Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels

Fabrice Danjou1,13, Magdalena Zoledziewska1,13, Carlo Sidore1–3, Maristella Steri1, Fabio Busonero1,2,4, Andrea Maschino1,2,4, Antonella Mulà1,3, Lucia Perseu1, Susanna Barella5, Eleonora Porcu1–3, Giorgio Pistis1–3, Maristella Pitfalls1, Mauro Pala1, Stephan Menzel6, Sarah Mettrudy7, Timothy D Spector7, Lidia Leoni8, Andrea Angius1,8, Manuela Uda1, Paolo Moi5,9, Swee Lay Thein6,10, Renzo Galanello5,9,12, Gonçalo R Abecasis2,14, David Schlessinger11,14, Serena Sanna1,14 & Francesco Cucca1,3,14

We report genome-wide association study results for the levels of A1, A2 and fetal hemoglobins, analyzed for the first time concurrently. Integrating high-density array genotyping and whole-genome sequencing in a large general population cohort from Sardinia, we detected 23 associations at 10 loci. Five signals are due to variants at previously undetected loci: MPHOSPH9, PLTP-PCIF1, ZFPM1 (FOG1), NFIX and CCND3. Among the signals at known loci, ten are new lead variants and four are new independent signals. Half of all variants also showed pleiotropic associations with different hemoglobins, which further corroborated some of the detected associations and identified features of coordinated hemoglobin species production.

The provision of oxygen to tissues depends on hemoglobin, requiring the coordinated expression of several globin chains that form functional tetramers. An index of the importance of hemoglobin function is the evolutionary duplication and divergence of regulation of globin gene copies to adapt to stages of development and buffer the effects of mutational loss. In particular, at birth, a switch occurs from fetal hemoglobin (HbF) to hemoglobin A2 (HbA2) and hemoglobin A1 (HbA1), such that during adult life the hemoglobin forms comprise ~1% HbF, ~3% HbA2 and ~96% HbA1. The different hemoglobins all contain α-globin chains, encoded by two eponymous genes (HBA1 and HBA2) on chromosome 16. These chains aggregate with non-α-globin chains encoded by HBG1 and HBG2 (for γ-globin; HBF), HBD (for δ-globin; HBA2) and HBB (for β-globin; HBA1) genes in the β-globin gene cluster on chromosome 11 (Fig. 1). The molecular switch between fetal and adult hemoglobin occurs via the binding of transcription factors to regulatory DNA sequences controlling the expression of globin genes. In particular, the various genes in the β-globin cluster are sequentially activated during ontogeny, such that time-specific expression patterns follow the genomic order.

Inherited disorders of hemoglobin, such as β-thalassemia, which is caused by mutations at the HBB (hemoglobin β) locus, represent the most common monogenic disorders worldwide. The prevalence of these disorders is highest in areas where malaria was or remains endemic. The severity of inherited hemoglobin disorders is also variable, from severe, lifelong transfusion-dependent anemia to mild anemia that does not require transfusion, depending on the molecular defect and genotype status as well as ameliorating variants in modifier genes. Therefore, studying the genetic regulation of hemoglobin levels might identify new factors involved and mechanisms to optimize strategies for therapy.

The large heritable contribution to phenotypic variance in HbA2 and HbF in the general population (0.728 and 0.633, respectively; Online Methods and a previous report) indicates that genetic analyses could lead to new insights. In genome-wide association studies (GWAS), two genomic regions, the β-globin gene cluster and the HBS1L-MYB locus, have been associated at a genome-wide significance level with variations in the amount of HbA2 (ref. 5), and only these loci and BCL11A have been associated with HbF levels. Variants at all four loci are powerful modifiers of the severity of β-thalassemia and sickle-cell disease. Notably, none of the variants associated with HbA2 or HbF levels have been found to be associated with total hemoglobin levels, even in the largest meta-analysis of over

Received 9 December 2014; accepted 23 April 2015; published online 14 September 2015; doi:10.1038/ng.3307

1Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche (CNR), Monserrato, Cagliari, Italy. 2Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA. 3Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA. 4Ospedale Regionale per le Microcitemie, ASL8, Cagliari, Italy. 5Department of Molecular Hematology, King’s College London, London, UK. 6Department of Twin Research and Genetic Epidemiology, King’s College London, London, UK. 7Center for Advanced Studies, Research and Development in Sardinia (CRS4), Parco Scientifico e Tecnologico della Sardegna, Pula, Italy. 8Department of Public Health and Clinical and Molecular Medicine, University of Cagliari, Cagliari, Italy. 9Department of Hematological Medicine, King’s College Hospital National Health Service (NHS) Foundation Trust, London, UK. 10Laboratory of Genetics, National Institute on Aging, US National Institutes of Health, Baltimore, Maryland, USA. 11Department of Twin Research and Genetic Epidemiology, King’s College London, London, UK. 12Deceased. 13These authors contributed equally to this work. 14These authors jointly supervised this work. Correspondence should be addressed to F.D. (fabrice.danjou@irgb.cnr.it) or F.C. (fcucca@uniss.it).
135,000 individuals\textsuperscript{11}. This indicates that, in analyses of total hemoglobin levels, the association signals for the subtypes are diluted and possibly obscured by opposite directions of effect. Currently, most of the heritability for variance in HbF and HbA2 levels also remains to be explained, and variation in HbA1 levels has not been specifically assessed by GWAS.

A promising source of genetic data to extend analyses is the founder Sardinian population, in which associations have previously been detected in a large cohort through the analysis of data from genotyping arrays bearing common and ubiquitous variants\textsuperscript{7}. Here we extend these analyses to rare and Sardinian-specific variants inferred from whole-genome sequencing of 2,120 Sardinians (Supplementary Fig. 1 and Supplementary Note). Furthermore, analyzing variants modulating HbA1, HbA2 and HbF levels concurrently in a single cohort provides a route to assess associations that overlap for different hemoglobin forms without the need to account for differences in study size, ancestry or measurements.

RESULTS

To test for genetic associations with the levels of HbA1, HbA2 and HbF, we interrogated ~10.9 million SNPs, genotyped or imputed in 6,602 general population volunteers of the SardiNIA longitudinal study\textsuperscript{4} (Online Methods and Supplementary Table 1).

Initial analyses showed a predominant role for the HBB c.118C>T mutation introducing a premature stop codon (p.Gln40\textsuperscript{*)}, better known as the $\beta^{039}$ mutation, a variant common in Sardinia (rs11549407: allele frequency = 4.8%). This mutation results in the classical significance threshold of $P = 5.9 \times 10^{-10}$). Because its effect had been established previously\textsuperscript{7,12}, we considered this mutation and other rarer $\beta^0$ mutations in $\beta$-thalassemia known in Sardinia as covariates (Online Methods and Supplementary Table 2). The assessed individuals in the cohort included 664 healthy heterozygous carriers of $\beta^0$ mutations but no individuals with $\beta$-thalassemia.

The genome-wide scan identified 23 unique variants at 10 loci with associated $P < 1.0 \times 10^{-200}$). Which gene is truly associated and how it affects hemoglobin production remain unclear, although, among the top associated SNPs, a variant in an intron of ARL6IP4 (chr12:123465483) falls in a highly conserved region rich in putative transcription factor binding sites and had the highest score for in silico prediction of deleterious impact on function (combined annotation-dependent depletion (CADD) score\textsuperscript{13}, as detailed in Supplementary Table 5. Although this association was just below the more stringent empirical threshold of significance, it was further strengthened by independent association with another hemoglobin form (HbA2, $P = 5.9 \times 10^{-10}$) (Table 1).

For HbA2, we identified three new signals. One, rs141006889, is a missense variant located in ZFPMT1, a gene also known as FOG1 that encodes a cofactor of the hematopoietic transcription factors GATA1 and GATA2 (ref. 15; Supplementary Fig. 4). The complexes formed by FOG1 and GATA proteins are essential for normal erythroid differentiation\textsuperscript{15}, as demonstrated by pathogenic mutations that abrogate the FOG1-GATA interaction to cause familial dyserythropoietic anemia and thrombocytopenia\textsuperscript{16}. Another signal was defined by a pair of statistically indistinguishable variants, rs113267280 and rs112236323 (P = 1.11 \times 10^{-29} and 1.29 \times 10^{-29}, respectively).

New associations at newly associated loci

New associations were detected for all three hemoglobin forms. For HbA1, we observed a signal led by chr12:123681790 (in an intron of MPHOSPH9), encompassing several SNPs in complete linkage disequilibrium (LD) in a region with several genes (Supplementary Fig. 3). Which gene is truly associated and how it affects hemoglobin production remain unclear, although, among the top associated SNPs, a variant in an intron of ARL6IP4 (chr12:123465483) falls in a highly conserved region rich in putative transcription factor binding sites and had the highest score for in silico prediction of deleterious impact on function (combined annotation-dependent depletion (CADD) score\textsuperscript{13}, as detailed in Supplementary Table 5. Although this association was just below the more stringent empirical threshold of significance, it was further strengthened by independent association with another hemoglobin form (HbA2, $P = 5.9 \times 10^{-10}$) (Table 1).

For HbA2, we identified three new signals. One, rs141006889, is a missense variant located in ZFPMT1, a gene also known as FOG1 that encodes a cofactor of the hematopoietic transcription factors GATA1 and GATA2 (ref. 15; Supplementary Fig. 4). The complexes formed by FOG1 and GATA proteins are essential for normal erythroid differentiation\textsuperscript{15}, as demonstrated by pathogenic mutations that abrogate the FOG1-GATA interaction to cause familial dyserythropoietic anemia and thrombocytopenia\textsuperscript{16}. Another signal was defined by a pair of statistically indistinguishable variants, rs113267280 and rs112236323 (P = 1.11 \times 10^{-29} and 1.29 \times 10^{-29}, respectively),

Figure 1 Association at the globin gene clusters. (a, b) Schematics of the association results in the genomic context of the $\beta$-globin gene cluster (a) and $\alpha$-globin gene cluster (b) gene clusters. For each hemoglobin form, the positions of the associated markers are indicated, with plus and minus signs indicating an increase or decrease in the levels of the corresponding hemoglobin form with the effect allele (as in Table 1). The symbol is larger if the marker is associated at the genome-wide level or smaller if the association results from the analysis of pleiotropic effects. The $\beta^{039}$ mutation and $\alpha$-3.7 type I deletion as well as relevant genes and the locus control region hypersensitivity (HS) sites are indicated. Finally, at the bottom of each panel is presented the LD ($r^2$) profile for the region in Sardinia, with colors representing a range of LD from high (red) to intermediate (green) to low (blue).
located in the CCND3 gene, whose product, cyclin D3, is thought to be critical for erythropoiesis\(^1\). Knockdown of Cend3 correlates with a reduction in the number of cell divisions during terminal erythropoiesis, thereby resulting in the production of fewer red blood cells that are larger in size\(^1\). These variants are also in partial LD with rs9349205 \((r^2 = 0.40)\), a SNP previously associated with mean red blood cell volume and number (Supplementary Table 6), which is located 160 bp away from rs112233623 in the same
determination analysis including as covariates that are the second category.

Table 1: Most significant independent association results from single-variant tests for A1, A2 and fetal hemoglobin

| Trait (units) and locus number | Candidate gene | Chromosome: position | rsID from dbSNP | Alleles (EA/OA) | RSQR | EAF | Effect (SE) | P value |
|--------------------------------|----------------|----------------------|-----------------|-----------------|------|-----|-------------|--------|
| HbA1 (g/dl)                    | α-globin gene cluster (p, o, b)| 16:149539\(^{a,d}\) | rs57001378\(^{1}\) | A/G             | 0.98 | 0.136 | −0.1995 (0.023) | 5.86 × 10\(^{-18}\) |
| HbA2 (g/dl)                    | α-globin gene cluster (p, o, b) | 16:391593\(^{a,c,e}\) | – | T/C | 0.94 | 0.012 | −0.4028 (0.058) | 3.28 × 10\(^{-12}\) |
| HbF (g/dl)                     | FAM3A (p); G6PD (p, c, o, b); INBK (p) | 15:3762634\(^{a,d}\) | rs5030868 | A/G | Genotyped | 0.085 | −0.1256 (0.019) | 2.78 × 10\(^{-11}\) |
| Locus 3\(^{a}\)                | MPhosph9 (p) | 12:123681790\(^{1}\) | – | A/C | 0.96 | 0.010 | −0.3606 (0.064) | 1.68 × 10\(^{-8}\) |
| HbA2                           | α-globin gene cluster (p, o, b) | 16:216593\(^{a,c,e}\) | rs141494605 | C/T | 0.97 | 0.149 | −0.3080 (0.025) | 3.94 × 10\(^{-35}\) |
| Locus 2\(^{a,c,e}\)            | α-globin gene cluster (p, o, b); HBG1 (p), LUC7L (p) | 16:391593\(^{a,c,e}\) | – | T/C | 0.94 | 0.012 | −0.5112 (0.063) | 6.48 × 10\(^{-16}\) |
| HbA2                           | α-globin gene cluster (p, o, b); AXIN1 (p) | 16:342218\(^{a,c,e}\) | rs148706947 | T/C | 0.93 | 0.021 | 0.2892 (0.051) | 1.04 × 10\(^{-8}\) |
| Locus 3\(^{b}\)                | CCND3 (p, b) | 6:41952511\(^{1}\) | rs113267280 | G/T | 0.99 | 0.101 | 0.2932 (0.026) | 1.11 × 10\(^{-29}\) |
| Locus 4\(^{b}\)                | MYB (b) | 6:135418916 | rs7776054 | G/A | Genotyped | 0.210 | 0.1762 (0.020) | 3.71 × 10\(^{-19}\) |
| Locus 5\(^{b}\)                | CTSA (p); PHF1 (p, c); PLTP (p, e); MMP9 (p); TNCC2 (p) | 20:44547672\(^{1}\) | rs59329875 | C/T | 1.00 | 0.134 | −0.1399 (0.024) | 3.64 × 10\(^{-9}\) |
| Locus 6\(^{b}\)                | FGFI1 (p, b, c); C16orf85 (p) | 16:88601281\(^{b}\) | rs141006889 | G/A | Genotyped | 0.007 | −0.5074 (0.087) | 5.33 × 10\(^{-9}\) |
| HbF (g/dl)                     | BCL11A (p, o, b) | 2:60720951 | rs4671393 | A/G | 1.00 | 0.136 | 0.578 (0.023) | 2.60 × 10\(^{-130}\) |
| Locus 2\(^{d}\)                | MYB (b) | 6:135419018 | rs9399137 | C/T | Genotyped | 0.205 | 0.4202 (0.020) | 1.09 × 10\(^{-93}\) |
| HbA2                           | HBS11 (p, c, e); ALDH2A1 (a) | 6:135356216 | rs11754265 | C/G | 1.00 | 0.367 | −0.1421 (0.021) | 0.04 × 10\(^{-12}\) |
| Locus 3\(^{d}\)                | α-globin gene cluster (p, o, b); HBG1 (p, b); HBG2 (a) | 11:5290370\(^{a}\) | rs87365838 | G/C | 1.00 | 0.236 | 0.2038 (0.019) | 1.09 × 10\(^{-25}\) |
| HbA2                           | α-globin gene cluster (p, o, b); HBG1 (p, b); HBG2 (a) | 11:5277236\(^{a}\) | rs2855122 | C/T | 1.00 | 0.395 | −0.1458 (0.022) | 2.57 × 10\(^{-11}\) |
| Locus 4\(^{a,e}\)              | NFX1 (p) | 19:13121899\(^{a}\) | rs183437571 | T/C | 0.97 | 0.010 | 0.4607 (0.081) | 1.61 × 10\(^{-8}\) |

The table shows the most significant association results from single-variant tests for A1, A2 and fetal hemoglobin. New results are shown in BOLD. Each locus is indicated in the column (EA/OA) RSQR EAF Effect (SE) P value HbA1 HbA2 HbF

\(\alpha\) Association results locally corrected for −3.7 deletion type i (g.34164,37967del3804, NG_000006.1) (Supplementary Note).\n
\(\beta\) First time associated with the trait and in a new locus.

\(\gamma\) Signal refinement at a previously reported signal.\n
\(\delta\) Result not found using the 1000 Genomes Project reference panel.

\(\varepsilon\) Variant refining previously reported signals.
**Figure 2** Diagrams of genome-wide associated loci. (a,b) Representations of genome-wide significant findings for hemoglobin levels in relation to their contribution to the phenotypic variation (variance explained) (a) and their individual impact (effect size) (b). At each level, the length of the black bar represents the magnitude of the variance explained (a) or the effect size (b) for each trait, locus, gene and variant. The bars are connected by colored bands to their subcomponents (loci for each trait, genes for each locus, variants for each gene). Three colors (yellow, green and blue) represent the three hemoglobin forms (HbA1, HbA2 and HbF, respectively); for loci or genes affecting more than one hemoglobin, gray indicates effects shared by HbA1 and HbA2, cyan indicates effects shared by HbA2 and HbF, and light gray indicates effects shared by all three hemoglobin forms. Each panel is drawn to show loci in order of their importance, from the largest to smallest amount of explained phenotypic variance (a) or effect size (b). The variance explained by each locus was calculated by fitting a regression model including all variants at that locus, and the effect size for each locus is the sum of the effect sizes of all the variants in that locus (supplementary table 3 reports the effect sizes for such joint models). For variants associated with more than one trait, the maximum value was used. Markers are reported using chromosome:position notation when an rsID was not available; when an intergenic region was involved, we show nearby genes in parentheses instead of indicating a single gene for the locus.
erythroid-specific enhancer functionally associated with CCND3 (refs. 18–20). rs112233623 was also the associated variant with the highest CADD score (Supplementary Table 5).

An additional variant related to HbA2, rs59329875, was observed for the first time in this study. It is situated between PLTP, which has been associated with several plasma lipoprotein and triglyceride levels21–24, and PCIF1, which is thought to negatively regulate gene expression by RNA polymerase II (ref. 25).

For HbF, we identified one new variant associated with its levels: rs183437571, located on chromosome 19 in an intron of NFIX, which encodes a CCAAT-binding transcription factor. The association for this variant was just below the empirical significance threshold of $P = 1.4 \times 10^{-8}$ but is supported by considerable biological evidence implicating the gene and the surrounding region in hemoglobin regulation. Specifically, rs183437571 falls in a CpG region that is differentially methylated in fetal and adult red blood cell progenitors26. In mice, Nfix was recently identified as one of the regulatory factors with relatively restricted expression in hematopoietic stem cells27 and was required for the survival of hematopoietic stem and progenitor cells during stress hematopoiesis28. Intriguingly, NFIX is situated in a region of ~300 kb that encompasses a number of genes involved in erythropoiesis (DNASE2 and KLF1)29–33 or otherwise associated with red blood cell traits, including mean corpuscular hemoglobin levels (SYCE2, FARS3 and CALR)31 (Supplementary Fig. 5 and Supplementary Table 6). KLF1 is a particularly interesting candidate gene33,34, but the mutations observed in previous studies35 were not found and the gene itself is situated in an LD block distinct from the one corresponding to our association signal. However, long-distance regulatory interactions remain a possibility.

Of the five new signals, the discovery of chr12:123681790 for HbA1, rs141006889 for HbA2 and rs183437571 for HbF was strongly influenced by the assessment of variants from Sardinian whole-genome sequencing. Specifically, chr12:123681790 was missing from 1000 Genomes Project phase 3 data36, and, when using this public reference panel, the signal was misplaced at another variant ~1 Mb away; rs141006889 was included in the design of one genotyping array (ExomeChip) after it was identified through our sequencing effort in the TwinsUK cohort enrolled from the UK general population43. For two loci, both associated with HbA2 levels, we successfully replicated the associations seen in Sardinia. Specifically, we observed a $P$ value of $6.98 \times 10^{-6}$ for rs59329875 in the PLTP-PCIF1 intergenic region (MAF = 0.18) and a $P$ value of $1.73 \times 10^{-4}$ for rs113267280 in CCND3 (MAF = 0.01). The rarity of other variants precluded replication. The MPHOSPH9 and FOG1 variants associated with HbA1 and HbA2 levels, respectively, are missing in publicly available imputation panels, and rs183437571 in NFIX, which associated with HbF levels in Sardinia, was imputed as monomorphic in the TwinsUK cohort (Table 2 and Online Methods).

Fine mapping at known loci

The integration of variants from whole-genome sequencing in the scan was also instrumental in refining signals at previously known loci, either identifying a better lead variant or indicating new independent signals. Specifically, we refined the associations within the α- and β-globin gene clusters for all three hemoglobins; the association of the HBS1L-MYB intergenic region with HbA2 and HbF; and the association of the BCL11A gene with HbF.

### Table 2 Replication of new loci

| Trait (units) | SNP | Candidate gene | INFO score | Alleles (EA/OA) | EAF | Effect (SE) | $P$ value |
|--------------|-----|----------------|------------|----------------|-----|-------------|---------|
| HbA1 (g/dl) | Locus 3 | chr12:123681790 | MPHOSPH9 | – | – | – | – |
| HbA2 (%)    | Locus 3 | rs113267280 | CCND3 | 0.843 | G/T | 0.011 | 0.442 (0.118) | $1.73 \times 10^{-4}$ |
|             | Locus 5 | rs59329875 | PLTP-PCIF1 | 0.994 | C/T | 0.185 | 0.132 (0.029) | $6.98 \times 10^{-6}$ |
|             | Locus 6 | rs141006889 | FOG1 | – | – | – | – |
| HbF (%)     | Locus 4 | rs183437571 | NFIX | 0.294 | T/C | 0.000 | – |

The table describes association in the TwinsUK cohort ($N = 4,131$ individuals). For each SNP, we indicate the associated hemoglobin tested, the imputation accuracy according to the IMPUTE INFO metric, the effect allele tested for association (EA) and the other allele at the SNP (OA), the SNP effect allele frequency (EAF) and the regression coefficients (with standard error, SE). The last column explains why some SNPs were not tested. The information in the first column is presented as in Table 1.
The associations within the β-globin gene cluster were intricate. As reported above, the strongest modifier in this region is the HBB β39 variant, which acts on all three hemoglobin types (Fig. 1, Online Methods and Supplementary Table 2). Multiple additional independent signals were observed in conditional analyses for HbA2 and Hbf, but they were distinct for each hemoglobin type, highlighting different regulatory patterns within the β-globin gene cluster. Specifically, for HbA2, we confirmed two known independent associations at missense mutations in the HBD gene (rs35152987 and rs35406175); the latter was perfectly tagged by our lead signal (Supplementary Table 5). In addition, we identified three new independent signals (rs12793110, rs11036338 and rs7936823) within an LD block encompassing the HBB gene, confirming a controlling role for this region in HbA2 production (Fig. 1 and Supplementary Fig. 4). For Hbf, two new independent signals were detected in a separate LD block of the β-globin gene cluster (Fig. 1 and Supplementary Fig. 5). The first, situated in an intron of the HBE1 gene (rs67385838), remained associated even when taking into account 43 other variants in the β-globin gene cluster associated with hemoglobin variation (Supplementary Note). The second was located in a cyclic AMP response element upstream of HBG2 (rs2855122) already implicated in drug-mediated Hbf induction by butyrate46; the different features of this marker make it a strong candidate in modulating fetal to adult hemoglobin switching (Supplementary Note).

At the α-globin gene cluster, two variants were associated with HbA1 and three variants were associated with HbA2, of which one affected both traits (Fig. 1 and Table 1). All results at this locus were corrected for any effect of the most frequent α-globin gene deletion present in Sardinia (g.34164_37967del3804 (NG_000006.1), known as α-3.7 deletion type I), directly genotyped in a subset of the volunteers and imputed for the rest of the cohort (Online Methods). This deletion was associated at the genome-wide level with both HbA1 and HbA2 levels but only nominally with Hbf levels (Table 1 and Supplementary Table 2). The most strongly associated signals (rs570013781 and rs141494605) were situated within the NRP3 and HBM genes, affecting HbA1 and HbA2 levels, respectively. NRP3L3 contains several hypersensitive sites involved in regulation of the α-globin genes. HBM encodes a globin member of the avian α-D family45; its expression is highly regulated in human erythroid cells, although the protein has not been detected in human erythroid tissues. These observations suggest a possible regulatory function for which high-level protein expression would not be required45. An independent variant associated with HbA1 and HbA2 (chr16:391593) was observed in the AXIN1 gene, in which a further independent SNP (rs148706947) was found to be associated with HbA2 alone (Supplementary Figs. 3 and 4).

We also examined variants in the HBS1L-MYB intergenic region known to be associated with Hbf and HbA2 levels5. We confirmed the role of the known variant (rs66650371, a TAC deletion) on the expression of both forms of hemoglobin46,47 (Supplementary Note). A further new independent signal for Hbf was found at rs11754265 in an intron of HBS1L, which has been shown to be a much stronger expression quantitative trait locus (eQTL) than rs66650371 for HBS1L and the neighboring ALDH8A1 gene in monocytes48.

In line with previous studies49,50, the second intron of BCL11A gave multiple signals of association with Hbf levels. These can be explained by the joint action of variants from each of two independent groups of statistically indistinguishable SNPs: one group constitutes rs4671393, rs766432 and rs1427407, with P values between 2.6 × 10^-130 and 5.6 × 10^-129, and the other group constitutes rs13019832 and rs7606173, with P values of 6.1 × 10^-33 and 9.1 × 10^-33, respectively, in our cohort. The most likely causal candidate in the first group is rs1427407, a variant already associated with Hbf levels in other population cohorts and functionally associated with BCL11A regulation51. In the second group, we can point to rs13019832 as the candidate causal variant, as this SNP had the highest functional CADD score (Supplementary Table 5). This variant has also been correlated in adipose tissue with the methylation of a CpG site (cg23678058) in a region that is functionally associated with BCL11A expression52 and shows evidence of an effect on GATA-1 binding in peripheral blood–derived erythroblasts53,54.

Pleiotropic effects

Among our 23 lead variants, 6 were associated (with at least P < 0.01) with a second hemoglobin type, and another 6 (including β39 and −α 3.7 deletion type I) were associated with all 3 hemoglobin forms (Fig. 1 and Table 1). Overall, all but three pleiotropic variants modulated different hemoglobins in the same manner, that is, with the same allele increasing the levels of all associated hemoglobins. The three exceptions included the β39 variant, which decreased HbA1 levels while increasing HbA2 and Hbf levels, and two SNPs mapping to the β-globin gene cluster, both affecting HbA2 and Hbf levels but in opposite directions (Fig. 1 and Table 1). Moreover, many of the additional suggestive signals were associated with more than one hemoglobin type, increasing the likelihood that they are true signals (Online Methods). In fact, 14 of these variants—all sharing effects on HbA1 and HbA2 levels but none having effects on Hbf levels—showed between-trait combined P values that were genome-wide significant (Supplementary Table 9) and hint at additional pathways of potential interest in hemoglobin dynamics.

In general, the extended number of genetic variants showing joint association with HbA1 and HbA2 rather than Hbf is consistent with high correlations of the levels of the adult hemoglobins HbA1 and HbA2 but only partial correlations of these hemoglobin forms with the levels of Hbf (Online Methods).

Given the central role of hemoglobin in providing oxygen to body tissues and the substantial fraction of total body cells accounted for by circulating red blood cells, factors influencing hemoglobin production and red blood cell count unsurprisingly have pleiotropic effects on other non-hematological traits. This is exemplified by the strong impact of the major β39 mutation on cholesterol and low-density lipoprotein (LDL) cholesterol levels (see the companion paper13). Here we extended the analysis for this mutation to 69 non-hematological quantitative traits selected from among those assessed in the SardiNIA cohort (Supplementary Note). We found that this variant was also significantly associated with increased total white blood cell counts ($P = 3 \times 10^{-7}$), with the major contribution coming from neutrophil counts ($P = 1 \times 10^{-6}$), and platelet counts ($P = 9 \times 10^{-5}$) (Supplementary Table 10).

DISCUSSION

We provide evidence for 23 associated variants at 10 loci influencing the levels of one or more of the 3 hemoglobin species measurable in postnatal life. Our results are based on a cohort from the Sardinian founder population that is much larger than the samples in previously described GWAS for Hbf and HbA2, and the analysis interrogates a high-resolution genetic map based on population sequencing that expanded the assessed spectrum of allelic variants by tenfold in comparison to previous studies. The finding that two of the five newly reported loci were not detectable without using the SardiNIA reference panel and that the others were mislocalized (Table 1 and Supplementary Table 7) further highlights how large-scale sequencing efforts in this founder population can identify
functionally relevant variants that may be very rare and hence missed in other populations.

For the same reasons, however, replication of the results for such variants or translation of findings directly into other populations is difficult. For example, the other currently reported sample of comparable size, from the UK, could provide replication only for the two variants present there. Similar limitations will likely be found in other GWAS designed to detect the effects of rare and founder variants. However, additional corroboration of our findings for such variants comes from their independent associations with other hemoglobin species and hematological traits in Sardinians and also from the biological functions of the genes involved. For instance, variant chr1:123681790 in MPHOSP9, associated with HbA1 levels, also shows suggestive evidence of association with HbA2 levels. The variant in FOG1, which is very rare in Europeans (MAF = 0.4%), is a missense variant in a gene implicated in erythropoiesis, and the variant in NFIX, which is absent in other European populations, falls within a cluster of genes involved in erythropoiesis and in a CpG region differentially methylated in fetal and adult red blood cell progenitors.

By carrying out GWAS analysis for HbA1, HbA2 and HbF levels assessed, to our knowledge, for the first time in the same individuals, we see a wide range of pleiotropic effects of variants across the three hemoglobin types (Table 1). Strikingly, HbA2 is associated with more than half of the loci discovered here (Fig. 2), with many of the loci having pleiotropic effects on HbA1 and some having such effects on HbF. Thus, although HbA2 has a minor role in the transport of oxygen to tissues, variations in its levels participate in pathways that regulate the amounts of the other hemoglobins active in postnatal life.

The direction of the pleiotropic effects among the different hemoglobin types provides some additional clues to mechanism. Within the α-globin gene cluster, in agreement with the presence of α-globin chains in HbA1, HbA2 and HbF, all variants affecting more than one hemoglobin show the same direction of effect on all. The regulation of globin chains from the β-globin gene cluster, however, is more complicated. It involves variants with the same direction of effect for all hemoglobins (rs7936823) and other variants most likely involved in switching mechanisms that affect fetal and adult hemoglobins in opposite directions (rs2855122). Still other variants change the kinetics of competition among α- and δ-globin chains; for example, the β39 mutation decreases δ-globin levels and thereby increases the availability of α-globin chains for combination with δ- and γ-globins, leading to higher levels of HbA2 and HbF.

Variants influencing only two forms of hemoglobin acted mainly in the same direction and never jointly affected HbA1 and HbF levels. Among these variants, those shared only by HbA2 and HbF can be attributed to specific cis-regulatory mechanisms in the β-globin gene cluster (rs12793110 and rs7944544) or to loci with a role in erythroid differentiation (CCND3 and MYB). By contrast, variants shared by HbA2 and HbA1 either acted in trans (in MPHOSPH9) or were localized to the α-globin gene cluster but with effect sizes probably too small to influence HbF production. Consistent with the latter possibility, α-3.7 deletion type I, which has strong genome-wide significant effects on HbA1 and HbA2, had much smaller only suggestive effects on HbF (Supplementary Table 2).

Our analyses also detected broader pleiotropic effects, most strikingly for the β39 variant. In addition to the effects on LDL cholesterol described in the companion paper13, we report for the first time, to our knowledge, that the β39 mutation is also significantly associated with increased total counts of white blood cells (and some subsets thereof) as well as platelet counts. These findings suggest that in heterozygous carriers this variant drives a broader increase in bone marrow-derived blood cell counts. Speculatively, some of these effects, such as augmented leukocyte and neutrophil counts, might have provided protection against pathogens in addition to those causing malaria, thus increasing selection for the balanced polymorphism.

The detected variants provide candidate modifiers influencing the clinical status of patients with monogenic hemoglobin disorders. For example, we carried out a preliminary analysis of a small sample of 306 patients with β-thalassemia homozygous for the β39 stop-gain mutation but showing very great heterogeneity in disease presentation and course. In addition to the variants described previously7–10, some variants detected in this study showed possible effects as modifiers of disease severity (Supplementary Note). However, the potential of these variants in helping to predict disease severity remains tentative without studies of larger sample sets. Nevertheless, the variants already add to the candidate targets for therapeutic intervention in the widely prevalent inherited β-thalassemia and other hemoglobinopathies.

URLs. SardNIA project, https://sardinia.irp.nia.nih.gov/; 1000 Genomes Project, http://www.1000genomes.org/; HumanExome BeadChip design, http://genome.sph.umich.edu/wiki/Exome_Chip_Design; Immunochip, http://www.t1dbase.org/poster/the-immunochip-custom-genotype-array/; Cardio-MetaboChip, http://csg.sph.umich.edu/kang/MetaboChip/; Human OmniExpress BeadChip, http://www.illumina.com/applications/genotyping/human-genotyping-arrays/omni-arrays.html; GenomeStudio software, http://www.illumina.com/applications/microarrays/microarray-software/genomestudio.html; MACH software, http://csg.sph.umich.edu/abecasis/MACH/; Minimac software, http://genome.sph.umich.edu/wiki/Minimac; zCall software, https://github.com/jigold/zCall/; IMPUTE v2 software, http://mathgen.stats.ox.ac.uk/impute/impute_v2.1.0.html; Merlin (including Merlin-regress and Merlin-offline), http://csg.sph.umich.edu/abecasis/merlin/; EPACTS software, http://csg.sph.umich.edu/abecasis/epacts/; Metal software, http://csg.sph.umich.edu/abecasis/metal/; GWAS catalog, http://www.genome.gov/gwastudies/; GRAIL software, https://www.broadinstitute.org/mpg/grail/; Gene Ontology, http://geneontology.org/; Human Expression Atlas at BioGPS, http://biogps.org/; Pritchard eQTL browser, http://eqtl.uchicago.edu/; R project, http://www.r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work is dedicated to Antonio Cao, Renzo Galanello and Maurizio Longinotti, who devoted their scientific lives to understanding, preventing and treating hematological diseases in Sardinia. We are also grateful to M.S. Ristaldi and M.G. Marini for knowledge and insight that they freely shared with us. Finally, we thank all the volunteers who generously participated in this study and made this research possible. The SardNIA study was funded in part by the US National Institutes of Health (National Institute on Aging, National Heart, Lung, and Blood Institute, and National Human Genome Research Institute). This research was supported by National Human Genome Research Institute grants HG005581, HG005552, HG006513 and HG007022; by National Heart, Lung, and Blood Institute grant HL117626; by the Intramural Research Program of the US National Institutes of Health, National Institute on Aging, contracts N01–AG–1-2109 and HHSN271201100005C; by Sardinian Autonomous Region (L.R. number 7/2009) grant cRP3-154; by grant FallBio2011 “Farmaci e Reti Biotecnologiche di Qualità”; and by the P05 InterOMics MIUR Flagship Project. The TwinsUK study was funded by the Wellcome Trust; the European Community’s Seventh Framework Programme
AUTHOR CONTRIBUTIONS
G.R.A., D.S. and F.C. conceived the study. F.D., D.S., S.S. and F.C. drafted the manuscript. F.D., M.Z., M.U., P.M., S.L.T., G.R.A., D.S., S.S. and F.C. revised the manuscript. F.B., A. Maschio and A.A. performed sequencing experiments. L.P. performed genotyping of α3.7 deletion type 1. M. Pala created an automated pipeline to query the public eQTL repositories. P.M. and R.G. provided genotypes and phenotypic data for patients with β-thalassemia. S.B. and R.G. supervised the characterization of the hemoglobin in the SardiniA cohort. F.D. analyzed the cohort of patients with β-thalassemia. S. Menzel, T.D.S. and S.L.T. provided replication and clinical samples. M.Z. analyzed replicated samples. J.L. provided FT support for sequencing and genotype data processing and analyses. D.S. and F.C. supervised the study. All authors reviewed and approved the final manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.
ONLINE METHODS

Sample description. The population studied here included 6,921 individuals, representing >60% of the adult population of 4 villages in the Lanusei valley in Sardinia, Italy. These individuals are part of the SardiNIA project, a longitudinal study including genetic and phenotypic data for 1,257 multigenerational families with more than 37,000 relative pairs. Details of phenotype assessments for these samples have been published previously. All participants gave informed consent to study protocols, which were approved by the Sardinian local research ethic committees: Comitato Etico di Azienda Sanitaria Locale 8, Lanusei (2009/0016600) and Comitato Etico di Azienda Sanitaria Locale 1, Sassari (2171/CE)) and by the NIH Office of Human Subject Research as governed by Italian institutional review board approval.

For whole-genome sequencing, we selected 1,122 individuals from the SardiNIA study and 998 individuals enrolled in case-control studies of multiple sclerosis and type 1 diabetes in Sardinia. Genomes were sequenced to an average coverage of 4.16-fold. Details on the sequencing protocol, data processing and variant calling can be found elsewhere and in the companion paper. The 2,120 sequenced samples consisted of 695 complete and incomplete parent-offspring trios; to avoid over-representation of rare haplotypes during the imputation process, we considered only the parents from each trio—totaling 1,488 samples—when building our reference panel and in the companion paper for further details.

Part of the sequencing data used in this study are available through the database of Genotypes and Phenotypes (dbGaP), under the SardiNIA Medical Sequencing Discovery Project, study accession phs000313.v3.p2.

Genotyping and imputation. The four microarrays used to genotype the entire SardiNIA cohort were the Illumina Infinium HumanExome BeadChip, Immunochip, Cardio-Metabochip and HumanOmniExpress BeadChip. Genotyping was carried out according to the manufacturer’s protocols at the SardiNIA Project Laboratory (Lanusei, Italy), at the Technological Center—Porto Conte Ricerche (Alghero, Italy) and at the National Institute on Aging Intramural Research Program Laboratory of Genetics (Baltimore, Maryland, USA). Genotypes were called using GenoStudio (version 1.9.4) and refined using zCall (version 3) when considered using the q.emmax test in EPACTS, which implements a linear mixed-model procedure to correct for cryptic relatedness and population stratification by incorporating a genomic-based kinship matrix. The associations reported in Table 1 refer to the best P value obtained with either percentage or original measurement units for HbA2 and HbF. Notably, HbF signals always resulted in lower P values when considering measurements in grams per deciliter, whereas, for HbA2 analysis, this was the only case for rs141494605. All loci passed the genome-wide significance threshold for primary signals was also considered for independent signals. For loci where different independent signals were found, we also report the model parameters of the jointly associated variants in Supplementary Table 3. Finally, the lead variants and their surrogates \( r^2 > 0.90 \) were annotated with CADD scores and are reported in Supplementary Table 5.

Heritability and variance explained. We estimated heritability for the three hemoglobins using Merlin-regress on the same sample used for the GWAS. The estimates for normalized hemoglobin levels were 0.520 g/dl for HbA1, 0.728% for HbA2 (0.700 g/dl) and 0.633% for HbF (0.624 g/dl). We then calculated for each hemoglobin form the proportion of phenotypic variance explained by the associated lead variants. We measured this proportion as the difference in the \( R^2 \)-adjusted parameter estimated for the full and the basic models, where the basic model included only phenotypic covariates (age, age$^2$ and sex) and the full model also included all the independent SNPs associated with either percentage or original measurement units for HbA2 and HbF. The \( R^2 \)-adjusted values were calculated using a linear mixed-model procedure from the im kin() function in the kinship R package. The estimates were 0.240 g/dl for HbA1, 0.492% for HbA2 and 0.383% for HbF.

Characterization of \( \beta \) mutations. For the present study, we designed a TaqMan custom assay for the \( HBB \) c.118C>T nonsense mutation (rs11549407), also known as \( \beta^39 \) and genotyped 6,602 samples. Comparison of the TaqMan genotypes and the imputation results (rs11549407, RSQR = 0.92) showed an overall concordance rate of 98.8%. Also, we further sequenced all samples discordant between red blood cell index-based diagnosis (using mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), HbF percentage and HbA2 percentage) and TaqMan genotypes, using Sanger sequencing to determine any additional \( \beta \)-globin mutations different from \( \beta^39 \), thus identifying three carriers for the \( HBB \) c.20delA (rs63749819) mutation and one carrier for the \( HBB \) c.315+1G>A (rs33945777) mutation.

Characterization of the deletion in the \( \alpha \)-globin gene cluster. In Sardinia, three variants are known to be mainly responsible for \( \alpha \)-thalassemia: SNPs

fully correlated with HbA2 (g/dl) \( (p = 0.662; P < 1 \times 10^{-20}) \) and poorly correlated with HbF (g/dl) \( (p = 0.055; P = 3.44 \times 10^{-5}) \). Likewise, HbA2 and HbF were weakly positively correlated as percentage measures \( (p = 0.108; P = 4.08 \times 10^{-3}) \) and were even less correlated when measured in grams per deciliter \( (p = 0.066; P = 5.81 \times 10^{-3}) \), consistent with previous findings. Measurements were available for a subset of 6,305 individuals; descriptive statistics are reported in Supplementary Table 1. Association results were considered genome-wide significant when the \( P \) value was \( < 5 \times 10^{-8} \); however, we also note in the text variants that would not meet a threshold of \( P = 1.4 \times 10^{-8} \) that we introduced for sequencing-based GWAS carried out in Sardinians for variants with MAF >0.5% (see the companion paper).

Before association analyses, trait measures were normalized using inverse normal transformation; we also removed for all assessed traits outliers with values above 5% for HbF. Analyses were adjusted for age, age$^2$ and sex as well as for the presence of at least one of the three \( \beta \) mutations (\( \beta^39 \) (rs11549407), \( HBB \) c.20delA (rs63749819) and \( HBB \) c.315+1G>A (rs33945777)), all directly genotyped or sequenced. Regression coefficients for the \( \beta^39 \) variant—the most common in Sardinia, with 10.3% of the population being a carrier—are reported in Supplementary Table 2.
rs11033603 and rs41474145 and the deletion g.34164_37967del3804 (NG_000006.1); the latter was known as −α.3.7 deletion type I, is by far the most common. We did not observe the rarer rs11033603 and rs41474145 SNPs in our sequencing effort. To establish genotypes at the deletion site in the full cohort, we used an inference strategy combined with experimental data. Specifically, we first characterized the structural variant by PCR in 260 unrelated sequenced individuals randomly selected from the SARDINIA cohort. We calculated the relative coverage of the deleted region in the whole genome–sequenced samples by considering the ratio of read counts in the potentially deleted region (223,450–226,953 bp, excluding 150-bp boundaries) with read counts in the nearby region not subject to deletion (227,254–230,757 bp). We then identified coverage ratio thresholds that best predicted PCR genotypes for the deletion and used these thresholds to infer genotypes for the 2,120 sequenced individuals. We inserted the genotypes in the Sardinian reference panel and imputed the deletion genotype for the total SARDINIA cohort.

To assess the accuracy of imputation, we considered the best-guess genotypes and searched for Mendelian errors in families. The observed error rate was 0.58% over 1,193 parent-offspring pairs, consistent with high imputation precision. The association results reported in the manuscript for this locus are corrected for inferred −α.3.7 deletion type I dosages.

**Variant validation.** We validated all variants that showed genome-wide significant P values in the primary or conditional analysis that were not directly genotyped or had no surrogates (r² > 0.90) that were directly genotyped. We did not validate variant rs3019832 in BCL11A for HbF, which was highly linked with findings from previous reports (rs7606173). Validation was performed using Sanger sequencing or TaqMan genotyping, depending on the variant frequency, for five variants. We selected for each variant all individuals carrying the minor allele (heterozygous and homozygous) plus a random subset of individuals homozygous for the other allele (in all, 3,084 subjects were genotyped), except for rs141494605 and chr16:391593, for which we specifically selected worse imputation dosages (borderline RSQR values). In addition, for rs17525396, we used independent genotypes available for a subset of the cohort, derived from Affymetrix 6.0 arrays (Supplementary Table 4).

**Replication of variant effects.** Replication was performed in the TwinsUK cohort. Genotyping was performed using a combination of Illumina arrays (HumanHap300, HumanHap610Q, 1M-Duo and 1.2M-Duo 1M), and imputation was performed using the IMPUTE software package (v2) and 1000 Genomes Project haplotypes released on 16 June 2014 in the phase 1 integrated variant set release. Details on quality control filters are provided in the Supplementary Note. HbA2 concentrations and HbF percentages were obtained by HPLC, and F cells were enumerated by staining for intracellular HbF and subsequent flow cytometry. Measurements were available for 4,131 samples. Association analyses were performed with the Merlin-offline package in Merlin, to account for relatedness. To be consistent with analyses performed in the SARDINIA study, age, age² and sex were used as covariates and the trait measures were transformed using quantile normalization.

**Selection of candidate genes.** At each locus, we generated a list of genes to be considered as plausible candidates, including those that satisfied one of the following criteria: (i) genes that were located within 25 kb of the lead SNP (indicated by p in Table 1); (ii) genes with exonic variants (frameshift, stop codon, nonsynonymous or synonymous) along with splice-site and 5′ or 3′ UTR variants in LD (r² ≥ 0.8) with the lead SNP (indicated by c in Table 1); (iii) genes whose expression was modulated by the SNP itself or by an eQTL in LD (r² ≥ 0.8) with the top SNP (indicated by e in Table 1); (iv) genes with a clear biological function connected with the traits (indicated by b in Table 1); or (v) genes harboring variants responsible for Mendelian diseases, as reported in OMIM (indicated by o in Table 1). We searched for candidate genes in eQTL data using an automated pipeline querying 16 public eQTL repositories, including the Pritchard eQTL browser; only top SNP eQTLs or SNPs with a false discovery rate (FDR) < 0.05 were considered.

**Pleiotropy and gene connection analysis.** To characterize genome-wide significant results and to identify suggestively significant ones, we searched for shared effects on the different hemoglobin forms as well as evidence of connections between them. Specifically, for genome-wide significant markers, we have simply reported the effect direction for all traits with P < 0.01 when a marker was associated at the genome-wide level for one trait (Table 1). To identify candidates with suggestive P values between 1.00 × 10⁻⁸ and 5.00 × 10⁻⁸, we selected among the following: (i) markers with MAF > 0.5% and showing two-trait combined P values < 5 × 10⁻³ (P values were combined using inverse variance–weighted meta-analysis, as implemented in Metal software) and (ii) markers located in or near genes that demonstrated evidence of connections with genome-wide significant loci, either in PubMed (using a set with data from 2006 or earlier to avoid confounding by subsequent GWAS discoveries) or the Human Expression Atlas and Gene Ontology databases, as determined using GRAIL, and considering genes reported with multiple hypothesis–corrected P < 0.05. Using these criteria, we identified 21 further genes with biological connections to the genome-wide significant loci, reported in Supplementary Table 8, and 14 variants with combined P values between 2.68 × 10⁻⁸ and 1.18 × 10⁻¹, reported in Supplementary Table 9.