	0.53	176	11	8.49	10.39	−0.44
HBD	1,879		4	0.31	3.73	−1.67		10	6.53	9.26	−0.70
HBBP1	1,935		11	8.75	9.96	−0.29		14	6.15	12.59	−1.30
HBG1	1,926		10	11.79	9.09	0.71		18	21.78	16.27	0.90
HBG2	1,933		11	17.47	9.97	1.83		16	22.09	14.41	1.39
HBE	1,862		4	5.39	3.76	0.79		13	7.39	12.15	−0.98

[a]Total number of sites surveyed.
[b]Number of chromosomes.
[c]Polymorphic sites.
[d]Nucleotide diversity ($\times 10^4$) (Nei and Li 1979).
[e]Watterson θ per site ($\times 10^4$) (Watterson 1975).
[f]Tajima's *D* statistic (Tajima 1989).

humans (Chakravarti et al. 1984; Smith et al. 1998; Wall et al. 2003). Then, we evaluated the polymorphism levels for the five *β-globin* cluster genes comprised in the LD region 2: *HBD, HBE*, *HBG2*, *HBG1,* and *HBBP1*, and for *HBB,* included in LD region 1. As shown in table 2, summary statistics of CEU and YRI populations are in agreement with the results from our resequencing study. In the complete 1000 Genomes data set, higher nucleotide diversity levels were observed in African populations (YRI, LWK, and ASW) (table 2 and supplementary file S1, Supplementary Material online). Notwithstanding, *HBD* consistently displayed reduced levels of nucleotide diversity and strongly negative Tajima's *D* values when compared with the remaining *β-globin* cluster genes, independently of the population studied (table 2 and supplementary file S1, Supplementary Material online). Interestingly, *HBBP1* sequence, 7 kb upstream of *HBD,* is also one of the less diverse regions in the cluster, only surpassed by *HBD* and *HBE*, the latter encoding the extremely conserved embryonic globin. Thus, diversity patterns observed for *HBD* and *HBBP1* suggest strong evolutionary constraints not related to protein function, as both are either marginally or not transcribed at all. Finally, we used the 1000 Genomes data to perform a sliding-window analysis of variation over the genomic region covering the entire *β-globin* cluster. As shown in figure 3, the genomic regions corresponding to *HBD* and *HBBP1* presented the lowest values of both nucleotide diversity and Tajima's *D*, only comparable with those obtained for the LCR region. Noteworthy, the intergenic region flanked by *HBBP1* and *HBD* shows the opposite trend, with high levels of nucleotide diversity and positive Tajima's *D* values, suggesting a complex evolutionary history possibly shaped by a noncanonical gene function.

HBBP1, *HBD*, and *HBB* Haplotype Analysis

We next focused on the comparison of the haplotype structure of the two adult genes, *HBD* and *HBB*. *HBBP1* was also included in the analysis because of the unusual low levels of diversity presented by this pseudogene. We reconstructed haplotype genealogies for *HBB*, *HBD,* and *HBBP1* using the full data set of the 1000 Genomes project. For *HBD,* a single common haplotype with an 88% frequency was identified. The star-shaped structure observed in the *HBD* network (supplementary figs. S2 and S3, Supplementary Material online) is in agreement with the atypical genetree of *HBD* built with our own resequencing data (CEU, PT, and YRI). Overall, these results give further support to the hypothesis of strong purifying selection operating on *HBD*, irrespectively of the diverse population demographic histories. The haplotype network for *HBBP1* might be consistent with purifying selection as well, even though its footprints are somewhat weaker than those at *HBD*.

Interspecies Comparisons

To gain further insight into the possible functional constraints that have been shaping the evolutionary history of this genomic region, we used human and chimpanzee sequences to calculate and compare the rate of divergence, in exons and introns, across *HBD*, *HBB,* and *HBBP1* (table 3). In *HBB*, introns display an evolutionary rate ~7 times higher than exons. Although higher divergence rates are expected for introns, in the case of *HBB* this likely reflects an increased intronic mutation rate as reported for *β-globin* genes (1.89%), when compared with the intronic average (1.03%) (Chen and Li 2001). Remarkably, *HBD* and *HBBP1 Homo–Pan* nucleotide differences are more homogeneously distributed between exons and introns, and the overall substitution rate is nearly half of that observed for *HBB* introns (supplementary table S2, Supplementary Material online), indicating again that these genes are under higher evolutionary constraints.

We next evaluated the haplotype diversity for these 3 genes in chimpanzees, by using sequence variation data from 10 Western chimpanzees (*Pan troglodytes verus*)

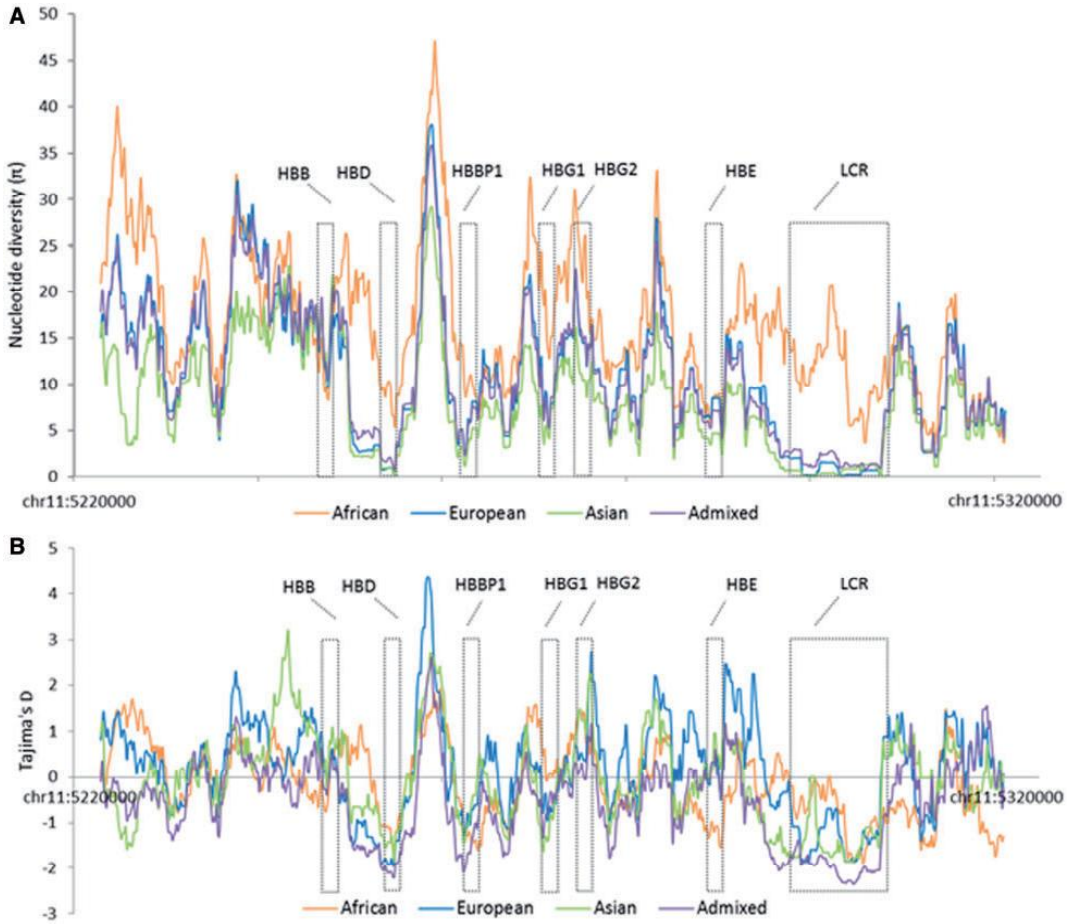


Fig. 3.—Sliding window analysis of the genomic region encompassing the *β-globin* gene cluster. Data were obtained from 1000 Genomes Project, representing 1,092 individuals from 14 populations: three African (ASW, LWK, and YRI), five European (CEU, FIN, GBR, IBS, and TSI), three Asian (CHB, CHS, and JPT) and three American-admixed populations (CLM, MXL, and PUR). Both π (*A*) and Tajima's *D* (*B*) were calculated in 2 kb windows with increments of 150 bp.

Table 3
Evolutionary Rates Based on Jukes–Cantor Distance

	Exons			Introns		
	Div[a]	Mutation Rate[b]		Div[a]	Mutation Rate[b]	
		Year	Generation		Year	Generation
HBD	0.80	0.74×10^{-9}	1.84×10^{-8}	0.99	0.91×10^{-9}	2.28×10^{-8}
HBB	0.26	0.24×10^{-9}	0.60×10^{-8}	1.75	1.62×10^{-9}	4.06×10^{-8}
HBBP1	1.45	1.34×10^{-9}	3.36×10^{-8}	1.04	0.96×10^{-9}	2.41×10^{-8}

[a]Average nucleotide divergence between human and chimpanzee.
[b]Mutation rate per site.

generated by the PanMap Project (supplementary table S2, Supplementary Material online), which has revealed that *HBD* and *HBBP1* presented lower nucleotide diversity relative to *HBB*. These results resemble those obtained for the different human populations, and therefore it seems likely that *HBD* and *HBBP1* are under purifying selection also in chimpanzees.

Scan for Regulatory Elements in *HBD–HBBP1* Region

The low levels of diversity found in *HBD* and *HBBP1* suggest that both sequences are evolving under purifying selection which cannot be attributed to constraints on a functional protein. An alternative explanation is that *HBD* and *HBBP1* lie within crucial regions for the regulation of gene transcription, in which selective pressures would be acting to maintain the nucleotide sequence. To explore this possibility, we analyzed the data generated by the Encyclopedia of DNA Elements (ENCODE) Consortium (Myers et al. 2011), available through the UCSC Genome Browser (http://genome.ucsc.edu/). A large number of transcription factor binding sites have been experimentally detected by ChIP-seq throughout the *β-globin* cluster, some of which with an established role in chromatin remodeling, namely the zinc finger protein (CTCF–CCCTC-binding factor), the C2H2 type zinc-finger protein (BCL11A) and the globin transcription factor 1 (Gata-1) (Vakoc et al. 2005; Splinter et al. 2006; Hou et al. 2010; Xu et al. 2010). Previous studies have also analyzed the long-range chromatin interactions in the *β-globin* cluster by Chromosome Conformation Capture Carbon Copy (5C) (Dostie et al. 2006; Sanyal et al. 2012) (supplementary fig. S4*A*, Supplementary Material online). Noteworthy, significant interactions were detected between a segment comprising both *HBD* and *HBBP1*, and different regions upstream *HBE*, which overlap the LCR. Moreover, these interactions were detected specifically in the erythropoietic cell line K562, in which *β-globin* genes are actively transcribed. Interestingly, in this cell line both the LCR and *HBD* lie in regions of open chromatin, as determined by DNAse I hypersensitivity and FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) assays, two different methods to identify nucleosome-depleted regions of the genome (supplementary fig. S4*B*, Supplementary Material online). The specific conformation observed in K562 cells suggests that this structure has a role in maintaining an active transcriptional state of the *β-globin* cluster in this cell line. In fact, both *HBD* and *HBBP1* may act as anchor regions in LCR-driven chromatin looping, a crucial mechanism for temporal coordination of gene expression in the human *β-globin* cluster (Bulger and Groudine 1999; Patrinos et al. 2004a).

Discussion

Unusual low levels of diversity have been described for human *HBD,* which are difficult to reconcile with the negligible function of HbA_2 in oxygen transport (Webster et al. 2003). In primates *HBD* exhibits a surprisingly high level of sequence conservation relative to functional paralogs (Steinberg and Adams 1991), suggesting an important role of *HBD* in those lineages. Here, we gain insight into the evolutionary history of *HBD* and its potential function in gene regulation by performing a detailed analysis of the sequence diversity and divergence of the *β-globin* cluster in humans and chimpanzees. Our results demonstrate that the *HBD* lack of diversity is a common trend across human populations that cannot be due neither to a systematic bias of clinical studies nor to an effect of population history. Even though recent evidence point to a human explosive growth over the past 10,000 years leading to an increment in very low frequency variants (Keinan and Clark 2012), the significance of the negative Tajima's *D* values obtained under the best-fit models discourage a demographic interpretation of *HBD* diversity. Furthermore, the findings of strong negative Tajima's *D* values in the absence of other features of positive selection, pinpoint *HBD* as a target of purifying selection like previously reported for innate-immunity genes under strong functional constraints (Barreiro et al. 2009; Mukherjee et al. 2009; Wlasiuk and Nachman 2010).

An atypical tree topology was identified for *HBD* consistent with the loss of mild deleterious variants by negative selective pressures and a low T_{MRCA} of 0.73 ± 0.33 Myr (or 0.58 ± 0.26 Myr for a generation time of 20 years), amongst the most recent estimates obtained for autosomal genes (0.20–0.31 Myr for a generation time of 20 years), was confirmed (Fullerton et al. 2000; Martinez-Arias et al. 2001; Excoffier 2002; Tishkoff and Verrelli 2003; Webster et al. 2003; Garrigan and Hammer 2006; Kim et al. 2010). Conversely, the polymorphism levels and T_{MRCA} estimate obtained for *HBB* are in agreement with the expectations under a neutral model and do not differ from previous studies (Harding et al. 1997).

Taken together, the results obtained for the *β-globin* cluster cannot be reconciled with other explanatory hypotheses rather than purifying selection: the low values of nucleotide diversity are confined to *HBD* and *HBBP1* and do not extend into the flanking regions which display contrastingly high levels of variation; the haplotype structure of *HBD* and *HBBP1* are similar across worldwide populations; furthermore, chimpanzee *HBD* and *HBBP1* also exhibit lower haplotype diversity than *HBB* and their mutation rates, as inferred from human–chimpanzee divergence, are considerably reduced when compared with *HBB* and to other *β-globin* noncoding regions (Chen and Li 2001). Evidence here presented suggests a long-term effect of purifying selection probably predating modern human origins (200,000 years), with the same strong functional constrains acting across different primate species for at least 5 Myr.

Several decades ago, a hypothesis was formulated holding an important regulatory role of *HBD* and *HBBP1* in the Hb

fetal-to-adult switch that matches quite well the assumption of strong negative selective forces acting on these sequences (Ottolenghi et al. 1979; Bank et al. 1980; Chang and Slightom 1984; Goodman et al. 1984). Over the past years, the *β-globin* cluster has been regarded as a complex genetic system and a paradigm of gene expression regulation. More recently, a boost of studies on the *β-globin* cluster have contributed to a better understanding of the mechanisms underlying the regulation of each gene in the cluster (Harju et al. 2002; Chakalova et al. 2005; Noordermeer and de Laat 2008; Sankaran et al. 2010). Remarkably, chromosome conformation (3C and 5C) analyses for the *β-globin* locus disclosed strong interactions between the LCR and the region encompassing both *HBD* and *HBBP1* (Dostie et al. 2006; Sanyal et al. 2012). Furthermore, distinct spatial interactions of the LCR in fetal and adult stages were uncovered by another study based only in 3C assay in which *HBD* sequence was proposed to be enrolled in the maintenance of a transcriptionally competent structure at the adult stage (Beauchemin and Trudel 2009). These recent findings suggest that *HBD* and *HBBP1* might be involved in chromatin looping in the human *β-globin* cluster, a crucial mechanism for temporal coordination of gene expression (Holwerda and De Laat 2012). Importantly, one SNP (rs10128556) in *HBBP1* has been also identified as a modulator of HbF levels reinforcing the idea that this genomic region is indeed involved in the Hb fetal-to-adult switch (Galarneau et al. 2010). Considering that the mechanism of Hb switch is common to all simian primates (Johnson et al. 2002), we might expect to find similar patterns of conservation and diversity in orthologous sequences of *HBD* and *HBBP1* and to detect signatures of purifying selection at the *β-globin* cluster over 40 Myr of primate evolution.

In contrast to the low diversity levels of *HBD* and *HBBP1* are the high levels found in the *HBD-HBBP1* intergenic region. This pattern has been previously observed in a study involving 23 individuals from the Luo population (Kenya, Africa) (Webster et al. 2003). There, the authors found a significant positive Tajima's *D* and two divergent haplotypes "R" and "T" regarded either as relics of a subdivided ancestral hominid population or as a signature of balancing selection (Maeda et al. 1983; Webster et al. 2002, 2003). It is known that the *HBD–HBBP1* intergenic region contains a small segment of 1 kb flanked by two opposite orientated Alu elements, which is expected to increase genetic instability and local mutation rate (Wang and Vasquez 2006), therefore providing an explanation for the observed high diversity levels. Also of note, is the fact that a 3.5 kb segment upstream of *HBD* was proven to be necessary for *HBG1* and *HBG2* gene silencing since its deletion causes an increment in HbF production throughout adulthood (Sankaran et al. 2011). This segment overlaps a binding region of BCL11A, which is a biochemically validated and fundamental switching factor (Sankaran et al. 2008, 2009), and contains other functional elements, such as a polypyrimidine tract that can serve as a binding site for multiprotein chromatin remodeling complexes (Bank et al. 2005). Furthermore, it has been demonstrated that the binding of protein complexes in this region induces conformational changes in the *β-globin* cluster, which in turn mediate long-range physical interactions between distal regulatory elements located in the LCR (Xu et al. 2010). Therefore, we propose that the two alternative haplotype configurations "R" and "T" may represent nucleotide combinations that preserve specific motifs for binding of protein complexes involved in transcriptional regulation and chromatin remodeling.

Collectively, these findings illustrate the complexity of the regulatory mechanisms within the cluster and provide a rationale for the heterogeneity in diversity levels observed within the *HBD–HBBP1* region. In the yeast genome, higher conservation was observed in nucleosome free regions, compared with regions with high nucleosome occupancy, irrespective of their overlap with coding sequences, which underscores the correlation between chromatin structure and evolutionary constraints (Nikolaou et al. 2010). Accordingly, a recent study integrating ENCODE and 1000 Genomes data uncovered a widespread signature of purifying selection on regulatory regions in humans, including promoters, enhancers, insulators and other regions with different chromatin conformational states (Ward and Kellis 2012). The authors also noticed that many of these regions are lineage specific and a fraction of these arose in the common ancestor of primates. We hypothesize that the clear signature of purifying selection at *HBD* and *HBBP1* may reflect constraints on local chromatin conformation and the maintenance of a nucleosome free region available for frequent interactions with the LCR.

In summary, the results here presented indicate that purifying selection is driving not only *HBD* evolution but also its neighbor pseudogene, *HBBP1*. In the light of recent advances in the characterization of the *β-globin* cluster, we propose that the complex patterns of diversity observed in this genomic region arose from distinct functional constraints related with the intricate process of chromatin and protein interactions coordinating the differential expression of genes at the *β-globin* cluster during development.

Supplementary Material

Supplementary tables S1 and S2, file S1, and figures S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Dr Letícia Ribeiro for her collaboration in providing the Portuguese samples. This work was supported by the Portuguese Foundation for Science and Technology

(FCT) fellowship (SFRH/BD/73508/2010 and SFRH/BPD/73366/2010) to A.M. and A.M.L., respectively, and by the POPH-QREN – Promotion of scientific employment, the European Social Fund, and national funds of the Ministry of Education and Science grants to S.S. IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Education and Science and is partially supported by FCT.

Literature Cited

Akey JM, et al. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. 2:7.
Altshuler DM, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.
Bamshad M, Wooding SP. 2003. Signatures of natural selection in the human genome. Nat Rev Genet. 4:99–111.
Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 16:37–48.
Bank A. 2006. Regulation of human fetal hemoglobin: new players, new complexities. Blood 107:435–443.
Bank A, Mears J, Ramirez F. 1980. Disorders of human hemoglobin. Science 207:486–493.
Bank A, et al. 2005. Role of intergenic human γ-δ-globin sequences in human hemoglobin switching and reactivation of fetal hemoglobin in adult erythroid cells. Ann N Y Acad Sci. 1054:48–54.
Barreiro LB, et al. 2009. Evolutionary dynamics of human toll-like receptors and their different contributions to host defense. PLoS Genet. 5: e1000562.
Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.
Beauchemin H, Trudel M. 2009. Evidence for a bigenic chromatin subdomain in regulation of the fetal-to-adult hemoglobin switch. Mol Cell Biol. 29:1635–1648.
Borg J, Georgitsi M, Aleporou-Marinou V, Kollia P, Patrinos GP. 2009. Genetic recombination as a major cause of mutagenesis in the human globin gene clusters. Clin Biochem. 42:1839–1850.
Bulger M, Groudine M. 1999. Looping versus linking: toward a model for long-distance gene activation. Genes Dev. 13:2465–2477.
Cao A, Moi P. 2000. Genetic modifying factors in β-thalassemia. Clin Chem Lab Med. 38:123–132.
Chakalova L, et al. 2005. Developmental regulation of the β-globin gene locus. In: Jeanteur P, editor. Epigenetics and chromatin. Berlin (Germany): Springer. p. 183–206.
Chakravarti A, et al. 1984. Nonuniform recombination within the human beta-globin gene cluster. Am J Hum Genet. 36:1239–1258.
Chang LYE, Slightom JL. 1984. Isolation and nucleotide sequence analysis of the β-type globin pseudogene from human, gorilla and chimpanzee. J Mol Biol. 180:767–783.
Chen FC, Li WH. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet. 68: 444–456.
de Bruin SH, Janssen LHM. 1973. Comparison of the oxygen and proton binding behavior of human hemoglobin A and A2. Biochim Biophys Acta. 295:490–494.
Dostie J, et al. 2006. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 16:1299–1309.
Excoffier L. 2002. Human demographic history: refining the recent African origin model. Curr Opin Genet Dev. 12:675–682.
Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413.
Frisse L, et al. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet. 69:831–843.
Fullerton SM, et al. 2000. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am J Hum Genet. 67: 881–900.
Gabriel SB, et al. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.
Galanello R, Origa R. 2010. Beta-thalassemia. Orphanet J Rare Dis. 5:11.
Galarneau G, et al. 2010. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nat Genet. 42:1049–1051.
Garrigan D, Hammer MF. 2006. Reconstructing human origins in the genomic era. Nat Rev Genet. 7:669–680.
Giambona A, Passarello C, Renda D, Maggio A. 2009. The significance of the hemoglobin A2 value in screening for hemoglobinopathies. Clin Biochem. 42:1786–1796.
Goodman M, Koop BF, Czelusniak J, Weiss ML, Slightom JL. 1984. The eta-globin gene: its long evolutionary history in the beta-globin gene family of mammals. J Mol Biol. 180:803–823.
Griffiths RC, Tavare S. 1994. Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci. 344: 403–410.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5: e1000695.
Hardies SC, Edgell MH, Hutchison CA. 1984. Evolution of the mammalian beta-globin gene cluster. J Biol Chem. 259:3748–3756.
Harding RM, et al. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet. 60:772–789.
Harju S, McQueen KJ, Peterson KR. 2002. Chromatin structure and control of β-like globin gene switching. Exp Biol Med. 227: 683–700.
Harpending HC, et al. 1998. Genetic traces of ancient demography. Proc Natl Acad Sci U S A. 95:1961–1967.
Holwerda S, De Laat W. 2012. Chromatin loops, gene positioning and gene expression. Front Genet. 3:217.
Hou C, Dale R, Dean A. 2010. Cell type specificity of chromatin organization mediated by CTCF and cohesin. Proc Natl Acad Sci U S A. 107: 3651–3656.
Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.
Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–159.
Johnson RM, Gumucio D, Goodman M. 2002. Globin gene switching in primates. Comp Biochem Physiol A Mol Integr Physiol. 133: 877–883.
Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740–743.
Kim HL, Igawa T, Kawashima A, Satta Y, Takahata N. 2010. Divergence, demography and gene loss along the human lineage. Philos Trans R Soc B: Biol Sci. 365:2451–2457.
Kwiatkowski DP. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet. 77:171–192.
Lacerra G, et al. 2008. Molecular evidences of single mutational events followed by recurrent crossing-overs in the common delta-globin alleles in the Mediterranean area. Gene 410:129–138.
Liu L, et al. 2009. High-density SNP genotyping to define beta-globin locus haplotypes. Blood Cells Mol Dis. 42:16–24.

Maeda N, Bliska JB, Smithies O. 1983. Recombination and balanced chromosome polymorphism suggested by DNA sequences 5′ to the human delta-globin gene. Proc Natl Acad Sci U S A. 80: 5012–5016.
Martin SL, Zimmer EA, Kan YW, Wilson AC. 1980. Silent delta-globin gene in Old World monkeys. Proc Natl Acad Sci U S A. 77:3563–3566.
Martinez-Arias R, et al. 2001. Sequence variability of a human pseudogene. Genome Res. 11:1071–1085.
Mathias RA, et al. 2012. Adaptive evolution of the FADS gene cluster within Africa. PLoS One 7:e44926.
Meyer M, et al. 2012. A high-coverage genome sequence from an Archaic Denisovan individual. Science 338:222–226.
Morgado A, et al. 2007. Mutational spectrum of delta-globin gene in the Portuguese population. Eur J Haematol. 79:422–428.
Mosca A, Paleari R, Ivaldi G, Galanello R, Giordano PC. 2009. The role of haemoglobin A2 testing in the diagnosis of thalassaemias and related haemoglobinopathies. J Clin Pathol. 62:13–17.
Mukherjee S, Sarkar-Roy N, Wagener DK, Majumder PP. 2009. Signatures of natural selection are not uniform across genes of innate immune system, but purifying selection is the dominant signature. Proc Natl Acad Sci U S A. 106:7073–7078.
Myers RM, et al. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 9:19.
Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 76:5269–5273.
Nikolaou C, Althammer S, Beato M, Guigo R. 2010. Structural constraints revealed in consistent nucleosome positions in the genome of *S. cerevisiae*. Epigenetics Chromatin 3:1756–8935.
Noordermeer D, de Laat W. 2008. Joining the loops: beta-globin gene regulation. IUBMB Life 60:824–833.
Opazo JC, Hoffmann FG, Storz JF. 2008. Genomic evidence for independent origins of β-like globin genes in monotremes and therian mammals. Proc Natl Acad Sci U S A. 105:1590–1595.
Orkin SH, Higgs DR. 2010. Sickle cell disease at 100 years. Science 329: 291–292.
Ottolenghi S, et al. 1979. Globin gene deletion in HPFH, δ° β° thalassaemia and Hb Lepore disease. Nature 278:654–657.
Pagnier J, et al. 1984. Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. Proc Natl Acad Sci U S A. 81: 1771–1773.
Papadakis MN, Patrinos GP. 1999. Contribution of gene conversion in the evolution of the human β-like globin gene family. Hum Genet. 104:117–125.
Patrinos GP, et al. 2004a. Multiple interactions between regulatory regions are required to stabilize an active chromatin hub. Genes Dev. 18: 1495–1509.
Patrinos GP, et al. 2004b. Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. Nucleic Acids Res. 32: D537–D541.
Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103–1108.
Phylipsen M, Gallivan MVE, Arkesteijn SGJ, Harteveld CL, Giordano PC. 2011. Occurrence of common and rare δ-globin gene defects in two multiethnic populations: thirteen new mutations and the significance of δ-globin gene defects in β-thalassemia diagnostics. Int J Lab Hematol. 33:85–91.
Reich D, et al. 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet. 89: 516–528.
Rozas J. 2009. DNA sequence polymorphism analysis using DnaSP. Methods Mol Biol. 537:337–350.
Sabeti PC, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.
Sankaran VG, et al. 2008. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. Science 322: 1839–1842.
Sankaran VG, et al. 2011. A functional element necessary for fetal hemoglobin silencing. N Engl J Med. 365:807–814.
Sankaran VG, Xu J, Orkin SH. 2010. Advances in the understanding of haemoglobin switching. Br J Haematol. 149:181–194.
Sankaran VG, et al. 2009. Developmental and species-divergent globin switching are driven by BCL11A. Nature 460:1093–1097.
Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. Nature 489:109–113.
Schaffner SF, et al. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15:1576–1583.
Schechter AN. 2008. Hemoglobin research and the origins of molecular medicine. Blood 112:3927–3938.
Smith RA, Ho PJ, Clegg JB, Kidd JR, Thein SL. 1998. Recombination breakpoints in the human beta-globin gene cluster. Blood 92: 4415–4421.
Splinter E, et al. 2006. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev. 20: 2349–2354.
Spritz R, Giebel L. 1988. The structure and evolution of the spider monkey delta-globin gene. Mol Biol Evol. 5:21–29.
Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. Mol Biol Evol. 22:63–73.
Steinberg M, Adams JG 3rd. 1991. Hemoglobin A2: origin, evolution, and aftermath. Blood 78:2165–2177.
Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet. 73:1162–1169.
Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 68: 978–989.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.
Thein S. 2005. Genetic modifiers of beta-thalassemia. Haematologica 90: 649–660.
Tishkoff SA, Verrelli BC. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet. 4:293–340.
Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell. 10:1453–1465.
Vakoc CR, et al. 2005. Proximity among distant regulatory elements at the β-globin locus requires GATA-1 and FOG-1. Mol Cell. 17: 453–462.
Voight BF, et al. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci U S A. 102:18508–18513.
Wall JD, Frisse LA, Hudson RR, Di Rienzo A. 2003. Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. Am J Hum Genet. 73:1330–1340.
Wall JD, Przeworski M. 2000. When did the human population size start increasing? Genetics 155:1865–1874.
Wang G, Vasquez KM. 2006. Non-B DNA structure-induced genetic instability. Mutat Res. 598:103–119.
Ward LD, Kellis M. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 337: 1675–1678.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7: 256–276.

Webster MT, Clegg JB, Harding RM. 2003. Common 5′ ß-globin RFLP haplotypes harbour a surprising level of ancestral sequence mosaicism. Hum Genet. 113:123–139.
Webster MT, Wells RS, Clegg JB. 2002. Analysis of variation in the human β-globin gene cluster using a novel DHPLC technique. Mutat Res. 501: 99–103.
Wlasiuk G, Nachman MW. 2010. Adaptation and constraint at Toll-like receptors in primates. Mol Biol Evol. 27:2172–2186.
WHO. 2011. World malaria report. Geneva (Switzerland): World Health Organization.
Xu J, et al. 2010. Transcriptional silencing of {gamma}-globin by BCL11A involves long-range interactions and cooperation with SOX6. Genes Dev. 24:783–798.
Zeng K, Fu Y, Shi S, Wu C. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174: 1431–1439.
Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18:292–298.

Associate editor: Ross Hardison

Supplementary Data

Supplementary file S1 - Summary Statistics of Population Variation for the β-globin cluster genes using the 1000 Genomes Project data

Population		N[a]	HBB					HBD					HBBP1					HBG1					HBG2					HBE				
			L[b]	S[c]	π[d]	θ_w[e]	TD[f]	L[b]	S[c]	π[d]	θ_w[e]	TD[f]	L[b]	S[c]	π[d]	θ_w[e]	TD[f]	L[b]	S[c]	π[d]	θ_w[e]	TD[f]	L[b]	S[c]	π[d]	θ_w[e]	TD[f]	L[b]	S[c]	π[d]	θ_w[e]	TD[f]
European	FIN	186	1843	9	10,62	8,42	0,59	1879	3	0,5	2,75	-1,32	1935	9	8,23	8,02	0,06	1926	14	13,43	12,53	0,18	1933	15	17,45	13,38	0,78	1862	7	4,45	6,48	-0,67
	GBR	178		9	9,36	8,48	0,24		4	0,47	3,7	-1,58		8	8,05	7,18	0,27		14	14,44	12,63	0,37		12	17,7	10,78	1,58		3	5,19	2,8	1,39
	IBS	28		6	9,53	8,37	0,4		2	0,76	2,74	-1,51		6	7,49	7,97	-0,17		14	19,34	18,68	0,12		13	20,05	17,28	0,54		3	5,94	4,14	1,04
	TSI	196		6	7,96	5,56	0,88		5	0,38	4,55	-1,76		12	8,27	10,6	-0,54		12	11,08	10,65	0,1		16	17,75	14,14	0,66		6	4,8	5,51	-0,26
Asian	CHB	194		8	10,42	7,43	0,89		1	0,05	0,91	-0,96		6	5,09	5,31	-0,08		11	7,41	9,78	-0,58		8	12,65	7,08	1,74		2	2,75	1,84	0,68
	JPT	178		8	10,28	7,54	0,81		2	0,24	1,85	-1,21		5	6,28	4,49	0,78		11	8,49	9,92	-0,35		8	13,62	7,19	2		2	2,69	1,87	0,62
	CHS	200		11	10,7	10,16	0,13		3	0,16	2,72	-1,51		5	4,86	4,4	0,2		9	6,77	7,96	-0,34		9	12,11	7,93	1,2		3	2,5	2,74	-0,14
African	ASW	122			10,15	10,09	0,01		12	4,74	11,88	-1,55		15	6,66	14,42	-1,45		18	20,81	17,38	0,55		17	24,35	16,36	1,35		13	6,17	12,98	-1,38
	LWK	194		11	7,17	10,22	-0,72		14	5,54	12,75	-1,43		15	6,45	13,27	-1,31		16	21,81	14,22	1,38		18	23,25	15,94	1,21		13	7,16	11,95	-1
American	CLM	120		7	9,27	7,09	0,7		2	0,18	1,99	-1,35		10	6,55	9,64	-0,8		14	10,18	13,56	-0,67		16	17,3	15,44	0,33		4	4,42	4,01	0,2
	MXL	132		9	10,77	8,95	0,49		4	0,48	3,9	-1,66		11	7,97	10,42	-0,59		15	11,53	14,27	-0,51		14	17,18	13,27	0,78		5	4,08	4,92	-0,35
(Admixed)	PUR	110		10	10,29	10,29	0		11	1,72	11,1	-2,18		12	7,66	11,76	-0,91		14	12,78	13,78	-0,2		15	19,22	14,72	0,84		6	4,78	6,11	-0,48

[a] number of chromosomes

[b] total number of sites surveyed

[c] polymorphic sites

[d] nucleotide diversity ($x10^{-4}$)

[e] Watterson θ per site ($x10^{-4}$)

[f] Tajima's D statistic

Evolutionary constraints in the *β-globin* cluster: the signature of purifying selection at the *δ-globin* (*HBD*) locus and its role in developmental gene regulation

Supplementary Material

Ana Moleirinho, Susana Seixas, Alexandra M. Lopes, Celeste Bento, Maria J. Prata , António Amorim

Corresponding author: Ana Moleirinho, IPATIMUP, Rua Dr Roberto Frias s/n, 4200-465 Porto, Portugal; Phone: (+351) 22 5570700; Fax: (+351) 22 5570799; E-mail: amoleirinho@ipatimup.pt

Keywords: *β-globin* cluster, Hemoglobin switch, gene diversity, chromatin interactions

Running head: **Signature of purifying selection at *δ-globin***

Table S1 – Populations sampled by the 1000 Genomes Project

Population		**Sample size**
European	CEU - Utah residents (CEPH) with Northern and Western European ancestry (CEU)	85
	FIN-Finnish from Finland	93
	GBR - British from England and Scotland	89
	IBS - berian populations in Spain	14
	TSI - Toscani in Italia	98
African	ASW - African Ancestry in Southwest US	61
	LWK - Luhya in Webuye, Kenya	97
	YRI - Yoruba in Ibadan, Nigeria	88
Asian	CHB - Han Chinese in Beijing, China	97
	CHS - Han Chinese South	100
	JPT - Japanese in Toyko, Japan	89
American (admixed)	CLM - Colombian in Medellin, Colombia	60
	MXL - Mexican Ancestry in Los Angeles, CA	66
	PUR - Puerto Rican in Puerto Rico	55

Table S2 – Summary Statistics of Population Variation sequence variation data from 10 Western chimpanzees (Pan troglodytes verus), generated by the PanMap Project

	N[a]	L[b]	S[c]	NH[d]	Hd[e]
HBB		1843	6	8	0.84
HBD	20	1877	9	3	0.57
HBBP1		1925	5	5	0.69

[a] Number of chromosomes

[b] Total number of sites surveyed

[c] Polymorphic sites

[d] Number of haplotypes

[e] Haplotype diversity

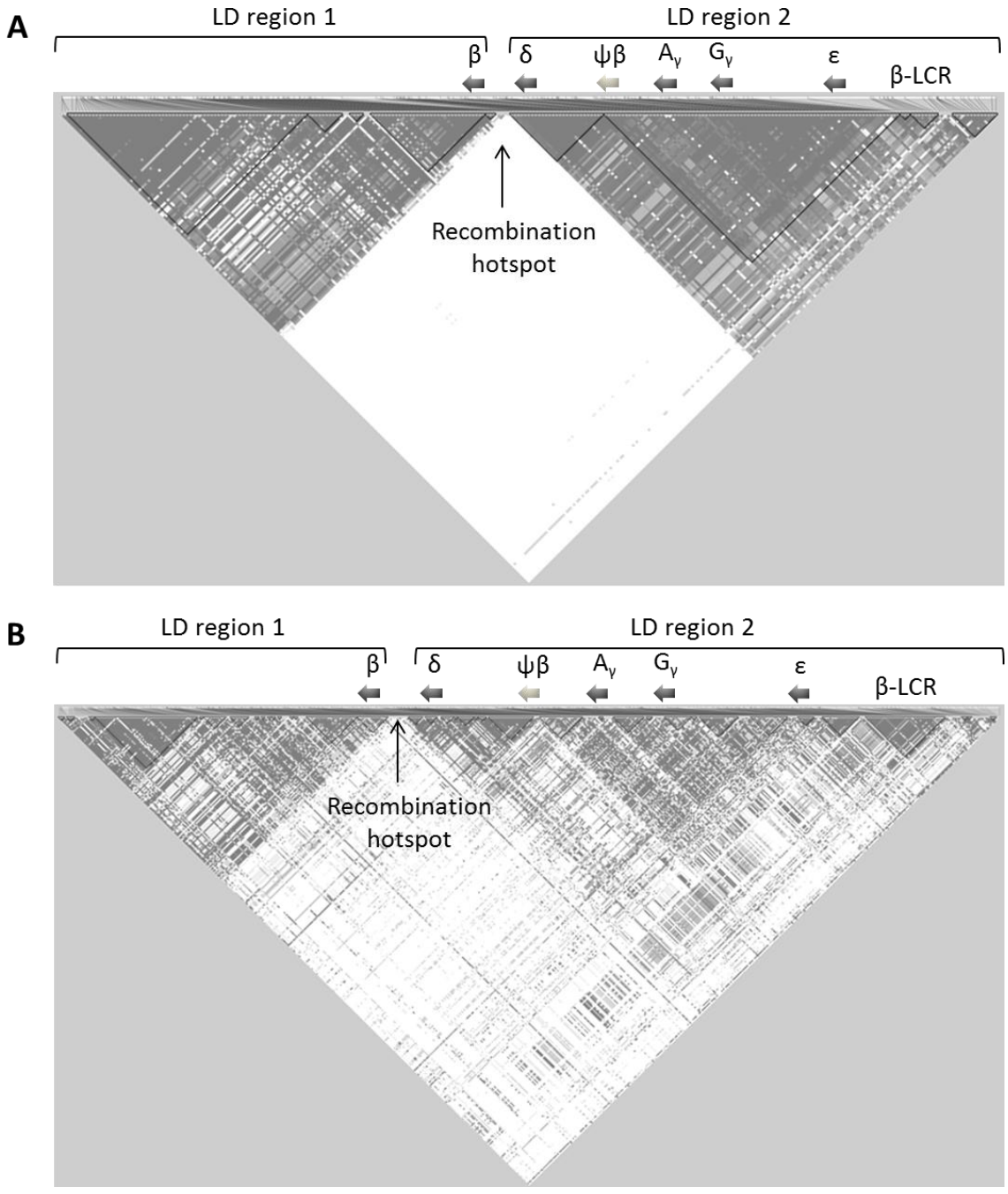


Figure S1 – LD plot of the β-globin cluster for the data from 1000 Genomes phase 1 release v3 for A) CEU and B) YRI. The image was constructed using Haploview 4.1 software. The triangles represent LD blocks. Two distinct regions with strong LD are identified: one containing *HBB* (LD region 1) and the other extending from *HBD* to the LCR (LD region 2). In this analysis only variants with a frequency ≥ 0.5% were considered.

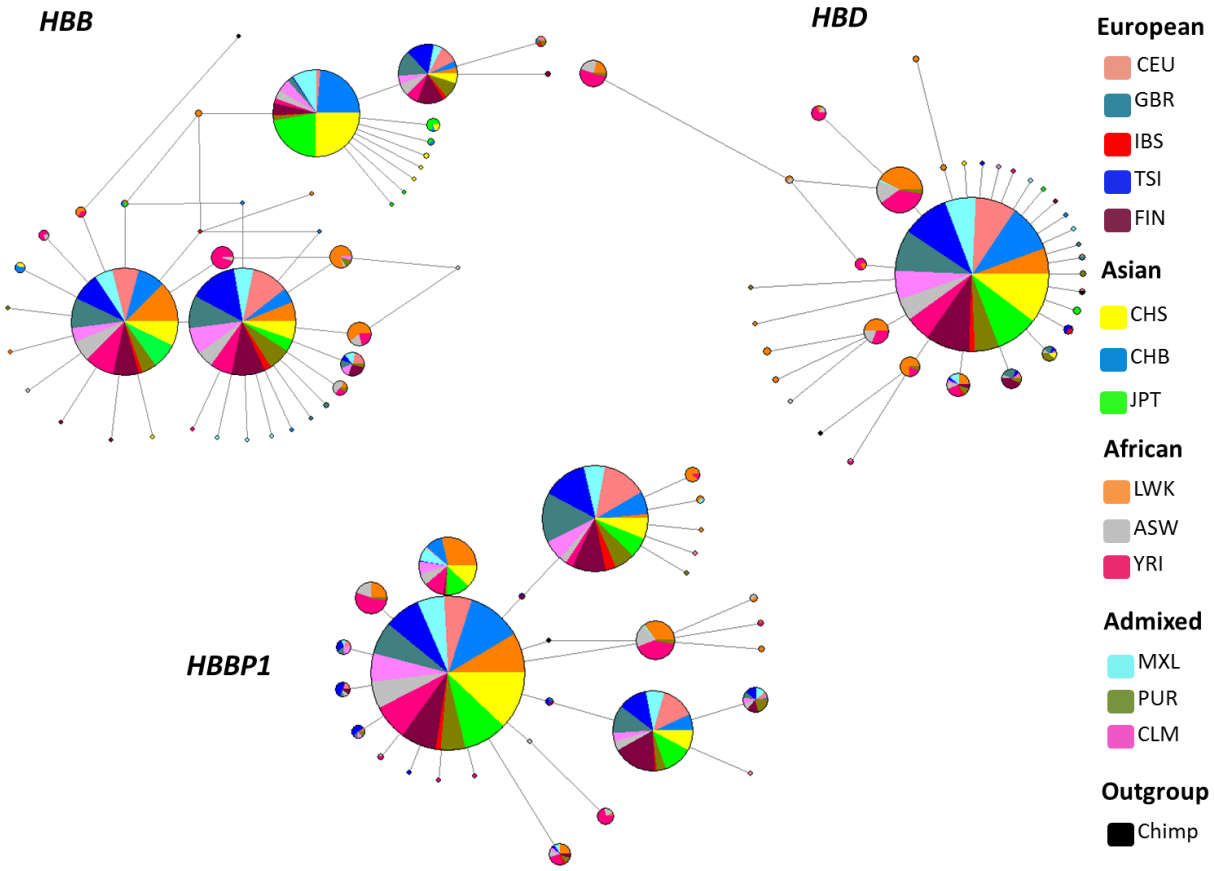
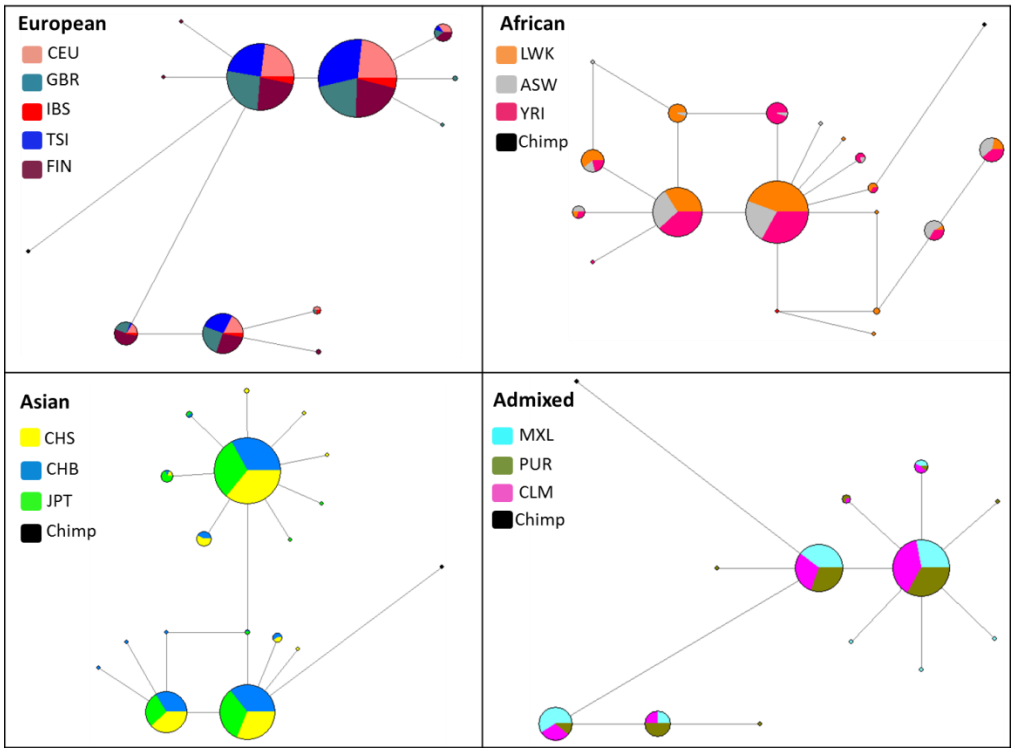
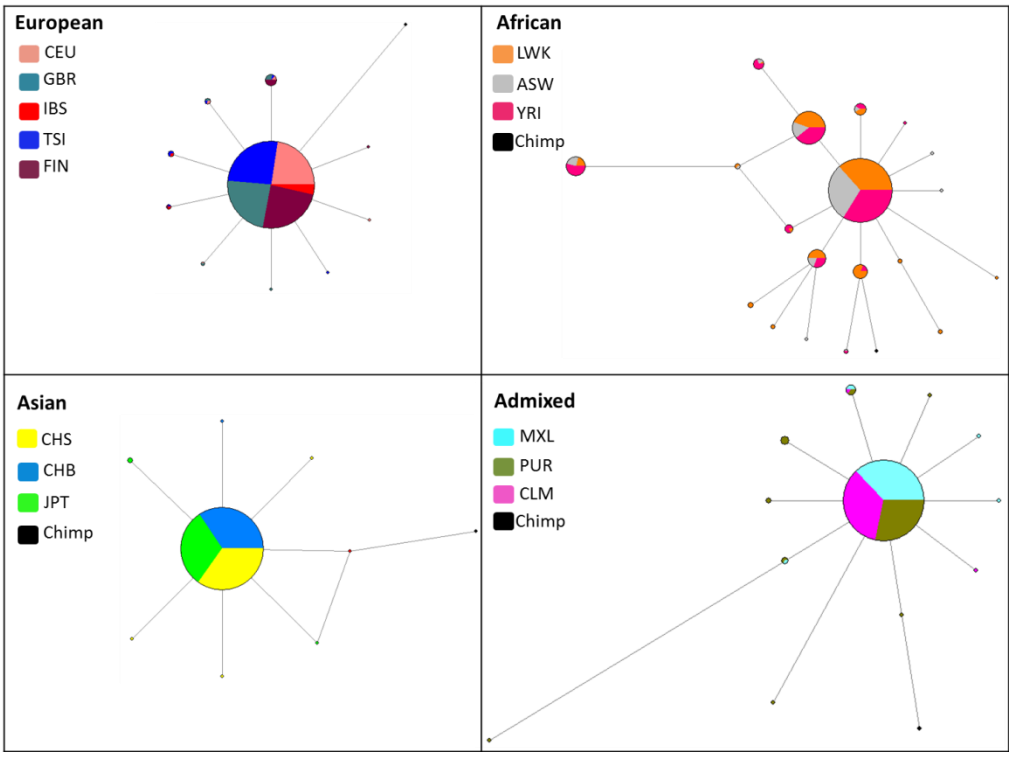


Figure S2 – Haplotype genealogies for *HBB*, *HBD* and *HBBP1*. Networks were built using haplotypes from 1000 Genomes Project (14 populations combined). Haplotypes are shown as circles with an area proportional to their frequency; lines connect different haplotypes and the number of mutations is proportional to their length. Full description of population's acronyms is available in table S1.

A)


European
CEU
GBR
IBS
TSI
FIN
African
LWK
ASW
YRI
Chimp
Asian
CHS
CHB
JPT
Chimp
Admixed
MXL
PUR
CLM
Chimp

B)


European
CEU
GBR
IBS
TSI
FIN
African
LWK
ASW
YRI
Chimp
Asian
CHS
CHB
JPT
Chimp
Admixed
MXL
PUR
CLM
Chimp

C)

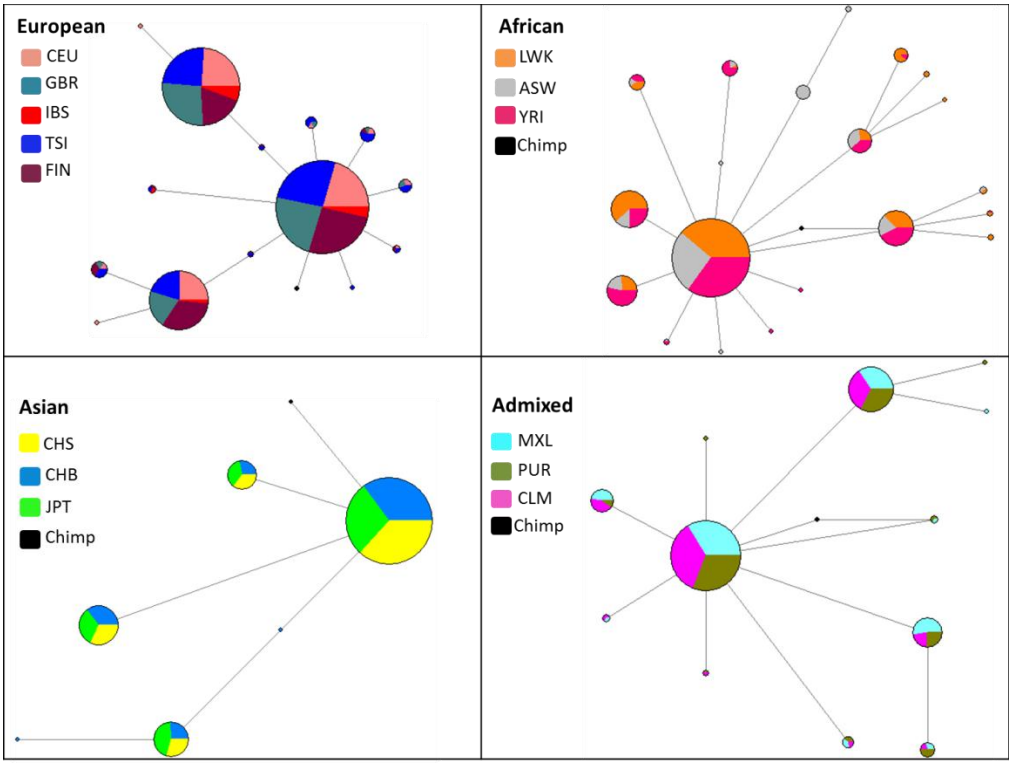
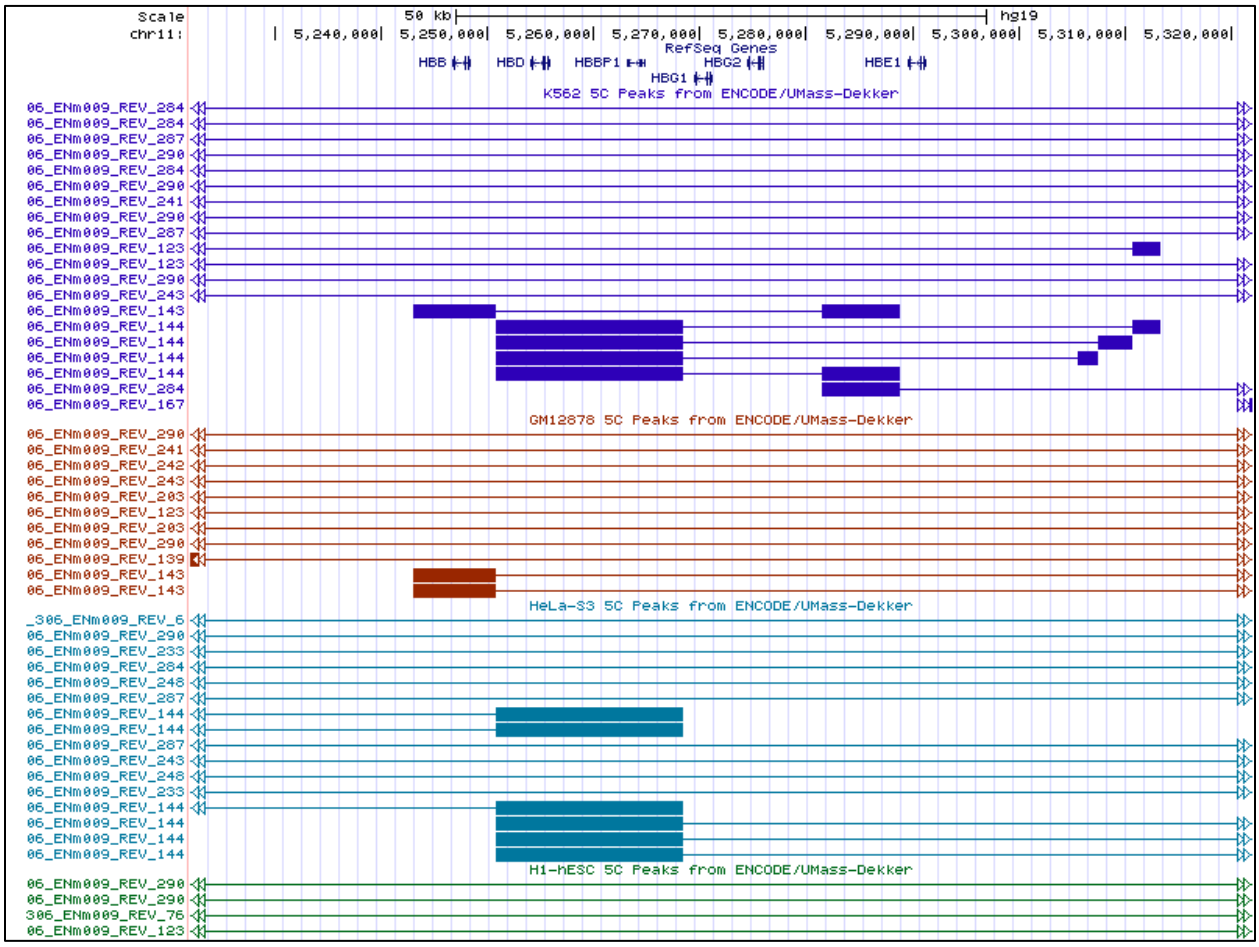


Figure S3 – Haplotype genealogies for A) *HBB*, B) *HBD* and C) *HBBP1*. Networks were built using haplotypes from 1000 Genomes Project, representing 1092 individuals from 14 populations: three African (ASW, LWK and YRI), five European (CEU, FIN, GBR, IBS and TSI), three Asian (CHB, CHS and JPT) and 3 American-admixed populations (CLM, MXL and PUR). Haplotypes are shown as circles with an area proportional to their frequency; lines connect different haplotypes and the number of mutations is proportional to their length. Full description of population's acronyms is available in table S1.

A)


Scale
50 kb
hg19
chr11:
5,240,000 5,250,000 5,260,000 5,270,000 5,280,000 5,290,000 5,300,000 5,310,000 5,320,000
RefSeq Genes
HBB HBD HBBP1 HBG2 HBE1
HBG1
K562 SC Peaks from ENCODE/UMass-Dekker
GM12878 SC Peaks from ENCODE/UMass-Dekker
HeLa-S3 SC Peaks from ENCODE/UMass-Dekker
H1-hESC SC Peaks from ENCODE/UMass-Dekker

B)

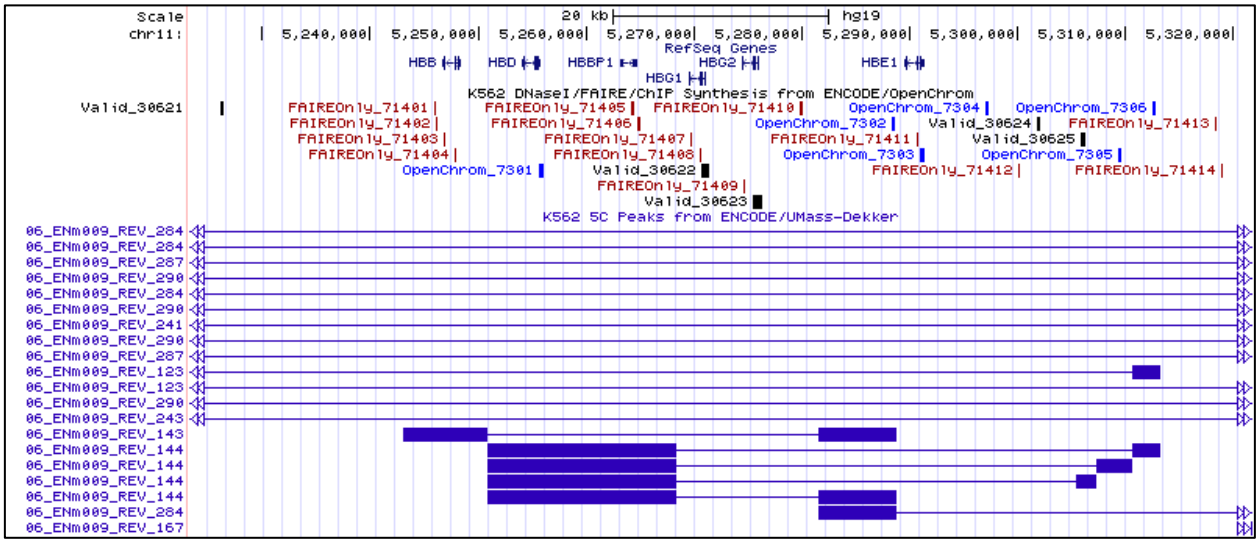

Scale
20 kb
hg19
chr11:
5,240,000 5,250,000 5,260,000 5,270,000 5,280,000 5,290,000 5,300,000 5,310,000 5,320,000
RefSeq Genes
HBB HBD HBBP1 HBG2 HBE1
HBG1
K562 DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom
Valid_30621
FAIREOnly_71401
FAIREOnly_71405
FAIREOnly_71410
OpenChrom_7304
OpenChrom_7306
FAIREOnly_71402
FAIREOnly_71406
OpenChrom_7302
Valid_30624
FAIREOnly_71413
FAIREOnly_71403
FAIREOnly_71407
FAIREOnly_71411
Valid_30625
FAIREOnly_71404
FAIREOnly_71408
OpenChrom_7303
OpenChrom_7305
OpenChrom_7301
Valid_30622
FAIREOnly_71412
FAIREOnly_71414
FAIREOnly_71409
Valid_30623
K562 SC Peaks from ENCODE/UMass-Dekker

Figure S4 – Chromatin features in the β-globin gene cluster as displayed in the UCSC Genome Browser. Human RefSeq β-globin genes are labeled and the genomic coordinates of the region displayed are shown (hg19). **A) Chromatin interactions determined by 5C (Chromatin Conformation Capture Carbon Copy).** Each color represents a different cell line (dark blue – K562; red – GM12878; light blue – HeLa-S3; green- H1-hESC). The letters to the right represent the primers used to detect each of the interactions. The regions involved in significant interactions in cis (i.e., from the same ENCODE pilot regions) are represented by blocks and connected by a horizontal line. Interactions displayed were filtered using a z-score interval from 500-1000. **B) Open chromatin regions and/or transcription factor binding sites identified in K562 cells by one or more complementary methodologies: DNaseI hypersensitivity (HS) (Duke DNaseI HS), Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) (UNC FAIRE), and chromatin immunoprecipitation (ChIP) for select regulatory factors (UTA TFBS).** Each color represents the assay(s) by which it was detected and its level of validation. Regions that overlap between methodologies identify regulatory elements that are cross-validated indicating high confidence regions (black). In addition, multiple lines of evidence suggest that regions detected by a single assay (e.g., DNase-only or FAIRE-only) are also biologically relevant: high significance regions that indicate open chromatin (blue); low significance regions (red).

3.2. Research Article:

"Distinctive patterns of evolution of the *δ-globin* gene (*HBD*) in primates"

PLOS ONE, 2015

(doi: 10.1371/journal.pone.0123365)

PLOS ONE

RESEARCH ARTICLE

Distinctive Patterns of Evolution of the *δ-Globin* Gene (HBD) in Primates

Ana Moleirinho[1,2,3]*, Alexandra M. Lopes[1,2], Susana Seixas[1,2], Ramiro Morales-Hojas[4], Maria J. Prata[1,2,3], António Amorim[1,2,3]

1 Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal, **2** IPATIMUP–Institute of Molecular Pathology and Immunology, University of Porto, Porto, Portugal, **3** Department of Biology, Faculty of Sciences, University of Porto, Porto, Portugal, **4** Genetics and Genomics Group, The Pirbright Institute, Compton Laboratory, Compton, Berkshire, United Kingdom

* amoleirinho@ipatimup.pt

OPEN ACCESS

Citation: Moleirinho A, Lopes AM, Seixas S, Morales-Hojas R, Prata MJ, Amorim A (2015) Distinctive Patterns of Evolution of the *δ-Globin* Gene (HBD) in Primates. PLoS ONE 10(4): e0123365. doi:10.1371/journal.pone.0123365

Academic Editor: Francesc Calafell, Universitat Pompeu Fabra, SPAIN

Received: June 6, 2014

Accepted: March 2, 2015

Published: April 8, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the Portuguese Foundation for Science and Technology (FCT) (PTDC/SAU-MET/110323/2009). AM and AML are supported by fellowships from FCT (SFRH / BD / 73508 / 2010 and SFRH / BPD / 73366 / 2010, respectively). SS is supported by POPH-QREN – Promotion of scientific employment, by the European Social Fund, and by national funds of the Ministry of Education and Science. IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Education and Science and is partially supported by FCT. The

Abstract

In most vertebrates, hemoglobin (Hb) is a heterotetramer composed of two dissimilar globin chains, which change during development according to the patterns of expression of α- and β-globin family members. In placental mammals, the β-globin cluster includes three early-expressed genes, ε(*HBE*)-γ(*HBG*)-ψβ(*HBBP1*), and the late expressed genes, δ (*HBD*) and β (*HBB*). While *HBB* encodes the major adult β-globin chain, *HBD* is weakly expressed or totally silent. Paradoxically, in human populations *HBD* shows high levels of conservation typical of genes under strong evolutionary constraints, possibly due to a regulatory role in the fetal-to-adult switch unique of Anthropoid primates. In this study, we have performed a comprehensive phylogenetic and comparative analysis of the two adult *β-like globin* genes in a set of diverse mammalian taxa, focusing on the evolution and functional divergence of *HBD* in primates. Our analysis revealed that anthropoids are an exception to a general pattern of concerted evolution in placental mammals, showing a high level of sequence conservation at *HBD*, less frequent and shorter gene conversion events. Moreover, this lineage is unique in the retention of a functional GATA-1 motif, known to be involved in the control of the developmental expression of the *β-like globin* genes. We further show that not only the mode but also the rate of evolution of the *δ-globin* gene in higher primates are strictly associated with the fetal/adult β-cluster developmental switch. To gain further insight into the possible functional constraints that have been shaping the evolutionary history of *HBD* in primates, we calculated dN/dS (ω) ratios under alternative models of gene evolution. Although our results indicate that *HBD* might have experienced different selective pressures throughout primate evolution, as shown by different ω values between apes and Old World Monkeys + New World Monkeys (0.06 versus 0.43, respectively), these estimates corroborated a constrained evolution for *HBD* in Anthropoid lineages, which is unlikely to be related to protein function. Collectively, these findings suggest that sequence change at the *δ-globin* gene has been under strong selective constraints over 65 Myr of primate evolution, likely due to a regulatory role in ontogenic switches of gene expression.

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Hemoglobin (Hb), found in the circulating red blood cells of all vertebrates, is the major oxygen-transporting molecule, playing a key role in the cellular aerobic metabolism [27]. In mammals, Hb is a heterotetramer composed of two α-like and two β-like globin chains that are differentially expressed during development, such that functionally distinct Hb isoforms are synthesized in embryonic and adult erythroid cells [27–29].

These globin chains are encoded by members of the *α-* and *β-globin* gene families, which arose via tandem duplication of an ancestral, single-copy globin gene approximately 450–500 Mya, in the common ancestor of jawed vertebrates [14,23,25,35,78]. The two paralogous gene families exhibit a number of significant differences in gene content among jawed vertebrate taxa. These differences are especially pronounced in the case of the *β-globin* cluster, in which distinct repertoires of mammalian *β-like globin* genes originated by independent lineage-specific duplications followed by functional divergence [33–35,52–54,57,78,79]. In both monotremes and marsupials, the *β-globin gene* cluster contains a single pair of genes, the early expressed ε-globin and the late expressed *β-globin* [53]. In contrast, within the eutherian stem, further tandem duplications gave rise to a cluster of five *β-like globin* genes, containing early-expressed genes, located at the 5' end of the cluster ε-(*HBE*)-γ(*HBG*)-ψβ(*HBBP1*), and late expressed genes, δ (*HBD*) and β (*HBB*), at the 3' end, consistent with the orientation in contemporary species [26,31,53]. The fine tuning of the level and timing of expression of each of these genes relies on interactions with the locus control region (LCR), located from approximately 6 to 18 kb upstream of *HBE* [4,11,84].

Over the course of eutherian evolution the structure of the *β-globin gene* cluster has been dynamic and the late-expressed *HBD* and *HBB* paralogs have experienced different evolutionary fates. In the majority of mammals, the adult form of Hb ($\alpha_2\beta_2$) contains similar β-chain subunits which are encoded by one or more copies of the *HBB* gene [52]. Contrastingly, the *δ-globin* gene, although present in almost all eutherian species examined to date, is frequently pseudogenized [19,24,26,30,54]. In a few species, a transcriptionally active but weakly expressed copy of the *δ-globin* gene was maintained, encoding the δ-globin chain of the minor fraction of the adult Hb ($\alpha_2\delta_2$), known as HbA_2, which is thus assumed to be physiologically irrelevant [45,46,76]. In fact, the δ-globin chain is absent in Old World Monkeys (OWM) [45,46] and ranges from 1% concentration in hominoids [10] to 40% in the galago [80], reaching 6% in New World Monkeys (NWM) [74] and 18% in tarsiers [40]. Surprisingly, in some eutherians *HBD* shows a level of sequence conservation typical of genes under strong evolutionary constraints [91]. In humans for example, *HBD* was found to have lower nucleotide diversity than *HBB*, suggesting that purifying selection has shaped the evolutionary history of *HBD* [49,87] through an unrecognized role not associated with oxygen transport [46,74,86].

The involvement of *HBD* in the fetal/adult Hb switch was proposed decades ago [5,55] and since then some studies have provided evidence supporting this hypothesis [6,49,67]. In fact, the fetal to adult Hb switch of anthropoid primates is unique. Furthermore, while both Anthropoids and Prosimians possess a γ-globin gene, its switch after birth only takes place in the major anthropoid branch, the catarrhines, occurring earlier in NWM, whereas in Prosimians it is only expressed at the embryonic stage [38]. Therefore, phylogenetic and comparative genomic analysis across placental mammals with distinct repertoires of β-like genes and corresponding expression programs should provide clues to the evolution and putative functional divergence of the *δ-globin* gene.

However, the evolutionary history of the eutherian *HBD* is quite complex due to unusually frequent sequence exchanges through extensive gene conversion and unequal recombination with its neighbor, *β-globin* [19,31,40,54,80], resulting in extensive sequence homogenization

and hampering the assignment of orthologous relationships among *HBD* and *HBB* genes. Indeed, *HBD* was initially thought to be the result of a recent duplication in primate evolution, approximately 40 MY [18], but recently an older origin has been proposed [30,53]. Under such scenario of controversy, a revisit of the evolutionary history of the adult *δ-globin* gene can help elucidate the origin of this gene family, which in spite of many efforts is not yet fully understood [31,37,40,45,46,53,60,74,80].

Here, we perform a comprehensive phylogenetic and comparative analysis of the two adult *β-like globin* genes in a wide range of mammalian taxa, with a special focus on primates. Our results further document reticulation in the topology of the evolutionary history of *δ-globin* gene, demonstrating that it has behaved as an evolutionary palimpsest, with repeated and partially overlapping β / δ sequence transfers obscuring orthology. Additionally, we show that the *δ-globin* gene is highly conserved in Anthropoids, with a particularly strong signal of purifying selection in Great Apes. Sequence conservation at this locus is unlikely related to protein function and may reflect mutational constraints on regulatory regions involved in the fetal-to-adult developmental expression switch of the *β-globin* cluster.

Materials and Methods

DNA Sequence data and gene identification

To obtain DNA sequences spanning the entire *HBD* and *HBB* genes, we used Blat queries to interrogate the genome assemblies of several mammalian species available in the UCSC Genome Browser website (http://genome-euro.ucsc.edu/index.html). Whenever the *HBB* sequence was available, we used it to identify its paralogous sequences (the first hit); alternatively we used the human *HBB* and *HBD* sequences to identify their corresponding orthologs. Due to a history of concerted evolution, in some cases high sequence identity between *HBD* and *HBB* produced ambiguous results. In these cases the genome coordinates were used to distinguish between the two genes, since the order of the *β-like globin* genes has been maintained throughout mammalian evolution. Additional genomic data was obtained either from the High Throughput Genomic Sequences database (HTSG), Trace Archives or by direct sequencing to complete sequence gaps and to include further mammalian species. Detailed information on sampling is listed in S1 Table. Following all these steps of manual curation, our sample included 29 sequences from the three major subclasses of mammals: 1 Prototheria (*Ornithorhynchus anatinus*), 1 Metatheria (*Monodelphis domestica*) and 27 Eutheria, including representatives of the following superordinal groups (1 Xenarthra, 7 Laurasiatheria and 19 Euarchontoglires). We also included one avian species (*Gallus gallus*) as outgroup.

Evolutionary analysis

The reconstruction of the evolutionary history of the *HBD* and *HBB* genes was carried out across the mammalian phylogenetic tree with the chicken sequence as outgroup. The identified coding sequences were translated into the corresponding protein and aligned using the Expresso mode of T-Coffee in order to take into account any structural information of the protein available in the databases [2]. Each of the non-coding sequences, including the 5' and 3' untranslated regions (UTRs) plus 2 introns, were aligned separately as nucleotide sequences using the accurate mode of T-Coffee [51]. Phylogenetic reconstruction was performed using Maximum Likelihood (ML) and Bayesian Inference (BI) as the optimality criteria. Trees were estimated using the complete gene sequence and including only the coding sequence (CDS). The partition strategies and models of evolution implemented in these analyses were identified using PartitionFinder v 1.1.1 [41]. PartitionFinder was run specifying the raxml and mrbayes models of evolution for computational capacity reasons.

ML analyses were performed with RAxML [75] run in the CIPRES Science Gateway [48]. Phylogenetic analysis of the complete gene was run with the following partitions: 1) 5' UTR + Exon 1, 2 and 3 + Intron 1; and 2) Intron 2 + 3' UTR. The GTR+G+I model was implemented for each of these. ML analysis of the CDS was performed with three partitions corresponding to the three codon positions. The GTR+G model was implemented for each partition. The resulting trees were evaluated with 1000 bootstrap replicates.

BI analyses were run in MrBayes 3.2 [64] in the CIPRES Science Gateway. The complete gene analysis was performed with three partitions: 1) 5' UTR + Exon 1, 2 and 3 + Intron 1; 2) Intron 2; and 3) 3' UTR. The implemented models of evolution were the K80+G, GTR+G+I and HKY+G+I for the first, second and third partitions, respectively. BI with the CDS sequence was performed with the data matrix partitioned according to codon position; the models of evolution implemented were the K80+G for the first codon position and the GTR+G for the second and third positions. Two independent runs of 10 million (complete gene analysis) and 5 million (CDS analysis) generations with 8 chains each (7 heated and one cold) were set up. Trees were sampled every 200 (complete gene) or 100 (CDS) generation and the first 12500 trees (25% of the sample) were discarded as burn-in. Convergence and burn-in were assessed using Tracer 1.6 [61], MCMC Trace Analysis Package http://tree.bio.ed.ac.uk/software/tracer/).

Additionally, a third phylogenetic analysis was performed using models of codon substitution with the coding sequence only as input. This method allows us to take into consideration any potential divergence in codon usage across the mammalian lineage and differences in selective pressure among gene copies, which included pseudogenised copies. Analyses were run using CodonPhyML [21] with the GY CF3x4 [22] model of codon substitution and specifying the M3 model of selective pressure [90]. An initial tree was obtained using BioNJ + GYECMK07 and the topology was searched using the subtree pruning and regrafting (SPR) heuristic search. Branch support was obtained using the approximate Likelihood Ration Test (aLRT) as implemented in the software. To identify potential recombination events in our primate data set we used the recombination detection package (RDP3) [43], applying a set of seven statistical methods, which includes RDP [42], GENECONV [56], Bootscan [44], Maxchi [72], Chimaera [59], SiSscan [20] and 3Seq [8]. Briefly, two of these are phylogenetic methods, which infer recombination when different parts of the genome result in discordant topologies (RDP and Bootscan), while the other five are nucleotide substitution methods, which examine the sequences either for a significant clustering of substitutions or for a fit to an expected statistical distribution: MaxChi, Chimaera, Geneconv, 3Seq and SiScan. The latter primarily uses genetic similarity estimates but also takes some phylogenetic information into account. To uncover both species specific and ancient events of gene conversion, we conducted the analysis with the two paralogous genes for all species simultaneously. Settings for all methods executed in RDP3 were as follows: Sequences were considered to be linear, the p-value cutoff was set to 0.05, and the standard Bonferroni correction was used. In addition, phylogenetic relationships were recovered for each fragment showing signs of gene conversion, in Anthropoids and Catarrhines, following the same procedure described above. ML analyses were performed using the GTR+G model without sequence partitioning. The resulting trees were evaluated with 1000 bootstrap replicates.

Search for open reading frames and promoter analysis

The genomic sequence of the *HBB* and *HBD* genes was used to predict the locations and exon-intron structures using the program Genscan [12], followed by sequence alignment of known exon sequences and manual inspection of homology. Protein sequence was obtained using the translation tool implemented in the ExPASy web server (http://www.expasy.org) [3]. Sequence

alignment for eutherian *HBB* proteins and *HBD* open reading frame were performed using ClustalW [83] implemented in Geneious version 5.5 created by Biomatters (available from http://www.geneious.com/). To identify conserved motifs previously shown to be essential for *HBB* and *HBD* expression, promoter sequences located 5' to the two *β-like globin* genes (~200bp) were aligned using the same approach as above. To confirm transcription factor binding site matches identified within the alignments, we used MatInspector implemented in the Genomatix Software Suite (http://www.genomatix.de/index.html).

Evolutionary rate estimates and selection tests

Pairwise sequence divergence was deduced from Jukes and Cantor distance calculated with DnaSP v.5.10 [65]. In our estimates we scored each insertion or deletion, regardless of length, as one difference as in [13]. Divergence times between species were obtained with TimeTree [32]. Maximum-likelihood estimates of dN/dS (ω; dS—synonymous substitution rate and dN—non-synonymous substitution rate) were carried out using the codeml program from the software package Phylogenetic Analysis by Maximum Likelihood—PAML version 4.8 [89]. To investigate the selective pressures that have shaped the evolution of *HBB* and *HBD* genes we first calculated dN/dS ratios (M0 model) for each gene separately in the entire mammalian phylogeny. Next, to test the hypothesis of variable selective pressures among *HBD* in primates, we performed nested branch models using either the one-ratio model calculated for the whole anthropoid phylogeny, the two-ratio estimated for Great Apes and other primates and three-ratio inferred for Great Apes, OWM and NWM [7,88]. Although ω values below 1 ($\omega < 1$) are generally considered as an evidence of purifying selection, to reject the hypothesis of neutral evolution all models were compared with a null model where ω was fixed to 1 ($\omega = 1$). The significance was obtained with likelihood ratio tests (LRT) which were calculated as twice the variation of the likelihoods ($-2\Delta l$) with a χ^2 distribution. For the calculation of *HBD* ω values, pseudogene sequences were only included after the removal of positions affected by premature stop codons and frameshift mutations. In the specific case of lemur species, the *HBD* sequences were excluded from the analysis given their hybrid *ψβ/δ* nature [37].

Results

Evolutionary history of the two adult β-like genes in mammals

We conducted a phylogenetic analysis of the adult *β-like globin* genes, *HBD* and *HBB*, in a diverse dataset, including monotremes, marsupials and placental mammals, and one avian species, which was used as outgroup. The phylogenies obtained with ML, BI and the codon model approach were similar when either the complete gene sequence or the CDS were used (Fig 1A–1D and S4 Fig).

However, trees obtained using the complete gene were different from those estimated with the CDS only (Fig 1 and S4 Fig). In the phylogenies based on the CDS the two *β-like* paralogous genes from the same species cluster together, consistent with a process of interparalogue gene conversion, referred to as "concerted evolution" and previously described in several taxa [19,31,40,52,80]. The exceptions to this general pattern occur in Anthropoid primates (monkeys and apes) where *HBD* and *HBB* are grouped into two reciprocally monophyletic groups, and in the lemur species. This could lead to the erroneous interpretation that *HBD* arose recently through a duplication in primate evolution, as initially thought [18], but more likely reflects a gene conversion event in the common ancestor of those primate lineages. In the lemur species, the exception is easily explained by the fact that in the ancestry of lemurs a hybrid ψβ/δ pseudogene was created by unequal crossing-over between misaligned *HBD* and *HBBP1* sequences [37]. Finally, an *HBD* ortholog is absent from rat, and in mouse we found

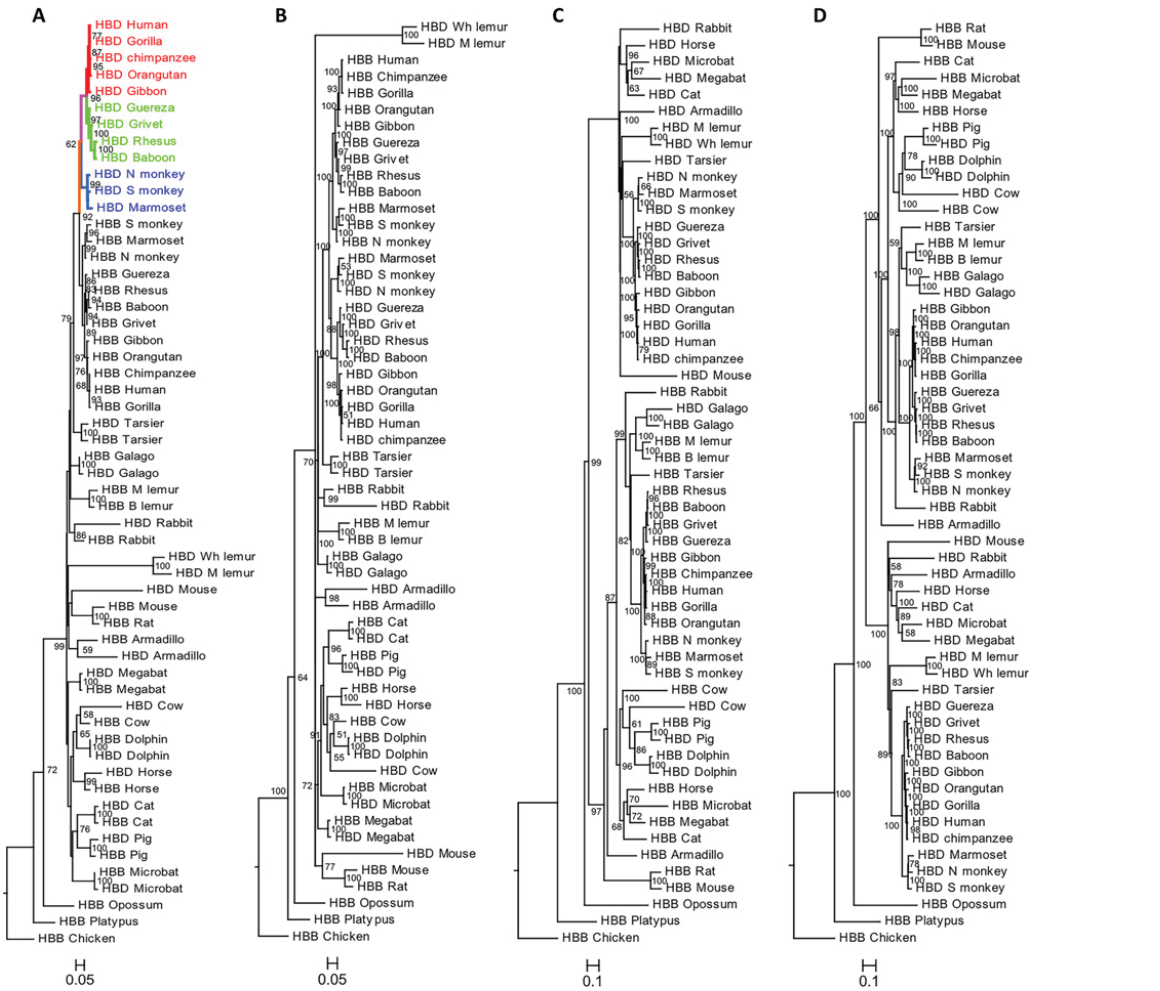


Fig 1. Phylograms depicting relationships among adult *β-like* genes in mammals. The phylogeny reconstructions were performed using two methods: A), C) Maximum Likelihood and B), D) Bayesian Inference; trees were estimated using the A), B) coding sequence and C), D) complete gene sequence. Branch support values are given on the internodes. Red branches represent the Great Apes, green the OWM and blue the NWM; the pink and orange branches represent the common branch of Catarrhines and Anthropoids, respectively.

doi:10.1371/journal.pone.0123365.g001

no evidence for interparalog gene conversion between *HBD* and *HBB* orthologs (*HBD-T1* and *HBB-T1*), in agreement with previous results [33]. Phylogenetic analyses based on the entire gene sequence lead to better-supported trees, which more reliably replicate the species tree, recovering the true evolutionary history of gene duplication. This phylogeny is consistent with previous results supporting a duplication event of *HBB* and *HBD* genes after the marsupial/eutherian split [53]. Both ML and BI trees contain two major sister clades, one comprising the *HBD* gene copies of most species and a second clade mainly with the *HBB* genes. Within the *HBB* clade, the *HBD* genes from galago, cow, pig and dolphins are clustered

with each of their *HBB* paralogs. In galago, it is well documented that the ancestral *HBD* gene was almost completely converted by the *HBB* gene, explaining the monophyletic pattern observed [80]. According to a recent analysis the phylogenetic incongruence seen in the latter three species is likely due to independent gene conversion events that followed an extensive unequal cross-over spanning *HBD* intron 2 in the stem lineage of cetartiodactyls [19]. Our results also confirm that the interparalog conversion is largely restricted to the coding regions of globin genes [19,26,30,31,40,45,53], because the incorporation of the phylogenetic information of non-coding sequence, including intronic and 5' and 3' flanking sequence, proved to be especially useful in assigning orthologous relationships.

Analysis of gene structure

In all species examined the *HBB* gene retains an intact ORF with conserved donor/acceptor splice sites encoding a polypeptide of 144–146 amino acids (S1 Fig). By contrast, the *HBD* gene has accumulated several inactivating mutations throughout the mammalian phylogeny. These include the introduction of premature stop codons, small insertions and deletions and mutations in the consensus donor (GT) or acceptor (AG) splice sites (S2 Fig). Moreover, *HBB* and *HBD* also display different across species conservation patterns (S3 Fig). The *HBB* promoter is highly conserved and contains conserved consensus TATA, CAAT and EKLF (Erythroid Krüppel-like factor) binding motifs in most species examined. Only cat and cow show substitutions in the EKLF consensus sequence (S3A Fig) but present an upstream EKLF binding element, identified by MatInspector, which is likely to replace the possible disrupted EKLF motif. A lower conservation in *HBD* promoter region is readily apparent from the difficulty in obtaining a good multiple alignment for all species. We detected three species, galago, cat and microbat, in which *HBD* has a β-like promoter, acquired through independent gene conversion events (S3A Fig), as previously shown [19,80]. In anthropoids, high sequence homology was observed at the *HBD* promoter region (S3B Fig), with a conserved consensus TATA binding motif, a functional GATA-1 motif [47] close to the mutated CAAC box, and lack of the EKLF binding element in all these species. These features of the *HBD* promoter region have been shown to be responsible for the low expression of the adult *δ-globin* gene [63,81,82]. The remaining species (S3C Fig) all share the major defect in the proximal δ-promoter, the absence of a consensus EKLF-binding motif, but do not have the GATA-1 motif common to all other δ-like promoters. The lack of various conserved motifs in the *HBD* promoters that are crucial for β-like globin expression suggests that they are transcriptionally inefficient in tarsier, mouse, rabbit, dolphin, cow, pig, horse, megabat and armadillo.

Recombination events in primates

In the phylogenetic analyses we did not detect evidence of recombination events between *HBD* and *HBB* in anthropoid primates. This result is in agreement with other studies that proposed a δ-β gene conversion in the anthropoid stem [40]. However, it has been suggested that further gene conversions occurred independently in catarrhine and platyrrhine lineages [37,45,60]. Although we did not find phylogenetic evidence for gene conversions within these primate lineages, we cannot exclude the possibility of short-tract gene conversion events not detectable by phylogenetic analysis. Therefore, we sought to re-examine the possibility of lineage-specific gene conversion taking advantage of a more comprehensive sample of primate species and the use of multiple methods implemented in the software package RDP3 [43] for detecting recombination signals and putative recombinant sequences. We found robust signals for four independent interparalog gene conversion events, in which portions of *HBB* were copied onto *HBD* (summarized in Table 1).

Table 1. Summary of gene conversion analysis for primate *HBD* and *HBB* paralogues.

Gene conversion event ID	Recombinant sequences	Major Parental Sequence [b]	Minor Parental Sequence [b]	Breakpoint [a] Positions Begin	Breakpoint [a] Positions End	Conversion Tract Length	Detection Methods [a]	P-value [a]
1	HBD_Tarsier	unknown	HBB_Tarsier	96	574	72 bp 5' flanking -183 bp into exon 2	RDP, **GENECONV**, BootScan, MaxChi, Chimaera, SiScan, 3Seq	3,475 x 10^{-25}
2	HBD_Galago	unknown	HBB_Galago	12	2036	149 bp 5' flanking—55 bp into exon 3	RDP, GENECONV, BootScan, MaxChi, Chimaera, SiScan, **3Seq**	3,665 x 10^{-18}
3	HBD_Human HBD_Chimpanzee HBD_Gorilla HBD_Gibbon HBD_Guereza HBD_Grivet HBD_Rhesus HBD_Babbon	unknown	HBB_Babbon	97	385	71 bp 5' flanking -123 into intron 1	RDP, **MaxChi**, Chimaera	3,198 x 10^{-4}
4	HBD_Human HBD_chimpanzee HBD_Gorilla HBD_Gibbon HBD_Guereza HBD_Grivet HBD_Rhesus HBD_Babbon HBD_N.monkey HBD_Marmoset HBD_S.monkey	unknown	HBB_N.monkey	197/367 [c]	632	106 bp 5' flanking -215 bp into exon 2	RDP	1,115 x 10^{-5}

[a] In cases where multiple methods detected the same or a similar conversion event, we reported the breakpoint positions and the method yielding the lowest average Bonferroni corrected p-value, which is shown in bold; breakpoint positions refer to the nucleotide positions in the full alignment of the Primate *HBB* and *HBD* sequences.

[b] The major parental and minor parental sequences correspond to the parent contributing to the larger fraction and to the minor fraction of the recombinant sequence, respectively. In all 4 events the major parental sequence is unknown given that the presence of a parent and a recombinant in the alignment is sufficient for a recombination event to be detected by these methods.

[c] The breakpoint for the recombination event number 4 varies depending on the species in which it was detected: 197 in Platyrrhines (N.monkey, Marmoset and S.monkey) and 367 in Catarrhines (Human, Chimpanzee, Gorilla, Gibbon, Guereza, Grivet, Rhesus, Babbon), leading to different gene conversion tract length predictions for these groups.

doi:10.1371/journal.pone.0123365.t001

Two of these gene conversion events, 1 and 2, were detected by all seven methods, and corroborate independent gene conversions previously identified in tarsier and galago [40,80]. A third event was detected, corresponding to the conversion of the first *HBD* exon and intron by *HBB* sequences, in catarrhines (OWM and Great Apes). Since this event was detected in orthologous *HBD* copies of all catarrhines represented, it most likely took place before the divergence of OWM and Great Apes, along the pink branch in Fig 1A. The event number 4 corresponds to a more extensive conversion tract present in all anthropoids (platyrrhines and catarrhines), suggesting that those sequences have all descended from an anthropoid ancestor sequence in which the recombination event occurred (orange branch in Fig 1A). This δ-β conversion

appears to extend from the 5' promoter region till the end of the second exon, however in catarrhines the signal for this older gene conversion is restricted to the second exon, because a subsequent event (identified as number 3) overprinted part of the older one.

In order to assess whether the conversion events 3 and 4 have been correctly identified, we constructed and compared two phylogenetic trees, one with the region containing evidence for an older gene conversion in both platyrrhines and catarrhines (nucleotide 367–632), and a second one with the portion of the alignment between the inferred breakpoints in event 3 (nucleotide 97–385) where a second, more recent conversion event occurred in the catarrhine stem (S5 Fig). The topology of the first tree (S5A Fig) is consistent with a gene conversion in the common ancestor of Anthropoids, in agreement with results obtained with other methods [31,40,45,73]. In contrast, in the second tree (S5B Fig) the *HBD* and *HBB* genes group together within the catarrhine and platyrrhine lineages, as expected under the hypothesis of a conversion event restricted to catarrhines. This tree topology could also indicate that parallel gene conversion events have occurred in the stem of catarrhines and platyrrhines; however, no gene conversion event in platyrrhines has been detected with any of the methods used in our analysis. Nevertheless, this alternative hypothesis remains disputable given that an older conversion event occurring between still very closely related sequences would be very difficult to detect. Therefore, although our results do not fully support previous evidence that *HBD* has been involved in a conversion in platyrrhines [60], it cannot be ruled out. It is noteworthy that albeit two independent gene conversion events occurred in different Anthropoid lineages, the GATA-1 motif in the *HBD* promoter region remained intact.

Evolutionary rates and functional constraints in primates

From our previous analysis, it is apparent that while *HBD* has diverged at markedly different rates in different primate lineages, as shown by the variable branch lengths in the phylogenetic trees (Fig 1), anthropoid *HBD*s share a high sequence identity not only in their coding region but also at the promoter. As a first measure of the rate of *HBD* evolution, we compared the genetic distance between humans and 13 other primate species, for both adult *β-like globin* genes. *HBBP1* was also included in the analysis due to the unusually slow substitution rates previously reported for this pseudogene estimated by comparison of human, gorilla and chimpanzee sequences [13]. Genetic distances were then plotted against the corresponding divergence times for each pairwise comparison, and the linear regression trend line was estimated for each group, as shown in Fig 2.

From the slope of the trend lines and the r^2 values we are able to compare the rate of intron and exon evolution and its constancy over time, respectively. The results presented in Fig 2 and S2 Table show that, overall, exons evolved at a lower rate than introns, except for *HBBP1*, in which nucleotide differences are more homogeneously distributed between exons and introns, as expected for a pseudogene. In Prosimians, there is a trend towards increasing evolutionary rates, more pronounced in non-coding regions (introns and *HBPP1* exons). Remarkably, the rate of evolution of *HBD* exons has remained relatively constant across primate evolution ($r^2 = 0.95$) and is comparable to that of *HBB* exons, even though in higher primates *HBD* is either silent or contributes to only a very small fraction of adult Hb. To gain further insight into the possible functional constraints that have shaped the evolutionary history of *HBD* in primates, we calculated dN/dS (ω) ratios under alternative models of gene evolution (table 2).

First, we estimated single ω values for the entire mammalian phylogeny (M0 model), for either *HBB* or *HBD* genes. The observed ω values were significantly lower than 1 ($\omega_{HBB} = 0.26$ and $\omega_{HBD} = 0.40$) pointing to an overall conservation of *HBB* and *HBD*. The ω obtained for *HBB* is the expected for a functional gene and is in agreement with previous estimates [1]. In

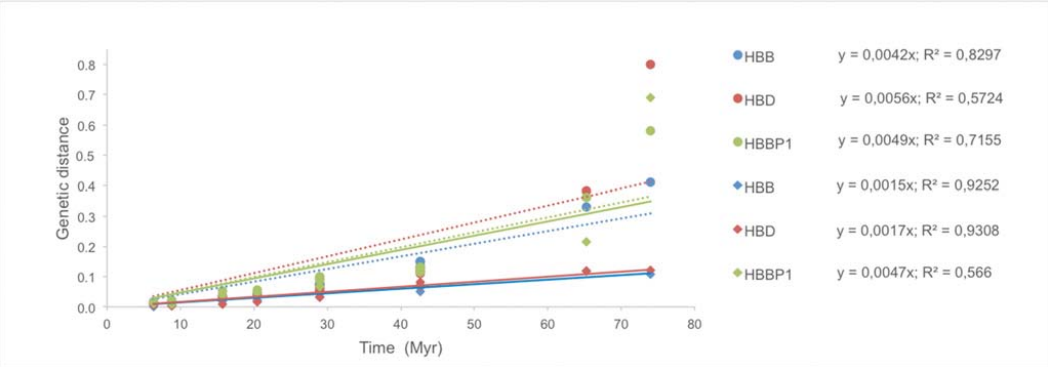


Fig 2. Genetic distance vs divergence times between human and different primate species (Anthropoids and Prosimians) for *β-like* genes. Circles and dotted lines correspond to introns while diamonds and solid lines correspond to exons. Divergence times between humans and other species were obtained with TimeTree [32] and are as follows: Ptr: 6,3 Myr; Ggo 8,8 Myr; Ppy: 15.7 Myr; Nle 20,4 Myr; OWM (Mcc, Panu, Cgue and Caa): 29 Myr; NWM (Sbol, Cjac and Anan): 42,6 Myr; Tsyr: 65,5 Myr; Ogar: 74 Myr. Linear regression trend lines were set to intercept the origin.

doi:10.1371/journal.pone.0123365.g002

the case of *HBD* orthologs, the detected signal of purifying selection may be an outcome of gene conversion by the *HBB* gene in multiple lineages, mostly in the coding region. As illegitimate recombination between *HBD* and *HBB* has been noticeably reduced since the last common ancestor of Anthropoids 65.5 Mya, we examined the extent of the selective pressures exerted in this specific clade. Overall, the significance of the low value obtained under the one-ratio model ($\omega_{HBD\_Anthropoids} = 0.31$) rejects the hypothesis of a neutral evolution of *HBD* in anthropoids. Then, to examine whether *HBD* has been subject to variable selective constraints

Table 2. Parameter Estimates and Likelihood Scores under Different Branch Models.

Model	Parameters for branches	Likelihood (*l*)
One ratio	$\omega_{HBB} = 0.26$**	-4397.16
	$\omega_{HBD} = 0.40$**	-3509.38
	$\omega_{HBD\_Anthropoids} = 0.31$**	-1019.71
Two ratio	$\omega_{HBD\_Apes} = 0.06$**	-1015.62
	$\omega_{HBD\_OWM+NWM} = 0.43$*	
Three ratio	$\omega_{HBD\_Apes} = 0.06$**	-1015.45
	$\omega_{HBD\_OWM} = 0.53$	
	$\omega_{HBD\_NWM} = 0.38$*	
Models Compared	**-2Δ*l***	
One vs. Two ratios	8.18 (df = 1) *	
Two vs. Three ratios	0.34 (df = 1)	

NOTE— ω_{HBB} and ω_{HBD}, ω for all *HBB* and *HBD* lineages, respectively; $\omega_{HBD\_Anthropoids}$, ω for all Anthropoid *HBD* lineages; ω_{HBD_Apes}, ω for Great ape *HBD* lineages; $\omega_{HBD\_OWM+NWM}$, ω for all OWM and NWM *HBD* lineages; ω_{HBD_OWM} and ω_{HBD_NWM}, ω for OWM and NWM *HBD* lineages, respectively; df—degrees of freedom.
*Significant $P < 0.01$
**Significant $P < 0.001$

doi:10.1371/journal.pone.0123365.t002

among different anthropoid clades (Fig 1A), we applied the two-ratio model separating the phylogenetic group of Great Apes from all remaining primates (OWM and NWM). In addition, we also applied the three-ratio model, in which ω was allowed to vary across Great Apes, OWM and NWM clades. The comparison of the one-ratio and the two-ratio models showed that the two-ratio presents an improved fit to anthropoid *HBD* evolution and that it did not differ significantly from the three-ratio model. Nevertheless, in both two-ratio and three-ratio models the Great Apes clade showed an extremely low ω value ($\omega_{HBD\_Apes} = 0.06$), while the OWM+NWM branches present a higher ω ($\omega_{HBD\_OWM+NWM} = 0.43$), but still suggesting a constrained evolution. Although our results indicate that *HBD* might have experienced different selective pressures throughout primate evolution, these estimates corroborated a high conservation of *HBD* in Anthropoid lineages that is unlikely related to protein function, since in most primate species this gene is either weakly expressed or not transcribed at all.

Discussion

In humans and in chimpanzees, unusually high levels of *HBD* sequence conservation, when compared to functional paralogs, have been described [49]. Such pattern of conservation has been difficult to reconcile with the negligible expression of HbA_2. Moreover, the evolutionary history of *HBD* is complex and orthologous relationships among *HBD* and its paralog gene (*HBB*) have been obscured by a history of recurrent gene conversion and unequal crossing overs, throughout eutherian evolution [19,52–54]. Here we gained insight into the evolutionary history of *HBD* and its likely regulatory role in the fetal-to-adult switch unique of Anthropoids, by performing a comprehensive phylogenetic and comparative analysis of the two adult *β-like globin* genes in a wide range of mammalian taxa. The results from our phylogenetic reconstruction are in agreement with previous findings which demonstrated that *HBD* duplication occurred before the radiation of Eutheria [53]. The obtained tree topology is also consistent with a history of concerted evolution between *HBD* and *HBB* that has created chimeric β/δ fusion genes in multiple, independent lineages [19,31,40,54,80]. However, our results show that primates represent an exception to this common trend, given that phylogenetic relationships are maintained throughout this lineage, suggesting that illegitimate recombination between *HBD* and *HBB* has been noticeably reduced since the last common ancestor of Anthropoids 65,5 Mya. Indeed, the recombination analyses here presented demonstrate that even though a certain level of gene conversion has occurred in some Anthropoid lineages, these events have taken place in the stem of the major branches (platyrrhines and catarrhines), as previously inferred from smaller datasets [31,40,45,60]. Moreover, gene conversion in Anthropoids was restricted to shorter regions (exon and intron 1) than those most frequently identified in other species (exons 1, 2, and 3 and intron 1) [19,30,31,40,80]. The distinct phylogenetic patterns between Anthropoid primates and nearly all other species, obtained when using the coding sequence, reflect the relative extent of these events. Although most mammals retain both adult *β-like* gene copies, only the *HBB* gene is functional and essential. *HBD* has apparently become dispensable and in several lineages has pseudogenized. In fact, we confirmed that *HBD* inactivation occurred in a wide range of eutherian species by the accumulation of loss-of-function mutations that disrupted either the ORF or the promoter region. It is noteworthy that when considering extant mammalian species it is only possible to establish bona-fide homology relationships between *HBD* orthologs among Anthropoid primates. Moreover, evolutionary tests suggest that *HBD* is evolving under selective constraints across different Anthropoid species, as hypothesized several decades ago [13,24,37,74]. Such high level of conservation is at odds with the variable expressivity of the δ-globin chain, which is absent in OWM [45,46] and ranges from 1% concentration in hominoids [10] to 6% in NWM [74]. Selective constraints on protein

evolution do not seem a plausible explanation for the signal of purifying selection detected, since to date, HBA_2 has no recognized physiological function [62,76,77].

Interestingly, a role for *HBD* and *HBBP1* in the regulatory mechanisms coordinating the fetal-to adult switch has been proposed in early independent studies [5,13,24,26,55]. Taking into account that the mechanism of Hb switch is common to all simian primates [38], we might expect to find similar patterns of conservation and diversity in ortholog *HBD* sequences for a 65.5 Myr time frame. Accordingly, the patterns of conservation we have now uncovered perfectly overlap with *HBG* duplication and the acquisition of a fetally expressed hemoglobin in anthropoid primates. Noteworthy, we detected in all anthropoid primates a conserved functional GATA-1 motif in the promoter of *HBD*, which has remained intact despite recurrent gene conversion events overlapping the promoter region among these lineages. Considering that *HBD* has very low expression levels in anthropoids, the conservation of a functional GATA-1 binding motif suggests other functional constraints rather than positive regulation of *δ-globin* gene expression. Indeed Gaudry, et al. [19] demonstrated that only late expressed *β-like globin* genes retaining an *HBB*-like promoter are efficiently transcribed. Developmental regulation of gene expression at the *β-globin* cluster involves the formation of chromatin loops mediated by several transcription factors and cofactors [66,68]. It has been shown that GATA-1, along with other cofactors, is required for efficient long-range chromatin interactions between LCR and *β-like globin* genes, namely at the time of γ- to β- globin switch [9,39,85]. Importantly, is has also been demonstrated that the *HBD* upstream region harbors a binding site for BCL11A, which is a biochemically validated and fundamental switching factor necessary for fetal hemoglobin silencing [67,69]. Remarkably, strong interactions between the LCR and the region encompassing both *HBD* and *HBBP1* were uncovered by chromosome conformation (3C and 5C) analyses at the *β-globin* locus [6,17,70]. Collectively, these findings suggest that *HBD* and *HBBP1* might be involved in chromatin looping in the human-globin cluster, a crucial mechanism for temporal coordination of gene expression [16,36]. Interestingly, the observed differences in the rate of evolution between the branches leading to Great Apes and the common branch of OWM and NWM suggest different selective pressures, which may reflect alternative mechanisms of controlling expression in the *β-globin* cluster among Anthropoids. Selective constraints on the protein function cannot be completely ruled out, although evidence of functional relevance of HbA_2 is lacking. HbA_2 has features that are nearly identical to those of HbA [15] but, even though in the absence of β-chain production in β-thalassemia major it becomes the predominant oxygen carrier, it never effectively replaces HbA function. Concordantly, in humans mutations in *HBD* are *per se* clinically silent [71,76]. The evolutionary history of *HBD* in mammals has been shaped by concerted evolution, however our results show that in the extant species of Anthropoids gene conversion events have not been frequent, and when they do occur the exchanged sequence tract is short. The low frequency of gene conversion as well as the conservation of a motif involved in chromatin remodeling could be an outcome of stronger selective constraints acting on *HBD* in anthropoids, as previously reported for other functionally important regions [50,58,92]. These results are also consistent with previous evidence of purifying selection reducing *HBD* genetic diversity in human populations and in chimpanzees [49]. We have now characterized the evolution of *HBD* in mammals and found that the unusual high levels of conservation in this genomic regions are shared across several primate species. In the light of recent advances in the understanding of the *β-globin* cluster regulation, we propose that the similar evolutionary trajectory of *HBD* in Anthropoids is due to functional constraints related with the intricate process of chromatin and protein interactions coordinating the developmental expression of β-*like globin* genes.

Supporting Information

S1 Fig. Sequence alignment for eutherian HBB proteins.
(PDF)

S2 Fig. Sequence alignment of eutherian *HBD* open reading frame. Blue and purple filled boxes mark the exons and donnor/acceptor splice sites, respectively. Dots represent nucleotide identities to the human sequence that was set as reference. Coloured nucleotides indicate changes to the human sequence and aminoacid alterations are marked by filled coloured boxes. The lemur species were excluded from the analysis given that their hybrid ψβ/δ pseudogene [37] generates multiple misalignments.
(PDF)

S3 Fig. Sequence alignment of eutherian *HBB* and *HBD* promoters. A) *HBB* and *HBB*-like *HBD* promoters; B) Anthropoid *HBD* promoters and C) *HBD* promoters lacking the TF binding motifs which are conserved in *HBB*-like and *HBD*-like promoters. Conserved binding motifs are indicated in grey boxes. Again, the lemur species were excluded from the analysis.
(PDF)

S4 Fig. Phylograms depicting relationships among adult *β-like* genes in mammals. The phylogenic tree, based on the coding sequence, was constructed using the Goldman–Yang codon model. Branch support values, obtained using the approximate Likelihood Ration Test (aLRT), are given on the internodes.
(PDF)

S5 Fig. Maximum Likelihood phylograms depicting relationships among *β-like* genes of Anthropoids. The phylogeny reconstructions were based on A) the portion of the alignment that contain evidence of the anthropoid gene conversion (nucleotide 367–632) and B) the portion of the alignment between the inferred breakpoint in event 3 (nucleotide 97–385). Bootstrap branch support (1000 replicates) are given on the internodes.
(PDF)

S1 Table. Listing of *HBB* and *HBD* sequences used for phylogeny reconstructions.
(PDF)

S2 Table. Evolutionary rates based on Jukes Cantor distance.
(PDF)

Author Contributions

Conceived and designed the experiments: AM AA. Performed the experiments: AM. Analyzed the data: AM AML SS MJP AA RMH. Contributed reagents/materials/analysis tools: AM AML SS RMH. Wrote the paper: AM AML SS MJP AA RMH.

References

1. Aguileta G, Bielawski JP, Yang Z. Gene conversion and functional divergence in the beta-globin gene family. J Mol Evol. 2004; 59: 177–189. PMID: 15486692
2. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, et al. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res. 2006; 34: W604–608. PMID: 16845081
3. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, et al. ExPASy: SIB bioinformatics resource portal. Nucleic Acids Research. 2012; 40: W597–W603. doi: 10.1093/nar/gks400 PMID: 22661580

4. Bank A. Regulation of human fetal hemoglobin: new players, new complexities. Blood. 2006; 107: 435–443. PMID: 16109777
5. Bank A, Mears JG, Ramirez F. Disorders of human hemoglobin. Science. 1980; 207: 486–493. PMID: 7352255
6. Beauchemin H, Trudel M. Evidence for a bigenic chromatin subdomain in regulation of the fetal-to-adult hemoglobin switch. Mol Cell Biol. 2009; 29: 1635–1648. doi: 10.1128/MCB.01735-08 PMID: 19114559
7. Bielawski JP, Yang Z. Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics. 2003; 3: 201–212. PMID: 12836699
8. Boni MF, Posada D, Feldman MW. An Exact Nonparametric Method for Inferring Mosaic Structure in Sequence Triplets. Genetics. 2007; 176: 1035–1047. PMID: 17409078
9. Bottardi S, Ross J, Bourgoin V, Fotouhi-Ardakani N, Affar el B, Trudel M, et al. Ikaros and GATA-1 combinatorial effect is required for silencing of human gamma-globin genes. Mol Cell Biol. 2009; 29: 1526–1537. doi: 10.1128/MCB.01523-08 PMID: 19114560
10. Boyer S, Crosby E, Noyes A, Fuller G, Leslie S, Donaldson L, et al. Primate hemoglobins: Some sequences and some proposals concerning the character of evolution and mutation. Biochemical Genetics. 1971; 5: 405–448. PMID: 4999925
11. Bulger M, Groudine M. Looping versus linking: toward a model for long-distance gene activation. Genes Dev. 1999; 13: 2465–2477. PMID: 10521391
12. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997; 268: 78–94. PMID: 9149143
13. Chang LY, Slightom JL. Isolation and nucleotide sequence analysis of the beta-type globin pseudogene from human, gorilla and chimpanzee. J Mol Biol. 1984; 180: 767–784. PMID: 6098690
14. Czelusniak J, Goodman M, Hewett-Emmett D, Weiss ML, Venta PJ, Tashian RE. Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. Nature. 1982; 298: 297–300. PMID: 6178039
15. de Bruin SH, Janssen LHM. Comparison of the oxygen and proton binding behavior of human hemoglobin A and A2. Biochimica et Biophysica Acta (BBA)—Protein Structure. 1973; 295: 490–494.
16. Dean A. On a chromosome far, far away: LCRs and gene expression. Trends Genet. 2006; 22: 38–45. PMID: 16309780
17. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006; 16: 1299–1309. PMID: 16954542
18. Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, et al. The structure and evolution of the human beta-globin gene family. Cell. 1980; 21: 653–668. PMID: 6985477
19. Gaudry MJ, Storz JF, Butts GT, Campbell KL, Hoffmann FG. Repeated Evolution of Chimeric Fusion Genes in the β-Globin Gene Family of Laurasiatherian Mammals. Genome Biology and Evolution. 2014. doi: 10.1093/gbe/evu097
20. Gibbs MJ, Armstrong JS, Gibbs AJ. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics. 2000; 16: 573–582. PMID: 11038328
21. Gil M, Zanetti MS, Zoller S, Anisimova M. CodonPhyML: Fast Maximum Likelihood Phylogeny Estimation under Codon Substitution Models. Molecular Biology and Evolution. 2013. doi: 10.1093/molbev/mst034
22. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994; 11: 725–736. PMID: 7968486
23. Goodman M, Czelusniak J, Koop BF, Tagle DA, Slightom JL. Globins: a case study in molecular phylogeny. Cold Spring Harb Symp Quant Biol. 1987; 52: 875–890. PMID: 3454296
24. Goodman M, Koop BF, Czelusniak J, Weiss ML. The eta-globin gene. Its long evolutionary history in the beta-globin gene family of mammals. J Mol Biol. 1984; 180: 803–823. PMID: 6527390
25. Goodman M, Moore GW, Matsuda G. Darwinian evolution in the genealogy of haemoglobin. Nature. 1975; 253: 603–608. PMID: 1089897
26. Hardies SC, Edgell MH, Hutchison CA 3rd. Evolution of the mammalian beta-globin gene cluster. J Biol Chem. 1984; 259: 3748–3756. PMID: 6706976
27. Hardison R. Evolution of hemoglobin and its genes. Cold Spring Harb Perspect Med. 2012; 2: a011627. doi: 10.1101/cshperspect.a011627 PMID: 23209182
28. Hardison R. Hemoglobins from bacteria to man: evolution of different patterns of gene expression. J Exp Biol. 1998; 201: 1099–1117. PMID: 9510523

29. Hardison R. Organization, evolution and regulation of the globin genes. In: Steinberg MH, Forget BG, Higgs DR, Nagel RL, editors. Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management. Cambridge: Cambridge University Press; 2001. PMID: 12779271

30. Hardison RC. Comparison of the beta-like globin gene families of rabbits and humans indicates that the gene cluster 5'-epsilon-gamma-delta-beta-3' predates the mammalian radiation. Mol Biol Evol. 1984; 1: 390–410. PMID: 6599973

31. Hardison RC, Margot JB. Rabbit globin pseudogene psi beta 2 is a hybrid of delta- and beta-globin gene sequences. Mol Biol Evol. 1984; 1: 302–316. PMID: 6599969

32. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 2006; 22: 2971–2972. PMID: 17021158

33. Hoffmann FG, Opazo JC, Storz JF. New genes originated via multiple recombinational pathways in the beta-globin gene family of rodents. Mol Biol Evol. 2008; 25: 2589–2600. doi: 10.1093/molbev/msn200 PMID: 18780876

34. Hoffmann FG, Storz JF. The alphaD-globin gene originated via duplication of an embryonic alpha-like globin gene in the ancestor of tetrapod vertebrates. Mol Biol Evol. 2007; 24: 1982–1990. PMID: 17586601

35. Hoffmann FG, Storz JF, Gorr TA, Opazo JC. Lineage-specific patterns of functional diversification in the alpha- and beta-globin gene families of tetrapod vertebrates. Mol Biol Evol. 2010; 27: 1126–1138. doi: 10.1093/molbev/msp325 PMID: 20047955

36. Holwerda S, de Laat W. Chromatin loops, gene positioning, and gene expression. Front Genet. 2012; 3: 217. doi: 10.3389/fgene.2012.00217 PMID: 23087710

37. Jeffreys AJ, Barrie PA, Harris S, Fawcett DH, Nugent ZJ, Boyd AC. Isolation and sequence analysis of a hybrid delta-globin pseudogene from the brown lemur. J Mol Biol. 1982; 156: 487–503. PMID: 6214636

38. Johnson RM, Gumucio D, Goodman M. Globin gene switching in primates. Comp Biochem Physiol A Mol Integr Physiol. 2002; 133: 877–883. PMID: 12443943

39. Keys JR, Tallack MR, Zhan Y, Papathanasiou P, Goodnow CC, Gaensler KM, et al. A mechanism for Ikaros regulation of human globin gene switching. Br J Haematol. 2008; 141: 398–406. doi: 10.1111/j.1365-2141.2008.07065.x PMID: 18318763

40. Koop BF, Siemieniak D, Slightom JL, Goodman M, Dunbar J, Wright PC, et al. Tarsius delta- and beta-globin genes: conversions, evolution, and systematic implications. J Biol Chem. 1989; 264: 68–79. PMID: 2491855

41. Lanfear R, Calcott B, Ho SY, Guindon S. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol. 2012; 29: 1695–1701. doi: 10.1093/molbev/mss020 PMID: 22319168

42. Martin D, Rybicki E. RDP: detection of recombination amongst aligned sequences. Bioinformatics. 2000; 16: 562–563. PMID: 10980155

43. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics. 2010; 26: 2462–2463. doi: 10.1093/bioinformatics/btq467 PMID: 20798170

44. Martin DP, Posada D, Crandall KA, Williamson C. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses. 2005; 21: 98–102. PMID: 15665649

45. Martin SL, Vincent KA, Wilson AC. Rise and fall of the delta globin gene. J Mol Biol. 1983; 164: 513–528. PMID: 6188843

46. Martin SL, Zimmer EA, Kan YW, Wilson AC. Silent delta-globin gene in Old World monkeys. Proc Natl Acad Sci U S A. 1980; 77: 3563–3566. PMID: 6251467

47. Matsuda M, Sakamoto N, Fukumaki Y. Delta-thalassemia caused by disruption of the site for an erythroid-specific transcription factor, GATA-1, in the delta-globin gene promoter. Blood. 1992; 80: 1347–1351. PMID: 1515647

48. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees; 2010. pp. 1–8.

49. Moleirinho A, Seixas S, Lopes AM, Bento C, Prata MJ, Amorim A. Evolutionary constraints in the beta-globin cluster: the signature of purifying selection at the delta-globin (HBD) locus and its role in developmental gene regulation. Genome Biol Evol. 2013; 5: 559–571. doi: 10.1093/gbe/evt029 PMID: 23431002

50. Noonan JP, Grimwood J, Schmutz J, Dickson M, Myers RM. Gene conversion and the evolution of protocadherin gene cluster diversity. Genome Res. 2004; 14: 354–366. PMID: 14993203

51. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000; 302: 205–217. PMID: 10964570
52. Opazo JC, Hoffmann FG, Storz JF. Differential loss of embryonic globin genes during the radiation of placental mammals. Proceedings of the National Academy of Sciences. 2008; 105: 12950–12955. doi: 10.1073/pnas.0804392105 PMID: 18755893
53. Opazo JC, Hoffmann FG, Storz JF. Genomic evidence for independent origins of β-like globin genes in monotremes and therian mammals. Proceedings of the National Academy of Sciences. 2008; 105: 1590–1595. doi: 10.1073/pnas.0710531105 PMID: 18216242
54. Opazo JC, Sloan AM, Campbell KL, Storz JF. Origin and Ascendancy of a Chimeric Fusion Gene: The β/δ-Globin Gene of Paenungulate Mammals. Molecular Biology and Evolution. 2009; 26: 1469–1478. doi: 10.1093/molbev/msp064 PMID: 19332641
55. Ottolenghi S, Giglioni B, Comi P, Gianni AM, Polli E, Acquaye CT, et al. Globin gene deletion in HPFH, delta (o) beta (o) thalassaemia and Hb Lepore disease. Nature. 1979; 278: 654–657. PMID: 450068
56. Padidam M, Sawyer S, Fauquet CM. Possible emergence of new geminiviruses by frequent recombination. Virology. 1999; 265: 218–225. PMID: 10600594
57. Patel VS, Cooper SJ, Deakin JE, Fulton B, Graves T, Warren WC, et al. Platypus globin genes and flanking loci suggest a new insertional model for beta-globin evolution in birds and mammals. BMC Biol. 2008; 6: 34. doi: 10.1186/1741-7007-6-34 PMID: 18657265
58. Petronella N, Drouin G. Purifying selection against gene conversions in the folate receptor genes of primates. Genomics. 2014; 103: 40–47. doi: 10.1016/j.ygeno.2013.10.004 PMID: 24184359
59. Posada D, Crandall KA. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci U S A. 2001; 98: 13757–13762. PMID: 11717435
60. Prychitko T, Johnson RM, Wildman DE, Gumucio D, Goodman M. The phylogenetic history of New World monkey beta globin reveals a platyrrhine beta to delta gene conversion in the atelid ancestry. Mol Phylogenet Evol. 2005; 35: 225–234. PMID: 15737593
61. Rambaut A, Suchard MA, Xie D, Drummond AJ. Tracer v1.6; 2014. Available: http://beast.bio.ed.ac.uk/Tracer.
62. Ranney HM, Lam R, Rosenberg G. Some properties of hemoglobin A2. American Journal of Hematology. 1993; 42: 107–111. PMID: 8416283
63. Ristaldi MS, Casula S, Porcu S, Marongiu MF, Pirastu M, Cao A. Activation of the delta-globin gene by the beta-globin gene CACCC motif. Blood Cells Mol Dis. 1999; 25: 193–209. PMID: 10575545
64. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012; 61: 539–542. doi: 10.1093/sysbio/sys029 PMID: 22357727
65. Rozas J. DNA sequence polymorphism analysis using DnaSP. Methods Mol Biol. 2009; 537: 337–350. doi: 10.1007/978-1-59745-251-9_17 PMID: 19378153
66. Sankaran VG, Orkin SH. The switch from fetal to adult hemoglobin. Cold Spring Harb Perspect Med. 2013; 3: a011643. doi: 10.1101/cshperspect.a011643 PMID: 23209159
67. Sankaran VG, Xu J, Byron R, Greisman HA, Fisher C, Weatherall DJ, et al. A Functional Element Necessary for Fetal Hemoglobin Silencing. New England Journal of Medicine. 2011; 365: 807–814. doi: 10.1056/NEJMoa1103070 PMID: 21879898
68. Sankaran VG, Xu J, Orkin SH. Advances in the understanding of haemoglobin switching. Br J Haematol. 2010; 149: 181–194. doi: 10.1111/j.1365-2141.2010.08105.x PMID: 20201948
69. Sankaran VG, Xu J, Ragoczy T, Ippolito GC, Walkley CR, Maika SD, et al. Developmental and species-divergent globin switching are driven by BCL11A. Nature. 2009; 460: 1093–1097. doi: 10.1038/nature08243 PMID: 19657335
70. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012; 489: 109–113. doi: 10.1038/nature11279 PMID: 22955621
71. Schechter AN. Hemoglobin research and the origins of molecular medicine. 3927–3938 p; 2008
72. Smith JM. Analyzing the mosaic structure of genes. J Mol Evol. 1992; 34: 126–129. PMID: 1556748
73. Song G, Hsu CH, Riemer C, Zhang Y, Kim HL, Hoffmann F, et al. Conversion events in gene clusters. BMC Evol Biol. 2011; 11: 226. doi: 10.1186/1471-2148-11-226 PMID: 21798034
74. Spritz RA, Giebel LB. The structure and evolution of the spider monkey delta-globin gene. Mol Biol Evol. 1988; 5: 21–29. PMID: 2833675
75. Stamatakis A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics. 2014.

76. Steinberg MH, Adams JG 3rd. Hemoglobin A2: origin, evolution, and aftermath. Blood. 1991; 78: 2165–2177. PMID: 1932737

77. Steinberg MH, et al. Disorders of Hemoglobin. Cambridge University Press; 2009.

78. Storz JF, Opazo JC, Hoffmann FG. Gene duplication, genome duplication, and the functional diversification of vertebrate globins. Mol Phylogenet Evol. 2013; 66: 469–478. doi: 10.1016/j.ympev.2012.07.013 PMID: 22846683

79. Storz JF, Opazo JC, Hoffmann FG. Phylogenetic diversification of the globin gene superfamily in chordates. IUBMB Life. 2011; 63: 313–322. doi: 10.1002/iub.482 PMID: 21557448

80. Tagle DA, Slightom JL, Jones RT, Goodman M. Concerted evolution led to high expression of a prosimian primate delta globin gene locus. J Biol Chem. 1991; 266: 7469–7480. PMID: 2019578

81. Tang DC, Ebb D, Hardison RC, Rodgers GP. Restoration of the CCAAT box or insertion of the CACCC motif activates [corrected] delta-globin gene expression. Blood. 1997; 90: 421–427. PMID: 9207479

82. Tang DC, Rodgers GP. Activation of the human delta-globin gene promoter in primary adult erythroid cells. Br J Haematol. 1998; 103: 835–838. PMID: 9858241

83. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22: 4673–4680. PMID: 7984417

84. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell. 2002; 10: 1453–1465. PMID: 12504019

85. Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, et al. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. Mol Cell. 2005; 17: 453–462. PMID: 15694345

86. Vincent KA, Wilson AC. Evolution and transcription of old world monkey globin genes. J Mol Biol. 1989; 207: 465–480. PMID: 2760921

87. Webster MT, Clegg JB, Harding RM. Common 5' beta-globin RFLP haplotypes harbour a surprising level of ancestral sequence mosaicism. Hum Genet. 2003; 113: 123–139. PMID: 12736816

88. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 1998; 15: 568–573. PMID: 9580986

89. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution. 2007; 24: 1586–1591. PMID: 17483113

90. Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 2000; 155: 431–449. PMID: 10790415

91. Zhang J. Evolution by gene duplication: an update. Trends in Ecology & Evolution. 2003; 18: 292–298.

92. Zhao Z, Hewett-Emmett D, Li WH. Frequent gene conversion between human red and green opsin genes. J Mol Evol. 1998; 46: 494–496. PMID: 9541545

Distinctive patterns of evolution of the *δ- globin* gene (*HBD*) in primates

Supplementary Material

Ana Moleirinho[1,2,3], Alexandra M. Lopes[1,2], Susana Seixas[1,2], Ramiro Morales-Hojas[4], Maria J. Prata[1,2,3], António Amorim[1,2,3]

1- Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal

2 - IPATIMUP- Institute of Molecular Pathology and Immunology, University of Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal;

3 - Department of Biology, Faculty of Sciences, University of Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

4 – Genetics and Genomics Group, The Pirbright Institute, Compton Laboratory, Compton, Berkshire RG20 7NN, UK

Corresponding author: Ana Moleirinho, IPATIMUP,Rua Dr Roberto Frias s/n,4200-465 Porto, Portugal; Phone: (+351) 22 5570700;Fax: (+351) 22 5570799; E-mail: amoleirinho@ipatimup.pt

Keywords: *β-globin* cluster, Hemoglobin switch, Gene conversion, Gata-1 motif, Gene evolution

Tables

S1 Table. Listing of *HBB* and *HBD* sequences used for phylogeny reconstructions.

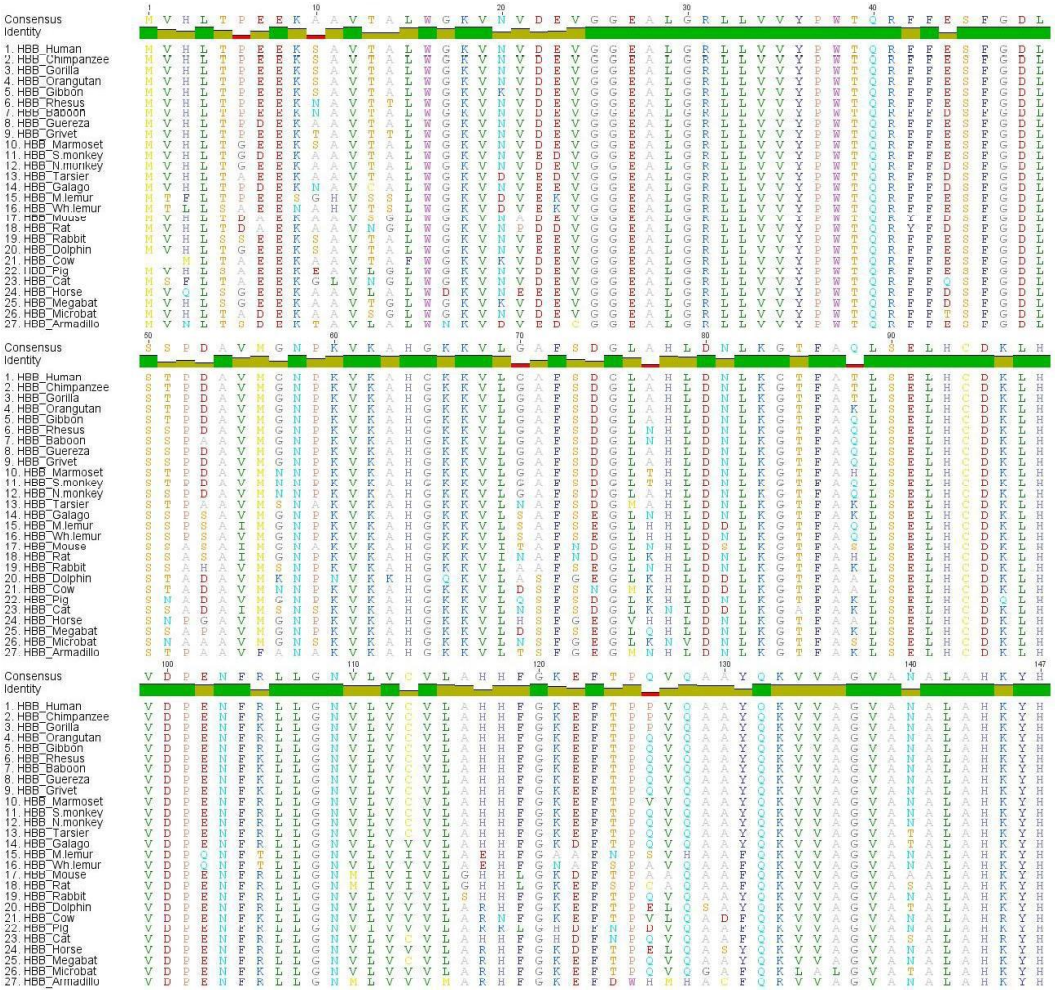
Order	SCIENTIFIC NAME	COMMON NAME	Assembly/ Accession Number	HBB	HBD
Primates	*Homo sapiens**	Human	GRCh37/hg19	chr11:5246600-5248430	chr11:5253981-5255837
	*Pan troglodytes**	Chimpanzee	CSAC 2.1.4/panTro4	chr11:4976349-4978179	chr11:4983729-4985583
	*Gorilla gorilla**	Gorilla	gorGor3.1/gorGor3	chr11:5181180-5183419	chr11:5189200-5190818
	*Pongo pygmaeus**	Orangutan	WUGSC 2.0.2/ponAbe2	chr11:65237578-65239448	chr11:65230079-65231933
	Nomascus leucogenys	Gibbon	GGSC Nleu3.0/nomLeu3	chr15:67092208-67094052	chr15:67099627-67101481
	Papio anubis	Baboon	Baylor Panu_2.0/papAnu2	chr14:60046909-60048748	chr14:60039529-60041379
	*Macaca mulatta**	Rhesus	BGI CR_1.0/rheMac3	chr14:68486458-68488296	chr14:68479060-68480908
	Colobus guereza	Guereza	AC175618.2	Range:43444-45290	Range:50789-52640
	Chlorocebus aethiops	Grivet	AC192680.2	Range:56560-58450	Range:64005-65870
	Callithrix jacchus	Marmoset	WUGSC 3.2/calJac3	chr11:68655845-68657643	chr11:68662011-68663826
	Aotus nancymaae	Night monkey	AC174399.2	Range:86520-88480	Range:93020-94850
	Saimiri boliviensis	Squirrel monkey	Broad/saiBol1	JH378113:894229-896044	JH378113:887920-889755
	Microcebus murinus	Mouse lemur	Broad/micMur1	scaffold_23051:7558-9396	scaffold_23051:11861-13574
	Eulemur macaco	Brown lemur	-	M15734.1	-
	Eulemur albifrons	White-headed lemur	-	-	V00644.1
	Tarsius syrichta	Tarsier	Broad/tarSyr1	J04429.1	scaffold_30223:10672-12718
	Otolemur garnettii	Galago	U60902.1	Range:50897-53130	Range:46123-48125
Rodentia	*Rattus norvegicus*	Rat	Baylor 3.4/rn4	chr1:161618682-161620327	-
	Mus musculus	Mouse	GRCm38/mm10	chr7:103826432-103828057	chr7:103838921-103840610
Lagomorpha	*Oryctolagus cuniculus*	Rabbit	Broad/oryCun2	chr1:146236925-146238433	chr1:146245465-146247150
Carnivora	*Felis catus*	Cat	AC129072.3	Range:29610-31430	Range:33950-35770
Chiroptera	*Myotis lucifugus*	Microbat	Myoluc2.0/myoLuc2	GL429905:2514138-2515941	GL429905:2519498-2521193
	Pteropus vampyrus	Megabat	Broad/pteVam1/ AC216164.2	scaffold_3459:3976-2676	scaffold_3459:9145-11021

				Range:179400-179563	
Cetartiodactyla	*Bos Taurus*	Cow	Baylor Btau_4.6.1/bosTau7	chr15:47792823-47794689	chr15:47782257-47784127
	Sus scrofa	Pig	SGSC Sscrofa10.2/susScr3	chr9:5632884-5634490	chr9:5640805-5642408
	Tursiops truncatus	Dolphin	Baylor Ttru_1.4/turTru2 gnl|ti|2241352997 XM_004330598.1	JH496320:15248-16841 Range:366-500	JH496320:8718-10536
Perissodactyla	*Equus caballus*	Horse	Broad/equCab2	chr7:73936358-73937959	chr7:73943584-73945347
Cingulata	*Dasypus novemcinctus*	Armadillo	Baylor/dasNov3	JH576106:328717-330317	JH576106:324817-326487
Didelphimorphia	*Monodelphis domestica*	Opossum	Broad/monDom5	chr4:352263811-352265844	-
Monotremata	*Ornithorhynchus anatinus*	Platypus	WUGSC 5.0.1/ornAna1	Contig7843:24151-25617	-
Galliformes	*Gallus gallus*	Chicken	ICGSC Gallus_gallus-4.0/galGal4	chr1:193728454-193730403	-

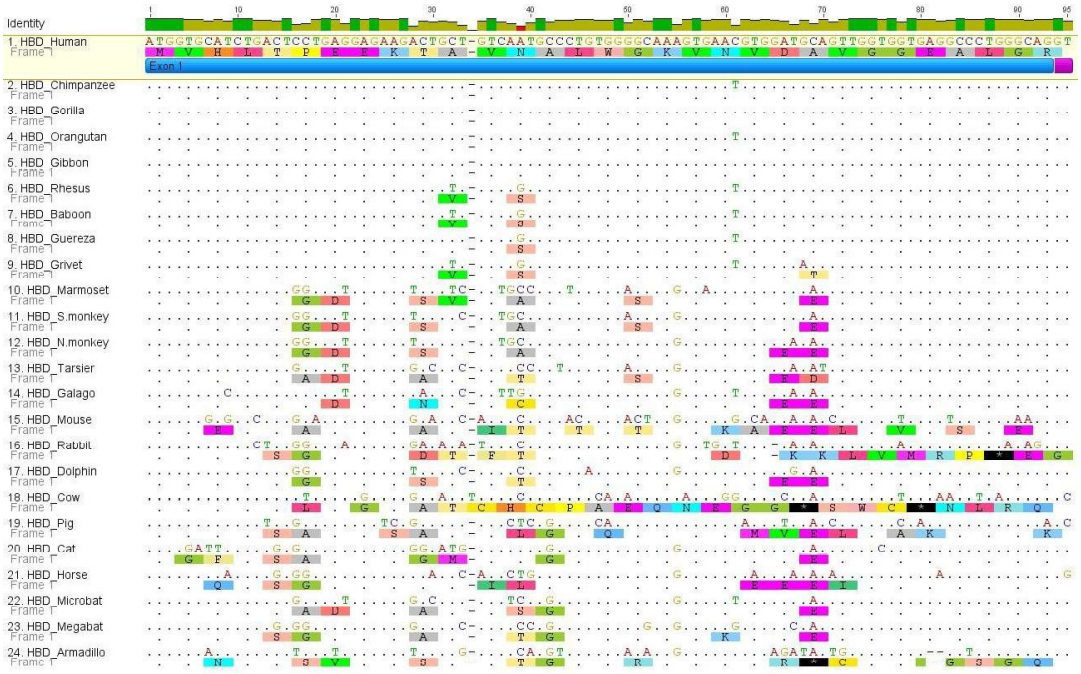
*sequences confirmed by sequencing

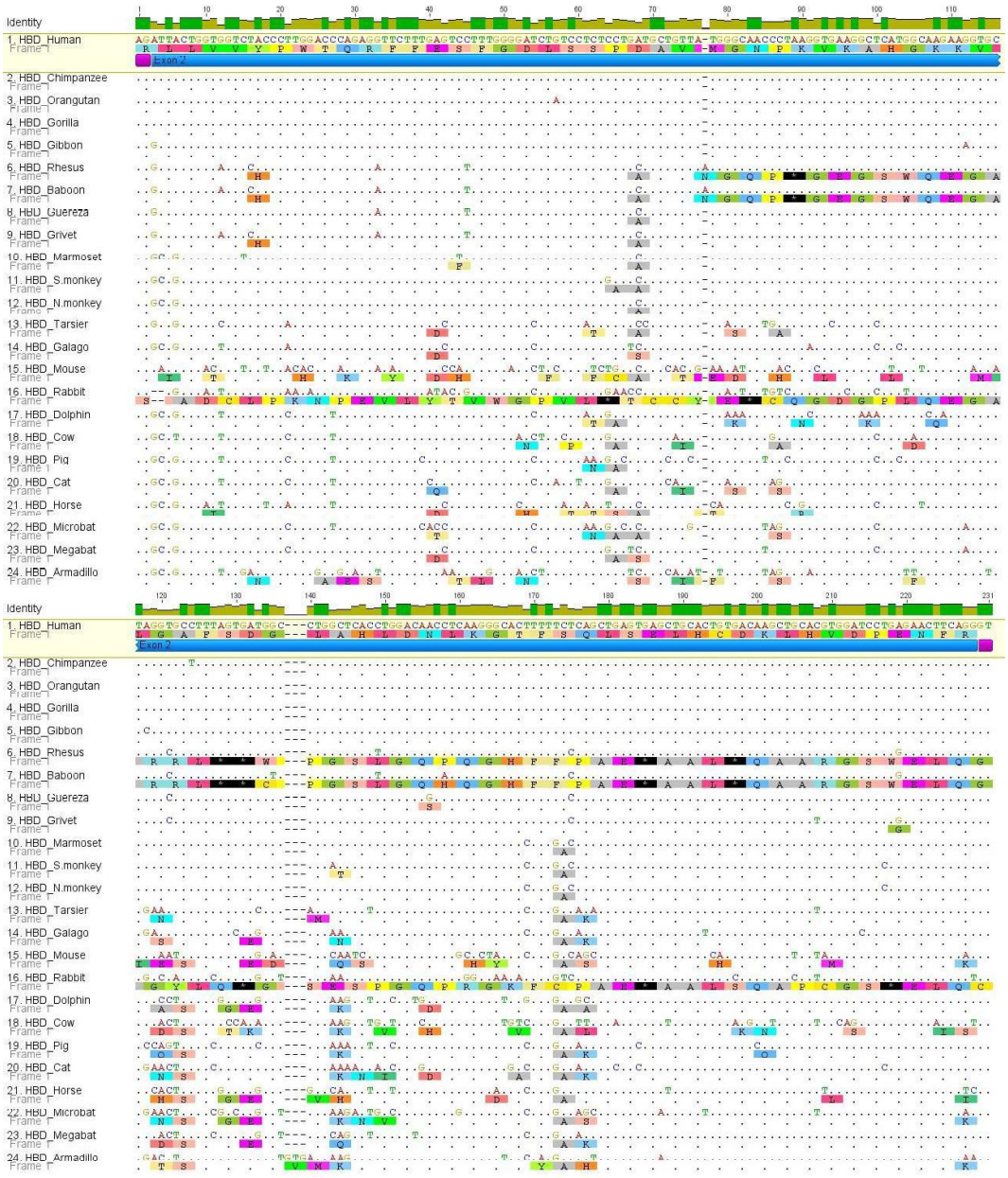
S2 Table. Evolutionary rates based on Jukes Cantor distance.

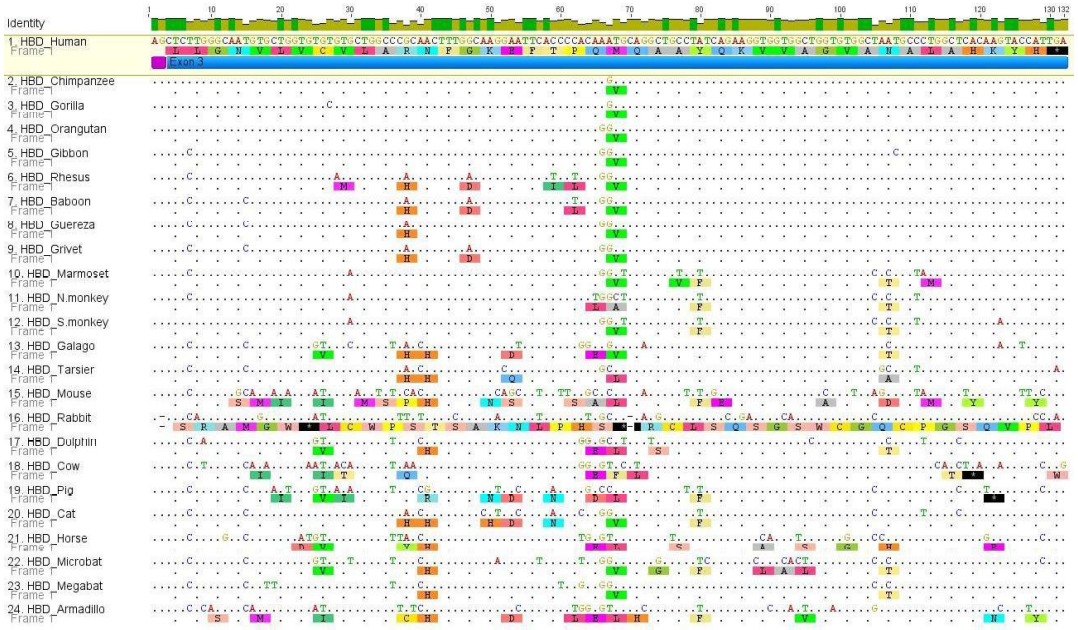
	Time (Myr)	*Exons*			*Introns*		
		HBB	**HBD**	**HBBP1**	**HBB**	**HBD**	**HBBP1**
Hum-Ptr	6,3	0,18	0,54	0,90	1,39	0,94	1,15
Hum-Ggo	8,8	0,26	0,26	0,52	0,88	0,90	1,18
Hum-Ppy	15,7	0,58	0,29	1,18	1,13	1,08	1,71
Hum-Nle	20,4	0,56	0,39	1,15	0,98	1,13	1,40
Hum-Mcc	29	0,80	0,88	1,18	1,25	0,89	1,53
Hum-Panu		0,84	0,92	1,18	1,29	0,86	1,61
Hum-Cgue		0,55	0,55	1,48	1,24	1,08	1,67
Hum-Caa		0,55	0,80	1,14	1,35	1,04	1,71
Hum-Sbol		0,80	0,83	1,39	1,75	1,30	1,54
Hum-Cjac	42,6	0,74	0,98	1,54	1,77	1,53	1,38
Hum-Anan		0,60	0,80	1,44	1,56	1,38	1,36
Hum-Tsyr	65,2	0,92	0,92	1,66	2,54	2,94	2,77
Hum-Ogar	74	0,74	0,82	4,66	2,78	5,41	3,92



S1 Figure. Sequence alignment for eutherian HBB proteins.

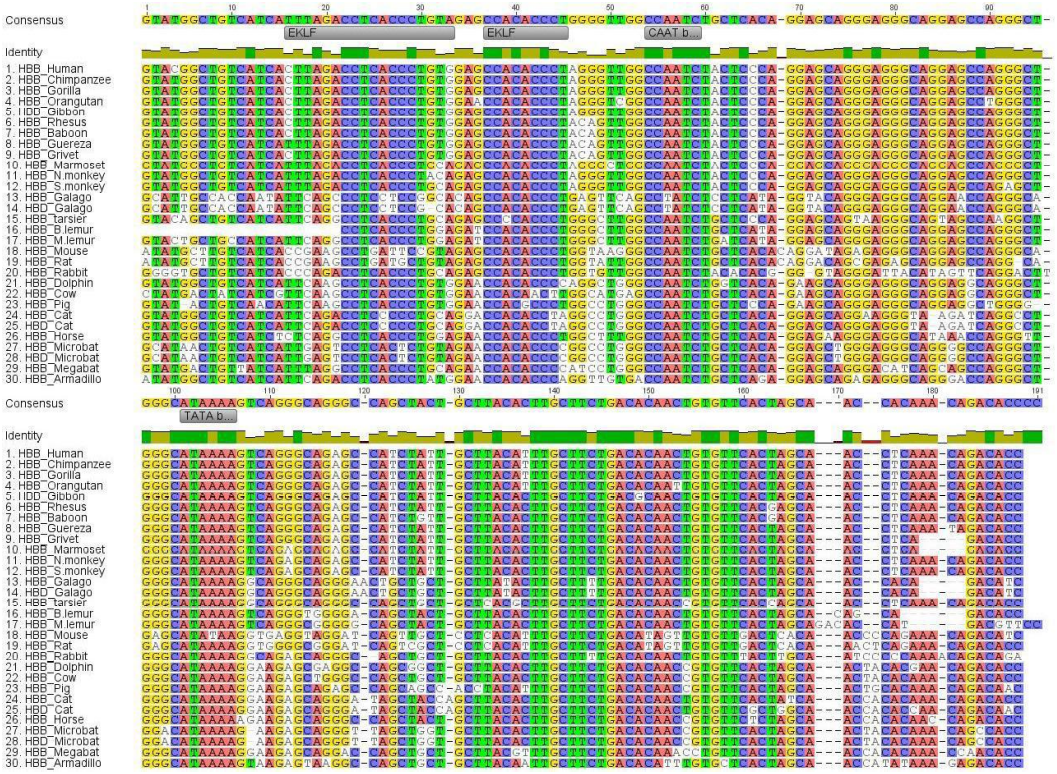

Identity
1. HBD_Human
Frame
Exon 1
2. HBD_Chimpanzee
3. HBD_Gorilla
4. HBD_Orangutan
5. HBD_Gibbon
6. HBD_Rhesus
7. HBD_Baboon
8. HBD_Guereza
9. HBD_Grivet
10. HBD_Marmoset
11. HBD_S.monkey
12. HBD_N.monkey
13. HBD_Tarsier
14. HBD_Galago
15. HBD_Mouse
16. HBD_Rabbit
17. HBD_Dolphin
18. HBD_Cow
19. HBD_Pig
20. HBD_Cat
21. HBD_Horse
22. HBD_Microbat
23. HBD_Megabat
24. HBD_Armadillo


Identity
1. HBD_Human
Exon 2
2. HBD_Chimpanzee
3. HBD_Orangutan
4. HBD_Gorilla
5. HBD_Gibbon
6. HBD_Rhesus
7. HBD_Baboon
8. HBD_Guereza
9. HBD_Grivet
10. HBD_Marmoset
11. HBD_S.monkey
12. HBD_N.monkey
13. HBD_Tarsier
14. HBD_Galago
15. HBD_Mouse
16. HBD_Rabbit
17. HBD_Dolphin
18. HBD_Cow
19. HBD_Pig
20. HBD_Cat
21. HBD_Horse
22. HBD_Microbat
23. HBD_Megabat
24. HBD_Armadillo

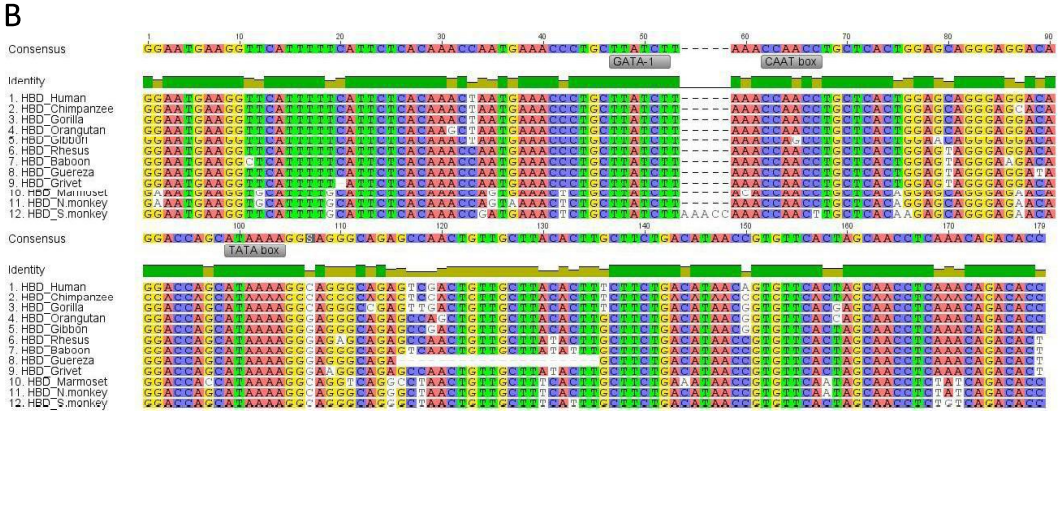


S2 Figure. Sequence alignment of eutherian *HBD* open reading frame. Blue and purple filled boxes mark the exons and donnor/acceptor splice sites, respectively. Dots represent nucleotide identities to the human sequence that was set as reference. Coloured nucleotides indicate changes to the human sequence and aminoacid alterations are marked by filled coloured boxes. The lemur species were excluded from the analysis given that their hybrid ψβ/δ pseudogene [37] generates multiple misalignments.

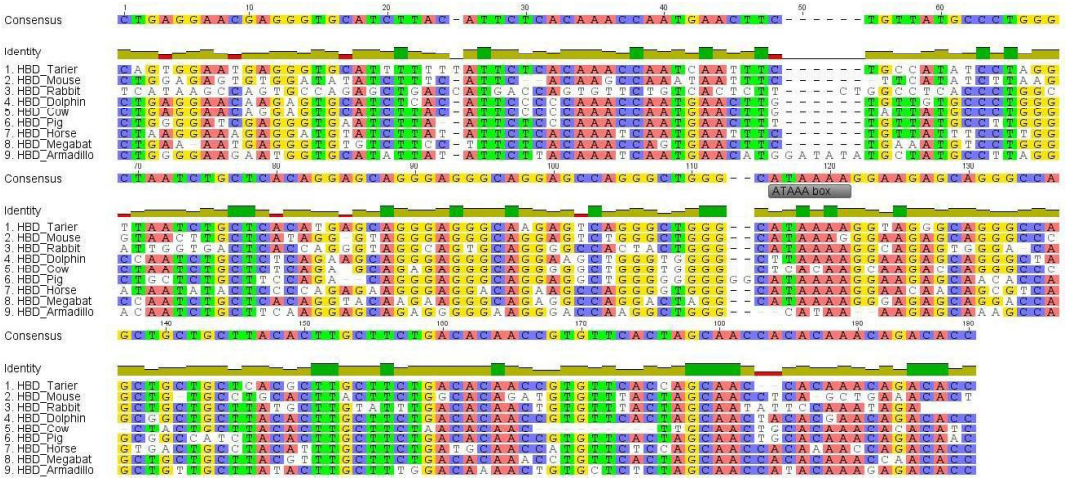
A


Consensus
EKLF
EKLF
CAAT b...
Identity
1. HBB_Human
2. HBB_Chimpanzee
3. HBB_Gorilla
4. HBB_Orangutan
5. HBD_Gibbon
6. HBB_Rhesus
7. HBB_Baboon
8. HBB_Guereza
9. HBB_Grivet
10. HBB_Marmoset
11. HBB_N.monkey
12. HBB_S.monkey
13. HBB_Galago
14. HBD_Galago
15. HBB_Tarsier
16. HBB_B.lemur
17. HBB_M.lemur
18. HBB_Mouse
19. HBB_Rat
20. HBB_Rabbit
21. HBB_Dolphin
22. HBB_Cow
23. HBB_Pig
24. HBB_Cat
25. HBD_Cat
26. HBB_Horse
27. HBB_Microbat
28. HBD_Microbat
29. HBB_Megabat
30. HBB_Armadillo
Consensus
TATA b...
Identity
1. HBB_Human
2. HBB_Chimpanzee
3. HBB_Gorilla
4. HBB_Orangutan
5. HBD_Gibbon
6. HBB_Rhesus
7. HBB_Baboon
8. HBB_Guereza
9. HBB_Grivet
10. HBB_Marmoset
11. HBB_N.monkey
12. HBB_S.monkey
13. HBB_Galago
14. HBD_Galago
15. HBB_Tarsier
16. HBB_B.lemur
17. HBB_M.lemur
18. HBB_Mouse
19. HBB_Rat
20. HBB_Rabbit
21. HBB_Dolphin
22. HBB_Cow
23. HBB_Pig
24. HBB_Cat
25. HBD_Cat
26. HBB_Horse
27. HBB_Microbat
28. HBB_Microbat
29. HBB_Megabat
30. HBB_Armadillo

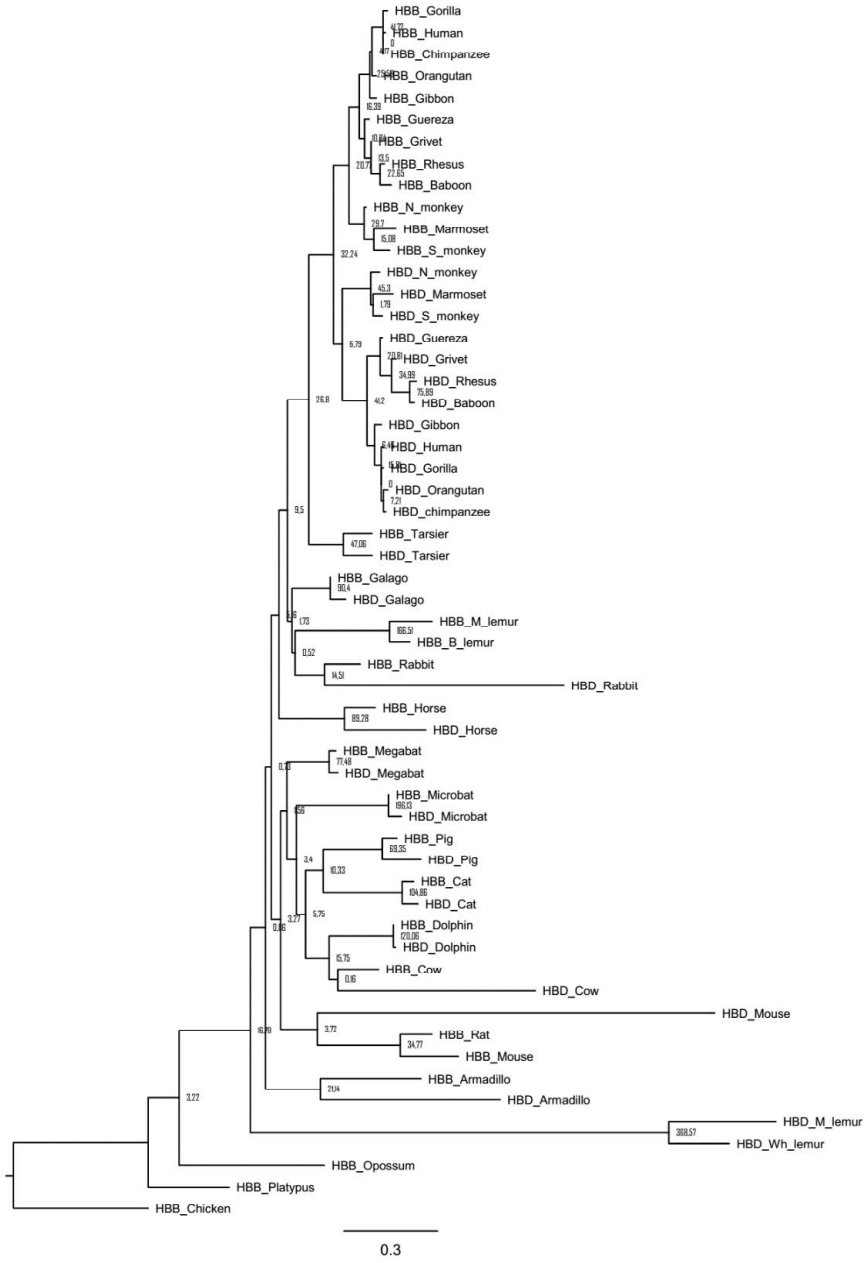
B


Consensus
GATA-1
CAAT box
Identity
1. HBD_Human
2. HBD_Chimpanzee
3. HBD_Gorilla
4. HBD_Orangutan
5. HBD_Gibbon
6. HBD_Rhesus
7. HBD_Baboon
8. HBD_Guereza
9. HBD_Grivet
10. HBD_Marmoset
11. HBD_N.monkey
12. HBD_S.monkey
Consensus
TATA box
Identity
1. HBD_Human
2. HBD_Chimpanzee
3. HBD_Gorilla
4. HBD_Orangutan
5. HBD_Gibbon
6. HBD_Rhesus
7. HBD_Baboon
8. HBD_Guereza
9. HBD_Grivet
10. HBD_Marmoset
11. HBD_N.monkey
12. HBD_S.monkey

C

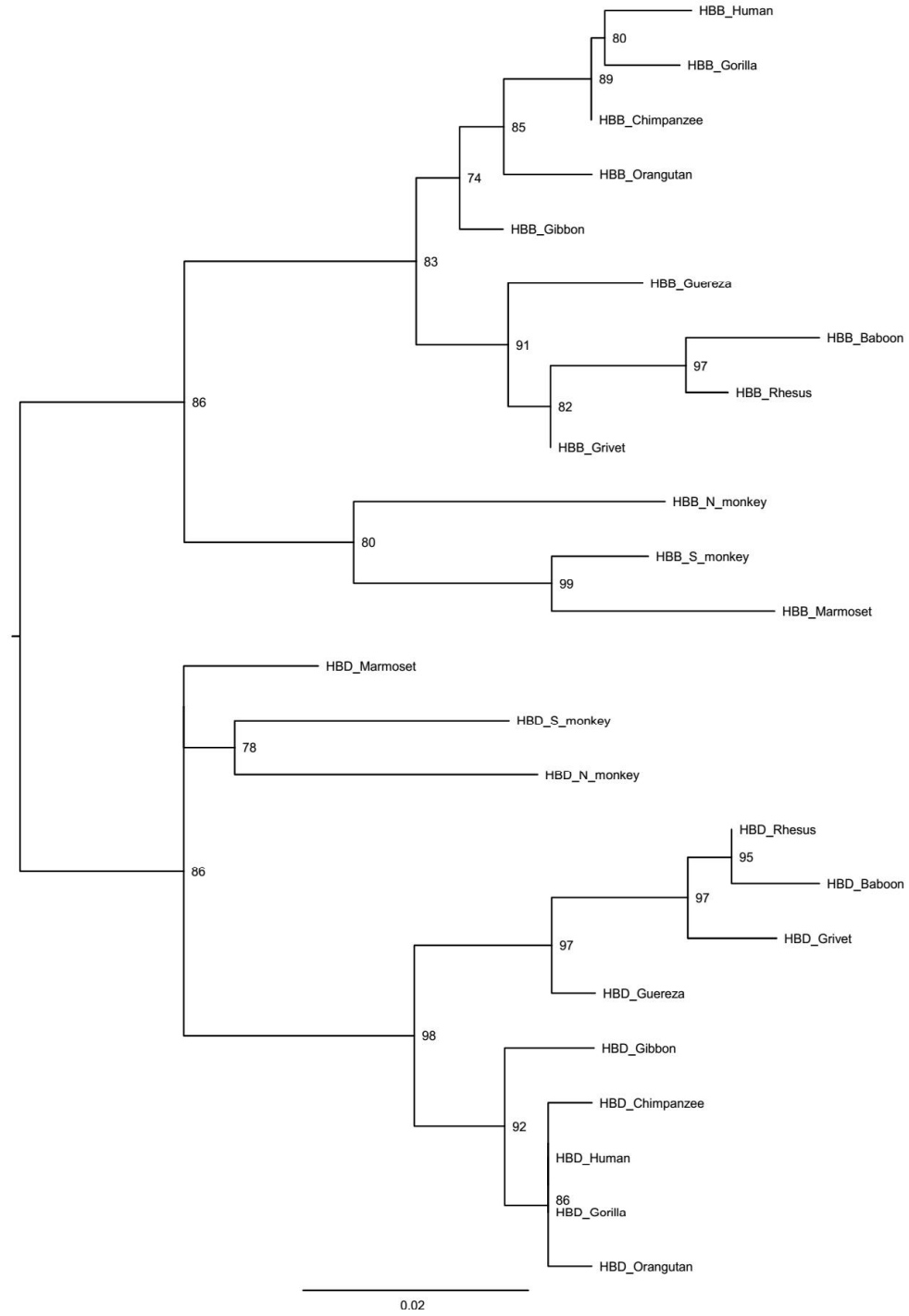


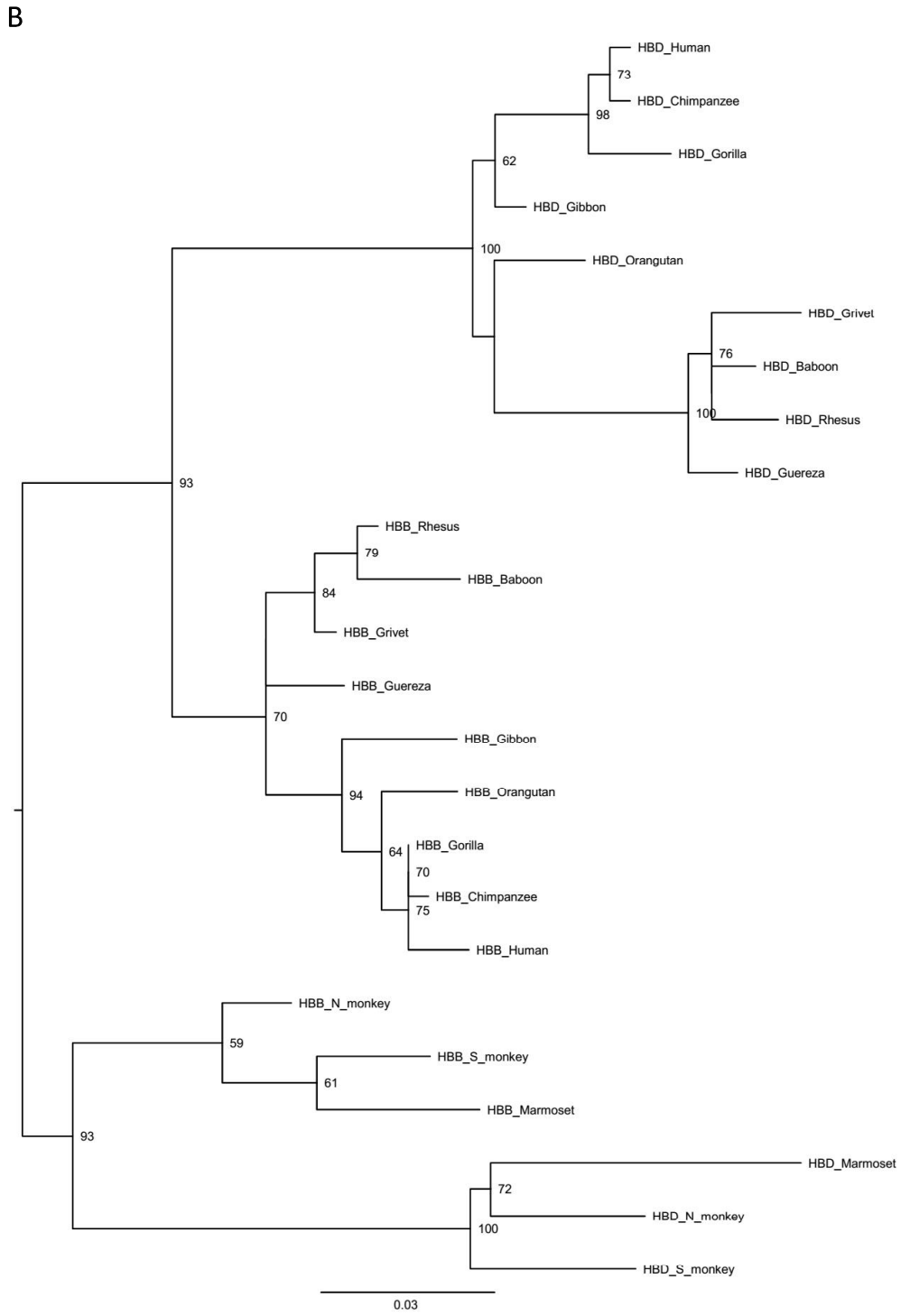
S3 Figure. Sequence alignment of eutherian *HBB* and *HBD* promoters. A) *HBB* and *HBB*-like *HBD* promoters; B) Anthropoid *HBD* promoters and C) *HBD* promoters lacking the TF binding motifs which are conserved in *HBB*-like and *HBD*-like promoters. Conserved binding motifs are indicated in grey boxes. Again, the lemur species were excluded from the analysis.



S4 Figure. Phylograms depicting relationships among adult *β-like* genes in mammals. The phylogenic tree, based on the coding sequence, was constructed using the Goldman–Yang codon model. Branch support values, obtained using the approximate Likelihood Ration Test (aLRT), are given on the internodes.

A


HBB_Human
80
HBB_Gorilla
89
HBB_Chimpanzee
85
HBB_Orangutan
74
HBB_Gibbon
83
HBB_Guereza
91
HBB_Baboon
97
HBB_Rhesus
82
HBB_Grivet
86
HBB_N_monkey
80
HBB_S_monkey
99
HBB_Marmoset
HBD_Marmoset
HBD_S_monkey
78
HBD_N_monkey
HBD_Rhesus
95
HBD_Baboon
97
HBD_Grivet
97
HBD_Guereza
86
98
HBD_Gibbon
HBD_Chimpanzee
92
HBD_Human
86
HBD_Gorilla
HBD_Orangutan
0.02



S5 Figure. Maximum Likelihood phylograms depicting relationships among *β-like* genes of Anthropoids. The phylogeny reconstructions were based on A) the portion of the alignment that contain evidence of the anthropoid gene conversion (nucleotide 367-632) and B) the portion of the alignment between the inferred breakpoint in event 3 (nucleotide 97-385). Bootstrap branch support (1000 replicates) are given on the internodes.

3.3. Research Article:

"DivStat: a user-friendly tool for single nucleotide polymorphism analysis of genomic diversity"

PLOS ONE, 2015

(doi: 10.1371/journal.pone.0119851)

PLOS ONE

RESEARCH ARTICLE

DivStat: A User-Friendly Tool for Single Nucleotide Polymorphism Analysis of Genomic Diversity

Inês Soares[1]*, Ana Moleirinho[1,2], Gonçalo N. P. Oliveira[2,3], António Amorim[1,2]

1 IPATIMUP, Institute of Molecular Pathology and Immunology of the University of Porto, Rua Dr. Roberto Frias s/n, 4200-465, Porto, Portugal, **2** Faculty of Sciences, University of Porto, Rua do Campo Alegre s/n, 4169-007, Porto, Portugal, **3** IFIMUP and IN—Institute of Nanoscience and Nanotechnology, Rua do Campo Alegre, 687, 4169-007, Porto, Portugal

* isoares@ipatimup.pt

OPEN ACCESS

Citation: Soares I, Moleirinho A, Oliveira GNP, Amorim A (2015) DivStat: A User-Friendly Tool for Single Nucleotide Polymorphism Analysis of Genomic Diversity. PLoS ONE 10(3): e0119851. doi:10.1371/journal.pone.0119851

Academic Editor: Swarup Kumar Parida, National Institute of Plant Genome Research (NIPGR), INDIA

Received: June 19, 2014

Accepted: January 18, 2015

Published: March 10, 2015

Data Availability Statement: DivStat software is freely available to download for MACINTOSH, UNIX and WINDOWS systems at the websites https://www.mediafire.com/folder/za5wjlcoc5oi1/DivStat and http://www.portugene.com/DivStat.html, accompanied by tutorials, example files and installation details.

Funding: This work was supported by the Portuguese Foundation for Science and Technology (FCT) fellowship (SFRH/BD/73508/2010) to A. M. IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education and is partly supported by FCT. The funders had no role in study design, data

Abstract

Recent developments have led to an enormous increase of publicly available large genomic data, including complete genomes. The 1000 Genomes Project was a major contributor, releasing the results of sequencing a large number of individual genomes, and allowing for a myriad of large scale studies on human genetic variation. However, the tools currently available are insufficient when the goal concerns some analyses of data sets encompassing more than hundreds of base pairs and when considering haplotype sequences of single nucleotide polymorphisms (SNPs). Here, we present a new and potent tool to deal with large data sets allowing the computation of a variety of summary statistics of population genetic data, increasing the speed of data analysis.

Introduction

The most widely-used software packages, such as DnaSP [1] and Arlequin [2] cannot handle the data formats adopted by massive re-sequencing projects. The development of potent tools to analyze the genetic variation of large scale data stored in the variant call format (VCF) developed by the 1000 Genomes Project that has been adopted by other projects, such as UK10K, dbSNP and the NHLBI Exome Project, became imperative [3,4,5]. Recently a program package was designed to provide a number of methods for working with VCF files (VCFtools): validating, merging, comparing and calculate some basic population genetic statistics [6]. It is a Perl based tool that uses the power of Linux/Unix environments through system calls. We have developed a new and robust algorithm, which runs on DivStat software, which uses the power of Linux/Unix, Macintosh and Windows environments, reducing the learning curve for those users less familiar with the shell commands. The program is implemented with a command line shell and also with a user-friendly graphical interface that facilitates algorithm use. This tool can be applied to either polymorphism data or DNA sequences. Moreover, it can compute sequentially a variety of summary statistics of population genetic data over a "sliding window". After each estimation, the window is slid across the surveyed area and new similar

collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

computations can be done. This type of analysis is used in several population genetic studies, allowing for the inspection of variation patterns along genomic fragments. In this paper we describe the DivStat software and demonstrate its usefulness. Furthermore, we also compare it with other tools and illustrate their capabilities. The algorithm details, tutorial, software, clear and explicit examples that users can use to test both versions of DivStat software and also output files showing the results form hypothetic studies are freely available as supplementary material of this manuscript at https://www.mediafire.com/folder/za5wjlcoc5oi1/DivStat and http://www.portugene.com/DivStat.html.

Material and Methods

We here describe a new tool that allows a simultaneous feed and analysis of large data sets. The developed algorithm was implemented into a program that allows for the computation of different statistics of population genetic data, comprising a number of DNA haplotype sequences encompassing more than a thousand of base pairs. The designed program accepts input files in VCF or fasta format, including both complete DNA sequences or SNP haplotypes. We designed a user-friendly interface in order to facilitate the use by the research community, allowing the upload of a file in the VCF format or a text file with the genetic data in the fasta format. Moreover, a command line version was also developed, allowing the upload of a folder with more than a VCF or text file.

Algorithm

An overview of the software, showing the algorithm details is presented in Fig. 1. The modified waterfall model was the process adopted to develop the DivStat approach and the full step-wise procedure is described in S1 Table.

Program features

The statistics that DivStat calculates can be performed for a single population sample or for multiple population samples. Furthermore, this tool allows for the computation of several statistics within a window with a user definable length. The full list of statistics is givenbelow. First, the users should define a set of parameters, namely, the start and end positions of the segment, the window size and the window increment. When using polymorphism data, the numbering of site positions within the file must be consistent with the numbering used to define the segments. Thus, for instance, defining a window size of n and considering p as its start position, the program calculates the statistics within the window $[p.p+n-1]$. If the window increment is v, it means that the next computations are done after sliding the window of v base pairs, i.e., in the window $[p+v.p+v+n-1]$.

The algorithm starts by assigning the digits 1, 2, 3 and 4 to bases A, C, G and T, respectively (similarly to the methodology adopted in [7]), and 5 to missing data, and then each sequence is converted into a vector according this numerical correspondence. Considering a dataset encompassing N haplotype sequences, each with M sites, a matrix X with M rows and N columns is constructed after the numerical correspondence. Based on X, the program allows for the quick computation of six statistics (for details, please see S1 File):

1. *S*: The number of polymorphic sites that are contained within the window.
2. *Haplotype number*: The number of different haplotypes within the window.

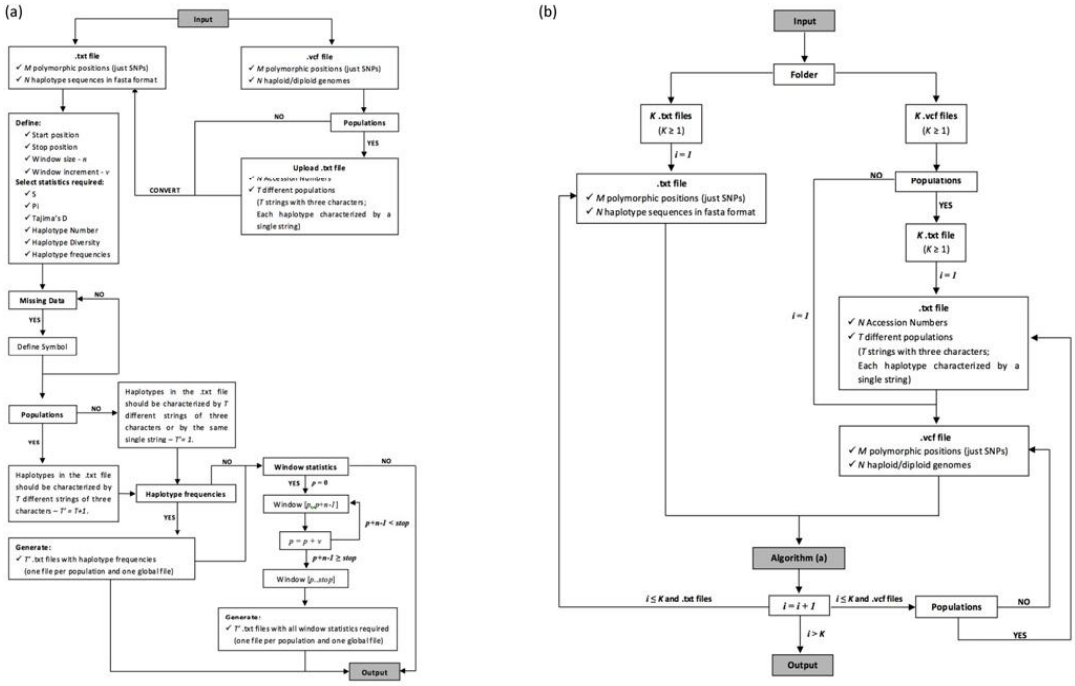


Fig 1. Design of DivStat algorithm: (a) GUI version; (b) cmd version.

doi:10.1371/journal.pone.0119851.g001

3. *Haplotype diversity*: The diversity of each haplotype within the window, which is given by the following as in [8]:

$$Hd = \frac{N}{N-1}(1 - \sum_j Hf[j]^2)$$

where $Hf[j]$ is the frequency of the j^{th} haplotype in the window.

4. π: The nucleotide diversity within the window. This statistic is computed based on the nucleotide diversity of each single polymorphic site i in the window, with k different alleles, usually represented as π_i (the formula for π_i is given in the Supplementary Information). The nucleotide diversity for a window containing S polymorphic positions is then computed according with [9]:

$$\Pi = \sum_{i=1}^{S} \pi_i$$

And, being n the size of the window, the nucleotide diversity per base pairs is calculated as:

$$\pi = \Pi / n$$

5. *Tajima's D*: Tajima's D is a statistic that allows for assessing the evidence or not of selection in the data set. The computation of this statistic is based on the Watterson's θ_w and $var(\pi\text{-}\theta_w)$ (formulas given in the Supplementary Information), being computed by the following formula as in [9]:

$$D = (\Pi - \theta_w)/\sqrt{Var(\pi - \theta_w)}$$

In the case of $var(\pi\text{-}\theta_w)$ is null, the value of D cannot be computed (in this case, in the output file produced by DivStat, the corresponding value of Tajima's D statistic will be represented by "?" symbol).

6. *Haplotype frequencies*: we developed an algorithm that starts by determining all different haplotypes in the dataset and then computes the haplotype frequencies. This calculation is independent of the window size and the increment defined.

Data

Our approach aims to analyze small and large data sets of DNA sequences and haplotypes of polymorphic sites in the text format as well as in the VCF format, easily and quickly converted into the accepted text format before computing the population statistics. In fact, the software was trained, tested and validated by a large collection of data sets that were obtained by combinations of a small or great amount of sequences or haplotypes with a few or more than a thousand of base pairs or polymorphisms.

Software

The algorithm was written in Python, version 2.7.6, and tested on Macbook pro, Mac OS X 10.9 system with an Intel(R) Core i7 CPU @ 2.3 GHz and 8b RAM, Linux—Ubuntu 12.10 and Windows 7–32bit system with an Intel(R) Core(TM)2 Duo CPu, T9300 @ 2.50 GHz and 4Gb of RAM and 64bit system with an Intel(R) Pentium (R) Dual CPU, T3400 @ 2.16 GHz and 4Gb of Ram. The interface, created using a tool prepared to interact with the Python programming language—VisualWx, was designed to run in Macintosh, Unix and Windows 7 systems.

Results

The developed algorithm starts by a pre-analysis of the inputted data in order to verify whether the information was correctly inserted. If some inconsistency is detected, namely files and/or fields outside the standard accepted format, the software aborts and alerts the user to check the data. The program identifies the field that is not filled according to the DivStat standard rules, facilitating the correction of the data and the restarting of the analysis. When the data have the correct format and the fields are filled correctly, DivStat efficiently computes the desired statistics. Examples of the specific file formats accepted in DivStat are available with the setup of the software. Furthermore, a tutorial for the software was added and can be accessed directly on the software or be consulted as supplementary information.

The software was satisfactorily tested with the genomic data of several hundreds of haplotype sequences. To demonstrate the usage and efficacy of DivStat, we provide two examples (S1 Data). The developed algorithm was tested in a training set of haplotype sequences from 1,092 individuals belonging to 14 populations. The data set contains 1707 SNPs spread along 100000 base pairs [10]. It was also applied to a larger data set of 40 DNA haplotype sequences from 20 individuals belonging to the same population, containing 77754 SNPs spread along 3800000 base pairs.

Program performance

The developed algorithm for statistics estimation proved to be computationally very fast, allowing for the simultaneous upload and analysis of large datasets. When performing the analysis of the first training set per populations, considering a window size of 2000 base pairs and an increment of 150 base pairs each time, the computation of all statistics was accomplished in 26 minutes and 25 seconds; when considering all haplotypes as a single group, and assuming the same parameters, the calculation of all statistics was carried out in 13 minutes and 23 seconds. Using the second training set, a window size of 10000 base pairs and an increment of 9000 base pairs was defined and, the computation of all statistics was accomplished in 2 hours, 43 minutes and 44 seconds. (Just the running times for the windows–32bits system were considered here. For other tested platforms, the algorithm showed a similarly high performance, exhibiting running times of the same order of magnitude.) Irrespectively of the type of the data set used, the developed algorithm showed a good performance, being fast and presenting the expected results.

Performance Comparison between DivStat and available tools

The features and performance of DivStat software was compared to other available tools with similar purposes, namely, SLIDER and VCFtools (http://genapps.uchicago.edu/slider/index.html and [6]). A thoroughly and fair comparison of our and these previous tools is unfeasible, since they handle different types and amounts of data. Therefore, we were only able to perform an in silico comparison, based on some features that are shared between Divstat and the other two tools. The capabilities and characteristics of each tool are illustrated in S2 Table and are described in the following subsections. The results obtained from comparison analyses are available in the (S2 Data).

DivStat VS VCFtools. VCFtools only accepts polymorphism data in the VCF format, contrarily to DivStat that can deal with both polymorphism data and DNA sequences, in the VCF or fasta format. The fasta format supported by Divstat is an asset to our approach, since it considerably compresses the size of the data set. For instance, considering the same polymorphism data, summarized in S2 Table, in the VCF format the file size is ~2.4Mb and in the fasta format supported by DivStat it is just ~200Kb. In terms of running time, considering the data set summarized in S2 Table, DivStat takes 27 seconds to generate the output file whereas VCFtools runs each command in 1 second. VCFtools is faster than DivStat, nevertheless, it requires the instruction of a command one by one to compute the desired statistics, whereas DivStat can compute all required statistics at the same time. Furthermore, only 3 statistics, S, Pi and Tajima's D are common to both VCFtools and Divstat, but we can only compare the results obtained over a sliding window, in which the window incrementally advances across the surveyed region, for S and Pi statistics. For tajima´s D computation, VCFtools does not allow the window to advance by adding an increment but only in contiguous non-overlapping windows. Additionally, DivStat is compatible with three operating systems, namely, Macintosh, Unix and Windows, while VCFtools only runs on the Linux/Unix platform.

DivStat VS SLIDER. SLIDER has a very strict maximum dataset size (1MB) and only accepts polymorphism data in file formats that increase considerably the size of the data set. For instance, considering the same polymorphism data, summarized in S2 Table, in the format accepted by SLIDER, the size of the file is ~1Mb and in the fasta format supported by DivStat it is just ~340Kb. Both tools can be configured to compute the statistics over a sliding window, but they could never be compared under the same computational requirements, since SLIDER is a web-based tool that runs on an online server. DivStat is compatible with three environments (Macintosh, Unix and Windows), while SLIDER just requires an internet connection.

Considering the data set summarized in S2 Table, to generate the same output data, DivStat only takes 2 minutes and 52 seconds whereas SLIDER needs 9 minutes and 28 seconds. It is worth noting that the running time of SLIDER could vary depending on the internet connection speed available.

Discussion

Most of the existing programs for computing a variety of summary statistics of population genetic data use DNA sequence. The software here described represents a new tool to efficiently use, not only DNA sequences but also polymorphism data, like those recently released in the VCF format. We have compared the improvement of our approach over other available methods having similar purposes, namely VCFtools and SLIDER. Unfortunately, from a practical perspective, it is quite difficult to perform a comparison between them because they handle different types and amounts of data. Nevertheless, we performed the reasonable comparisons taking into account some shared features. VCFtools only accepts polymorphism data in the VCF format, whereas DivStat can deal with both, polymorphism data and DNA sequences in VCF or fasta format. For the inspection of variation patterns along genomic fragments, DivStat can be configured, so that its window size is defined either in terms of a fixed number of base pairs or a fixed number of segregating sites, whereas VCFtools only allows the first option. In fact, when doing a whole-genome analysis, it is suitable to calculate some statistics, such as nucleotide diversity or Tajima's D, on a fixed number of segregating sites, because it would be hard to compare values across the genome in fixed bp windows, since some windows will have fewer sites while other will be densely populated. However, when genetic variation is used to investigate particular genomic regions, it is more accurate to compute those statistics per base pair, since the existence of big gaps with no diversity is a common feature of the eukaryotic genome landscape. Moreover, VCFtools only runs on a Linux/Unix environment whereas DivStat is compatible with three operating systems: Macintosh, Unix and Windows, and is implemented with a command line shell and also with a user-friendly graphical interface which does not require computational expertise. Concerning SLIDER (http://genapps.uchicago.edu/slider/index.html), it also computes a variety of summary statistics of population genetic data over a "sliding window", which size can be defined by number of base pairs and by number of polymorphic sites. Nonetheless, contrarily to DivStat, aiming to apply the second option in SLIDER, the user cannot input DNA sequences but only polymorphism data. Additionally, SLIDER does not support VCF format, so it is necessary first to convert the data into a Fasta format. Furthermore, it has a very strict dataset size (1MB) and is a web-based application that runs on an online server, requiring thus an internet connection, so the internet speed and also the availability of the server are the main factors influencing the running time of SLIDER. Our approach allows for the simultaneous feed and analyses of large data sets, and also provide a tool that allows users to upload more than a data file simultaneously, making the analysis quicker and more effective. Since in terms of accuracy the three tools are identical, choosing between them depends primarily on the type and size of data in hand.

Supporting Information

S1 Table. DivStat algorithm: (a) GUI algorithm; (b) cmd algorithm.
(DOCX)

S2 Table. Summary comparison between (a) DivStat and VCFtools and (b) DivStat and SLIDER performances.
(DOCX)

S1 File. Supplementary Information.
(DOC)

S2 File. Read Me.
(TXT)

S3 File. Tutorial.
(PDF)

S1 Data. Test Data.
(RAR)

S2 Data. Performance Comparisons.
(RAR)

S1 DivStat. DivStat Windows version.
(RAR)

S2 DivStat. DivStat Linux version.
(RAR)

S3 DivStat. DivStat MacOS-GUI version.
(DMG)

S4 DivStat. DivStat MacOS-cmd version.
(DMG)

Acknowledgments

The authors thank João Alves for his suggestions and the two reviewers for helpful comments that greatly improved the manuscript.

Author Contributions

Conceived and designed the experiments: IS AM GNPO AA. Performed the experiments: IS AM GNPO AA. Analyzed the data: IS AM GNPO AA. Contributed reagents/materials/analysis tools: IS AM GNPO AA. Wrote the paper: IS AM GNPO AA. Designed and developed the software: IS. Revised the software: IS GNPO.

References

1. Rozas J, Sanchez-DelBarrio J, Messeguer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 2003; 19(18): 2496–2497. PMID: 14668244
2. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evol Bioinform Online. 2005; 1: 47–50.
3. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491: 56–65. doi: 10.1038/nature11632 PMID: 23128226
4. Muddyman D, Smee C, Griffin H, Kaye J. Implementing a successful data-management framework: the UK10K managed access model. Genome Med. 2013; 5: 100. doi: 10.1186/gm504 PMID: 24229443
5. Sherry ST, Ward M- H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29: 308–311. PMID: 11125122
6. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–2158. doi: 10.1093/bioinformatics/btr330 PMID: 21653522
7. Soares I, Amorim A, Goios A. A new algorithm for mtDNA sequence clustering. Forensic Sci Int Genetics Sup Series. 2011; 3(1): e315–e316.

8. Nei M and Tajima F. DNA polymorphism detectable by restriction endonucleases. Genetics. 1981; 97:145–163. PMID: 6266912
9. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123(3): 585–95. PMID: 2513255
10. Moleirinho A, Seixas S, Lopes AM, Bento C, Prata MJ and Amorim A. Evolutionary Constraints in the β-Globin Cluster: The Signature of Purifying Selection at the δ-Globin (HBD) Locus and Its Role in Developmental Gene Regulation. Genome Biol Evol. 2013; 5: 559–571. doi: 10.1093/gbe/evt029 PMID: 23431002

Supporting Information

Tables

S1 Table. DivStat algorithm: (a) GUI algorithm; (b) cmd algorithm

(a) GUI algorithm

INPUTS
file: .txt or .vcf file storing the sequence data (N haplotype sequences (HapSeq), each with M sites)
s: start position
e: stop position
n: window size (defined by number of base pairs or number of segregating sites)
v: window increment
MD: missing data symbol (in case of data with missing data)
Stat: set storing the required statistics:
S: number of polymorphic sites
Hn:haplotype number
Hd: haplotype diversity
π: Pi
D: Tajima's D
hf: haplotype frequencies
Pop: population requirements:
Pop ← YES: if the computation should be performed per populations (T populations)
file_pop: .txt file storing the population information in case of .vcf file (in case of .txt file, each haplotype sequence should be characterized by a string of three characters).
Pop ← NO: if the computation should not be performed per populations

```
1: if file is .vcf:
2:    file ← converted into .txt format
3: X ← matrix(M,N)
4: for i in range(1,N):
5:    for j in range(1,M):
6:       if HapSeq[i,j] = A:
7:          X[i,j] ← 1
8:       if HapSeq[i,j] = C:
9:          X[i,j] ← 2
10:      if HapSeq[i,j] = G:
11:         X[i,j] ← 3
12:      if HapSeq[i,j] = T:
13:         X[i,j] ← 4
14:      if HapSeq[i,j] = MD:
15:         X[i,j] ← 5
16: if Pop = YES:
17:   T' ← T+1 (compute the statistics for each population separately and also for the global set of haplotype sequences)
18: if Pop = NO:
19:   T' ← 1 (compute the statistics just for the global set of haplotype sequences)
20: for p in range (1:T'):
21:   X' ← matrix(M,N') (a sub-matrix of X that just considers the haplotype sequences in population p; N' represents the number
                          of haplotype sequences in population p)
22:   if hf in Stat: (Computes the haplotype frequencies (hf) according to equation (6) in the Supplementary Information.)
23:      H ← # different columns of X'
24:      for i in range(1:H):
25:         h[i] ← # columns of X' equal to i^th different column
26:         hf[i] ← h[i]/N'
27:   if ((S or Hn or Hd or π or D) in Stat):
28:      p ← s
29:      while (p+n–1 ≤ e):
30:         Y ← X'[p...p+n-1,N']
31:         if ((S or π or D) in Stat): (Computes S according to (1) in the Supplementary Information.)
32:            S ← # non-conserved rows of Y
```

33: *if ((Hn or Hd) in Stat): (Computes Hn according to (2) in the Supplementary Information.)*
34: *Hn* ← *# different columns of Y*
35: *if (Hd in Stat): (Computes Hd according to equations (3) in the Supplementary Information.)*
36: *for j in range(1:Hn):*
37: $H[j]$ ← *# columns of Y equal to* j^{th} *haplotype.*
38: $Hf[j] \leftarrow H[j]/N'$
39: $Hd \leftarrow \frac{N'}{N'-1}\left(1-\sum_j Hf[j]^2\right)$
40: *if (π or D) in Stat): (Computes π according to equations (4) in the Supplementary Information.)*
41: *for i in range(1:S):*
42: k ← *# different entries in the non-conserved row i of Y.*
43: *for q in range(1:k):*
44: x_{iq} ← *# columns of Y with an entry equal to the* q^{th} *different entry at row i*
45: $\pi_i \leftarrow 1-\left(\sum_{q=1}^{k}\left(x_{iq}^2-x_{iq}\right)\Big/\left(\left(\sum_{q=1}^{k}x_{iq}\right)^2-\sum_{q=1}^{k}x_{iq}\right)\right)$
46: $\Pi \leftarrow \sum_{i=1}^{S}\pi_i$
47: $\pi \leftarrow \Pi/n$
48: *if D: (Computes D according to equations (5) in the Supplementary Information.)*
49: $a_1 \leftarrow \sum_{j=1}^{N-1}1/j$
50: $a_2 \leftarrow \sum_{j=1}^{N-1}1/j^2$
51: $b_1 \leftarrow (N'+1)/(3(N'-1))$
52: $b_2 \leftarrow (2(N'^2+N'+3))/(9N'(N'-1))$
53: $c_1 \leftarrow b_1-(1/a_1)$
54: $c_2 \leftarrow b_2-((N'+2)/(a_1N'))+a_2/a_1^2$
55: $e_1 \leftarrow c_1/a_1$
56: $e_2 \leftarrow c_2/(a_1^2+a_2)$
57: $\theta_w \leftarrow S/a_1$
58: $Var(\pi-\theta_w) \leftarrow e_1S+e_2S(S-1)$
59: $D \leftarrow (\Pi-\theta_w)/\sqrt{Var(\pi-\theta_w)}$
60: $p \leftarrow p+v$
61: *if (p+n–1>e and p<e):*
62: $Y \leftarrow X'[p\ldots e,N']$
63: *Repeat lines 31 – 59*

OUTPUT
.txt file (s) storing the required statistics

(b) cmd algorithm

INPUTS

folder: folder with K≥1 .txt or .vcf files, which stores the sequence data (each file f ($f \in \{1,2,\ldots,K\}$) contains N haplotype sequences (HapSeq), each with M sites)

$s_1 \ldots s_K$: start position in each file f

$e_1 \ldots e_K$: stop position in each file f

n: window size (defined by number of base pairs or number of segregating sites)

v: window increment

MD: missing data symbol (in case of data with missing data)

Stat: set storing the required statistics:

- *S: number of polymorphic sites*
- *Hn: Haplotype number*
- *Hd: Haplotype diversity*
- *π: Pi*
- *D: Tajima's D*
- *hf: haplotype frequencies*

Pop: population requirements:

- *Pop ← YES: if the computation should be performed per populations*
  - *file_pop$_1$... file_pop$_K$: .txt files storing the population information of each file f, in case of .vcf file (in case of .txt file, each haplotype sequence should be characterized by a string of three characters). (T populations in the file f)*
- *Pop ← NO: if the computation should not be performed per populations*

1:for f in folder:
2: file ← f
3: s ← s_f
4: e ← e_f
5: file_pop ← file_pop$_f$
6: follow the lines 1 – 63 of GUI algorithm (a)

OUTPUT

.txt files storing the required statistics for the K inputted files.

S2 Table. Summary comparison between (a) DivStat and VCFtools and (b) DivStat and SLIDER performances.

	Softwares			
	(a)		(b)	
Features	**DivStat**	**VCFtools**	**DivStat**	**SLIDER**
Data Size	208Kb	2.405Mb	342Kb	999Kb
Running Time	27sec	1sec/step*	2min52sec	9min28sec
Maximum Data Size	-	-	-	1Mb
Allowed Statistics:				
S	x	x	x	x
Pi	x	x	x	x
Tajima's D	x	x	x	x
Haplotype Number	x		x	x
Haplotye Diversity	x		x	x
Compatible Operative Systems:				
Windows	x		x	x+
Linux	x	x	x	x+
Max OS	x		x	x+

* The software computes only one statistic each time, and each step runs in 1sec.
+ The software runs in an online server.

S1 File: Supplementary Information:

1 METHODS

1.1 Program features

The statistics that DivStat calculates can be performed for a single population sample or for multiple population samples. Furthermore, this tool allows for the computation of several statistics within a window with a definable length. The full list of statistics is given below. First, the users should define a set of parameters, namely, the start and end positions of the segment, the window size and the window increment. When using polymorphism data, the numbering of site positions within the file must be consistent with the numbering used to define the segments. Thus, for instance, defining a window size of *n* base pairs and considering *p* as its start position, the program calculates the statistics within the window [*p..p+n-1*]. If the window increment is *v*, it means that the next computations are done after sliding the window of *v* base pairs, ie, in the window [*p+v..p+v +n-1*].
The algorithm starts by assigning the digits 1, 2, 3 and 4 to bases A, C, G and T, respectively (similarly to the methodology adopted in [1]), and 5 to missing data symbol (in the case it happens), and then each sequence is converted into a vector according this numerical correspondence. Considering a dataset encompassing *N* haplotype sequences, each with *M* sites, a matrix *X* with *M* rows and *N* columns is constructed after the numerical correspondence. Based on *X*, the program allows for the quick computation of six statistics. Considering, for instance, the sub-matrix *Y* of *X* in the window [*p..p+n-1*]:

$$X = \begin{bmatrix} a_{11} \cdots a_{1j} \cdots a_{1N} \\ a_{21} \cdots a_{2j} \cdots a_{2N} \\ \vdots \\ a_{i1} \cdots a_{ij} \cdots a_{iN} \\ \vdots \\ a_{M1} \cdots a_{Mj} \cdots a_{MN} \end{bmatrix} \qquad Y = \begin{bmatrix} a_{p1} & \cdots & a_{pj} & \cdots & a_{pN} \\ & & \vdots & & \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{iN} \\ & & \vdots & & \\ a_{(p+n-1)1} & \cdots & a_{(p+n-1)j} & \cdots & a_{(p+n-1)N} \end{bmatrix}$$

with $a_{ij} \in \{1,2,3,4\}$, the different statistics are determined as explained below:

(1) S: *S* is the number of polymorphic sites that are contained within the window. The computation of *S* is resumed to the computation of non-conserved rows of *Y* (which corresponds to the rows of *X* falling down within the window [*p..p+n-1*] and having more than one entry).

(2) Haplotype number: Haplotype number is the number of different haplotypes within the window. The algorithm developed computes this statistic by determining the number of different columns of *Y*.

(3) Haplotype diversity: Haplotype diversity is the diversity of each haplotype within the window. Considering the different haplotypes computed in the previous item, the haplotype frequency is given by the following:

$$Hf[j] = H[j]/N$$

where *H*[*j*] and *N* are the number of occurrences of the j^{th} haplotype in the window (ie, the number of columns of *Y* equals to haplotype *j*) and the total number of sequences in the data set, respectively. After this, the haplotype diversity is computed as in [2]:

$$Hd = \frac{N}{N-1}\left(1 - \sum_j Hf[j]^2\right)$$

(4) π: π is the nucleotide diversity within the window. This statistic is computed based on the nucleotide diversity of each single polymorphic site *i* in the window, with *k* different alleles, usually represented as π_i and given by the following:

$$\pi_i = 1 - \left(\sum_{q=1}^{k} \left(x_{iq}^2 - x_{iq}\right) \Big/ \left(\left(\sum_{q=1}^{k} x_{iq} \right)^2 - \sum_{q=1}^{k} x_{iq} \right) \right)$$

where, x_{iq} is the number of haplotypes within the window with the allele a_q at position *i*, ie, the number of columns of *Y* with a_{qj}. The nucleotide diversity for a window containing *S* polymorphic positions is computed according with [3]:

$$\Pi = \sum_{i=1}^{S} \pi_i$$

And, being *n* the size of the window, the nucleotide diversity per base pairs is calculated as the following:

$$\pi = \Pi / n$$

(5) Tajima's D: Tajima's D is a statistic that allows for assessing the evidence or not of selection in the data set. The computation of this statistic is based on the Watterson's θ_W and $Var(\pi - \theta_W)$, according to [3], which are given by:

$$\theta_w = S/a_1 \text{ and } Var(\pi - \theta_w) = e_1 S + e_2 S(S-1)$$

where, $Var(x)$ means the variation of *x*, $e_1 = c_1/a_1$, $c_1 = b_1 - (1/a_1)$, $b_1 = (N+1)/(3(N-1))$, $a_1 = \sum_{j=1}^{N-1} 1/j$, $a_2 = \sum_{j=1}^{N-1} 1/j^2$, $e_2 = c_2/(a_1^2 + a_2)$, $c_2 = b_2 - ((N+2)/(a_1 N)) + a_2/a_1^2$ and $b_2 = (2(N^2+N+3))/(9N(N-1))$.
The Tajima's D statistic is then computed by the following:

$$D = (\Pi - \theta_w) / \sqrt{Var(\pi - \theta_w)}$$

(6) Haplotype Frequencies: We developed an algorithm that starts by determining all different haplotypes in the dataset and then computes the haplotype frequencies according to:

$$hf\left[i\right] = h\left[i\right] / N$$

for the i^{th} haplotype, being $h[i]$ and N the number of occurrences of haplotype i and the total number of haplotype sequences in the dataset, respectively. This calculation is independent of the window size and the increment defined.

REFERENCES

1. Soares I, Amorim A, Goios A (2011) A new algorithm for mtDNA sequence clustering. Forensic Science International: Genetics Supplement Series 3 (1), e315-e316.

2. Nei M and Tajima F (1981) DNA polymorphism detectable by restriction endonucleases. Genetics 97:145-163.

3. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123 (3): 585–95.

S2 File: Read Me

DivStat Read Me
Version 2.0
December 2014

==========
CONTENTS:
==========

1.FEATURES

DivStat is a software designed for analysis of large sets of population genetic data stored in VCF (variant call format) files. It was developed in a graphical user interface version (GUI), in order to make easy the use by research community, and in a command line version (cmd). In GUI, the user can upload a VCF file or a text file with the genetic data, while the cmd version allows the upload of a folder with more than one VCF or text file.
The user should start by defining a set of parameters, namely the type of file (VCF or text - in the first case the number of polymorphic positions and the ploidy of the genome also should be indicated), the start and stop positions of inputted haplotype sequences in the complete genome, the window size (defined by number of base pairs or number of segregating sites) and the window increment.
The program needs to know if the data has or not missing data and, in affirmative case, the symbol used. Furthermore, the user should indicate whether the calculations must be done per population or globally. In the first case, and just if using VCF files, the corresponding populations should be inputted in separated files.
The program allows for the computation of 5 different statistics:

* number of polymorphic sites – S;
* Haplotype Number;
* Haplotype diversity;
* π;
* Tajima's D;
* haplotype frequencies (independent of the window size and the increment defined).

The output obtained corresponds to a text file containing all window statistic computations (the 5 first statistics indicated above) and/or a text file with the haplotype frequencies.

2.INSTALATION
---

*
To install DivStat GUI version on windows system, run the executable file named setupDivStat_GUI_win.exe on windows 32 or 64 bits, and follow the instructions on the screen.

*
To install DivStat cmd version on Windows system, run the executable file named setupDivStat_cmd_win.exe on windows 32 or 64 bits, and follow the instructions on the screen. After installation, to use this version, the user should first fill the parameters on file "call.py" and then do a double click over the application DivStat_cmd or indicate the directory where it is on the command line and then write DivStat_cmd. The program should start running automatically.

* For linux or mac OS GUI version, just do a double click over the application DivStat_GUI or indicate the directory where it is on the command line and then write "./DivStat_GUI". The software should open automatically.

* For linux cmd version, first, the user should fill the parameters on file "call.py" and then do a double click over the application DivStat_cmd or indicate the directory where it is on the command line and then write "./DivStat_cmd". The program should start running automatically.

* For mac OS cmd version, first, the user should fill the parameters on file "call.py". Then, open the command line window and go to the directory where the application DivStat_cmd is. Then, write "./DivStat_cmd.app/Contents/MacOS/DivStat_cmd" on the command line. The program should start running automatically.

3.MINIMUM REQUIREMENTS
---

* Windows 7
* Linux - Ubuntu 12
* Mac OS X

4.CONTACT INFORMATION
---

* Inês Soares
email address: isoares@ipatimup.pt

* Ana Moleirinho
email address: amoleirinho@ipatimup.pt

* Gonçalo N. P. Oliveira
email address: goliveira@fc.up.pt

* António Amorim
email address: aamorim@ipatimup.pt

Tutorial

DivStat is a new software developed in order to help users to analyze great sets of population genetic data stored in VCF (variant call format) files. It allows the computation of six statistics, namely, the number of polymorphic sites – S, Haplotype Number, Haplotype diversity, π, Tajima's D, and haplotype frequencies. They can be computed for the complete DNA fragment or over a "sliding window". First, the users should define a set of parameters, namely, the start and end positions of the segment, the window size (in base pairs or segregating sites) and the window increment in base pairs. When using polymorphism data, the numbering of site positions within the inputted file must be consistent with the numbering used to define the segments.

A user-friendly interface was developed in order to facilitate the use by the research community. The graphical interface allows the upload of a VCF file or a text file with the genetic data in the fasta format. Moreover, a command line version was developed, allowing the upload of a folder with more than a VCF or text file.

1. *Graphical User Interface version – GUI:*

   When the user opens the DivStat software GUI version, the following window appears on the screen:

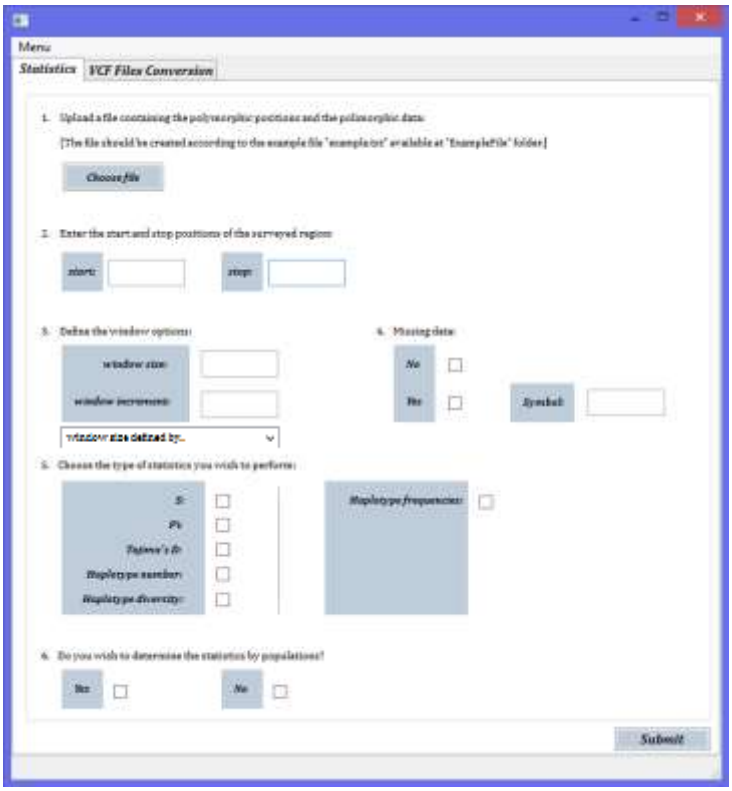


Figure 1. GUI of DivStat software.

First (1st point), the user should upload a file containing the polymorphism data, which could be a VCF or a text file containing the SNPs and the corresponding position number in the complete genome.

1.1. Uploading a text file

If the user uploads a text file containing the SNPs and the corresponding position number in the complete genome, it should be similar to the following examples:

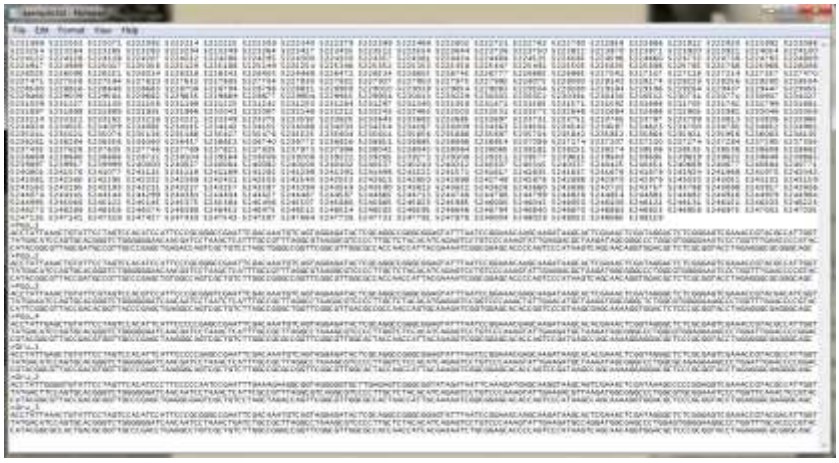


Figure 2. Example of an accepted input text file in the fasta format.

Figure 3. Example of an accepted input text file, in the fasta format, with missing data (represented by symbol "-").

Note that, the position numbers should be written at the first line of the document, the following ones corresponding to the SNP sequences.

On the 2nd and 3rd points, the user should define a set of parameters, namely, the start and end positions of the haplotype sequences, the window size, defined by number of base pairs or segregating sites, and the window increment. Defining, for example, a window size of *n* base pairs and considering *p* as its start position, the program computes the chosen statistics within the window [*p..p+n-1*], working just with the SNP positions that fall within this interval. If the window increment is *v*, it means that the next computations are computed after sliding the window of *v* base pairs, i.e., in the window [*p+v..p+v +n-1*].

On the 4th point, the user should indicate whether the data has or not missing data. In the affirmative case, the user should indicate the symbol used.

Finally (5th point), the user should select or deselect the statistics to be performed. Note that, the haplotype frequency computation does not take into consideration the window parameters defined.

The calculations can be done per populations or globally (6th point). The global estimation is computed in both cases, but population specific outputs are only obtained if indicated in this point. To obtain results by population, the first three characters of each line started by ">" should identify the population. For example, on the text file of Figure 2. there are two different populations, "Pop" and "Gru"; on the text file of Figure 3., all eight sequences belongs to the same population, "CEU".

On the following images, it is possible to see two different examples of fields filling:

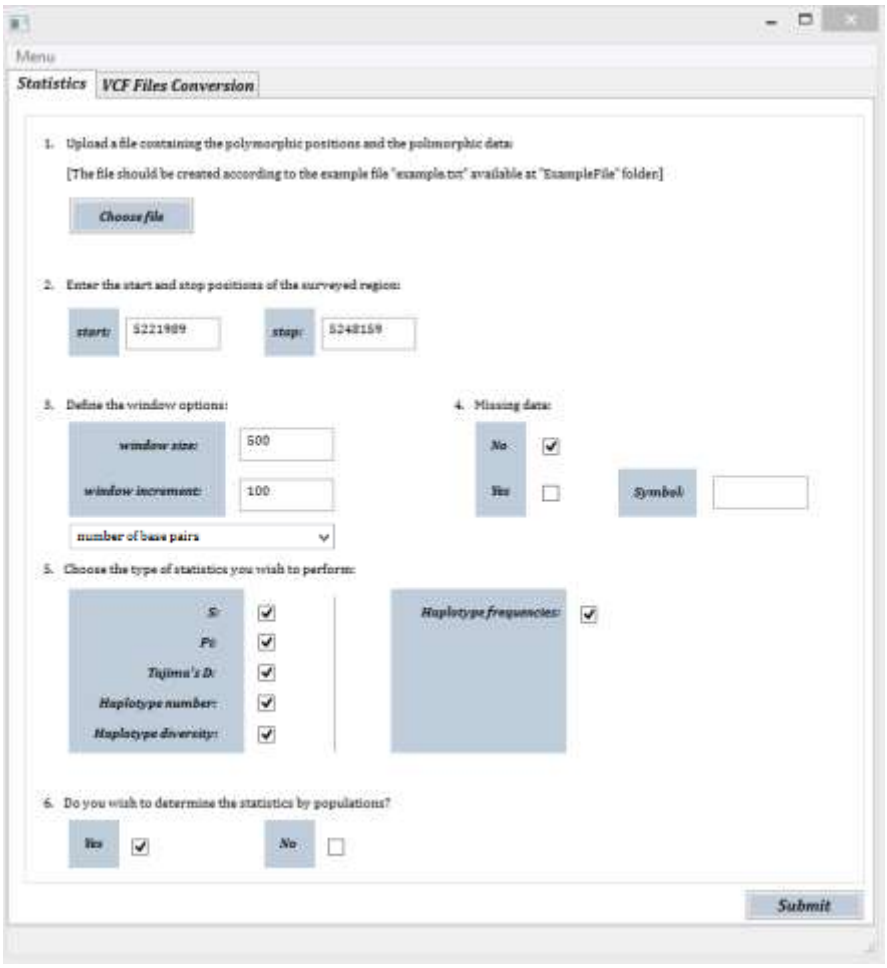


Figure 4. Example of possible preferences using the file of Figure 2.

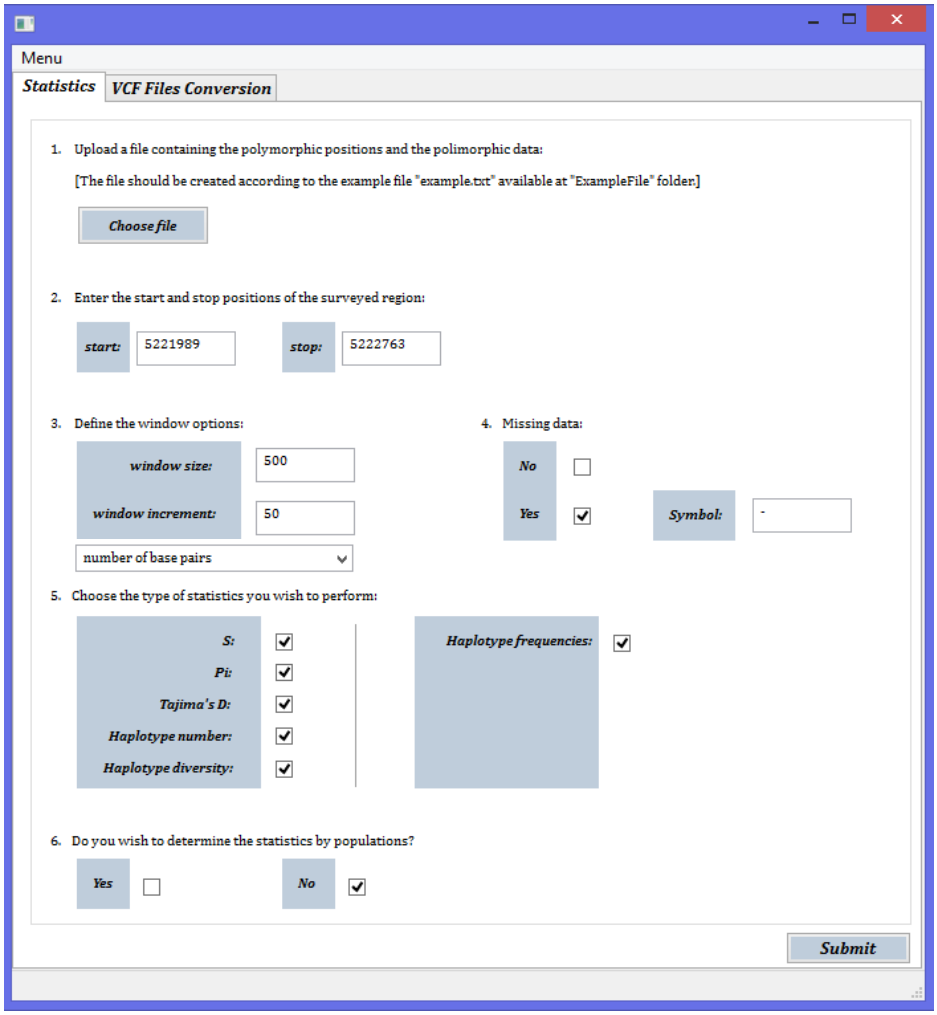


Figure 5. Example of possible preferences using the file of Figure 3.

On the first case (Figures 2. and 4.), the file has not missing data and the required statistics are calculated per population. On the second one (Figures 3. and 5.), the file has missing data, which is identified by symbol "-", and the required statistics are not computed per population. In both cases, the window is defined by number of base pairs.

The haplotype frequencies will be saved on independent files, while all window statistics will be saved on the same file. Furthermore, the output comprises a file per population and a global file (comprising the statistics computed for all sequences considered as a global group). Examples of the output can be seen on the following images:

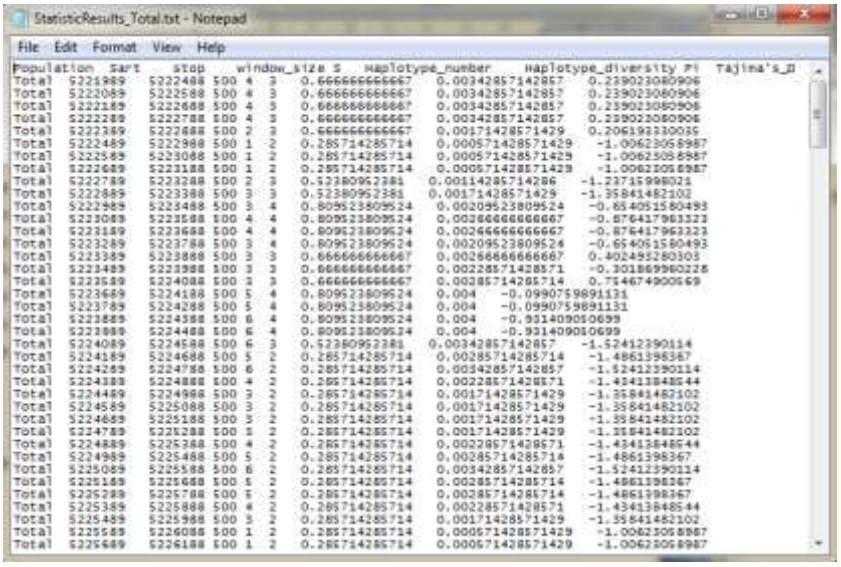


Figure 6. Example of an output (window statistic computation – S, Haplotype Number, Haplotype diversity; π and Tajima's D – computed globally) using the file of Figure 2. and fields filed according Figure 4.

Figure 7. Example of an output for haplotype frequencies using the file of Figure 2. and fields filed according Figure 4.

The shown example files correspond to statistics computed globally. The statistics calculated for specific populations produce similar output files.

1.2. Uploading a VCF file

If the user uploads a VCF file, the following window appears on the screen:

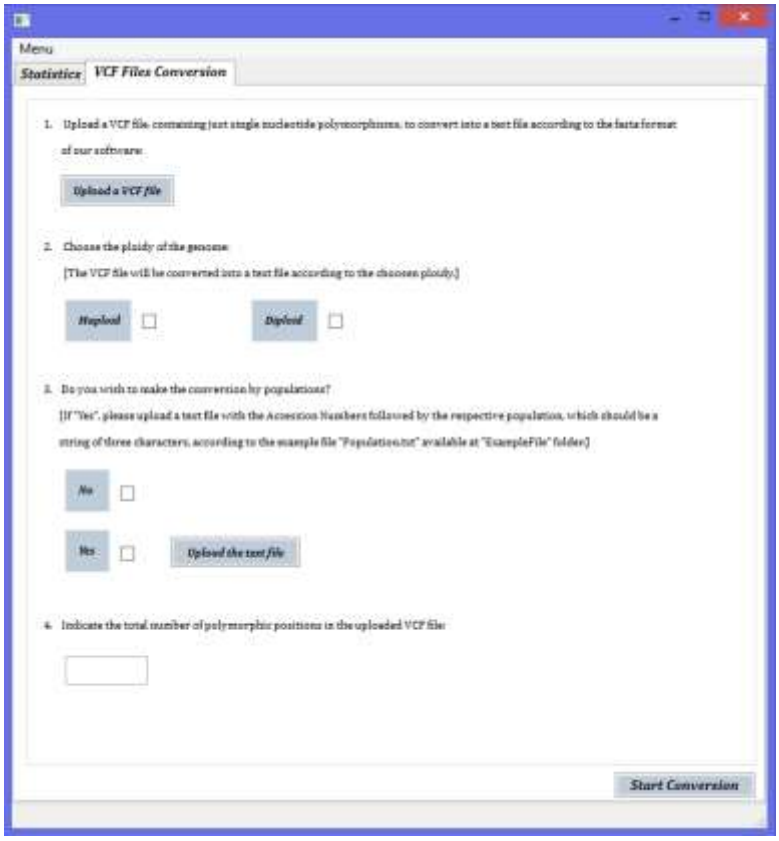


Figure 8. "VCF Files Conversion" tab from the GUI of DivStat software.

This tab is triggered when the user choose to upload a VCF file in the tab "Statistcs" to compute the statistics mentioned in the previous section. Nevertheless, it can be used whenever the user needs to convert a VCF file into a text file.

On the 2nd point of this tab, the user should identify the ploidy of the data, in order to enable a good reading and conversion of the VCF file into the text file.

On the 3rd point, the user should indicate whether the information about the population should be considered or not. In the affirmative case, the user should upload a text file with the information on the populations. More precisely, the file should contain the identification of the individual samples followed by the corresponding population indicated by a string of three characters, according to the following example:

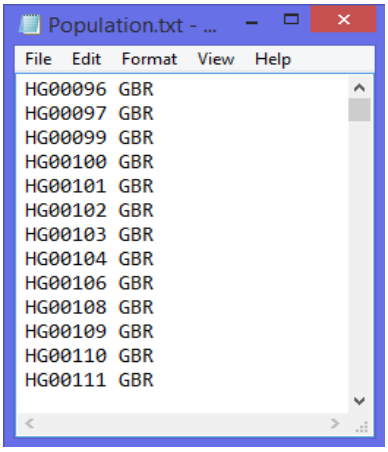


Figure 9. Example of an iutput text file with the information about the population (a string with three characters) of each genome (identified by its accession number).

Otherwise, all sequences are considered as belonging to the same population.

On the last point, the user should indicate the total number of polymorphic positions that are stored in the inputted file.

The output of this operation will be a text file with the data in the fasta format , which is similar to those shown in the figures 2. and 3. of the previous section (without and with missing data, respectively).

After file conversion, the user can proceed to the computation of the statistics.

2. *Command Line version – cmd:*

If the user needs to analyze more than one file, the cmd version is a most suitable option. In this case, the user should create a folder containing all text or VCF files to be analyzed, for instance "folder", and then open the file "call.py" to indicate the parameters to perform the analysis of all files. The parameters to be defined are:

- *type_file*: the type/extension of the inputted files;
- *positions_VCFs*: the number of polymorphic positions in each inputted VCF file. This parameter is just required in case of define type_file as ".vcf";
- *ploidy*: The ploidy of the data. The possibilities are "Haploid" and "Diploid";
- *starts*: start position in the complete genome of the inputted haplotype sequences in each file;
- *stops*: end position in the complete genome of the inputted haplotype sequences in each file;
- *window*: window size ;
- *increment:* window increment;

- *window_def:* to state whether the window is defined by "number of base pairs" or "number of segregating sites";
- *statistics:* dictionary with required statistics, which are marked as "YES" (those marked as "" are not computed);
- *input_way*: path to the folder with the text or VCF files to be analyzed by the software;
- *general_name*: general name of the text or VCF files on the inputted folder. All files on the folder should have the same prefix, which should be followed by consecutive numbers, for example, "example_1.txt", "example_2.txt", "example_3.txt", etc. In this case, the general name is "example_";
- *output_way*: path to the folder where output files should be saved;
- *Pop*: "YES" to compute the statistics by populations and "No" otherwise;
- *general_name_pop_file*: general name of the text files on the inputted folder with the populations information. All files on the folder should have the same prefix, which should be followed by consecutive numbers, for example, "Population_1.txt", "Population _2.txt", "Population _3.txt", etc. In this case, the general name is "Population _". Note that, this parameter is just required in the case of defining the parameters type_file as ".vcf" and Pop as "YES";
- *MD*: "YES" if the files have missing data and "NO" otherwise;
- *symbol:* symbol used on the files for missing data (just needed when MD="YES");
- *num_runs*: number of FASTA files on the inputted folder that should be analyzed. The software runs one time for each file.

The python file "call.py" is similar to the following:

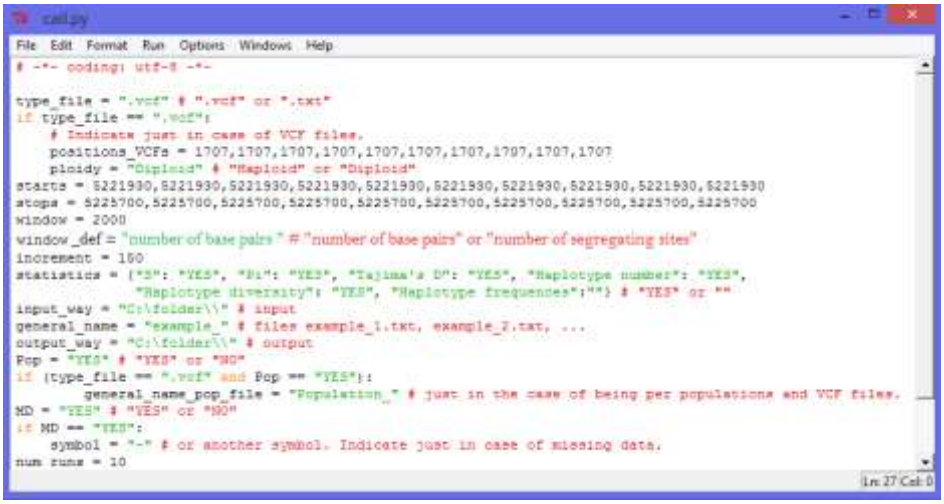


Figure 10. Python file named "call.py" in which the user should define the parameters to perform the analysis. Here, the user has VCF files with 1707

polymorphic positions and diploid genomes. The user defined the start and end positions of the inputted haplotype sequences of all files as being 5221930 and 5225700, respectively; the window size (defined by number of base pairs) and the window increment being 2000 and 150, respectively; the statistics required are S, Haplotype Number, Haplotype diversity, π and Tajima's D – window statistics; the files to be analyzed are in the folder named "folder" and each file has a name with prefix "example_"; the output should be in the same folder "folder"; the statistics should be computed per populations, being the information stored in files where the name has the prefix "Population_", and the files have missing data indicated by "-". Note that, the number of files on the inputted folder "folder" is 10, thus, DivStat should run 10 times with the defined parameters, one for each file in the folder.

CHAPTER 4. DISCUSSION

The assemblage of hemoglobins present in the erythrocytes of adult humans includes HbA_2, a hemoglobin without a recognizable physiological function because it is normally less than 3% of the total Hb (Steinberg and Adams 1991). Intriguingly, however, *HBD* shows a level of sequence conservation typical of genes under strong evolutionary constraints. This inconsistency represents the main question of this thesis, and therefore we have attempted to solve this conservation paradox, using population genetics and comparative genomics tools to gain insight into the relevance of *δ-globin* gene conservation for developmental and physiological processes in some placental mammals.

First, we aimed to accurately estimate diversity levels at the two adult *β-like* globin genes in human populations (results in section 3.1). Even though the *β-globin* cluster is among the most extensively studied regions in the human genome, the genetic diversity estimates for the *HBD* and *HBB* genes across human populations, available at the time this study was initiated, were likely biased because the analyses performed for these two genes were oriented for diagnostic purposes and often based on a set of pre-ascertained SNPs (Lacerra, et al. 2008; Liu, et al. 2009; Morgado, et al. 2007; Phylipsen, et al. 2011). A better characterization of the diversity patterns at the *HBD* and *HBB* genes in human populations was therefore needed to fully understand the evolutionary forces acting on both genes. Consequently, an unbiased characterization of the genetic diversity was performed in a large dataset that comprised individuals from multiple present-day populations. Furthermore, we also assessed the diversity levels in nonhuman primates using chimpanzee sequence data. We have found that purifying selection shaped the evolution of *HBD*, and to some extent *HBBP1*, suggesting that these two genomic regions are evolving under the same selective constraints across different primate species for at least 5 Myr. Since both genes are either weakly expressed or not transcribed at all, the constrained evolution is unlikely to be related to protein function. Therefore, we asked whether *HBD* and *HBBP1* lie within crucial regions for the regulation of gene transcription, in which selective pressures would be acting to maintain the nucleotide sequence. Interestingly, a segment comprising both *HBD* and *HBBP1* was found to interact with different regions upstream *HBE* which overlap the LCR. Such findings fit the previously formulated hypothesis that the *HBD* and *HBBP1* might have a regulatory role in the Hb fetal to adult switch, unique to Anthropoids. To date, however, this possibility has never been explicitly tested, and (to our knowledge) no one has ever established a link between

the recent advances in our understanding of the mechanisms coordinating *β-globin* gene expression and the decreased diversity observed at the *HBD* gene, which represents the most innovative aspect of our approach. To ascertain if the signal of purifying selection was also present in other primate species, particularly in Anthropoids, we intended to broaden the analyses of within-species genetic diversity to nonhuman primates, but the limited data available on genetic polymorphisms and the virtual lack of genetic maps for these species, except the chimpanzee data that we have included in our previous study, have hampered this strategy. Our tenet was that the study of *HBD* evolutionary trajectory across placental mammals with distinct repertoires of *β-like* genes and corresponding expression programs should provide clues into the evolution and putative functional divergence of the *δ-globin* gene. Despite the large number of studies on the history of gene duplication and evolution in the mammalian *β-globin* gene cluster, the evolutionary history of the eutherian *HBD* is not yet fully understood. This is mainly due to recurrent gene conversion and unequal recombination with *HBB*, which hamper the assignment of orthologous relationships among *HBD* and *HBB* genes, yielding conflicting explanations regarding *HBD* origin. Moreover, previous comparative genomic and evolutionary analyses of *β-globin* gene family did not include enough primate representatives to provide a clear picture of *HBD* evolution in these lineages. All this prompted us to re-examine the *HBD* evolutionary trajectory across eutherian mammals, focusing at Anthropoid primates, to clarify if the pattern of conservation at *HBD* gene extended to other primate species and to obtain evidence on whether a selective constraint for the maintenance of the fetal-to-adult Hb switch was modulating the evolution of *HBD* in primates. The results of our analyses (results in section 3.2) are consistent with a duplication event of *HBB* and *HBD* genes after the marsupial/eutherian split, and support the widely held notion that gene conversion drives the evolutionary history of *HBD* in eutherians. Our analyses revealed also that anthropoids are an exception to the general pattern of concerted evolution in placental mammals, and further documented that the *δ-globin* gene is highly conserved in Anthropoids. These results supported our previous findings of purifying selection reducing *HBD* genetic diversity in human populations and in chimpanzees, and indicate that sequence change at the *δ-globin* gene has been under strong selective constraints over 65 Myr of primate evolution. We further showed that not only sequence conservation but also the mode of evolution of the *δ-globin* gene in higher primates are strictly associated with the fetal/adult β-cluster developmental switch. Altogether, the results clearly indicate

that in some primate lineages *HBD* plays an essential and nonredundant role, unlikely associated with oxygen transport but possibly related to the ontogenic regulation of Hb synthesis. While our *in silico* approach has enabled functional inferences, experimental proof is still lacking to fully demonstrate the hypothesis that *HBD* and *HBBP1* have a regulatory role. Ideally, our study should be complemented with functional assays, but experimental design to disclose the putative regulatory role of *HBD* within the *β-globin* cluster remains a challenging task as the mechanism controlling *β-globin* gene expression is not completely understood. Another limitation relies on the choice of the model system. Up to now, the temporal switch from fetal to adult hemoglobin has been mostly studied using transgenic mice harboring the human β-globin locus, which proved to be inappropriate. The main reason is that the *γ-globin* gene is targeted largely as a mouse embryonic globin gene in the context of the mouse erythroid trans-acting environment since this species lacks the biological support that accompanies the developmental activation of the *β-globin* genes, which is present in humans and other primates (Sankaran, et al. 2009). Nevertheless, evidence favoring our hypothesis emerges from naturally occurring deletions in the human *β-globin* cluster that disturb the fetal to adult switch. A recent study of these unusual deletions has uncovered a 3.5-kb intergenic region near the 5' end of *HBD* necessary for this regulation (Sankaran, et al. 2011). Before, chromatin interactions in the *β-globin* gene cluster, determined by 5C, had also revealed strong interactions between the LCR and a region containing both *HBD* and *HBBP1* genes, suggesting that they might be involved in chromatin looping in the human *β-globin* cluster (Dostie, et al. 2006).

In conclusion, through a comprehensive assessment of *HBD* and *HBBP1* genetic diversity in human populations, this study convincingly showed that both genomic regions are under strong selective pressure in humans. Moreover, the evolutionary studies performed here contributed not only to a better understanding of the *HBD* evolutionary trajectory in primates, but also to tackle the *δ-globin* gene conservation inconsistency and to add support to the hypothesis that *HBD* has a role in the fetal-to-adult switch unique of Anthropoid species. Since the available *in vitro* models do not mimic the developmental *milieu* leading to mature human erythrocytes, indirect approaches as we advance here are still required to circumvent the existing technological and ethical limitations, in order to obtain a better comprehension of the structural and functional evolution of coding and non-

coding regions contained in the *β-globin* cluster, which ultimately will help elucidate the biological role of *HBD* and *HBBP1*. The handling of a large amount of data, as required for our analyses, prompted us to develop of a new informatic tool useful for general population genetic studies that is now available to the scientific community.

CHAPTER 5. REFERENCES

Akinsheye I, Alsultan A, Solovieff N, Ngo D, Baldwin CT, Sebastiani P, Chui DHK, Steinberg MH. 2011. Fetal hemoglobin in sickle cell anemia.

Altshuler, RM D, GR A, DR B, A C, AG C, P D, EE E, P F 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65. doi: 10.1038/nature11632

Asano H, Li XS, Stamatoyannopoulos G 1999. FKLF, a novel Kruppel-like factor that activates human embryonic and fetal beta-like globin genes. Mol Cell Biol 19: 3571-3579.

Bank A, Mears JG, Ramirez F 1980. Disorders of human hemoglobin. Science 207: 486-493.

Bank A, O'Neill D, Lopez R, Pulte D, Ward M, Mantha S, Richardson C 2005. Role of Intergenic Human γ-δ-Globin Sequences in Human Hemoglobin Switching and Reactivation of Fetal Hemoglobin in Adult Erythroid Cells. Annals of the New York Academy of Sciences 1054: 48-54. doi: 10.1196/annals.1345.057

Bauer DE, Kamran SC, Orkin SH. 2012. Reawakening fetal hemoglobin: prospects for new therapies for the β-globin disorders.

Bauer DE, Orkin SH 2011. Update on fetal hemoglobin gene regulation in hemoglobinopathies. Current opinion in pediatrics 23: 1-8. doi: 10.1097/MOP.0b013e3283420fd0

Bender MA, Bulger M, Close J, Groudine M 2000. β-globin Gene Switching and DNase I Sensitivity of the Endogenous β-globin Locus in Mice Do Not Require the Locus Control Region. Mol Cell 5: 387-393.

Bouva M, Harteveld C, van Delft P, Giordano P. 2006. Known and new delta globin gene mutations and their diagnostic significance.

Boyer S, Crosby E, Noyes A, Fuller G, Leslie S, Donaldson L, Vrablik G, Schaefer E, Jr., Thurmon T 1971. Primate hemoglobins: Some sequences and some proposals concerning the character of evolution and mutation. Biochemical Genetics 5: 405-448. doi: 10.1007/BF00487132

Bulger M, Groudine M 1999. Looping versus linking: toward a model for long-distance gene activation. Genes Dev 13: 2465-2477.

Calzolari R, McMorrow T, Yannoutsos N, Langeveld A, Grosveld F 1999. Deletion of a region that is a candidate for the difference between the deletion forms of hereditary persistence of fetal hemoglobin and deltabeta-thalassemia affects beta- but not gamma-globin gene expression. The EMBO Journal 18: 949-958. doi: 10.1093/emboj/18.4.949

Carter D, Chakalova L, Osborne CS, Dai YF, Fraser P 2002. Long-range chromatin regulatory interactions in vivo. Nat Genet 32: 623-626. doi: 10.1038/ng1051

Chakalova L, Carter D, Debrand E, Goyenechea B, Horton A, Miles J, Osborne C, Fraser P. 2005. Developmental Regulation of the β-Globin Gene Locus. In: Jeanteur P, editor. Epigenetics and Chromatin: Springer Berlin Heidelberg. p. 183-206.

Chan FY, Robinson J, Brownlie A, Shivdasani RA, Donovan A, Brugnara C, Kim J, Lau BC, Witkowska HE, Zon LI 1997. Characterization of adult alpha- and beta-globin genes in the zebrafish. Blood 89: 688-700.

Czelusniak J, Goodman M, Hewett-Emmett D, Weiss ML, Venta PJ, Tashian RE 1982. Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. Nature 298: 297-300.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Group GPA 2011. The variant call format and VCFtools. Bioinformatics 27: 2156-2158. doi: 10.1093/bioinformatics/btr330

de Bruin SH, Janssen LHM 1973. Comparison of the oxygen and proton binding behavior of human hemoglobin A and A2. Biochimica et Biophysica Acta (BBA) - Protein Structure 295: 490-494.

Dekker J 2003. A closer look at long-range chromosomal interactions. Trends in Biochemical Sciences 28: 277-280.

Dekker J, Rippe K, Dekker M, Kleckner N 2002. Capturing Chromosome Conformation. Science 295: 1306-1311. doi: 10.1126/science.1067799

Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 16: 1299-1309.

Drissen R, Palstra RJ, Gillemans N, Splinter E, Grosveld F, Philipsen S, de Laat W 2004. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. Genes Dev 18: 2485-2490. doi: 10.1101/gad.317004

Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ 1980. The structure and evolution of the human beta-globin gene family. Cell 21: 653-668.

Excoffier L, Laval G, Schneider S 2005. Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1: 47-50.

Fitch DH, Bailey WJ, Tagle DA, Goodman M, Sieu L, Slightom JL 1991. Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. Proc Natl Acad Sci U S A 88: 7396-7400.

Forget BG 1998. Molecular Basis of Hereditary Persistence of Fetal Hemoglobin. Annals of the New York Academy of Sciences 850: 38-44. doi: 10.1111/j.1749-6632.1998.tb10460.x

Forget BG 2011. Progress in Understanding the Hemoglobin Switch. New England Journal of Medicine 365: 852-854. doi: doi:10.1056/NEJMe1106969

Fuchs C, Burmester T, Hankeln T 2006. The amphibian globin gene repertoire as revealed by the Xenopus genome. Cytogenet Genome Res 112: 296-306. doi: 10.1159/000089884

Galarneau G, Palmer CD, Sankaran VG, Orkin SH, Hirschhorn JN, Lettre G 2010. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nat Genet 42: 1049-1051.

Gaudry MJ, Storz JF, Butts GT, Campbell KL, Hoffmann FG 2014. Repeated Evolution of Chimeric Fusion Genes in the β-Globin Gene Family of Laurasiatherian Mammals. Genome Biol Evol. doi: 10.1093/gbe/evu097

Giambona A, Passarello C, Renda D, Maggio A 2009. The significance of the hemoglobin A2 value in screening for hemoglobinopathies. Clinical Biochemistry 42: 1786-1796.

Giardine B, Borg J, Viennas E, Pavlidis C, Moradkhani K, Joly P, Bartsakoulia M, Riemer C, Miller W, Tzimas G, Wajcman H, Hardison RC, Patrinos GP 2014. Updates of the HbVar database of human hemoglobin variants and thalassemia mutations. Nucleic Acids Res 42: D1063-D1069. doi: 10.1093/nar/gkt911

Goodman M, Czelusniak J, Koop BF, Tagle DA, Slightom JL 1987. Globins: a case study in molecular phylogeny. Cold Spring Harb Symp Quant Biol 52: 875-890.

Goodman M, Koop BF, Czelusniak J, Weiss ML 1984. The eta-globin gene. Its long evolutionary history in the beta-globin gene family of mammals. J Mol Biol 180: 803-823.

Goodman M, Moore GW, Matsuda G 1975. Darwinian evolution in the genealogy of haemoglobin. Nature 253: 603-608.

Gribnau J, Diderich K, Pruzina S, Calzolari R, Fraser P 2000. Intergenic transcription and developmental remodeling of chromatin subdomains in the human beta-globin locus. Mol Cell 5: 377-386.

Gumucio DL, Shelton DA, Zhu W, Millinoff D, Gray T, Bock JH, Slightom JL, Goodman M 1996. Evolutionary Strategies for the Elucidation ofcisandtransFactors That Regulate the

Developmental Switching Programs of the β-like Globin Genes. Mol Phylogenet Evol 5: 18-32.

Hardies SC, Edgell MH, Hutchison CA, 3rd 1984. Evolution of the mammalian beta-globin gene cluster. J Biol Chem 259: 3748-3756.

Hardison R 2012a. Evolution of hemoglobin and its genes. Cold Spring Harb Perspect Med 2: a011627. doi: 10.1101/cshperspect.a011627

Hardison R. 2001. Organization, evolution and regulation of the globin genes. In: Steinberg MH, Forget, B. G., Higgs, D. R., and Nagel, R. L., eds., editor. Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management. Cambridge: Cambridge University Press.

Hardison R, Slightom JL, Gumucio DL, Goodman M, Stojanovic N, Miller W 1997. Locus control regions of mammalian β-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights. Gene 205: 73-94.

Hardison RC 1984. Comparison of the beta-like globin gene families of rabbits and humans indicates that the gene cluster 5'-epsilon-gamma-delta-beta-3' predates the mammalian radiation. Mol Biol Evol 1: 390-410.

Hardison RC 2012b. Evolution of Hemoglobin and Its Genes. Cold Spring Harb Perspect Med 2. doi: 10.1101/cshperspect.a011627

Hardison RC 2008. Globin genes on the move. Journal of Biology 7: 35-35. doi: 10.1186/jbiol92

Hardison RC, Margot JB 1984. Rabbit globin pseudogene psi beta 2 is a hybrid of delta- and beta-globin gene sequences. Mol Biol Evol 1: 302-316.

Harju S, McQueen KJ, Peterson KR 2002. Chromatin structure and control of beta-like globin gene switching. Exp Biol Med (Maywood) 227: 683-700.

Higgs DR, Engel JD, Stamatoyannopoulos G Thalassaemia. The Lancet 379: 373-383. doi: 10.1016/S0140-6736(11)60283-3

Hoffmann FG, Opazo JC, Storz JF 2008. New genes originated via multiple recombinational pathways in the beta-globin gene family of rodents. Mol Biol Evol 25: 2589-2600. doi: 10.1093/molbev/msn200

Hoffmann FG, Storz JF 2007. The alphaD-globin gene originated via duplication of an embryonic alpha-like globin gene in the ancestor of tetrapod vertebrates. Mol Biol Evol 24: 1982-1990. doi: 10.1093/molbev/msm127

Hoffmann FG, Storz JF, Gorr TA, Opazo JC 2010. Lineage-specific patterns of functional diversification in the alpha- and beta-globin gene families of tetrapod vertebrates. Mol Biol Evol 27: 1126-1138. doi: 10.1093/molbev/msp325

Hosbach HA, Wyler T, Weber R 1983. The Xenopus laevis globin gene family: chromosomal arrangement and gene structure. Cell 32: 45-53.

Jeffreys AJ, Barrie PA, Harris S, Fawcett DH, Nugent ZJ, Boyd AC 1982. Isolation and sequence analysis of a hybrid delta-globin pseudogene from the brown lemur. J Mol Biol 156: 487-503.

Jeffreys AJ, Wilson V, Wood D, Simons JP, Kay RM, Williams JG 1980. Linkage of adult alpha- and beta-globin genes in X. laevis and gene duplication by tetraploidization. Cell 21: 555-564.

Johnson KD, Grass JA, Boyer ME, Kiekhaefer CM, Blobel GA, Weiss MJ, Bresnick EH 2002a. Cooperative activities of hematopoietic regulators recruit RNA polymerase II to a tissue-specific chromatin domain. Proc Natl Acad Sci U S A 99: 11760-11765. doi: 10.1073/pnas.192285999

Johnson RM, Buck S, Chiu C-h, Schneider H, Sampaio I, Gage DA, Shen T-L, Schneider MPC, Muniz JA, Gumucio DL, Goodman M 1996. Fetal Globin Expression in New World Monkeys. Journal of Biological Chemistry 271: 14684-14691. doi: 10.1074/jbc.271.25.14684

Johnson RM, Buck S, Chiu CH, Gage DA, Shen TL, Hendrickx AG, Gumucio DL, Goodman M 2000. Humans and old world monkeys have similar patterns of fetal globin expression. J Exp Zool 288: 318-326. doi: 10.1002/1097-010x(20001215)288:4<318::aid-jez4>3.0.co;2-0

Johnson RM, Gumucio D, Goodman M 2002b. Globin gene switching in primates. Comp Biochem Physiol A Mol Integr Physiol 133: 877-883.

Kay RM, Harris R, Patient RK, Williams JG 1980. Molecular cloning of cDNA sequences coding for the major alpha- and beta-globin polypeptides of adult Xenopus laevis. Nucleic Acids Res 8: 2691-2707.

King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res 15: 1051-1060. doi: 10.1101/gr.3642605

Koop BF, Goodman M 1988. Evolutionary and developmental aspects of two hemoglobin beta-chain genes (epsilon M and beta M) of opossum. Proceedings of the National Academy of Sciences 85: 3893-3897.

Koop BF, Siemieniak D, Slightom JL, Goodman M, Dunbar J, Wright PC, Simons EL 1989. Tarsius delta- and beta-globin genes: conversions, evolution, and systematic implications. J Biol Chem 264: 68-79.

Lecrone CN. 1970. Brief Report: Absence of Special Fetal Hemoglobin in Beagle Dogs.

Lettre G, Sankaran VG, Bezerra MAC, Araújo AS, Uda M, Sanna S, Cao A, Schlessinger D, Costa FF, Hirschhorn JN, Orkin SH 2008. DNA polymorphisms at the BCL11A, HBS1L-MYB, and β-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. Proc Natl Acad Sci U S A 105: 11869-11874. doi: 10.1073/pnas.0804799105

Manchinu MF, Marongiu MF, Poddie D, Casu C, Latini V, Simbula M, Galanello R, Moi P, Cao A, Porcu S, Ristaldi MS. 2014. In vivo activation of the human δ-globin gene: the therapeutic potential in β-thalassemic mice.

Martin SL, Vincent KA, Wilson AC 1983. Rise and fall of the delta globin gene. J Mol Biol 164: 513-528.

Martin SL, Zimmer EA, Kan YW, Wilson AC 1980. Silent delta-globin gene in Old World monkeys. Proc Natl Acad Sci U S A 77: 3563-3566.

McMorrow T, Tewari R, Wai AWK, Burgtorf C, Drabek D, Ventress N, Langeveld A, Higgs D, Tan-Un K, Grosveld F, Philipsen S. 2003. Functional and comparative analysis of globin loci in pufferfish and humans.

Menzel S, Garner C, Gut I, Matsuda F, Yamaguchi M, Heath S, Foglio M, Zelenika D, Boland A, Rooks H, Best S, Spector TD, Farrall M, Lathrop M, Thein SL 2007. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. Nat Genet 39: 1197-1199.

Miller IJ, Bieker JJ 1993. A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Kruppel family of nuclear proteins. Mol Cell Biol 13: 2776-2786.

Modell B, Darlison M 2008. Global epidemiology of haemoglobin disorders and derived service indicators. Bulletin of the World Health Organization 86: 480-487.

Morgado A, Picanço I, Gomes S, Miranda A, Coucelo M, Seuanes F, Seixas MT, Romão L, Faustino P 2007. Mutational spectrum of delta-globin gene in the Portuguese population. European Journal of Haematology 79: 422-428. doi: 10.1111/j.1600-0609.2007.00949.x

Mosca A, Paleari R, Ivaldi G, Galanello R, Giordano PC 2009. The role of haemoglobin A2 testing in the diagnosis of thalassaemias and related haemoglobinopathies. Journal of Clinical Pathology 62: 13-17. doi: 10.1136/jcp.2008.056945

Musallam KM, Taher AT, Cappellini MD, Sankaran VG. 2013. Clinical experience with fetal hemoglobin induction therapy in patients with β-thalassemia.

Noordermeer D, de Laat W 2008. Joining the loops: beta-globin gene regulation. IUBMB Life 60: 824-833. doi: 10.1002/iub.129

Opazo JC, Hoffmann FG, Storz JF 2008a. Differential loss of embryonic globin genes during the radiation of placental mammals. Proceedings of the National Academy of Sciences 105: 12950-12955. doi: 10.1073/pnas.0804392105

Opazo JC, Hoffmann FG, Storz JF 2008b. Genomic evidence for independent origins of β-like globin genes in monotremes and therian mammals. Proceedings of the National Academy of Sciences 105: 1590-1595. doi: 10.1073/pnas.0710531105

Opazo JC, Sloan AM, Campbell KL, Storz JF 2009. Origin and Ascendancy of a Chimeric Fusion Gene: The β/δ-Globin Gene of Paenungulate Mammals. Mol Biol Evol 26: 1469-1478. doi: 10.1093/molbev/msp064

Ottolenghi S, Giglioni B, Comi P, Gianni AM, Polli E, Acquaye CT, Oldham JH, Masera G 1979. Globin gene deletion in HPFH, delta (o) beta (o) thalassaemia and Hb Lepore disease. Nature 278: 654-657.

Palstra RJ, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, de Laat W 2003. The beta-globin nuclear compartment in development and erythroid differentiation. Nat Genet 35: 190-194. doi: 10.1038/ng1244

Patel VS, Cooper SJ, Deakin JE, Fulton B, Graves T, Warren WC, Wilson RK, Graves JA 2008. Platypus globin genes and flanking loci suggest a new insertional model for beta-globin evolution in birds and mammals. BMC Biol 6: 34. doi: 10.1186/1741-7007-6-34

Patel VS, Ezaz T, Deakin JE, Graves JA 2010. Globin gene structure in a reptile supports the transpositional model for amniote alpha- and beta-globin gene evolution. Chromosome Research 18: 897-907. doi: 10.1007/s10577-010-9164-5

Patrinos G, Antonarakis S. 2010. Human Hemoglobin. In: Speicher M, Motulsky A, Antonarakis S, editors. Vogel and Motulsky's Human Genetics: Springer Berlin Heidelberg. p. 365-401.

Patrinos GP, de Krom M, de Boer E, Langeveld A, Imam AM, Strouboulis J, de Laat W, Grosveld FG 2004. Multiple interactions between regulatory regions are required to stabilize an active chromatin hub. Genes Dev 18: 1495-1509. doi: 10.1101/gad.289704

Pearson R, Fleetwood J, Eaton S, Crossley M, Bao S 2008. Krüppel-like transcription factors: A functional family. The International Journal of Biochemistry & Cell Biology 40: 1996-2001.

Pisano E, Cocca E, Mazzei F, Ghigliotti L, di Prisco G, William Detrich Iii H, Ozouf-Costaz C 2003. Mapping of α- and β-globin genes on Antarctic fish chromosomes by fluorescence in-situ hybridization. Chromosome Research 11: 633-640. doi: 10.1023/A:1024961103663

Prychitko T, Johnson RM, Wildman DE, Gumucio D, Goodman M 2005. The phylogenetic history of New World monkey beta globin reveals a platyrrhine beta to delta gene conversion in the atelid ancestry. Mol Phylogenet Evol 35: 225-234. doi: 10.1016/j.ympev.2004.11.002

Ranney HM, Lam R, Rosenberg G 1993. Some properties of hemoglobin A2. American Journal of Hematology 42: 107-111. doi: 10.1002/ajh.2830420121

Ristaldi MS, Casula S, Porcu S, Marongiu MF, Pirastu M, Cao A 1999. Activation of the δ-Globin Gene by the β-Globin Gene CACCC Motif. Blood Cells, Molecules, and Diseases 25: 193-209.

Rohrbaugh ML, Hardison RC 1983. Analysis of rabbit β-like globin gene transcripts during development. J Mol Biol 164: 395-417.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497. doi: 10.1093/bioinformatics/btg359

Sankaran VG 2011. Targeted Therapeutic Strategies for Fetal Hemoglobin Induction. ASH Education Program Book 2011: 459-465. doi: 10.1182/asheducation-2011.1.459

Sankaran VG, Menne TF, Xu J, Akie TE, Lettre G, Van Handel B, Mikkola HKA, Hirschhorn JN, Cantor AB, Orkin SH 2008. Human Fetal Hemoglobin Expression Is Regulated by the Developmental Stage-Specific Repressor BCL11A. Science 322: 1839-1842. doi: 10.1126/science.1165409

Sankaran VG, Nathan DG 2010. Reversing the Hemoglobin Switch. New England Journal of Medicine 363: 2258-2260. doi: doi:10.1056/NEJMcibr1010767

Sankaran VG, Orkin SH 2013. The switch from fetal to adult hemoglobin. Cold Spring Harb Perspect Med 3: a011643. doi: 10.1101/cshperspect.a011643

Sankaran VG, Xu J, Byron R, Greisman HA, Fisher C, Weatherall DJ, Sabath DE, Groudine M, Orkin SH, Premawardhena A, Bender MA 2011a. A functional element necessary for fetal hemoglobin silencing. N Engl J Med 365: 807-814. doi: 10.1056/NEJMoa1103070

Sankaran VG, Xu J, Byron R, Greisman HA, Fisher C, Weatherall DJ, Sabath DE, Groudine M, Orkin SH, Premawardhena A, Bender MA 2011b. A Functional Element Necessary for Fetal Hemoglobin Silencing. New England Journal of Medicine 365: 807-814. doi: doi:10.1056/NEJMoa1103070

Sankaran VG, Xu J, Orkin SH 2010a. Advances in the understanding of haemoglobin switching. Br J Haematol 149: 181-194. doi: 10.1111/j.1365-2141.2010.08105.x

Sankaran VG, Xu J, Orkin SH 2010b. Transcriptional silencing of fetal hemoglobin by BCL11A. Annals of the New York Academy of Sciences 1202: 64-68. doi: 10.1111/j.1749-6632.2010.05574.x

Sankaran VG, Xu J, Ragoczy T, Ippolito GC, Walkley CR, Maika SD, Fujiwara Y, Ito M, Groudine M, Bender MA, Tucker PW, Orkin SH 2009. Developmental and species-divergent globin switching are driven by BCL11A. Nature 460: 1093-1097. doi: 10.1038/nature08243

Satoh H, Inokuchi N, Nagae Y, Okazaki T 1999. Organization, Structure, and Evolution of the Nonadult Rat β-Globin Gene Cluster. J Mol Evol 49: 122-129. doi: 10.1007/PL00006525

Schechter AN. 2008. Hemoglobin research and the origins of molecular medicine.

Schimenti JC, Duncan CH 1985. Structure and organization of the bovine beta-globin genes. Mol Biol Evol 2: 514-525.

Shapiro SG, Schon EA, Townes TM, Lingrel JB 1983. Sequence and linkage of the goat εI and εII β-globin genes. J Mol Biol 169: 31-52.

Song G, Riemer C, Dickins B, Kim HL, Zhang L, Zhang Y, Hsu CH, Hardison RC, Nisc Comparative Sequencing P, Green ED, Miller W 2012. Revealing mammalian evolutionary relationships by comparative analysis of gene clusters. Genome Biol Evol 4: 586-601. doi: 10.1093/gbe/evs032

Song S-H, Hou C, Dean A 2007. A positive role for NLI/Ldb1 in long range β-globin locus control region function. Mol Cell 28: 810-822. doi: 10.1016/j.molcel.2007.09.025

Spritz RA, Giebel LB 1988. The structure and evolution of the spider monkey delta-globin gene. Mol Biol Evol 5: 21-29.

Stamatoyannopoulos G 2005. Control of globin gene expression during development and erythroid differentiation. Experimental hematology 33: 259. doi: 10.1016/j.exphem.2004.11.007

Steinberg MH, Adams JG, 3rd 1991. Hemoglobin A2: origin, evolution, and aftermath. Blood 78: 2165-2177.

Steinberg MH, et al. 2009. Disorders of Hemoglobin: Cambridge University Press.

Stockell A, Perutz M, Muirhead H, Glauser SC 1961. A comparison of adult and foetal horse haemoglobins. J Mol Biol 3: 112-116.

Storz JF, Opazo JC, Hoffmann FG 2013. Gene duplication, genome duplication, and the functional diversification of vertebrate globins. Mol Phylogenet Evol 66: 469-478. doi: 10.1016/j.ympev.2012.07.013

Storz JF, Opazo JC, Hoffmann FG 2011. Phylogenetic diversification of the globin gene superfamily in chordates. IUBMB Life 63: 313-322. doi: 10.1002/iub.482

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT 1988. Embryonic ε and γ globin genes of a prosimian primate (Galago crassicaudatus): Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203: 439-455.

Tagle DA, Slightom JL, Jones RT, Goodman M 1991. Concerted evolution led to high expression of a prosimian primate delta globin gene locus. J Biol Chem 266: 7469-7480.

Thein SL 2004. Genetic insights into the clinical diversity of β thalassaemia. Br J Haematol 124: 264-274. doi: 10.1046/j.1365-2141.2003.04769.x

Thein SL, Menzel S, Lathrop M, Garner C 2009. Control of fetal hemoglobin: new insights emerging from genomics and clinical implications. Human Molecular Genetics 18: R216-R223. doi: 10.1093/hmg/ddp401

Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell 10: 1453-1465.

Townes TM, Fitzgerald MC, Lingrel JB 1984. Triplication of a four-gene set during evolution of the goat beta-globin locus produced three genes now expressed differentially during development. Proc Natl Acad Sci U S A 81: 6589-6593.

Uda M, Galanello R, Sanna S, Lettre G, Sankaran VG, Chen W, Usala G, Busonero F, Maschio A, Albai G, Piras MG, Sestu N, Lai S, Dei M, Mulas A, Crisponi L, Naitza S, Asunis I, Deiana M, Nagaraja R, Perseu L, Satta S, Cipollina MD, Sollaino C, Moi P, Hirschhorn JN, Orkin SH, Abecasis GR, Schlessinger D, Cao A 2008. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and

amelioration of the phenotype of β-thalassemia. Proceedings of the National Academy of Sciences 105: 1620-1625. doi: 10.1073/pnas.0711566105

Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, Weiss MJ, Dekker J, Blobel GA 2005. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. Mol Cell 17: 453-462.

Weatherall DJ 2001. Phenotype[mdash]genotype relationships in monogenic disease: lessons from the thalassaemias. Nat Rev Genet 2: 245-255.

Weatherall DJ, Clegg JB. 2008. Distribution and Population Genetics of the Thalassaemias. In. The Thalassaemia Syndromes: Blackwell Science Ltd. p. 237-284.

Weatherall DJ, Clegg JB 2001. Inherited haemoglobin disorders: an increasing global health problem. Bulletin of the World Health Organization 79: 704-712.

Webster MT, Clegg JB, Harding RM 2003. Common 5' beta-globin RFLP haplotypes harbour a surprising level of ancestral sequence mosaicism. Hum Genet 113: 123-139. doi: 10.1007/s00439-003-0954-0

Welch JJ, Watts JA, Vakoc CR, Yao Y, Wang H, Hardison RC, Blobel GA, Chodosh LA, Weiss MJ 2004. Global regulation of erythroid gene expression by transcription factor GATA-1. Blood 104: 3136-3147. doi: 10.1182/blood-2004-04-1603

Whitelaw E, Tsai SF, Hogben P, Orkin SH 1990. Regulated expression of globin chains and the erythroid transcription factor GATA-1 during erythropoiesis in the developing mouse. Mol Cell Biol 10: 6596-6606. doi: 10.1128/mcb.10.12.6596

Wijgerde M, Gribnau J, Trimborn T, Nuez B, Philipsen S, Grosveld F, Fraser P 1996. The role of EKLF in human beta-globin gene competition. Genes Dev 10: 2894-2902.

Wilber A, Nienhuis AW, Persons DA 2011. Transcriptional regulation of fetal to adult hemoglobin switching: new therapeutic opportunities. Blood 117: 3945-3953. doi: 10.1182/blood-2010-11-316893

Williams TN, Weatherall DJ 2012. World Distribution, Population Genetics, and Health Burden of the Hemoglobinopathies. Cold Spring Harb Perspect Med 2. doi: 10.1101/cshperspect.a011692

Xu J, Sankaran VG, Ni M, Menne TF, Puram RV, Kim W, Orkin SH 2010. Transcriptional silencing of γ-globin by BCL11A involves long-range interactions and cooperation with SOX6. Genes Dev 24: 783-798. doi: 10.1101/gad.1897310


U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO