Correction notice

Nat. Genet. **43**, 295–301 (2011)

Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach

Belinda Giardine, Joseph Borg, Douglas R Higgs, Kenneth R Peterson, Sjaak Philipsen, Donna Maglott, Belinda K Singleton, David J Anstee, A Nazli Basak, Barnaby Clark, Flavia C Costa, Paula Faustino, Halyna Fedosyuk, Alex E Felice, Alain Francina, Renzo Galanello, Monica V E Gallivan, Marianthi Georgitsi, Richard J Gibbons, Piero C Giordano, Cornelis L Harteveld, James D Hoyer, Martin Jarvis, Philippe Joly, Emmanuel Kanavakis, Panagoula Kollia, Stephan Menzel, Webb Miller, Kamran Moradkhani, John Old, Adamantia Papachatzopoulou, Manoussos N Papadakis, Petros Papadopoulos, Sonja Pavlovic, Lucia Perseu, Milena Radmilovic, Cathy Riemer, Stefania Satta, Iris Schrijver, Maja Stojiljkovic, Swee Lay Thein, Jan Traeger-Synodinos, Ray Tully, Takahito Wada, John S Waye, Claudia Wiemann, Branka Zukic, David H K Chui, Henri Wajcman, Ross C Hardison & George P Patrinos

In the version of this paper originally published online, **Supplementary Table 1** was omitted. This error has been corrected as of 27 March 2011.

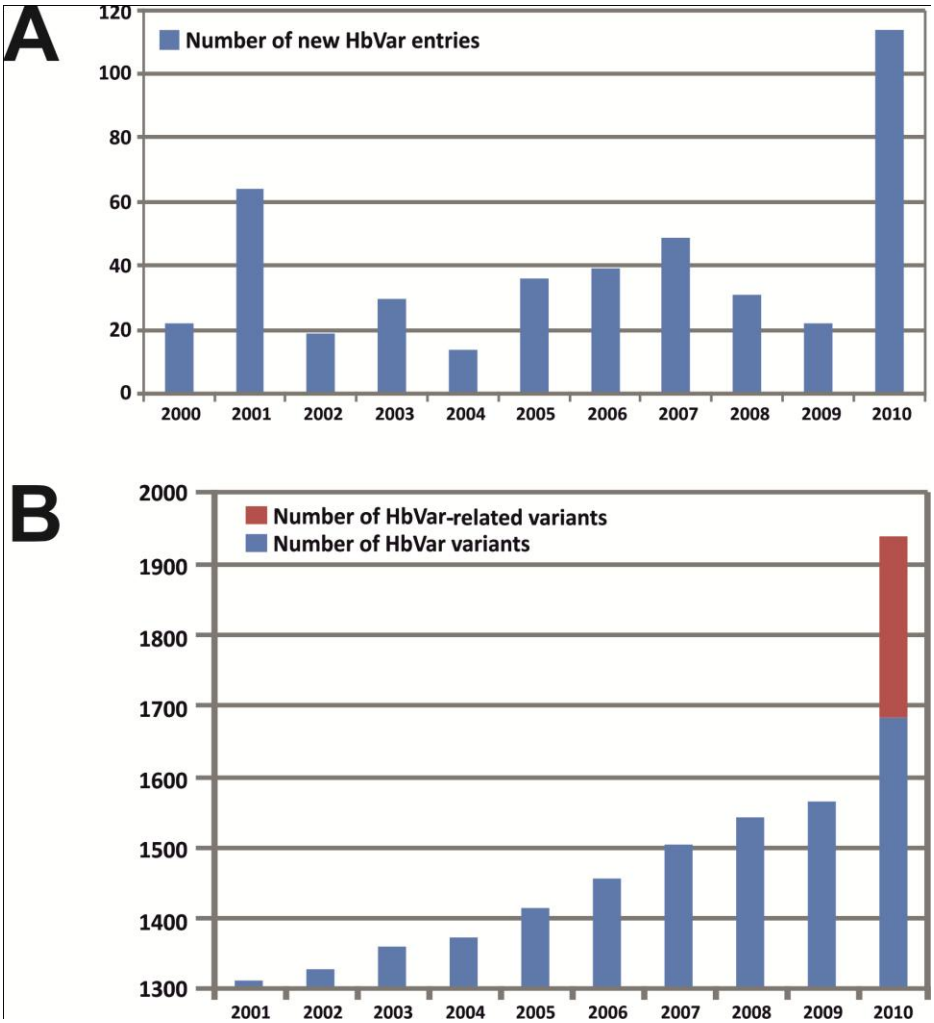
Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach

Belinda Giardine, Joseph Borg, Douglas R. Higgs, Kenneth R. Peterson, Sjaak Philipsen, Donna Maglott, Belinda K. Singleton, David J. Anstee, A. Nazli Basak, Barnaby Clark, Flavia C. Costa, Paula Faustino, Halyna Fedosyuk, Alex E. Felice, Alain Francina, Renzo Galanello, Monica V. E. Gallivan, Marianthi Georgitsi, Richard J. Gibbons, Piero C. Giordano, Cornelis L. Harteveld, James D. Hoyer, Martin Jarvis, Philippe Joly, Emmanuel Kanavakis, Panagoula Kollia, Stephan Menzel, Webb Miller, Kamran Moradkhani, John Old, Adamantia Papachatzopoulou, Manoussos N. Papadakis, Petros Papadopoulos, Sonja Pavlovic, Lucia Perseu, Milena Radmilovic, Cathy Riemer, Stefania Satta, Iris Schrijver, Maja Stojiljkovic, Swee Lay Thein, Jan Traeger-Synodinos, Ray Tully, Takahito Wada, John Waye, Claudia Wiemann, Branka Zukic, David H. K. Chui, Henri Wajcman, Ross C. Hardison, and George P. Patrinos

Supplementary Information

Supplementary Figures 1-5 and Supplementary Tables 1-2 and Supplementary Note

Nature Genetics: doi:10.1038/ng.785

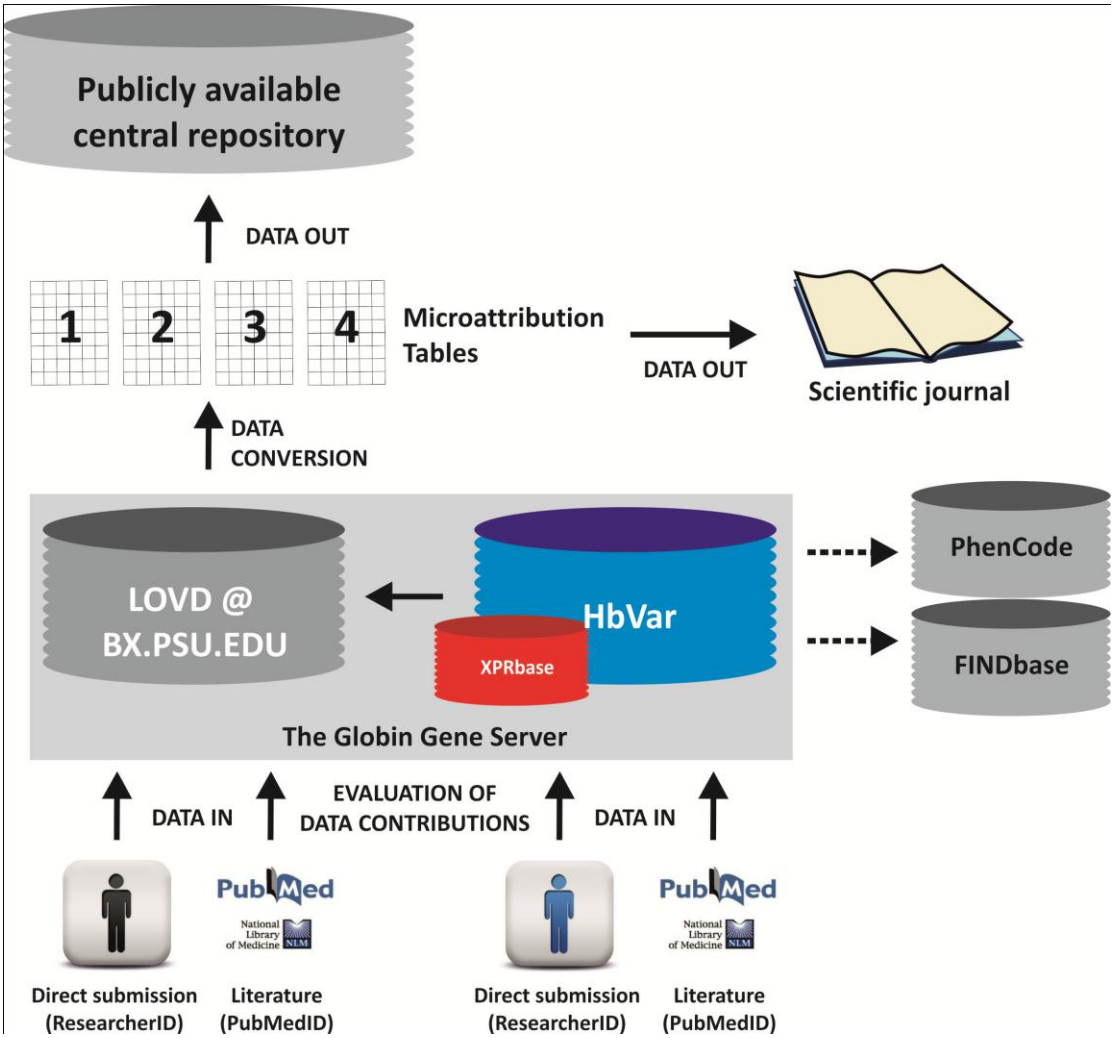


Supplementary Figure 1

A. Depiction of the number of new entries to HbVar database on an annual basis as of 2000 when the database started being developed. Notably, the number of new entries in 2010, when microattribution was first implemented, was the highest, compared even to 2001 when HbVar was officially launched. This underlines the impact of microatribution in genome variation data contribution.

B. Cumulative growth of the number of entries to HbVar (shown in blue) and the related LSDBs (shown in red).

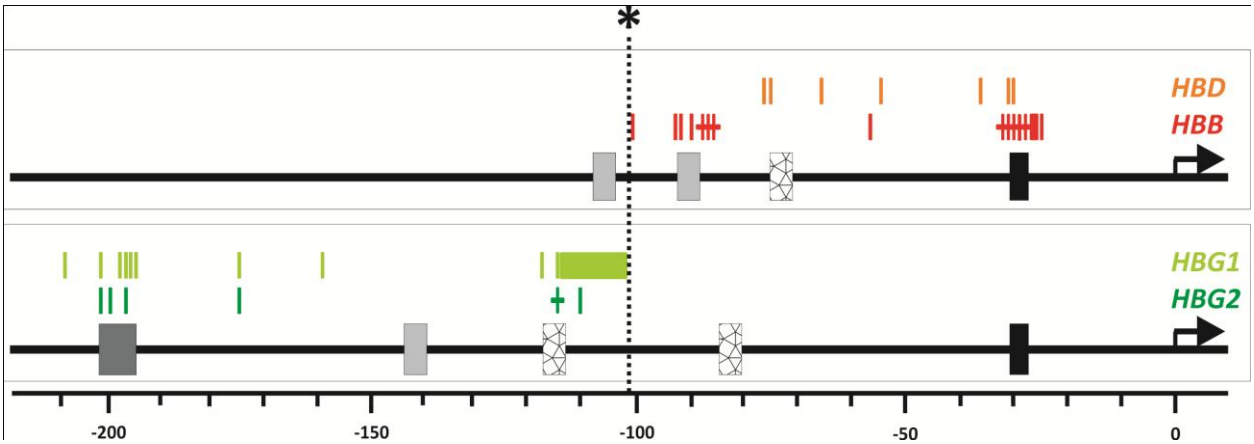
Nature Genetics: doi:10.1038/ng.785



Supplementary Figure 2

Outline of the microattribution approach implemented for documenting hemoglobinopathies-related genomic variation. Data contributors deposit data into HbVar either by directly contacting data curators or directly to the LSDBs and receive credit in the form of micro-citations vis-à-vis their ResearcherIDs. Data are thereafter forwarded to the central NCBI depository in the format shown in the microattribution tables (**Supplementary Tables 1**). Microattribution data flows are depicted by solid arrows. HbVar also intercommunicates with the related PhenCode project [30] and FINDbase [31,32] (dashed arrows).

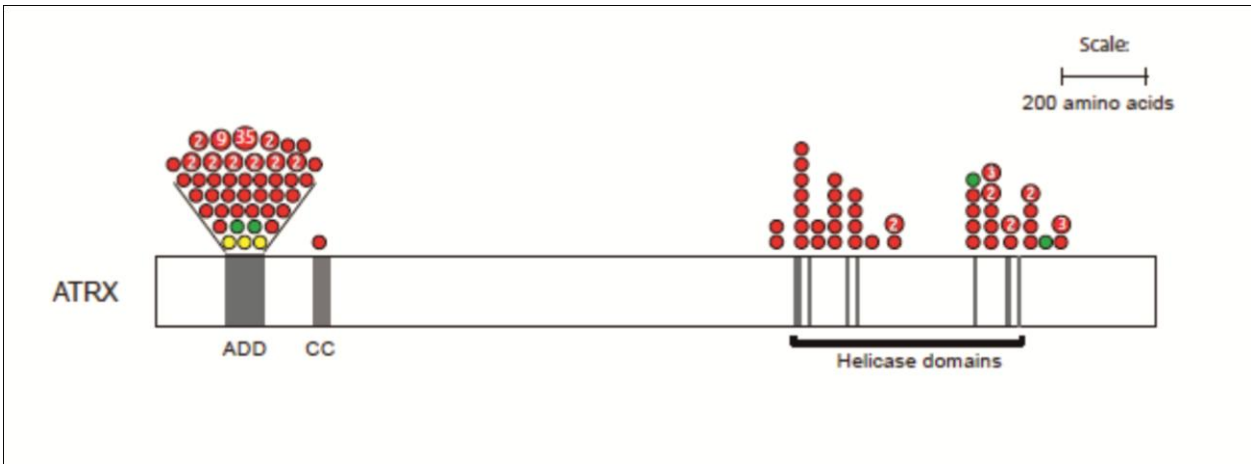
Nature Genetics: doi:10.1038/ng.785



Supplementary Figure 3

The globin gene promoter variants in the fetal (*HBG1*, *HBG2*) and adult (*HBD*, *HBB*) globin genes cluster in different locations in their promoters (separated by a dashed line and indicated by an asterisk). Vertical lines represent a single variant; crosses represent multiple variants in the same nucleotide, while solid boxes represent small deletions. The black arrow denotes the transcription initiation site, the black box represents the TATA box and the grey solid and hatched boxes represent the CACCC and CCAAT boxes respectively, while the dark grey solid box represents the response element. Numbers indicate distance in base pairs from the transcription start site.

Nature Genetics: doi:10.1038/ng.785



Supplementary Figure 4

Outline of the *ATRX* gene variants, resulting from missense mutations and small indels. Abbreviations: ADD: PHD related zinc finger seen in **A**TRX, **D**NMT3, **D**NMT3L, CC: coiled-coil domain, Helicase domain: SNF2 ATPase domain. Numbers in circles show number of independent mutations. Variants leading to ATR-X syndrome are depicted in red, variants leading to ATMDS are depicted in green, while variants that have been documented in both syndromes are depicted in yellow. For a complete list of the *ATRX* gene variants, see also Ref. 33.

Nature Genetics: doi:10.1038/ng.785

Supplementary Figure 5

Graphical output at the PSU genome browser of α-like globin genes variants deposited in HbVar (*HBA1*=262 variants (**A**), *HBA2*=302 variants) (**B**). The distribution of hemoglobin variants for the *HBA1* and *HBA2* genes is very similar due to the fact that for 372 α-globin chain variants there is uncertainty as to which globin gene is mutated (see text for details).

Nature Genetics: doi:10.1038/ng.785

Supplementary Table 1

Microattribution Tables 1-4 for hemoglobinopathies variants available separately online (please see **Supplementary Note** for details).

Nature Genetics: doi:10.1038/ng.785

Supplementary Table 2

Outline of the different *KLF1* gene variants deposited in HbVar and their corresponding HbF levels. See also **Fig. 3**.

DNA change	Deduced protein change	No of cases	Hb F (%)	References / Source
Heterozygotes for *KLF1* gene variants				
NC_000019.9:g.12998102G>A	-	1	11.5	Microattribution
NC_000019.9:g.12998078T>C	-	1	1.3	Microattribution
NM_006563.2:c.380T>A	p.L127X	1	1.1	Microattribution
NM_006563.2:c.544T>C	p.F182L	1	12.0	Microattribution
NM_006563.2:c.809C>A	p.S270X	25	1.9 ± 0.6	Microattribution
NM_006563.2:c.862A>T	p.K288X	10	8.4 ± 5.6	Ref. 7
NM_006563.2:c.874A>T	p.K292X	1	1	Microattribution
NM_006563.2:c.954dupG	p.R319EfsX34	4	2.0 ± 1.1	Microattribution
NM_006563.2:c.973G>A	p.E325K	2	37, 40	Refs. 19, 20
NM_006563.2:c.977T>G	p.L326R	1	0.2	Microattribution
NM_006563.2:c.983G>A	p.R328H	1	2.9	Microattribution
NM_006563.2:c.994A>C	p.K332Q	4	1.1 ± 0.5	Microattribution
Control samples				
-	-	22	0.34 ± 0.17	-

Nature Genetics: doi:10.1038/ng.785

Supplementary Note

Database structure

The HbVar database of hemoglobin variants and thalassemia mutations is a publicly available LSDB which provides timely information to interested users including the globin research community, providers of genetic services and counselling, patients and their families, and pharmaceutical industries. The database is designed to allow regular entry updates and corrections, and has a user-friendly query interface that provides easy access to this information as an aid-in-diagnosis. Variant pathogenicity is established either directly from clinical findings or indirectly from linkage analysis, association, conservation and/or quantitative functional assays.

According to the literature available, variations in genes residing outside the human globin gene clusters are also implicated in producing or modifying the phenotypes associated with the hemoglobinopathies and thalassemias. We have therefore developed separate LSDBs for such genes. We have based these LSDBs on the Leiden Open-Access Variation database (LOVD) management system. We have developed LOVD-based LSDBs for a total of 37 genes, documenting some 1,941 unique genetic variants. Thirty-one of these LSDBs (*ALOX5AP, AQP9, ARG2, ASS1, ATRX, BCL11A, CNTNAP2, CSNK2A1, EPAS1, ERCC2, FLT1, GATA1, GPM6B, HAO2, HBS1L, KDR, KL, KLF1, MAP2K1, MAP3K5, MAP3K7, MYB, NOS1, NOS2, NOS3, NOX3, NUP133, PDE7B, SMAD3, SMAD6,* and *TOX*) correspond to genes encoding erythroid-expressed proteins, including transcription factors (e.g. *GATA1, KLF1*), chromatin associated co-

Nature Genetics: doi:10.1038/ng.785

factors (e.g. *ATRX*) and genes implicated in modulating HbF levels (e.g. *AQP9, ARG2, BCL11A, HBS1L, MAP3K5, MAP3K7, MYB, PDE7B).*

In order to bridge the wealth of information stored in HbVar with information deposited in the newly developed LOVD-based LSDBs that document genetic variation in other genes (referred to hereafter as HbVar-related LSDBs), we needed to overcome the different architecture and design of these two databases. Therefore, we developed a LOVD copy of HbVar that serves as an intermediate between the HbVar-related LSDBs and HbVar. The LOVD copy of HbVar consists of 6 separate LSDBs for the functional globin genes (*HBA2*, *HBA1*, *HBG2*, *HBG1*, *HBD*, *HBB*), that intercommunicate with HbVar. This is necessary since the current capacity of the LOVD system cannot accommodate the wealth of information (particularly phenotypic data) presently deposited in HbVar.

Contents of the Microattribution Tables

All genetic variation data leading to hemoglobinopathies and thalassemias have been collected in 4 publicly available Microattribution Tables (**Supplementary Table 1**) that have been centrally deposited to NCBI. The contents of these tables are as follow:

(a) The first table contains the information required for **submission of SNPs to the central depository**, namely dbSNP. A blank form can be obtained by downloading from dbSNP along with the information specifying a SNP. When processing variant submissions in their official nomenclature, NCBI 'stabilizes' the corresponding rs number for the variant and maintains the strand and orientation of the allele state.

Nature Genetics: doi:10.1038/ng.785

(b) The second table contains information needed for **microattribution**. It lists the variant nucleotide in the official nomenclature and the conventional, or common, variant name. For each variant, it includes both a local ID (e.g. from a LSDB) and an ID for the central repository (rs ID for dbSNP), as well as an ID for other general databases (OMIM variant ID and Swiss-Prot variant ID), when available. Each person involved in generating the biological information in the table (authors of papers, data submitters) can be listed as a unique ResearcherID, from the ResearcherID system of Thomson ISI. This unique identifier is assigned to each variant and corresponds to a unique data contributing entity, the latter being from individual researchers to international research consortia. This approach allows instant views of an author's citation metrics. Other systems that can be employed for this purpose include OpenID or Research Identification Primer.

(c) The third table provides **phenotype information**. It depicts the phenotypic outcome of the genetic variant (hemoglobin variant, thalassemic condition, HbF modulation), hematological indices (Hb, MCH, MCV, HbA_2, HbF), and clinical features (e.g. morphological alterations of erythrocytes, associated organ pathology). Depending on different diseases, these columns will differ accordingly.

(d) The fourth table gives **allele frequency information** about each variant. Each record gives an ethnic group and the frequency at which the variant is found (either as a count or a percentage). This, as well as the former, table has more than one row for a variant.

Nature Genetics: doi:10.1038/ng.785