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

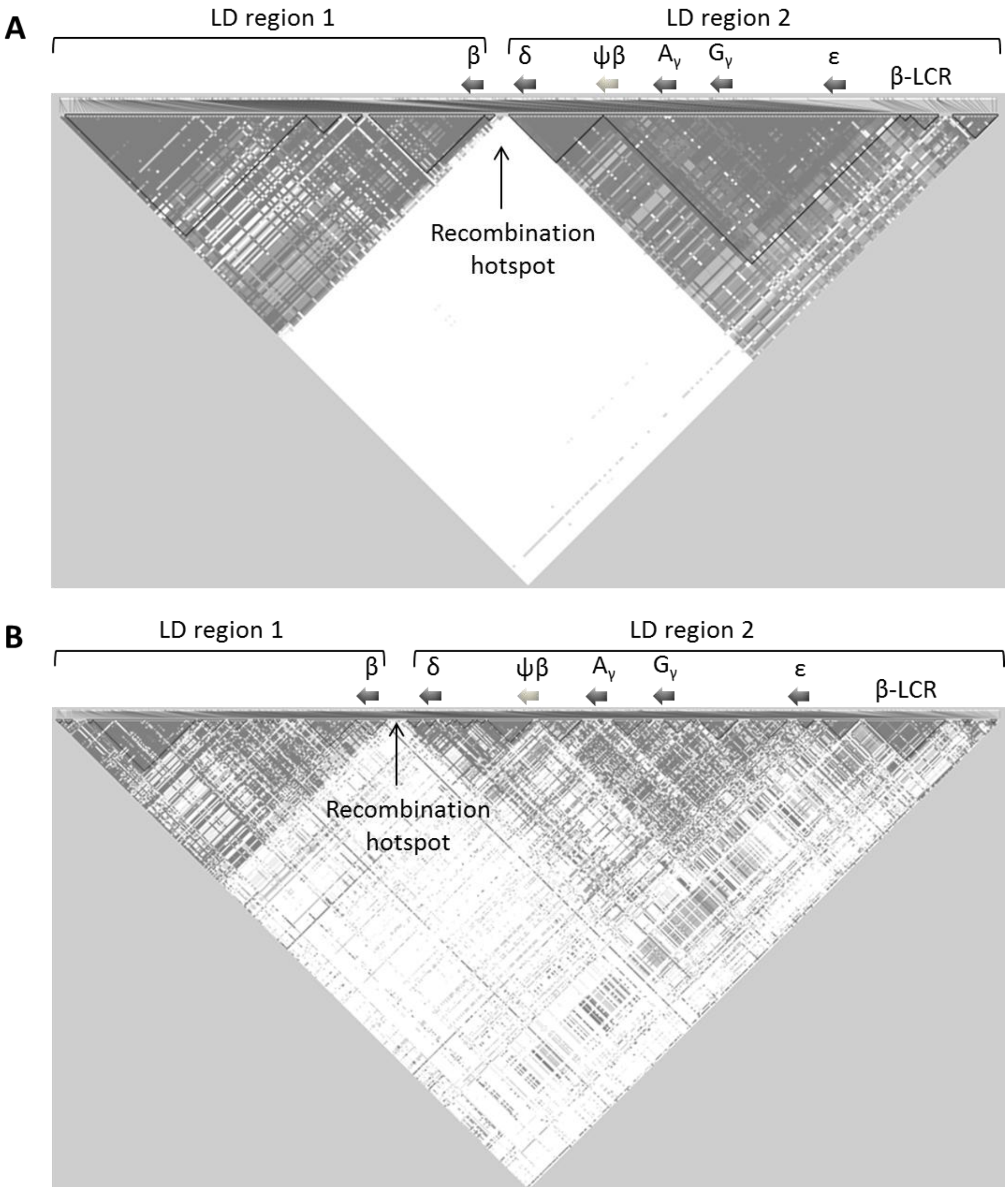


Figure S1 – LD plot of the β-globin cluster for the data from 1000 Genomes phase 1 release v3 for A) CEU and B) YRI. The image was constructed using Haploview 4.1 software. The triangles represent LD blocks. Two distinct regions with strong LD are identified: one containing *HBB* (LD region 1) and the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

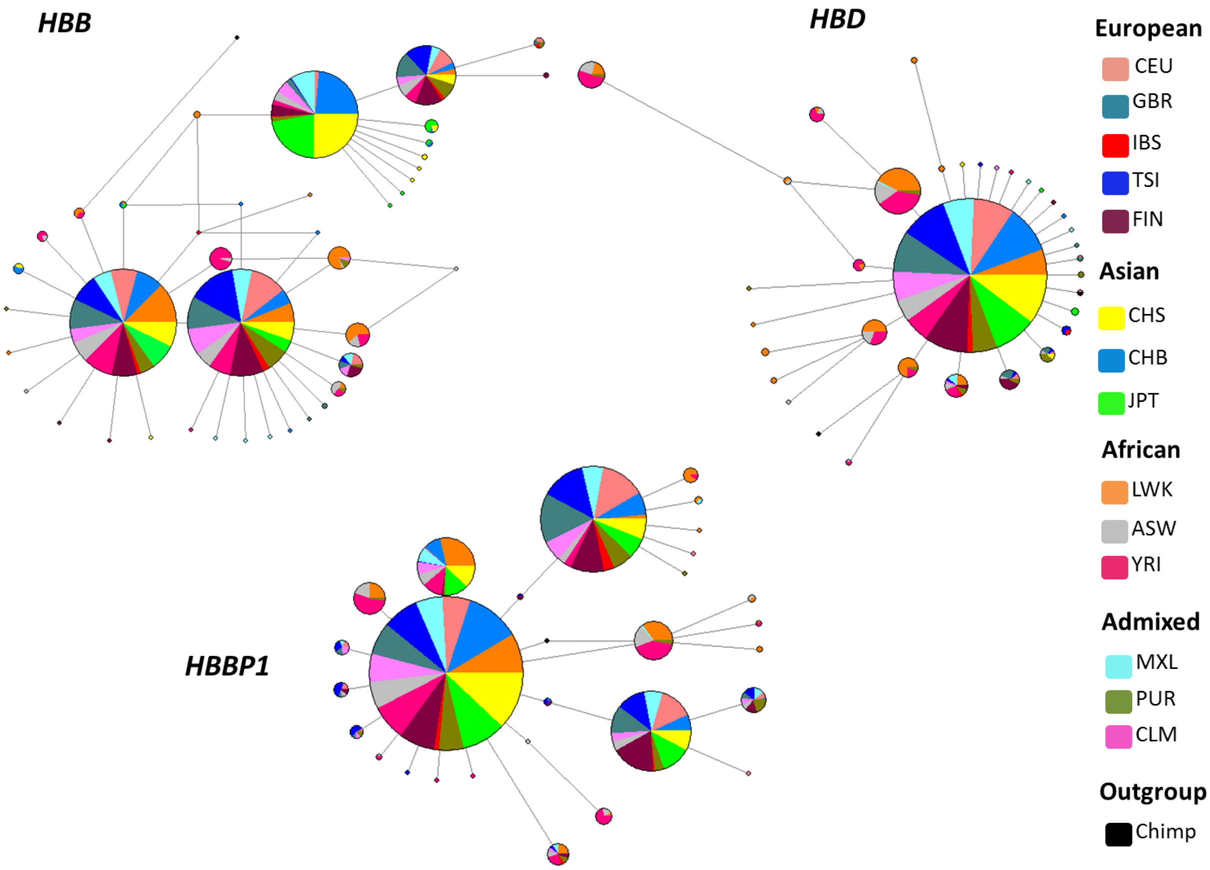
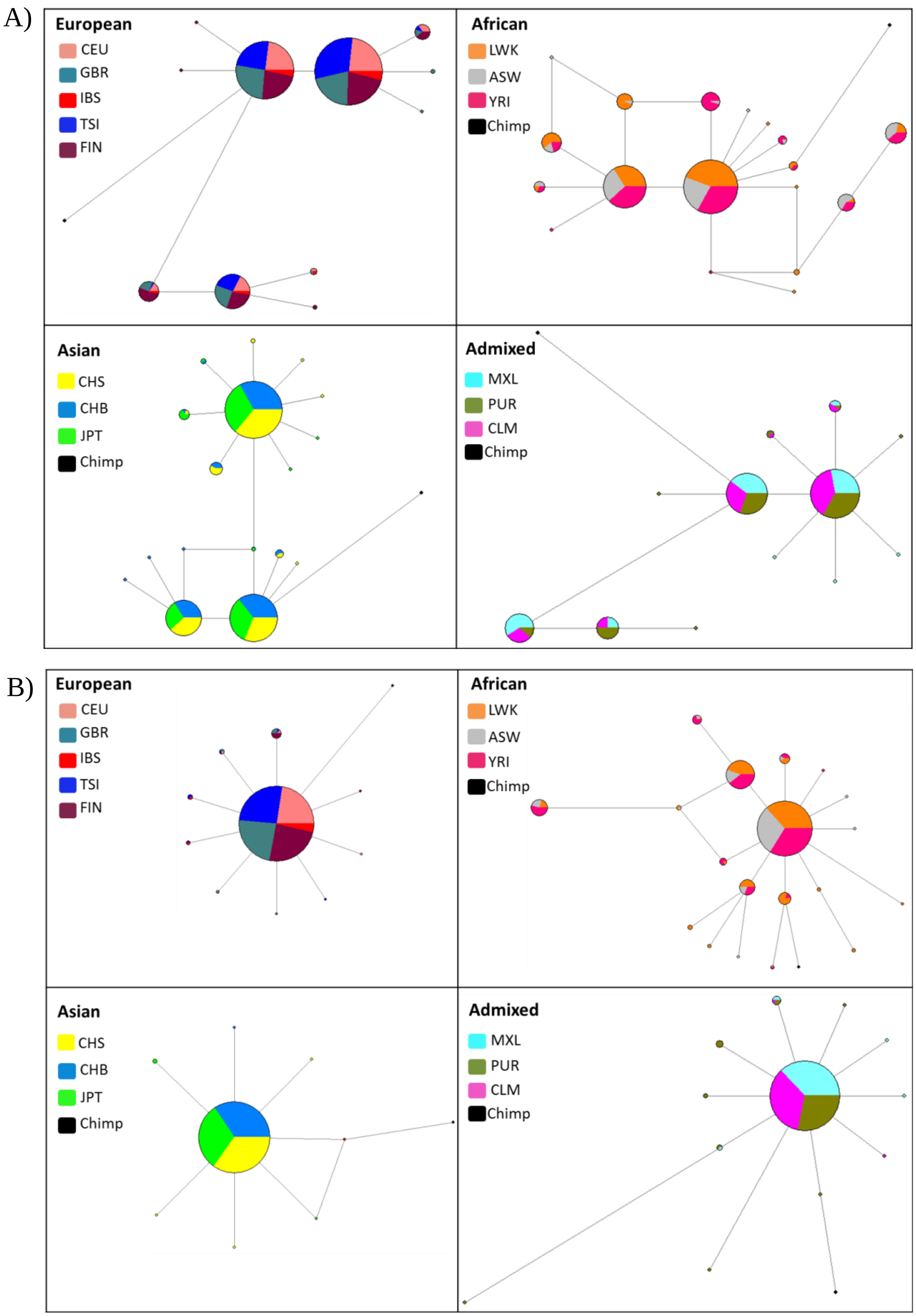


Figure S2 – Haplotype genealogies for *HBB*, *HBD* and *HBBP1*. Networks were built using haplotypes from 1000 Genomes Project (14 populations combined). Haplotypes are shown as circles with an area proportional to their frequency; lines connect different haplotypes and the number of mutations is proportional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

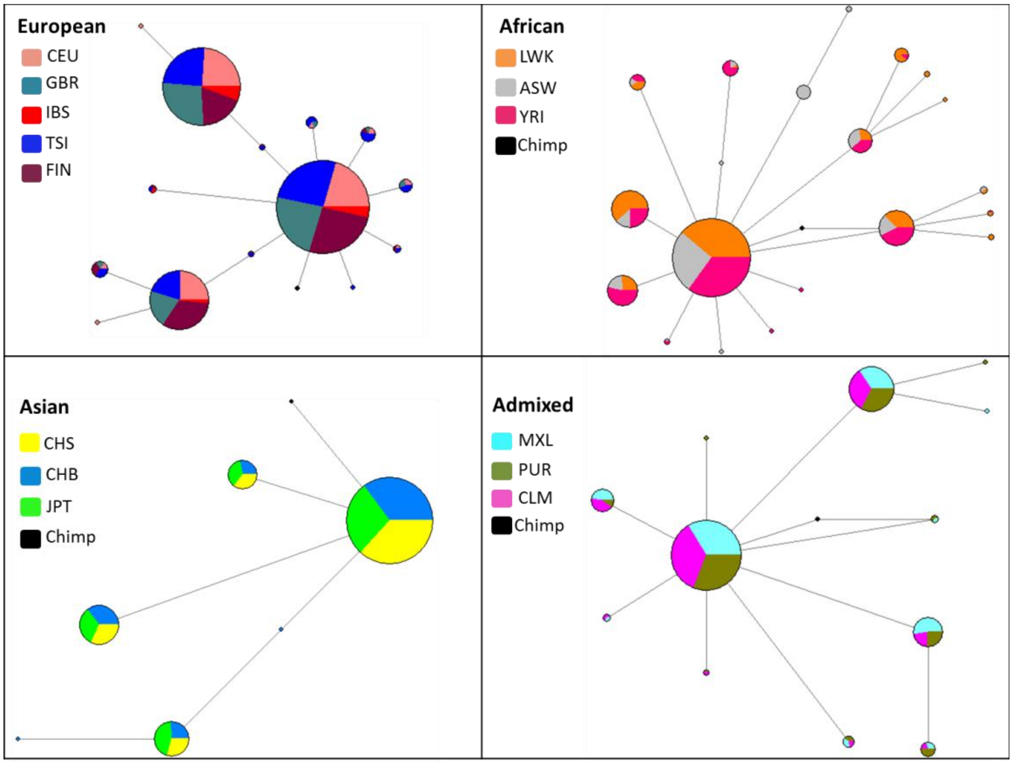
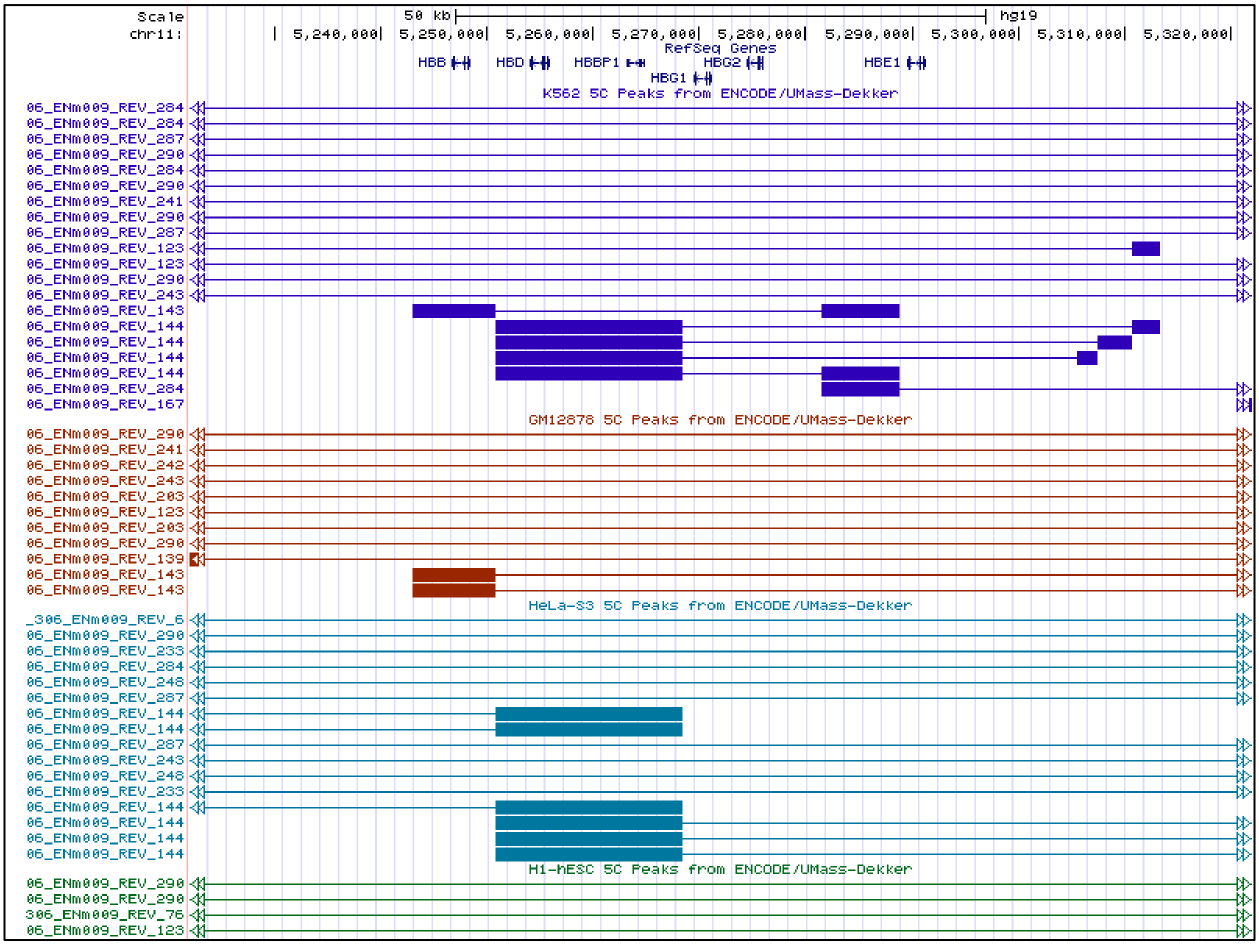


Figure S3 – Haplotype genealogies for A) *HBB*, B) *HBD* and C) *HBBP1*. Networks were built using haplotypes from 1000 Genomes Project, representing 1092 individuals from 14 populations: three African (ASW, LWK and YRI), five European (CEU, FIN, GBR, IBS and TSI), three Asian (CHB, CHS and JPT) and 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
chr11:
50 kb
hg19
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 5C Peaks from ENCODE/UMass-Dekker
GM12878 5C Peaks from ENCODE/UMass-Dekker
HeLa-S3 5C Peaks from ENCODE/UMass-Dekker
H1-hESC 5C Peaks from ENCODE/UMass-Dekker

B)

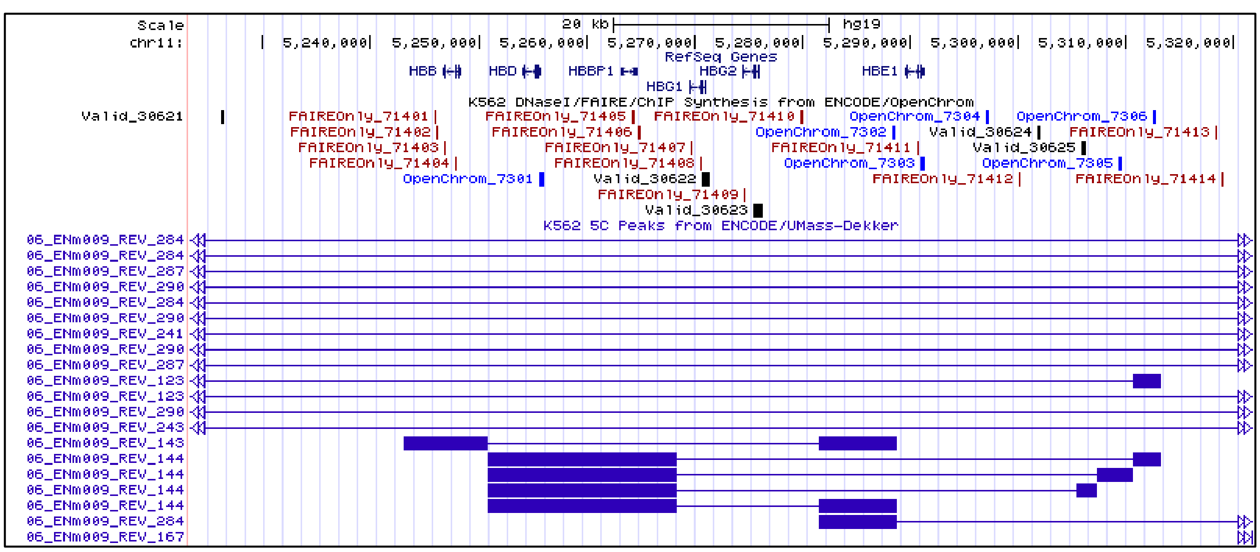

Scale
chr11:
20 kb
hg19
5,240,000
5,250,000
5,260,000
5,270,000
5,280,000
5,290,000
5,300,000
5,310,000
5,320,000
RefSeq Genes
HBB
HBD
HBBP1
HBG2
HBE1
HBG1
K562 DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom
Valid_30621
FAIREOnly_71401
FAIREOnly_71405
FAIREOnly_71410
OpenChrom_7304
OpenChrom_7306
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 5C 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).