EUROPEAN HEMATOLOGY ASSOCIATION
haematologica
Journal of the European Hematology Association

Gene panel sequencing improves the diagnostic work-up of patients with idiopathic erythrocytosis and identifies new mutations

by Carme Camps, Nayia Petousi, Celeste Bento, Holger Cario, Richard R. Copley, Mary Frances McMullin, Richard vanWijk, WGS500 Consortium, Peter J. Ratcliffe, Peter A. Robbins, and Jenny C. Taylor

Haematologica 2016 [Epub ahead of print]

Citation: Camps C, Petousi N, Bento C, Cario H, Copley RR, McMullin MF, vanWijk R, WGS500 Consortium, Ratcliffe PJ, Robbins PA, and Taylor JC. Gene panel sequencing improves the diagnostic work-up of patients with idiopathic erythrocytosis and identifies new mutations. Haematologica. 2016; 101:xxx doi:10.3324/haematol.2016.144063

Gene panel sequencing improves the diagnostic work-up of patients with idiopathic erythrocytosis and identifies new mutations

Carme Camps,[1,2] Nayia Petousi,[3] Celeste Bento,[4] Holger Cario,[5] Richard R. Copley,[1,2] Mary Frances McMullin,[6] Richard vanWijk,[7] WGS500 consortium,[8] Peter J. Ratcliffe,[3] Peter A. Robbins,[9] and Jenny C. Taylor[1,2]

[1]National Institute for Health Research (NIHR) Comprehensive Biomedical Research Centre, Oxford, UK;
[2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK;
[3]Nuffield Department of Medicine, University of Oxford, Oxford, UK;
[4]Hematology Department, Centro Hospitalar e Universitário de Coimbra, Coimbra, Portugal;
[5]Department of Pediatrics and Adolescent Medicine, University Medical Center, Ulm, Germany;
[6]Centre for Cancer Research and Cell Biology, Queen's University, Belfast, UK.
[7]Department of Clinical Chemistry and Hematology, University Medical Center Utrecht, Utrecht, Netherlands;
[8]A list of members and affiliations is provided in the supplementary information; and
[9]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK

CC and NP contributed equally to this work
PJR, PAR and JCT jointly supervised this work
Corresponding author: Nayia Petousi, Henry Wellcome Building, Roosevelt Drive, OX3 7BN, nayia.petousi@ndm.ox.ac.uk

Running head: Gene panel for idiopathic erythrocytosis

Word count for abstract: 248; Word count for text: 4000;
Figure/table count: 7; Supplemental files: 1

Acknowledgements

The authors would like to thank the patients and their families who consented to this study, Melissa M. Pentony for the support provided with the management of Ion Torrent data and the Core and administration services at the Wellcome Trust Centre for Human Genetics, which are funded by the Wellcome Trust Core Award [090532/Z/09/Z]. This work was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre Oxford with funding from the Department of Health's NIHR Biomedical Research Centre's funding scheme. The WGS500 study was funded by the Wellcome Trust Core Award (090532/Z/09/Z) and a Medical Research Council Hub grant (G0900747 91070) to Peter Donnelly (director of the Wellcome Trust Centre of Human Genetics), the NIHR Biomedical Research Centre Oxford, the UK Department of Health's NIHR Biomedical Research Centres funding scheme and Illumina. NP is funded via a NIHR Clinical Lectureship. PJR is a member of the Ludwig Institute for Cancer Research.

Abstract

Erythrocytosis is a rare disorder characterized by increased red cell mass and elevated hemoglobin concentration and hematocrit. Several genetic variants have been identified as causes for erythrocytosis in genes belonging to different pathways including oxygen sensing, erythropoiesis and oxygen transport. However, despite clinical investigation and screening for these mutations, the cause of disease cannot be found in a considerable number of patients, who are classified as having idiopathic erythrocytosis.

In this study, we developed a targeted next generation sequencing panel encompassing the exonic regions of 21 genes from relevant pathways (~79Kb) and sequenced 125 patients with idiopathic erythrocytosis. The panel effectively screened 97% of coding regions of these genes, with an average coverage of 450X. It identified 51 different rare variants, all leading to alteration of protein sequence, with 57 out of 125 cases (45.6%) having at least 1 of these variants. Ten of these were known erythrocytosis-causing variants, which had been missed following existing diagnostic algorithms. Twenty-two were novel variants in erythrocytosis-associated genes (*EGLN1*, *EPAS1*, *VHL*, *BPGM*, *JAK2*, *SH2B3*) and in novel genes included in the panel (e.g. *EPO*, *EGLN2*, *HIF3A, OS9*), some with high likelihood of functionality, for which future segregation, functional and replication studies will be useful in providing further evidence for causality. The rest were classified as polymorphisms. Overall, these results demonstrate the benefits of using a gene panel versus existing methods where focused genetic screening is performed depending on biochemical measurements: it improves diagnostic accuracy and provides the opportunity for novel variant discovery.

Introduction

Erythrocytosis* is a clinical condition characterized by increased red cell mass and typically elevated hemoglobin (Hb) concentration and hematocrit (Hct)[1]. It can be *congenital* (e.g. genetic) or *acquired* and classified as *primary* or *secondary*[1] (Figure 1A). Several causal genetic mutations have been identified. Heterozygous mutations in the Epo receptor (*EPOR*) gene cause primary congenital erythrocytosis[2, 3], while *JAK2* mutations are predominantly associated with primary acquired erythrocytosis i.e. polycythemia vera[4-6]. Homozygous germline mutations in *VHL* e.g. Chuvash Polycythemia and heterozygous germline mutations in *EGLN1* (*PHD2*) and *EPAS1* (*HIF2A*) have been found in patients with secondary congenital erythrocytosis[2, 7]. Regarding *EPAS1*, somatic gain-of-function mutations have been detected in pheochromocytomas and paragangliomas in patients with congenital erythrocytosis, attributed to tissue mosaicism[8]. Some patients, particularly those with polycythemia vera and some forms of genetic erythrocytosis, present with increased incidence of both arterial and venous thromboembolic events[9]. Other congenital lesions include high oxygen-affinity hemoglobinopathies or 2,3-bisphosphoglycerate (2,3-BPG) deficiency[10-12], through mutations in globin genes (*HBA1, HBA2*, *HBB*) or the *BPGM* gene, respectively. These genes belong to key pathways involved in the pathogenesis of erythrocytosis

* In this manuscript, the term erythrocytosis rather than polycythemia is used consistently throughout.

e.g. oxygen-sensing (Hypoxia-Inducible Factor, HIF) pathway, erythropoiesis and oxygen transport (Figure 1B). Briefly, HIFs are transcription factors composed of 2 subunits: HIFα, which is oxygen-sensitive, and HIFβ. There are 3 HIFα isoforms, but HIF2α (*EPAS1*) is *EPO*'s main transcriptional regulator[13,14]. In normoxia, HIFα is hydroxylated by oxygen-dependent prolyl hydroxylases (PHDs, encoded by *EGLN1*, *EGLN2* and *EGLN3*), binds to VHL and becomes ubiquitinated and degraded. In hypoxia, hydroxylation diminishes, HIFα stabilizes and initiates the transcription of target genes, including *EPO*[15]. Epo binds to the EPOR of erythroid progenitor cells in the bone marrow, stimulating proliferation and differentiation into red blood cells (RBC), through a JAK2-mediated signaling cascade. In RBC, BPGM promotes the release of oxygen to local tissues by producing 2,3-bisphosphoglycerate (2,3-BPG), which decreases the affinity of hemoglobin to oxygen.

Even if fully investigated (including screening for known mutations), there remains a considerable proportion of patients (~70%) without an identified cause, described as having idiopathic erythrocytosis[3, 9]. About two thirds of these patients have inappropriately normal or elevated Epo levels suggesting a defect in oxygen-sensing or oxygen delivery pathways. Most patients have early-onset disease and/or often a family history, suggesting a high probability of genetic etiology. Logically, further investigation of these patients should begin by fully sequencing genes in which genetic variants are already known to cause erythrocytosis as opposed to simply screening for particular known variants. As many of these are in the HIF pathway, sequencing other key genes in this pathway (in which variants have not yet been observed) and also other erythropoiesis-related genes, is likely to be fruitful in the effort to resolve functional variants.

Using traditional DNA sequencing methods, e.g. Sanger, to comprehensively sequence a large number of genes in a substantial number of patients with a relatively rare disease is time-consuming, labor-intensive and impractical. Conversely, high-throughput technology e.g. whole genome sequencing (WGS) has its own caveats with generation of huge data volumes, high cost and complex bioinformatic analysis. A way forward is the development of disease-relevant targeted next generation sequencing (NGS) gene panels.

We developed a NGS erythrocytosis gene panel, using an ultra-high multiplex PCR method (*AmpliSeq, Thermo Fisher*), which allows rapid high-throughput sequencing of the full length of multiple genes in multiple samples. We defined a custom-made panel of 21 candidate genes from key pathways involved in the pathogenesis of erythrocytosis, and used it to sequence 125 idiopathic erythrocytosis patients. We also included novel candidate genes suggested by an initial WGS study, the WGS500 project[16], in which 500 samples across a diverse spectrum of clinical disorders were sequenced, which included some idiopathic erythrocytosis cases with high suspicion of a genetic cause.

The aims of the study were to: create a targeted sequencing panel, as a research tool, for the genetic investigation of erythrocytosis; evaluate its diagnostic utility in a cohort of idiopathic erythrocytosis patients; search for novel variants in erythrocytosis-associated genes; and include new candidate genes identified in WGS500 to determine if they are mutated in additional patients.

Methods

Patient description

Patient DNA samples extracted from blood were acquired from 4 separate idiopathic erythrocytosis databases (UK, Portugal, Germany and the Netherlands). Participants gave informed consent and appropriate ethical approval was gained. The inclusion criteria were:

1. Confirmed absolute erythrocytosis with a red cell mass >125% predicted, and Hb >180 g/L and Hct >0.52% in adult males or Hb >160 g/L and Hct >0.48% in adult females, or Hb and Hct levels above the 99th centile of age-appropriate reference values in children.
2. Registered as idiopathic (unidentified cause of illness), following appropriate investigation at each Center (Figure S1).
3. Early-onset disease, or cases with long-standing idiopathic erythrocytosis.

Details are found in Supplementary Information (SI).

Ten samples were whole-genome sequenced as part of the WGS500 project. 125 samples were sequenced using our erythrocytosis gene panel alongside 10 positive controls.

Whole genome sequencing

Samples were sequenced at a 30X depth with Illumina HiSeq2000. Details are found in the SI.

Ion Torrent sequencing and analysis

A customized panel, encompassing the coding and untranslated (UTR) regions of the candidate genes (Table 1), was created using the Ion AmpliSeq Designer (*Thermo Fisher*), whereby 635 primer pairs generating amplicons of ~200bp were designed. This panel covered 90.3% of the target region (78.96Kb), with 97.4% average coverage of the coding regions. The primers, synthesized in two multiplex pools, were used with the Ion Ampliseq Library kit 2.0 and Ion Xpress barcode adapters (*Thermo Fisher*) to create libraries. Library quality and concentration were assessed using a 2100 Bioanalyzer (*Agilent Technologies*). Pools of 8 libraries were used for template preparation, loaded into an Ion 316 chip and sequenced on an Ion PGM instrument (500 flows).

The Torrent Suite Software (*Thermo Fisher*) was used for quality control and alignment of the sequencing data to the human genome (Hg19). Variants were called with the Ion Reporter Software v4.2 (*Thermo Fisher*), using the germline workflow for single samples and the default parameters, and annotated with ANNOVAR[17]. Only variants fulfilling all of the following conditions were selected for further analysis: confidence ≥40, read depth ≥20, frequency in 1000 Genomes (1000G) ≤3% and frequency in NHLBI ESP exomes (6500si) ≤3%. Provean and the SIFT and PolyPhen2-HDIV scores and cut-offs from the ANNOVAR LJB23 database were used to assess causality of non-synonymous variants. Synonymous variants were investigated for possible splicing effects using Human Splicing Finder, NetGene2 and FSPLICE. Details are found in the SI.

Sanger validation

All relevant variants identified by Ion Torrent were confirmed by Sanger sequencing. For protocol and primer details see the SI and Table S1.

Results

Novel candidate genes and variants were identified by WGS

The whole genomes of a small number of idiopathic erythrocytosis cases with high suspicion of a genetic cause were sequenced as part of the WGS500 project. Candidate variants were found in novel genes, not previously associated with erythrocytosis: *EPO*, *GFI1b*, *KDM6A* and *BHLHE41*. Details of the rationale and criteria used to select these genes as candidates are given in the SI and Table S2. On this basis, these genes were included in the NGS gene panel along with other erythrocytosis candidate genes (Table 1).

The erythrocytosis gene panel has high performance in sequencing and variant detection

Overall, 135 samples were sequenced on the Ion Torrent using the gene panel (125 undiagnosed patients, 10 positive controls). On average, 89% of mapped reads were on target regions, which indicates a successful custom panel according to the manufacturer's guidance. The average coverage depth of the amplicons generated was 450X (Figure 2A). Most samples (133 out of 135) had over 92% of amplicons with coverage above 20X (Figure 2B). Only two samples presented substantial failure across the panel (Figure 2B), which was related to DNA quality. Only 17 amplicons (2.6%) had an average coverage below 20X across samples, indicating a general poor amplification of these regions within the highly-multiplexed reactions (Table S3). Ten of these (1.6% of all amplicons) had completely failed (coverage <20X in all samples), probably due to sequence context issues. Therefore, the sequencing was generally successful across samples, with a high percentage of the target sequence included at a good depth for germline variant calling.

We compiled a list of all known erythrocytosis-associated variants from the literature[2, 3], including the variants identified in the WGS study, and cross-referenced their genomic coordinates with those of the generated amplicons. With the exception of two missense variants in *VHL*, all the other variants were within amplicons that performed well. The two *VHL* missense variants – c.235C>T and c.311G>T – fall within an amplicon in exon 1 that showed complete failure and therefore would not be detected.

Importantly, our panel reliably detected 10 known variants – in different genes and hence in different amplicons – in the positive control samples, in which mutations had previously been identified either through WGS or Sanger sequencing (Table S4).

51 exonic variants were identified across 57 patients by the erythrocytosis gene panel and validated by Sanger sequencing

We identified 98 different variants across the coding regions of the genes examined, of which 19 were insertions or deletions (INDELS), 49 non-synonymous single nucleotide variations (SNVs) and 30 synonymous SNVs (Figure 3). None of the synonymous SNVs is predicted to alter splicing according to Human Splicing Finder, NetGene2 and FSPLICE. Therefore, we focused on variants resulting in protein sequence alterations: following Sanger sequencing, 17 out of the 19 INDELS

appeared to be false positives but 2 were confirmed. All 49 non-synonymous SNVs were confirmed, although for one SNV there was a single base discrepancy: Ion Torrent detected a triple base change (CAA>ATT) in exon 12 of *JAK2* (chr9:5070025-5070027) but only a double change (AA>TT, chr9:5070026-5070027) was confirmed by Sanger. As a result, a total of 51 variants (49 SNVs, 2 INDELS) were detected (Table S5). Therefore, 57 out of 125 cases had at least 1 exonic variant (45.6%); of those, 38 patients had only 1 exonic variant detected (30.4%), while 19 had more than one (15.2%).

To investigate whether the variants discovered are unique to erythrocytosis patients (and therefore more likely to be disease-causing), we used *in silico* data from the 1000G project as control. For this, we examined the variant calls from the 1000G project after integrating both exome and low coverage data across 1041 individuals and extracted the SNVs identified within the coordinates of the amplicons generated by our gene panel. We found that of the 49 non-synonymous SNVs discovered, 30 were uniquely found in our erythrocytosis cohort and not in the 1000G *in silico* control cohort, whereas the other 19 were also found in the control cohort at similar or higher frequencies (Fisher's exact test and Benjamini and Hochberg false discovery correction[18]) (Figure 3). Those 19 SNVs (Table S6) are thus unlikely to be disease-causing mutations and most likely represent polymorphisms.

Out of the 30 uniquely identified variants in our patient cohort, 10 were previously reported in the literature as causing erythrocytosis and hence are classified here as disease-causing variants (Table 2). The remaining 20 had no previous clinical associations. No exonic variants were identified in *EGLN3, HIF1AN (FIH), HBA1, HBA2, GFI1B* and *ZNF197*.

Novel genes and variants identified by the erythrocytosis gene panel

The 22 novel variants (20 SNVs and 2 INDELs) identified (Table 3) are extremely rare: 9 were absent from both dbSNP142 and ExAC (Exome Aggregation Consortium) databases, the latter containing data from 60,706 unrelated individuals; 8 were reported only in ExAC at an extremely low allele frequency (≤0.0007), and only 5 were reported in both databases at very low allele frequency (≤0.005).

Fourteen of these novel or very rare variants were found in known erythrocytosis-associated genes, such as *VHL*, *EPAS1*, *JAK2, SH2B3* (*LNK*)*, EGLN1* and *BPGM*. Some of these variants have a high likelihood of causality based on the location and predicted effect of the protein coding change as well as on genetic evidence for causality, and are of particular physiological interest. For example, EPAS1 p.Y532H, a novel exon 12 mutation, is located one position downstream of residue 531, which is the prolyl hydroxylation site on HIF2α on the C-terminal oxygen-dependent degradation domain (ODD). Furthermore, it is part of a 6-residue domain which is highly conserved both across all HIFα isoforms and across species and which interacts with the VHL complex[19]. Thus, this mutation likely interferes with hydroxylation of HIF2α by PHDs and binding to the VHL complex, leading to upregulation of Epo. It was found in two related patients, father and son, both of whom had idiopathic erythrocytosis with raised Epo levels, and thus inherited in an autosomal dominant manner. Furthermore, EGLN1 p.L279P is affecting a conserved residue,

previously reported as altered (p.L279Tfs43, a frameshift variant) in a patient with erythrocytosis[20]. Structurally, this residue is located on helix 3, which interacts with both N-terminal and C-terminal ODD hydroxylation domains on HIFα[21]; a proline substitution may affect protein stability and diminish ODD binding, reducing HIFα hydroxylation. The VHL p.E52X variant introduces a stop codon, predicting translation termination of the long VHL isoform (p30) while allowing only the translation of the alternative form of VHL (p19) from a translation site at M54. To date, only a few variants upstream of the *VHL* internal start codon 54 have been described, associated with either pheochromocytomas (codon 25 and 38) or with VHL disease (p.E46X and p.E52K)[22-24]. The role of the heterozygous VHL p.E52X in producing erythrocytosis in the patient in our study is not clear and the patient will be advised screening for the presence of VHL disease; there is evidence that erythrocytosis is seen in about 5-20% of patients with VHL disease[25]. For the remaining variants, most were classified as deleterious by either SIFT, PolyPhen2 or Provean (Table 3), with high degree of agreement between tools, so further investigations are needed to elucidate their functional impact.

Eight variants were identified in novel genes included in the panel because of their association with the oxygen-sensing pathway but in which no previous erythrocytosis-associated mutation has been reported, such as *EGLN2, HIF3A* and *OS9* (Table 3). In addition, novel variants were also found in *EPO* and *BHLHE41*, two genes without previous genetic association with erythrocytosis that were revealed by WGS500. For *EPO*, the most striking variant found is a frameshift, p.P7fs, detected in heterozygous status in one patient. Although at present it is difficult to link an apparently inactivating mutation to the generation of erythrocytosis, the variant has since been confirmed in heterozygous status in the patient's father who also presents with high hematocrit and hemoglobin levels. Two other EPO SNVs were detected in other patients but these are most likely very rare polymorphisms (Table 3). Regarding *BHLHE41*, the novel missense variant identified (p.F149L) is classified as benign by Provean, PolyPhen2 and SIFT and is thus unlikely to be pathogenic, a notion supported by segregation analysis in the patient's family (Table 3).

Discussion

The technical advances in next generation sequencing, together with the increasing understanding of the biological pathways underlying the pathogenesis of erythrocytosis, provide new opportunity to advance the genetic investigation of patients with erythrocytosis.

Our approach allowed the creation of a NGS-based targeted gene panel with the capacity to process a large group of samples and simultaneously examine a large number of genes across several biological pathways in a systematic and efficient manner.

Our panel exhibited high performance and reliability. It produced high quality sequencing data with good target coverage. It accurately detected variants on 10 positive controls. It was excellent at reliably calling SNVs, with all SNVs identified subsequently validated in all samples by Sanger sequencing. A few limitations have, nevertheless, been recognized and should be taken into account when considering its future applications. For example, a few amplicons – including a region on *VHL* exon 1 – showed complete failure across samples and thus potential variants within

them would not be detected. Furthermore, there were some false positive INDELs, as previously reported by other Ion Torrent sequencing users[26-28]. These could be addressed by re-designing primers covering that particular *VHL* genomic region, optimizing the variant calling bioinformatics workflow and employing recently proposed strategies to increase the accuracy of INDEL detection[26, 28]. Another limitation of the panel – related to the nature of its technology – is that it can only identify SNVs and short INDELs but not other structural variants such as large INDELs or copy number variations. Also, variant detection in genes with high sequence similarity such as *HBA1* and *HBA2* is challenging and caution is needed for variant calling.

Currently, the clinical consensus for investigating erythrocytosis involves: establishing the diagnosis of absolute erythrocytosis, excluding systemic causes (e.g. hypoxic lung diseases or tumors) and then proceeding to focused genetic testing based on algorithms that attempt to predict the type of mutation that might be present. There is variability in procedures employed at different centers (Figure S1), but as a general rule: if Epo is low, variants in genes involved in erythropoiesis (*EPOR*, *JAK2*) are screened for. If Epo is high or normal, P50 (partial pressure of oxygen at which 50% of Hb is saturated with oxygen) is calculated and if low, Hb electrophoresis is performed and/or variants in oxygen delivery pathways (globin genes, *BPGM*) are screened for; if P50 is normal or not available, variants in the oxygen-sensing HIF pathway (*VHL*, *EPAS1*, *EGLN1*) are screened for[2, 29, 30].

Using our gene panel we were able to provide definitive genetic diagnoses in 9 patients that were previously missed. For example, a variant in *EPAS1*, p.G537R – a well-described gain-of-function mutation found in erythrocytosis patients[31, 32] – was detected. This was previously missed because the patient was not screened for *EPAS1* variants, owing to the fact that the Epo level was not high enough (and was thus directed to a different branch of the diagnostic algorithm). Similarly, we identified a homozygous *VHL* variant (p.H191D) known to cause erythrocytosis[33]. Interestingly, we found four variants in the *HBB* gene, all relating to high-affinity hemoglobinopathies associated with erythrocytosis: HBB p.H147P (Hb York), HBB p.H144Q (Hb Little Rock), HBB p.V110M (Hb San Diego) and HBB p.E102D (Hb Potomac)[34-38]. These were missed previously, either because conventional screening with Hb electrophoresis can miss hemoglobinopathies[38] or because of difficulties obtaining optimal fresh venous blood samples for P50 measurements in all patients. In addition, we identified a heterozygous variant in *JAK2* (p.K539L) and two in *SH2B3* (p.E208Q and p.E400K), all known to associate with erythrocytosis[5, 39, 40]. The patient with variant JAK2 p.K439L, originally classified as idiopathic erythrocytosis as the conventional criteria for Polycythemia Vera (PV) including JAK2 p.V617F screening were not met, should now be considered as PV with a *JAK2* exon 12 mutation. As highlighted in previous studies[5, 6], the clinical picture of this subtype of PV is indistinguishable from that of idiopathic erythrocytosis. This emphasizes that *JAK2* exon 12 mutations should actively be screened for in idiopathic erythrocytosis patients. Furthermore, the findings of *SH2B3* variants highlight that this gene should also be surveyed, which currently is not routinely done. The erythrocytosis gene panel can successfully do both. Thus, we demonstrated that the panel allows reliable detection of known erythrocytosis-causing mutations, avoiding pitfalls that may occur when following existing algorithms.

In this study, 4 out of the 125 patients were heterozygous for VHL p.R200W. This variant causes Chuvash polycythemia in the homozygous state[41, 42]. Congenital erythrocytosis also occurs in patients who are compound heterozygotes[43-45], but heterozygous carriers are usually unaffected. Nevertheless, VHL p.R200W heterozygous mutations feature significantly more frequently in erythrocytosis databases[3] than in general populations[46], suggesting a causal role. For one of the 4 patients here, the variant was newly identified. For the other 3, previous genetic tests had also identified it. Thus, within this study we aimed to detect additional genetic changes that might explain their clinical phenotype. We did not detect any other variants within *VHL*, except for 2 SNPs in the 3'UTR with high minor allelic frequencies (≥0.35 in dbSNP142). Alternatively, the co-occurrence of this heterozygous variant with another heterozygous variant in a separate gene of the same biological pathway could act in synergy to produce disease. We did not obtain conclusive evidence for this in our four patients: 2 did not have an additional variant; in the other 2, the VHL p.R200W co-occurred with heterozygous missense variants classified as polymorphisms (Table S6), i.e. with EGLN1 p.A157Q and EGLN2 p.T405M in one patient and with EPOR p.G46E and EGLN2 p.S58L in the other.

As this research panel provides full-gene sequencing instead of specific mutation screening, it allowed the detection of 22 novel variants. Some of these have strong likelihood of causality, based on the location of the mutated residues on functional or regulatory domains and the expected disturbance they would cause on protein structure and function (as explained in Results for EGLN1 p.L279P, EPAS1 p.Y532H and VHL p.E52X), and based on genetic evidence of familial segregation (e.g. EPAS1 p.Y532H and EPO p.P7fs, which are dominantly inherited). For other variants, mostly found in known erythrocytosis-associated genes, there is strong consensus in the *in silico* prediction of deleterious effect, whereas others have lesser evidence of functional candidacy (Table 3). While the functional significance of newly identified variants cannot at present be confirmed – and indeed clinical causation cannot be concluded –, future functional studies and screening of larger erythrocytosis patient cohorts for replication of findings will be needed to provide further evidence of causality.

Moreover, we explored other genes, not previously associated with erythrocytosis, because of their involvement in the HIF pathway or their discovery through WGS500 as potential candidates. Candidate variants were found in *EGLN2* and *HIF3A* but not in key HIF pathway genes such as *EGLN3, HIF1AN* and importantly, *HIF1A*. This is consistent with existing literature in which variation in *EPAS1,* but not *HIF1A,* is associated with erythrocytosis. The precise WGS-identified variants in *EPO, GFI1b, KDM6A* and *BHLHE41* were not found in this cohort of 125 cases, suggesting that larger patient cohorts need to be sequenced before the significance of variation in these genes can be properly interpreted. However, in the case of *EPO,* other variants were identified suggesting that *EPO* should be actively surveyed as an erythrocytosis-associated candidate gene. Accrued use of the panel in further patients will provide insight into which novel genes play a role in erythrocytosis and will allow refinement of any future diagnostic panels.

One limitation of our study is the lack of DNA from a source other than blood to determine germline or somatic status. This would only be a concern for *JAK2* and *SH2B3,* in which somatic

mutations are associated with polycythaemia vera and myeloproliferative diseases. Where variants in *JAK2* and *SH2B3* are found by the panel, further studies in skin/nail DNA are probably warranted. For all other genes, variants detected in blood with the panel are most likely germline. While somatic mutations in *EPAS1* can be found in tumors of patients with erythrocytosis[8], these would not be detectable in blood with our methodology.

Thus, despite the few technical limitations described, the erythrocytosis gene panel is useful in the genetic investigation of patients with erythrocytosis from a research perspective. Furthermore, following appropriate optimization and refinement, gene panel sequencing has the potential to improve the diagnostic work-up of erythrocytosis patients in clinical practice. A point to note is that the gene panel in our study was applied to a highly-selected group that had undergone significant "filtering", clinical and genetic (Figure S1) before inclusion in the study. Despite this, candidate variants – known causal and novel – were detected in 29% of patients. Thus, we propose that gene panel sequencing should be applied directly on "erythrocytosis cases where a genetic cause is suspected", i.e. after clinical exclusion of acquired systemic causes and at the point where genetic testing is considered (Figure 4). This would undoubtedly increase the diagnostic yield and, because genetic testing would be conducted in an unbiased manner, it would improve diagnostic accuracy by decreasing the number of missed diagnoses. In conclusion, we hope to demonstrate the immediate utility of a targeted gene panel in the investigation of erythrocytosis at a time when next-generation sequencing is revolutionizing clinical medicine.

References

1. McMullin MF. The classification and diagnosis of erythrocytosis. Int J Lab Hematol. 2008;30(6):447-459.
2. Hussein K, Percy M, McMullin MF. Clinical utility gene card for: familial erythrocytosis. Eur J Hum Genet. 2012;20(5):
3. Bento C, Percy MJ, Gardie B, et al. Genetic basis of congenital erythrocytosis: mutation update and online databases. Hum Mutat. 2014;35(1):15-26.
4. Percy MJ, Jones FG, Green AR, Reilly JT, McMullin MF. The incidence of the JAK2 V617F mutation in patients with idiopathic erythrocytosis. Haematologica. 2006;91(3):413-414.
5. Scott LM, Tong W, Levine RL, et al. JAK2 exon 12 mutations in polycythemia vera and idiopathic erythrocytosis. N Engl J Med. 2007;356(5):459-468.
6. Percy MJ, Scott LM, Erber WN, et al. The frequency of JAK2 exon 12 mutations in idiopathic erythrocytosis patients with low serum erythropoietin levels. Haematologica. 2007;92(12):1607-1614.
7. Lee FS, Percy MJ, McMullin MF. Oxygen sensing: recent insights from idiopathic erythrocytosis. Cell Cycle. 2006;5(9):941-945.
8. Zhuang Z, Yang C, Lorenzo F, et al. Somatic HIF2A gain-of-function mutations in paraganglioma with polycythemia. N Engl J Med. 2012;367(10):922-930.
9. McMullin MF. Idiopathic erythrocytosis: a disappearing entity. Hematology Am Soc Hematol Educ Program. 2009;629-635.
10. Percy MJ, Butt NN, Crotty GM, et al. Identification of high oxygen affinity hemoglobin variants in the investigation of patients with erythrocytosis. Haematologica. 2009;94(9):1321-1322.

11. Hoyer JD, Allen SL, Beutler E, Kubik K, West C, Fairbanks VF. Erythrocytosis due to bisphosphoglycerate mutase deficiency with concurrent glucose-6-phosphate dehydrogenase (G-6-PD) deficiency. Am J Hematol. 2004;75(4):205-208.
12. Petousi N, Copley RR, Lappin TR, et al. Erythrocytosis associated with a novel missense mutation in the BPGM gene. Haematologica. 2014;99(10):e201-204.
13. Gruber M, Hu CJ, Johnson RS, Brown EJ, Keith B, Simon MC. Acute postnatal ablation of Hif-2alpha results in anemia. Proc Natl Acad Sci U S A. 2007;104(7):2301-2306.
14. Rankin EB, Biju MP, Liu Q, et al. Hypoxia-inducible factor-2 (HIF-2) regulates hepatic erythropoietin in vivo. J Clin Invest. 2007;117(4):1068-1077.
15. Webb JD, Coleman ML, Pugh CW. Hypoxia, hypoxia-inducible factors (HIF), HIF hydroxylases and oxygen sensing. Cell Mol Life Sci. 2009;66(22):3539-3554.
16. Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet. 2015;47(7):717-726.
17. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
18. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289-300.
19. Min JH, Yang H, Ivan M, Gertler F, Kaelin WG, Jr., Pavletich NP. Structure of an HIF-1alpha -pVHL complex: hydroxyproline recognition in signaling. Science. 2002;296(5574):1886-1889.
20. Jang JH, Seo JY, Jang J, et al. Hereditary gene mutations in Korean patients with isolated erythrocytosis. Ann Hematol. 2014;93(6):931-935.
21. Chowdhury R, McDonough MA, Mecinovic J, et al. Structural basis for binding of hypoxia-inducible factor to the oxygen-sensing prolyl hydroxylases. Structure. 2009;17(7):981-989.
22. van der Harst E, de Krijger RR, Dinjens WN, et al. Germline mutations in the vhl gene in patients presenting with phaeochromocytomas. Int J Cancer. 1998;77(3):337-340.
23. Olschwang S, Richard S, Boisson C, et al. Germline mutation profile of the VHL gene in von Hippel-Lindau disease and in sporadic hemangioblastoma. Hum Mutat. 1998;12(6):424-430.
24. Dollfus H, Massin P, Taupin P, et al. Retinal hemangioblastoma in von Hippel-Lindau disease: a clinical and molecular study. Invest Ophthalmol Vis Sci. 2002;43(9):3067-3074.
25. Friedrich CA. Genotype-phenotype correlation in von Hippel-Lindau syndrome. Hum Mol Genet. 2001;10(7):763-767.
26. Costa JL, Sousa S, Justino A, et al. Nonoptical massive parallel DNA sequencing of BRCA1 and BRCA2 genes in a diagnostic setting. Hum Mutat. 2013;34(4):629-635.
27. Junemann S, Sedlazeck FJ, Prior K, et al. Updating benchtop sequencing performance comparison. Nat Biotechnol. 2013;31(4):294-296.
28. Yeo ZX, Chan M, Yap YS, Ang P, Rozen S, Lee AS. Improving indel detection specificity of the Ion Torrent PGM benchtop sequencer. PLoS One. 2012;7(9):e45798.
29. Cario H, McMullin MF, Bento C, et al. Erythrocytosis in children and adolescents-classification, characterization, and consensus recommendations for the diagnostic approach. Pediatr Blood Cancer. 2013;60(11):1734-1738.
30. Bento C, Almeida H, Maia TM, et al. Molecular study of congenital erythrocytosis in 70 unrelated patients revealed a potential causal mutation in less than half of the cases (Where is/are the missing gene(s)?). Eur J Haematol. 2013;91(4):361-368.
31. Percy MJ, Furlow PW, Lucas GS, et al. A gain-of-function mutation in the HIF2A gene in familial erythrocytosis. N Engl J Med. 2008;358(2):162-168.

32. Gale DP, Harten SK, Reid CD, Tuddenham EG, Maxwell PH. Autosomal dominant erythrocytosis and pulmonary arterial hypertension associated with an activating HIF2 alpha mutation. Blood. 2008;112(3):919-921.
33. Tomasic NL, Piterkova L, Huff C, et al. The phenotype of polycythemia due to Croatian homozygous VHL (571C>G:H191D) mutation is different from that of Chuvash polycythemia (VHL 598C>T:R200W). Haematologica. 2013;98(4):560-567.
34. Charache S, Jacobson R, Brimhall B, et al. Hb Potomac (101 Glu replaced by Asp): speculations on placental oxygen transport in carriers of high-affinity hemoglobins. Blood. 1978;51(2):331-338.
35. Gonzalez Fernandez FA, Villegas A, Ropero P, et al. Haemoglobinopathies with high oxygen affinity. Experience of Erythropathology Cooperative Spanish Group. Ann Hematol. 2009;88(3):235-238.
36. Bromberg PA, Alben JO, Bare GH, et al. High oxygen affinity variant of haemoglobin Little Rock with unique properties. Nat New Biol. 1973;243(127):177-179.
37. Misgeld E, Gattermann N, Wehmeier A, Weiland C, Peters U, Kohne E. Hemoglobinopathy York [beta146 (HC3) His==>Pro]: first report of a family history. Ann Hematol. 2001;80(6):365-367.
38. Wajcman H, Galacteros F. Abnormal hemoglobins with high oxygen affinity and erythrocytosis. Hematol Cell Ther. 1996;38(4):305-312.
39. Spolverini A, Pieri L, Guglielmelli P, et al. Infrequent occurrence of mutations in the PH domain of LNK in patients with JAK2 mutation-negative 'idiopathic' erythrocytosis. Haematologica. 2013;98(9):e101-102.
40. McMullin MF, Wu C, Percy MJ, Tong W. A nonsynonymous LNK polymorphism associated with idiopathic erythrocytosis. Am J Hematol. 2011;86(11):962-964.
41. Ang SO, Chen H, Hirota K, et al. Disruption of oxygen homeostasis underlies congenital Chuvash polycythemia. Nat Genet. 2002;32(4):614-621.
42. Smith TG, Brooks JT, Balanos GM, et al. Mutation of von Hippel-Lindau tumour suppressor and human cardiopulmonary physiology. PLoS Med. 2006;3(7):e290.
43. Bento MC, Chang KT, Guan Y, et al. Congenital polycythemia with homozygous and heterozygous mutations of von Hippel-Lindau gene: five new Caucasian patients. Haematologica. 2005;90(1):128-129.
44. Cario H, Schwarz K, Jorch N, et al. Mutations in the von Hippel-Lindau (VHL) tumor suppressor gene and VHL-haplotype analysis in patients with presumable congenital erythrocytosis. Haematologica. 2005;90(1):19-24.
45. Percy MJ, McMullin MF, Jowitt SN, et al. Chuvash-type congenital polycythemia in 4 families of Asian and Western European ancestry. Blood. 2003;102(3):1097-1099.
46. Liu E, Percy MJ, Amos CI, et al. The worldwide distribution of the VHL 598C>T mutation indicates a single founding event. Blood. 2004;103(5):1937-1940.
47. Panovska-Stavridis Iea. Familiar JAK2 G571S Variant Not Linked with Essential Trombocythemia. Blood, 2014:21.

Tables

Table 1. Genes included in the custom-made erythrocytosis gene panel

Candidate Gene	Position	No of exons	Transcript ID	Pathway	Candidacy	Inheritance
VHL	Chr3:10183319-10195354	3	NM_000551	Oxygen-sensing	Known erythrocytosis-causing variants	Recessive / Compound Heterozygous (based on reported cases)
EPAS1	Chr2:46524541-46613842	16	NM_001430	Oxygen-sensing	Known erythrocytosis-causing variants	Dominant (based on reported cases)
EGLN1	Chr1:231499497-231560790	4	NM_022051	Oxygen-sensing	Known erythrocytosis-causing variants	Dominant (due to haploinsufficiency, based on reported cases)
HIF1A	Chr14:62162119-62214977	15	NM_001530	Oxygen-sensing	Key gene of the HIF pathway	Unknown (no cases reported). Likely dominant by function similarity to *EPAS1*
HIF3A	Chr19:46800303-46846690	13	NM_022462	Oxygen-sensing	Key gene of the HIF pathway	Unknown (no cases reported). Likely dominant by function similarity to *EPAS1*
EGLN2	Chr19:41305048-41314346	5	NM_053046	Oxygen-sensing	Key gene of the HIF pathway	Unknown (no cases reported). Likely dominant by function similarity to *EGLN1*
EGLN3	Chr14:34393421-34420284	5	NM_022073	Oxygen-sensing	Key gene of the HIF pathway	Unknown (no cases reported). Likely dominant by function similarity to *EGLN1*
HIF1AN	Chr10:102295641-102313681	6	NM_017902	Oxygen-sensing	Key gene of the HIF pathway	Unknown (no cases reported). No function similarity to any known associated gene
EPO	Chr7:100318423-100321323	5	NM_000799	Erythropoiesis/ Oxygen-sensing	1. Key gene in erythropoiesis 2. Identified in WGS500	Dominant (based on WGS500 variant), but cannot discard other patterns of inheritance
EPOR	Chr19:11487881-11495018	8	NM_000121	Erythropoiesis	Known erythrocytosis-causing variants	Dominant (based on reported cases)

JAK2	Chr9:4985245-5128183	25	NM_004972	Erythropoiesis	Known erythrocytosis-causing variants	Somatic (based on reported cases)
SH2B3	Chr12:111843752-111889427	8	NM_005475	Erythropoiesis	Known erythrocytosis-causing variants	Dominant (somatic or germline, based on reported cases)
BPGM	Chr7:134331531-134364568	3	NM_001724	Oxygen transport	1. Known erythrocytosis-causing variants 2. Identified in WGS	Dominant / Compound Heterozygous / Recessive (based on reported cases)
HBB	Chr11: 5246696-5248301	3	NM_000518	Oxygen transport/ Hemoglobin synthesis	Known erythrocytosis-causing variants	Dominant (based on reported cases)
HBA1	Chr16:226679-227520	3	NM_000558	Oxygen transport/ Hemoglobin synthesis	Key gene in oxygen transport	Dominant (based on reported cases)
HBA2	Chr16:222846-223709	3	NM_000517	Oxygen transport/ Hemoglobin synthesis	Key gene in oxygen transport	Dominant (based on reported cases)
KDM6A	ChrX:44732423-44971747	29	NM_021140	Oxygen-regulated demethylase	Identified in WGS500	Recessive X-linked inheritance (based on WGS500 variant)
GFI1b	Chr9:135854098-135867084	11	NM_004188	Erythropoiesis	Identified in WGS500	Recessive (based on WGS500 variant), but cannot discard other patterns of inheritance
BHLHE41	Chr12:26272959-26278003	9	NM_030762	Factor associated with HIF	Identified in WGS500	Recessive (based on WGS500 variant), but cannot discard other patterns of inheritance
OS9	Chr12:58087738-58115340	15	NM_001261421	Factor associating with HIF	Related to the HIF pathway	Unknown (no reported cases). No function similarity to any known associated gene
ZNF197	Chr3: 44666511-44689963	5	NM_006991	Factor associating with HIF	Related to the HIF pathway	Unknown (no reported cases). No function similarity to any known associated gene

Official gene symbols according to HUGO Gene Nomenclature Committee are given here. Other gene symbols used frequently in the literature are: *HIF2A* (*EPAS1*), *PHD2* (*EGLN1*), *PHD1* (*EGLN2*), *PHD3* (*EGLN3*), *FIH* (*HIF1AN*), *LNK* (*SH2B3*), *DEC2* (*BHLHE41*).

Table 2. Variants detected by the erythrocytosis gene panel, known to cause erythrocytosis

Genomic location	Gene	cDNA/protein change	Genotype	No of cases	Patient Information	Type of Erythrocytosis	Mechanism of action	Previous Publication
chr2:46607420 G>A	*EPAS1*	c.G1609A p.G537R	Het	1	Female; age at diagnosis, 13; Hb, 196 g/L; Hct, 59.1%; Epo, 7.5 miU/mL; chronic headache; pulmonary hypertension	Secondary Erythrocytosis	Gain of function of HIF2A	Percy et al 2008[31], Gale et al 2008[32]
chr3:10191578 C>G	*VHL*	c.C571G p.H191D	Hom	1	Male; age at diagnosis, 12; Hb, 154 g/L; Hct, 59%; Epo, 23 miU/mL	Secondary Erythrocytosis	Loss of function (enhances HIF regulated gene expression)	Tomasic et al 2013[33]
chr3:10191605 C>T	*VHL*	c.C598T p.R200W	Het*	4	Patient 1: Male; age at diagnosis, 47; Hb, 182 g/L; Hct, 54%; Epo, 8 miU/mL; no family history Patient 2: Female; age at diagnosis, 48; Hb, 199 g/L; Hct, 67%; Epo, 22 miU/mL; no family history Patient 3: Male; age at diagnosis, 1; Hb, 179 g/L; Hct, 54%; Epo, 60 miU/mL; brother of patient 4 Patient 4: Male; age at diagnosis, 2; Hb, 146 g/L; Hct, 44.4%; Epo, high; brother of patient 3	Secondary Erythrocytosis	Loss of function (decreased HIF binding & hydroxylation, enhances HIF-regulated gene expression)	Ang et al 2002[41]
chr9:5070025 AA>TT	*JAK2*	c.1615_ 1616invAA p.K539L	Het	1	Male, age at diagnosis, 35; Hb, 155 g/L; Hct, 52%; RBC, 6.37x10^6/µL; WCC and Plats, normal range; Epo, 5-46 miU/mL (after venesection); first presentation with large stroke; splenomegaly; BM biopsy, erythropoietic hyperplasia; JAK2 V617F negative	Primary Erythrocytosis	Gain of function of JAK2 (K539L)	Scott et al. 2007[5]

chr11:5246832 T>G	*HBB*	c.A440C p.H147P	Het	1	Female; age at diagnosis, 27; Hb, 173 g/L; Hct, 52.4%; Epo, 24 miU/mL; family history, two brothers, mother and grandfather and great grand-mother affected with erythrocytosis (maternal line)	Secondary Erythrocytosis	High oxygen affinity Hb (Hb York)	Misgeld et al 2001[37]
chr11:5246840 G>C	*HBB*	c.C432G p.H144Q	Het	1	Male; age at diagnosis, 34; Hb, 200 g/L; Hct, 58%; Epo, 7.5-27 miU/mL; asymptomatic	Secondary Erythrocytosis	High oxygen affinity Hb (Hb Little Rock)	Bromberg et al 1973[36], Wajcman et al 1996[38]
chr11:5246944 C>T	*HBB*	c.G328A p.V110M	Het	1	Male; age at diagnosis, 49; Hb, 181 g/L; Hct, 53%; Epo, 25 miU/mL	Secondary Erythrocytosis	High oxygen affinity Hb (Hb San Diego)	Wajcman et al 1996[38], Gonzalez et al 2009[35]
chr11:5247816 C>G	*HBB*	c.G306C p.E102D	Het	1	Male; age at diagnosis, 25; Hb, 204 g/L; Hct, 58%; Epo, 5.5 miU/mL	Secondary Erythrocytosis	High oxygen affinity Hb (Hb Potomac)	Charache et al 1978[34]
chr12:111856571 G>C	*SH2B3*	c.G622C p.E208Q	Het	1	Male; age at diagnosis, 14; Hb, 210 g/L; Hct, not available; Epo, 20.2 miU/mL (after venesection)	Primary Erythrocytosis	Enhances JAK2 signalling	Spolverini et al 2013[39]
chr12:111885310 G>A	*SH2B3*	c.G1198A p.E400K	Het	1	Male; age at diagnosis, 40; Hb, 189 g/L; Hct, 52.7%; RBC, 5.78x10^6/µL; Epo, 2.5 miU/mL	Primary Erythrocytosis	Interacts with JAK2 signalling	McMullin et al 2011[40], Spolverini et al 2013[39]

Official gene symbols according to HUGO Gene Nomenclature Committee are given here. Other gene symbols used frequently in the literature are: *HIF2A* (*EPAS1*), *LNK* (*SH2B3*). Chr indicates chromosome; Het, heterozygous; Hom, homozygous; Hb, hemoglobin; Hct, hematocrit; WCC, white cell count; Plats, platelets; RBC, red blood count and BM, bone marrow. Typical normal ranges: Hb, 130-180 g/L (adult males) and 115-155 g/L (adult females); Hct, 45-52 % (adult males) and 37-48 % (adult females); RBC, 4.7-6.1 million cells/ul (adult males) and 4.2-5.4 million cells/ul (adult females); Epo, 3.3-15.8 (adult males and females, although this range can vary between laboratories).

* This variant causes Chuvash polycythemia in the homozygous state. In one of the patients, this variant was discovered in this study, whereas the other 3 had been detected in previous genetic tests.

Table 3. Novel variants detected by the erythrocytosis gene panel

Genomic location	Gene	cDNA/ protein change	Genotype	No of cases	Patient Information	Type of Erythrocytosis	SIFT/ Polyphen/ Provean	Allele freq.: dbSNP142 ExAc	DNA studies in family members	Evidence of causality
chr1:231556799 A>G	*EGLN1*	c.T836C p.L279P	Het	1	Male; age at diagnosis, 52; Hb, 198 g/L; Hct, 61.2%; RBC, 6.42x10^6/µL; Epo, 12.2 miU/mL; headaches and dizziness; family history	Putative Secondary Erythrocytosis	D/D/D	Not found Not found	Not available	Predicted structural/ functional effects
chr2:46607405 T>C	*EPAS1*	c.T1594C p.Y532H	Het	2	Patient 1: Male; age at diagnosis, 12; Hb, 190 g/L; Hct, 54%; Epo, not available; clinically well; no pulmonary hypertension; family history, father with congenital erythrocytosis (Patient 2) Patient 2: Male; age at diagnosis, 42	Putative Secondary Erythrocytosis	D/D/D	Not found Not found	Variant present in both affected father and son	Predicted structural/ functional effects & Segregation
chr3:10183685 G>T	*VHL*	c.G154T p.E52X	Het	1	Male; age at diagnosis, 28; Hb, 184 g/L; Hct, 54.6%; Epo, not available; headaches and dizziness; family history, affected brother	Putative Secondary Erythrocytosis	T/NA/NA	Not found 4.16E-05	Not available	Predicted structural/ functional effects
chr7:100319185 TC>T	*EPO*	c.19delC p.P7fs	Het	1	Female; age at diagnosis, 3; Hb, 194 g/L; Hct, 58%; Epo, 4.1 miU/mL; asymptomatic; family history, affected father and paternal grandmother with high Hb and Hct	Putative Secondary Erythrocytosis	NA/NA/NA	Not found Not found	Variant present in father, who has high Hb and Hct	Segregation
chr2:46574031 AAGG>A	*EPAS1*	c.47delAGG p.del17E	Het	1	Male; age at diagnosis, 47; Hb, 186 g/L; Hct, 58.6%; Epo, normal range; headaches and dizziness	Putative Secondary Erythrocytosis	NA/NA/D	Not found Not found	Not available	1. Deleterious by at least two prediction tools, 2.

chr9:5050747 A>T	*JAK2*	c.A530T p.E177V	Het	1	Male; age at diagnosis, 14; Hb, 158 g/L; Hct, 52%; Epo, 7.6 miU/mL; normal liver and spleen; normal cardiopulmonary function; no family history	Putative Primary Erythrocytosis	D/D/D	Not found Not found	Not available	Most in known erythrocytosis-causing genes 3. Most not found in large population databases
chr12:111856181 G>A	*SH2B3*	c.G232A p.E78K	Het	1	Female; age at diagnosis, not reported (current age 85); Hb, 196 g/L; Hct, 57%; Epo, 12.2 miU/mL; normal liver and spleen size; no family history	Putative Primary Erythrocytosis	D/P/D	Not found 7.40E-04	Not available	
chr12:111884812 G>A	*SH2B3*	c.G901A p.E301K	Het	1	Male; age at diagnosis, 46; Hb, 187 g/L; Hct, 52%; RBC, 6.16x10^6/μL; WCC and Plats, normal range; Epo, 10 miU/mL; fatigue; mild splenomegaly; normal liver and spleen (on ultrasound); BM biopsy, increased erythropoiesis and slight dyserythropoiesis, normal megakaryopoiesis, no myeloproliferative neoplasia; no family history	Putative Primary Erythrocytosis	D/D/Neutral	3.20E-05 3.30E-05	Not available	
chr12:111885466 C>T	*SH2B3*	c.C1243T p.R415C	Hom	1	Male; age at diagnosis, 18; Hb, 188 g/L; Hct, 57%; Epo, not available; splenomegaly; no family history	Putative Primary Erythrocytosis	D/D/D	1.00E-03 4.19E-05	Not available	
chr19:41313427 G>T	*EGLN2*	c.G1139T p.R380L	Het	1	Male; age at diagnosis, 16; Hb, 183 g/L; Hct, 53.7%; Epo, 2.8 miU/mL; family history	Putative Secondary Erythrocytosis	D/D/D	Not found Not found	Not available	
chr2:46611651 T>C	*EPAS1*	c.T2465C p.M822T	Het	1	Female; age at diagnosis, 9; Hb, 162 g/L; Hct, 48%; Epo, 5.8 miU/mL; no family history	Putative Secondary Erythrocytosis	D/B/Neutral	Not found 8.24E-06	Not available	Extremely rare variants

chr3:10183605 C>T	*VHL*	c.C74T p.P25L	Het	2	<u>Patient 1</u>: Male; age at diagnosis, 21; Gitelman syndrome (with positive SLC12A3 mutation); <u>Patient 2</u>: Male; age at diagnosis, 15; Hb, 190 g/L; Hct, 55%; Epo, not available; mild headaches; family history, father also affected	Putative Secondary Erythrocytosis	D/B/Neutral	4.00E-04 5.17E-03	Not available
chr7:134346563 C>A	*BPGM*	c.C304A p.Q102K	Het	1	Male; age at diagnosis, 52; Hb, 186g/dL; Hct, 52.5%; Epo, normal range; myocardial infarction; no family history	Putative Secondary Erythrocytosis	D/B/Neutral	Not found Not found	Not available
chr9:5022168 G>A	*JAK2*	c.G181A p.E61K	Het	1	Male; age at diagnosis, 41; Hb, 172g/L; Hct, 53%; WCC and Plats, normal range; Epo, 10-27 miU/mL (while venesected); no family history	Putative Primary Erythrocytosis	T/B/Neutral	Not found Not found	Not available
chr9:5054775 G>C	*JAK2*	c.G827C p.G276A	Het	1	Female; age at diagnosis, 33; Hb, 172 g/L; Hct, 53.3%; Epo, 6 miU/mL; headaches; dizziness; family history, affected brother	Putative Primary Erythrocytosis	T/B/Neutral	Not found 8.29E-06	Not available
chr12:58109559 G>A	*OS9*	c.G497A p.G166D	Het	1	Male; age at diagnosis, 37; Hb, 188g/L; Hct, 51.8%; Epo, 8.24miU/mL	Unknown	T/D/D	Not found 3.42E-05	Not available
chr19:46811511 A>C	*HIF3A*	c.A190C p.I64L	Het	1	Female; age at diagnosis, 19; Hb, 183 g/L; Hct, 58%; Epo, not available	Putative Secondary Erythrocytosis	D/B/Neutral	Not found 1.65E-05	Not available

chr19:46823777 C>A	*HIF3A*	c.C896A p.A299D	Het	1	Male; age at diagnosis, 53; Hb, 188 g/L; Hct, 58%; Epo, 23 miU/mL; coronary heart disease	Putative Secondary Erythrocytosis	T/B/Neutral	0.00066 7.93E-04	Not available	
chr9:5072561 G>A	*JAK2*	c.G1711A p.G571S*	Het	1	Male; age at diagnosis, 4; Hb, 180 g/L; Hct, not available; Epo, 12.4 miU/mL	Unknown	T/D/D	7.4 E-04 4.81E-04	Variant present in non-affected parent	Unlikely disease-causing in erythrocytosis
chr12:26276001 A>C	*BHLHE41*	c.T447G p.F149L	Het	1	Male; age at diagnosis, 23; Hb, 183 g/L; Hct, 57%; Epo, not available	Unknown	T/B/Neutral	Not found 1.54E-04	Variant present in non-affected paternal aunt; absent in non-affected mother and sibling	
chr7:100320336 A>G	*EPO*	c.A296G p.E99G**	Het	1	Male; age at diagnosis, 13; Hb, 149 g/L; Hct, 49.6%; Epo, 7.5 miU/mL; subsequent red cell mass measurement negative for absolute erythrocytosis despite high Hb	Not absolute erythrocytosis	D/D/D	Not found Not found	Not available	
chr7:100320290 G>C	*EPO*	c.G250C p.G84R***	Het	2	<u>Patient 1</u>: Male; age at diagnosis, 12; Hb, 190 g/L; Hct, 54%; clinically well; no pulmonary hypertension; family history, father with congenital erythrocytosis (Patient 2) <u>Patient 2</u>: Male; age at diagnosis, 42	Putative Secondary Erythrocytosis	T/D/Neutral	Not found 8.04E-05	Variant present in both affected father and son	

Official gene symbols according to HUGO Gene Nomenclature Committee are given here. Other gene symbols used frequently in the literature are: *HIF2A* (*EPAS1*), *PHD2* (*EGLN1*), *PHD1* (*EGLN2*), *LNK* (*SH2B3*), *DEC2* (BHLHE41). Chr indicates chromosome; Het, heterozygous; Hom, homozygous; Hb, hemoglobin; Hct, hematocrit; WCC, white cell count; Plats, platelets; RBC, red blood count; BM, bone marrow; D, deleterious (applicable to SIFT and Provean predictions) and probably damaging (applicable to Polyphen2 HDIV predictions); T, tolerated by SIFT; P, possibly damaging by Polyphen2 HDIV; B, benign by Polyphen2 HDIV and NA, non-applicable. Typical normal ranges: Hb, 130-180 g/L (adult males) and 115-155 g/L (adult females); Hct, 45-52 % (adult males) and 37-48 % (adult females); RBC, 4.7 -6.1 million cells/ul (adult males) and 4.2 - 5.4 million cells/ul (adult females); Epo: 3.3 - 15. 8 (adult males and females, although this range can vary between laboratories). ExAc: Exome Aggregation Consortium, Cambridge, MA (http://exac.broadinstitute.org). All variants except the ones included in the "unlikely disease-causing in erythrocytosis" section are submitted to a dedicated database (www.erythrocytosis.org).

* JAK2 p.G751S has been reported previously in myeloproliferative disorders[47] but it is thought to be a silent non-functioning polymorphism; ** predicted deleterious but although patient has high Hb and Hct was subsequently found to have normal red cell mass. *** Likely to be non-pathogenic: predicted benign/tolerated and present in patients with an identified variant in *EPAS1* with very high likelihood of causality.

Figure Legends

Figure 1. (A) Causes of erythrocytosis. Erythrocytosis can be congenital or acquired. It is classified as primary, where there is an intrinsic defect in erythropoietic cells and erythropoietin (Epo) levels are low, or secondary, where the increased red cell production is externally driven through increased Epo production and Epo levels are high or inappropriately normal. Note: in this manuscript, the term erythrocytosis rather than polycythemia is used consistently throughout **(B) Pathways involved in the pathogenesis of erythrocytosis.** (i) Hypoxia inducible factor (HIF) oxygen sensing pathway in renal EPO producing cells. HIFs are dimeric transcription factors composed of one α- and one β- subunit. In normoxia, HIFα subunits are hydroxylated by oxygen-dependent prolyl-hydroxylases (PHDs) and asparaginyl hydroxylase (HIF1AN). The hydroxylated prolines (P) are recognised by VHL, which mediates the ubiquitination and proteasomal degradation of HIFα. The hydroxylated asparagine (N) compromises the interaction of HIFα with cofactors necessary for transcriptional activity (p300/CBP). In hypoxia, PHDs and HIF1AN are less active, HIFα subunits stabilize and translocate into the nucleus where they interact with HIFβ subunit and cofactors and initiate transcription of target genes, including *EPO* (ii) Erythropoiesis in the bone marrow. It is triggered by the binding of EPO to the EPO receptor (EPOR) located in the surface of erythroid progenitor cells and the subsequent activation of JAK2-signalling cascade. The process is inhibited by the interaction of SH2B3 and JAK2. (iii) Hemoglobin (Hb) synthesis and oxygen transport. BPGM produces 2,3-BPG, which promotes the release of oxygen to local tissues by decreasing the affinity of deoxygenated Hb to oxygen. Alterations in the Hb chains (Hb-α and Hb-β) or BPGM could shift the Hb-oxygen dissociation curve and alter oxygen levels, which directly influences Epo production. (PV, Polycythemia Vera; ECYT 1-4, erythrocytosis type 1-4; Hb, hemoglobin; O_2, oxygen; 2,3-BPG, 2,3-bisphosphoglycerate; RBC, red blood cells; EPO, erythropoietin; PHDs, prolyl hydroxylases). PHDs: PHD1 (*EGLN2*), PHD2 (*EGLN1*) and PHD3 (*EGLN3*).

Figure 2. Coverage of the amplicons generated by the erythrocytosis gene panel across 135 samples. (A) Each boxplot represents the distribution of the number of reads obtained for all the amplicons generated by the panel within each sample. The horizontal line across the plot shows the average coverage (450X). (B) Each dot represents the percentage of amplicons with coverage over 20X within each sample.

Figure 3. Overview of the exonic variants detected with Ion Torrent sequencing among 125 erythrocytosis patients, validation and further classification.

Figure 4. Proposed use of a gene panel in the investigation of erythrocytosis. A gene panel would make genetic testing more efficient and streamlined. It enables the simultaneous survey of the full length of 21 candidate genes, in a systematic and unbiased manner, allowing the detection of known causal variants as well as novel variants in known and novel genes.

Figure 1

A

PRIMARY ERYTHROCYTOSIS (typically low Epo levels)

- **Congenital**
  - **Epo Receptor (*EPOR*) mutations (ECYT1)**
- **Acquired**
  - **Polycythemia vera (PV) (including *JAK2* mutations)**

SECONDARY ERYTHROCYTOSIS (high or normal Epo levels)

- **Congenital**
  - **Defects in oxygen-sensing pathway (ECYT 2-4):**
    - *VHL* mutations (e.g. Chuvash Polycythemia)
    - *EGLN1* (*PHD2*) mutations
    - *EPAS1* (*HIF2A*) mutations
  - **Left shift of Hemoglobin (Hb)-oxygen dissociation curve:**
    - High-oxygen affinity Hb: *HBB, HBA1, HBA2* mutations
    - 2,3-BPG deficiency: *BPGM* mutations
- **Acquired**
  - **Systemic hypoxia:** Chronic lung disease, right-to-left cardiopulmonary vascular shunts, smoker's erythrocytosis, sleep-disordered breathing, high altitude
  - **Local hypoxia:** Renal artery stenosis, renal cysts, post-renal transplant erythrocytosis
  - **Pathological Epo production (tumors):** Brain (cerebellar hemangioblastoma), renal cell carcinoma, hepatocellular carcinoma, pheochromocytoma
  - **Drug-associated:** Epo/androgen administration

B

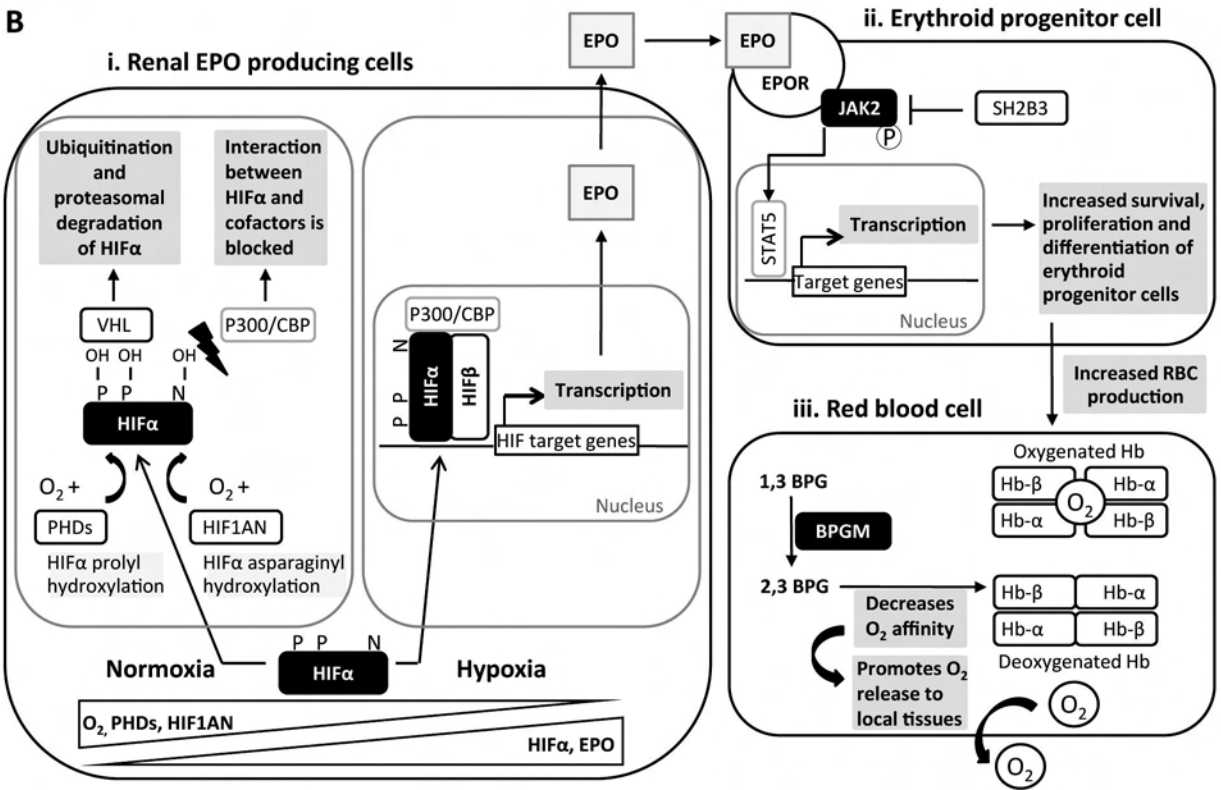
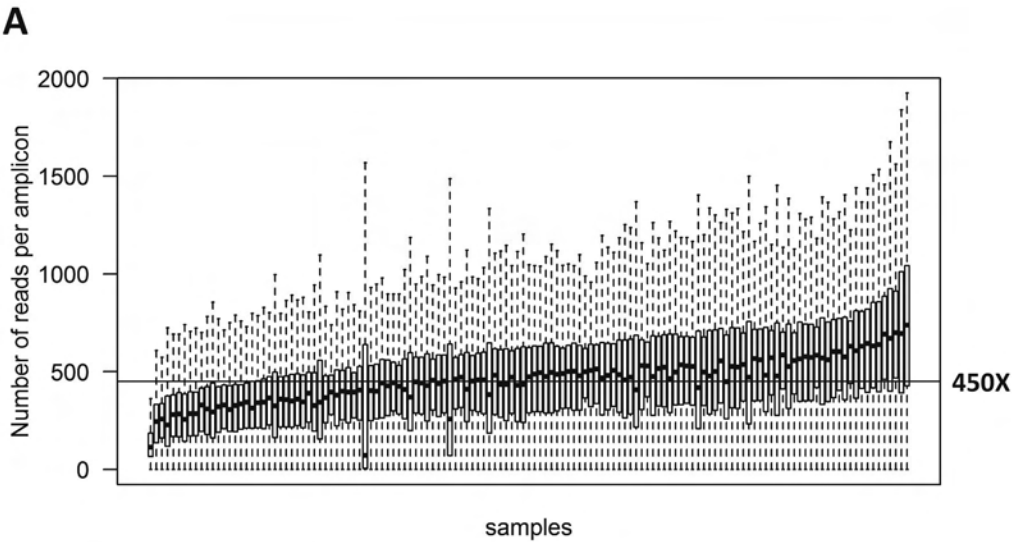


Figure 2


A
2000
1500
1000
500
0
Number of reads per amplicon
450X
samples

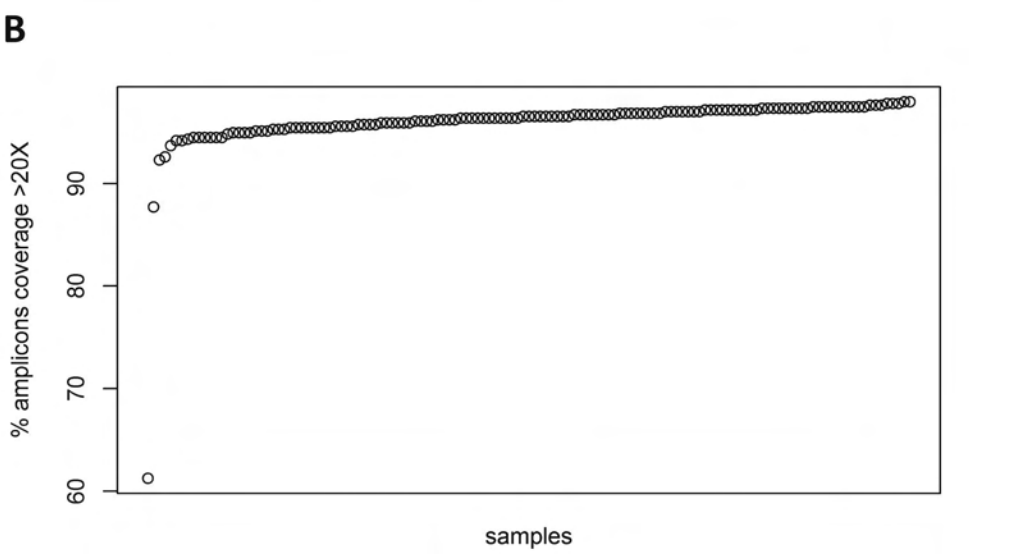

B
% amplicons coverage >20X
90
80
70
60
samples

Figure 3

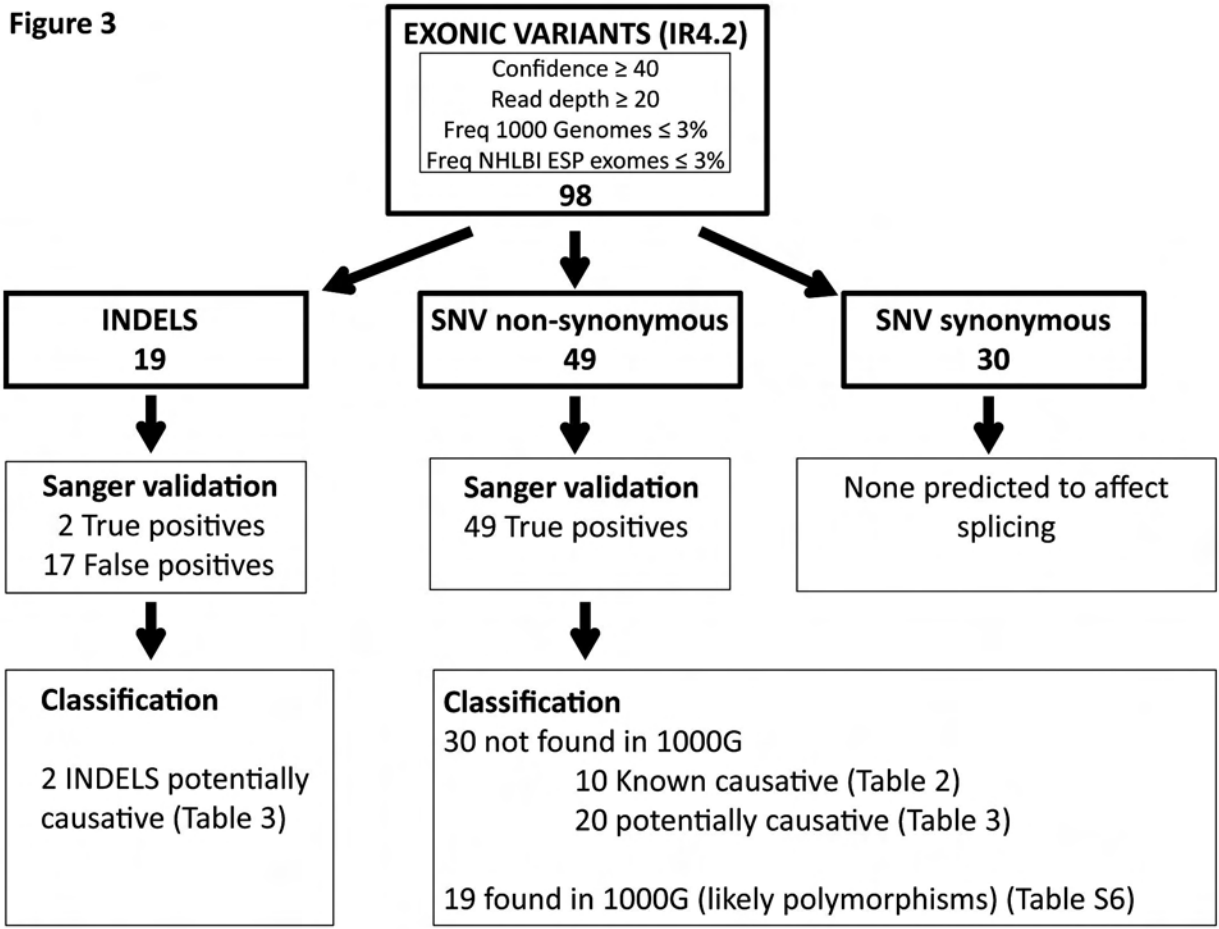
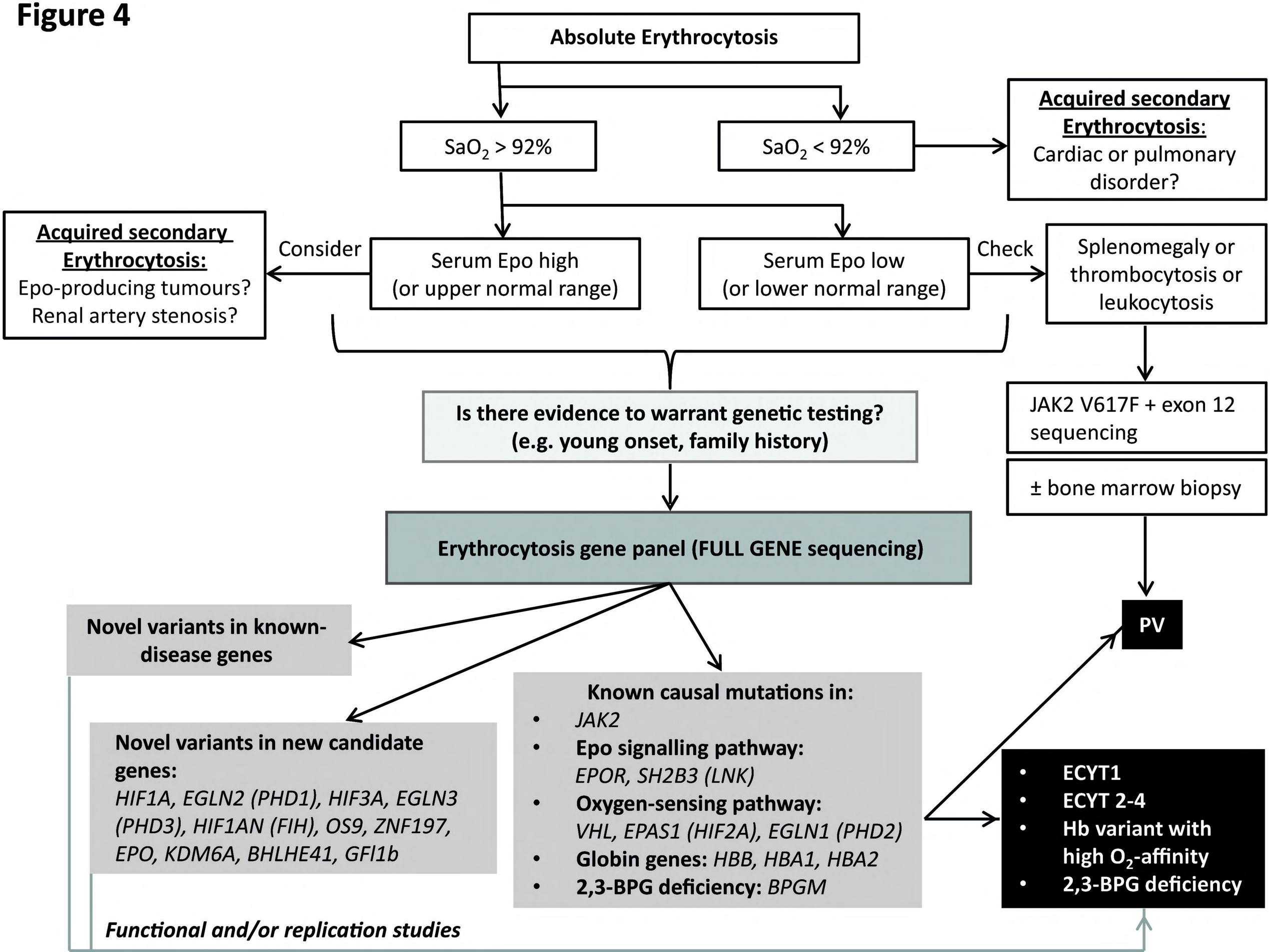

EXONIC VARIANTS (IR4.2)
Confidence ≥ 40
Read depth ≥ 20
Freq 1000 Genomes ≤ 3%
Freq NHLBI ESP exomes ≤ 3%
98
INDELS
19
SNV non-synonymous
49
SNV synonymous
30
Sanger validation
2 True positives
17 False positives
Sanger validation
49 True positives
None predicted to affect splicing
Classification
2 INDELS potentially causative (Table 3)
Classification
30 not found in 1000G
10 Known causative (Table 2)
20 potentially causative (Table 3)
19 found in 1000G (likely polymorphisms) (Table S6)

Figure 4


Absolute Erythrocytosis
SaO_2 > 92%
SaO_2 < 92%
Acquired secondary Erythrocytosis: Cardiac or pulmonary disorder?
Acquired secondary Erythrocytosis: Epo-producing tumours? Renal artery stenosis?
Consider
Serum Epo high (or upper normal range)
Serum Epo low (or lower normal range)
Check
Splenomegaly or thrombocytosis or leukocytosis
Is there evidence to warrant genetic testing? (e.g. young onset, family history)
JAK2 V617F + exon 12 sequencing
± bone marrow biopsy
Erythrocytosis gene panel (FULL GENE sequencing)
PV
Novel variants in known-disease genes
Novel variants in new candidate genes: HIF1A, EGLN2 (PHD1), HIF3A, EGLN3 (PHD3), HIF1AN (FIH), OS9, ZNF197, EPO, KDM6A, BHLHE41, GFI1b
Known causal mutations in:
• JAK2
• Epo signalling pathway: EPOR, SH2B3 (LNK)
• Oxygen-sensing pathway: VHL, EPAS1 (HIF2A), EGLN1 (PHD2)
• Globin genes: HBB, HBA1, HBA2
• 2,3-BPG deficiency: BPGM
• ECYT1
• ECYT 2-4
• Hb variant with high O_2-affinity
• 2,3-BPG deficiency
Functional and/or replication studies

Gene panel sequencing improves the diagnostic work-up of patients with idiopathic erythrocytosis and identifies new mutations

Carme Camps, Nayia Petousi, Celeste Bento, Holger Cario, Richard R. Copley, Mary Frances McMullin, Richard vanWijk, WGS500 consortium, Peter J. Ratcliffe, Peter A. Robbins, and Jenny C. Taylor

Supplementary Information:

Whole genome sequencing (WGS) – Supplementary Methods and Results:

WGS500 is a project aimed at evaluating the clinical utility of whole genome sequencing across a number of human diseases. In WGS500, the genomes of 500 patients and family members, spanning a range of diseases including rare (inherited) disorders, severe and early onset immunological conditions and cancer, were sequenced with the hope of identifying variants in novel genes or pathways to inform diagnosis, prognosis, and treatment.

Ten idiopathic erythrocytosis patient samples were whole-genome sequenced as part of the WGS500 project. The samples included 3 sporadic cases, 2 unrelated families exhibiting an autosomal pattern of inheritance of erythrocytosis and 2 cases that were distant relatives. These patients were specifically selected as having either elevated or inappropriately normal erythropoietin (Epo) levels.

Details about the patients and family pedigrees, as well as the methodology employed in the WGS500 project is described in detail elsewhere[1]. Briefly, samples were sequenced at a 30X depth using the Illumina HiSeq2000. Reads were mapped to the human reference genome (Hg19) using STAMPY[2], and variants were identified and annotated using Platypus[3] and ANNOVAR[4]. We first searched for heterozygous or homozygous variants in candidate genes: *HIF1A, EPAS1, HIF3A, HIF1B, FIH, EGLN1, EGLN2, EGLN3, VHL, EPO, EPOR, JAK2, HBB, HBA1, HBA2* and *BPGM* and then extended our search for rare coding variants in any gene. For the sporadic cases, a recessive model was favored, searching for rare homozygous variants that had to fulfil the following criteria: not reported in 1000G database or reported with a frequency <0.05, not reported in dbSNP and not present as homozygous in the WGS500 union file (file containing all variants called across all samples sequenced in the WGS500 project). The rare homozygous variants in protein coding regions were further filtered by requiring a polyphen2 score >0.5 and prioritized based on gene function. For families, a dominant inheritance pattern was assumed. Search was focused on rare variants (1000G frequency <0.05), giving priority to shared familial variants and genes with variants in common between different patients.

Candidate variants identified by WGS are shown in Table S 2. The variants found in *BPGM* and *EPO* have been reported in other publications from our group[1, 5]. *BPGM* has previously been shown to be affected in erythrocytosis[6]. The variant was detected in one of the sporadic cases and further experiments demonstrated that it was impairing BPGM function[5]. The variant in *EPO* was found in common among the patients of the two unrelated families and further studies demonstrated that

the variant segregated with the erythrocytosis phenotype in both families[1]. These results are further supported by segregation of the same variant in an additional family, reported at an international conference[7]. Overall, this is the first disease-causing variant reported in the *EPO* gene.

WGS also identified rare coding homozygous variants in other novel genes not previously associated with erythrocytosis, which are currently of unknown functional significance: *GFI1b, KDM6A and BHLHE41.* The variant in *GFI1b* was identified in a sporadic case and prioritized among 24 other rare homozygous protein coding variants found in the same patient due the function of this gene: *GFI1b* is an essential transcriptional regulator of erythroid and megakaryocyte development[8] which affects hematopoiesis as shown in knock-out mice studies[9]. The variant p.C168F found in this patient would remove a conserved cysteine of a zinc finger domain of this protein. The variant in *KDM6A*, an X-linked gene, was identified in hemizygous status in a male child who presented as a sporadic case and we subsequently showed that this variant was inherited from his mother. The variant was prioritized among 4 other rare homozygous protein coding variants due to the connection of *KDM6A* with oxygen sensing pathways. Indeed, KDM6A is a chromosome X-coded JmJC-domain-containing demethylase, which is oxygen-dependent and also upregulated in hypoxia[10], with variants affecting its function found in renal cancer[11, 12] and the congenital Kabuki syndrome[13, 14]. The two identified variants in *BHLHE41* (*DEC2*) are located in the 3'UTR and co-occur in the homozygous state in two patients with erythrocytosis who are distantly related (female patient: first cousin of the father of the male patient). They were not found as homozygous in any of the other WGS500 samples. There were no other candidate variants (rare homozygous or heterozygous variants) in common between both patients or located within the same gene in both patients. BHLHE41 (DEC2) is a hypoxia-regulated transcription factor which interacts with HIF and causes proteasomal inhibition of HIF and transcriptional suppression of HIF-target genes[15], and which has also been linked to renal cancer susceptibility[16] and Ethiopian high altitude adaptation [17]. The candidacy of this gene may not appear as strong as for most of the other genes, but its link with the HIF pathway prompted us to include *BHLHE41* together with *EPO*, *GFI1b* and *KDM6A* in the targeted NGS erythrocytosis gene panel and explore its variation across a larger cohort of erythrocytosis patients.

Erythrocytosis gene panel – Supplementary Methods:

Patient samples:

A hundred and twenty five samples were included, obtained from 4 idiopathic erythrocytosis databases (UK, Portugal, Germany and the Netherlands). Participants gave informed consent according to the declaration of Helsinki and appropriate ethical approval was gained for each center where samples were collected. Relevant ethics committee reference numbers have been provided to the journal editors. Of those, 90 (72.0%) were male and 35 (28.0%) were female. To the best of our knowledge, the age of diagnosis was known for 109 of the patients, mostly comprised between childhood and early adulthood (median age: 24; age range: 1 - 57). There were 16 patients also included, who at the time of study were 60 years old or older; these had long-standing erythrocytosis (of several decades) with no identifiable cause. For inclusion, patients had to have an elevated red cell mass of > 125% predicted, and a hemoglobin (Hb) > 180 g/L and

hematocrit (Hct) > 0.52 L/L in adult males or Hb > 160 g/L and Hct > 0.48 L/L in adult females, or Hb and Hct levels above the 99th centile of age-appropriate reference values in children, at the time of diagnosis. At the time of sampling for this study, some patients had normal Hb levels due to previous venesection. Epo reference levels vary from laboratory to laboratory and Epo levels can also vary within an individual with repeated measures, so patients were included irrespective of Epo levels, with the median Epo level being 12.4 miU/ml. The investigation algorithm followed at each Centre prior to registration as idiopathic is shown in Figure S1.

Ion Torrent sequencing:

The custom panel primer pool for idiopathic erythrocytosis was used together with the Ion Ampliseq Library kit 2.0 (*Thermo Fisher*) to create libraries suitable for sequencing on the Ion Torrent platform (*Thermo Fisher*). For each patient DNA sample, two amplification reactions were set up, one for each of the two multiplex pools. Each amplification reaction contained 10ng of genomic DNA, 1X primer pool and 1X Ion Ampliseq HiFi Master mix in a total volume of 10µl complemented with water. A peqSTAR 96 Universal Gradient Thermocycler (*Peqlab*) was used and cycling conditions included a first step to activate the enzyme (99°C for 2 minutes) followed by 17 cycles of amplification (99°C for 15 seconds, 60°C for 4 minutes). The two amplification products resulting from each DNA sample were subsequently combined in a single tube and library preparation was completed according to Ion Ampliseq Library kit 2.0 manual. Ion Xpress barcodes Adapters (*Thermo Fisher*) were used during adapter ligation to allow multiplexing during sequencing. The quality and concentration of the final libraries were assessed using a High sensitivity DNA kit (*Agilent Technologies*) and Agilent 2100 Bioanalyzer (*Agilent Technologies, Santa Clara, California, USA*). The concentration of each library was normalized to 100 pM and pools of 8 libraries were made by combining equal amounts. Each pool was further diluted to 10pM and used for template preparation, using the Ion PGM Template OT2 200 kit and the Ion OneTouch 2 instrument (*Thermo Fisher*), followed by enrichment on template-positive Ion Sphere Particles with the Ion OneTouch ES (*Thermo Fisher*). The template was further processed using the Ion PGM sequencing 200 kit v2, loaded onto an Ion 316 chip and sequenced on an Ion PGM instrument (500 flows), as per the manufacturer's protocol.

Analysis of Ion Torrent Sequencing data:

The Torrent Suite Software (*Thermo Fisher*) was used for basic quality control of the sequencing data generated by the Ion PGM instrument as well as for read alignment to the human genome (Hg19). The alignment was restricted to the genomic coordinates enclosed by our custom panel. An individual BAM file was generated for each sample and imported into the Ion Reporter Software v4.2 (*Thermo Fisher*) for variant calling, performed using the germline workflow for single samples and the default parameters. The resulting vcf files were further annotated with ANNOVAR[4]. Variants were subsequently filtered, selecting for further analysis only the variants fulfilling all of the following conditions: confidence ≥ 40, read depth ≥ 20, frequency in 1000 Genomes (1000G) ≤ 3% and frequency in NHLBI ESP exomes (6500si) ≤ 3%.

SIFT, PolyPhen-2 and Provean were used to evaluate the potential of causality of non-synonymous variants. For SIFT and Polyphen-2 HDIV, we used the scores and cut-offs obtained from the LJB23

database in ANNOVAR. According to this, a variant is considered deleterious (D) by SIFT when sift score ≤0.05 and tolerated (T) when sift score >0.05. For PolyPhen 2 HDIV, a variant is classified as probably damaging (D) when pp2_hdiv score ≥0.957, possibly damaging (P) when 0.453≤ pp2_hdiv score ≤0.956, or benign (B) when pp2_hdiv score ≤0.446. Regarding Provean (http://provean.jcvi.org), we used the default score threshold set at -2.5 for binary classification of the variants (i.e. deleterious vs neutral).

Synonymous variants were further investigated for possible splicing effects using Human Splicing Finder (http://www.umd.be), NetGene2 (http://www.cbs.dtu.dk) and FSPLICE http://linux1.softberry.com).

Sanger sequencing:

The FastStart Taq DNA polymerase kit (Roche) was used for setting up PCR reactions, each one containing 40ng of DNA, 1X buffer supplied with magnesium, 0.2mM dNTP (each), 1.25U of Taq polymerase and 0.4pM of forward and reverse primers. Some reactions aimed to amplify genomic regions with high GC content were complemented with 1X GC rich solution, as indicated in Supplementary Table 2. Cycling conditions included a first step at 95°C for 2 minutes followed by 35 cycles of amplification (95°C for 30 seconds, Ta for 30 seconds as specified in Supplementary Table 2 for each pair of primers, 72°C for 30 seconds) and a final amplification at 72°C for 6 minutes. All PCR products (5µl) were run on a 1% agarose gel. They were then cleaned in a reaction containing 15µl of PCR product, 0.1µl exonuclease I (*NEB*), 1µl shrimp alkaline phosphatase (SAP) (*Affymetrix*), 1µl 10X SAP buffer and 0.9µl of water, which was incubated at 37°C for 30 minutes and 80°C for 15 minutes. Sanger sequencing reactions were set up with 1µl of clean PCR product, 0.5µl of 3.3pM primer, 1.5µl of 5X buffer, 1µl of Big Dye (*Applied Biosystems*) and 6µl of water. Reactions were then incubated at 96°C for 1 minute, followed by 35 cycles of amplification (96°C for 30 seconds, 50°C for 15 seconds and 60°C for 4 minutes). Products were precipitated for 15 minutes at room temperature with a mixture containing 2µl of 125mM EDTA, 2µl of 3M sodium acetate and 50µl of ethanol and pelleted by centrifugation for 30 minutes at 3000 rcf and 4°C. Pellets were washed once with 70µl of 70% ethanol, centrifuged for 15 minutes at 1650 rcf at 4°C and allowed to dry at room temperature for 1 hour. Once dry, they were frozen and submitted to Oxford University Zoology department for final processing.

WGS500 consortium

Steering committee

Peter Donnelly (Chair)[1], John Bell[2] , David Bentley[3], Gil McVean[1], Peter Ratcliffe[1], Jenny Taylor[1,4], Andrew Wilkie[4, 5]

Operations committee

Peter Donnelly[1] (Chair) John Broxholme[1], David Buck[1], Jean-Baptiste Cazier[1], Richard Cornall[1], Lorna Gregory[1], Julian Knight[1], Gerton Lunter[1], Gilean McVean[1], Jenny Taylor[1,4], Ian Tomlinson[1, 4], Andrew Wilkie[4, 5]

Sequencing & experimental follow up
David Buck[1] (Lead) Christopher Allan[1], Moustafa Attar[1], Angie Green[1], Lorna Gregory[1], Sean Humphray[3,] Zoya Kingsbury[3], Sarah Lamble[1], Lorne Lonie[1], Alistair Pagnamenta[1], Paolo Piazza[1], Guadelupe Polanco[1], Amy Trebes[1]

Data analysis
Gil McVean[1] (Lead), Peter Donnelly[1], Jean-Baptiste Cazier[1], John Broxholme[1], Richard Copley[1], Simon Fiddy[1], Russell Grocock[3], Edouard Hatton[1], Chris Holmes[1], Linda Hughes[1], Peter Humburg[1], Alexander Kanapin[1], Stefano Lise[1], Gerton Lunter[1], Hilary Martin[1], Lisa Murray[3] , Davis McCarthy[1], Andy Rimmer[1], Natasha Sahgal[1], Ben Wright[1], Chris Yau[6]

[1] The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK.
[2] Office of the Regius Professor of Medicine, Richard Doll Building, Roosevelt Drive, Oxford, OX3 7LF, UK
[3] Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Essex, CB10 1XL, UK
[4] NIHR Oxford Biomedical Research Centre, Oxford, UK.
[5] Weatherall Inst of Molecular Medicine, University of Oxford; John Radcliffe Hospital Headington, Oxford OX3 9DS, UK
[6] Imperial College London, South Kensington Campus, London, SW7 2AZ. UK

Supplementary Figure Legends:

Figure S1. Algorithms used for the diagnosis of erythrocytosis in the different centers participating in this study: University Medical Center, Ulm (Germany), Queen's University, Belfast (UK), University Medical Center (The Netherlands) and Centro Hospitalar e Universitário de Coimbra (Portugal). Patients for whom no diagnosis was reached following these algorithms were classified as having idiopathic erythrocytosis. (PV: Polycythemia Vera; ECYT 1-4: erythrocytosis type 1-4; Hb: hemoglobin; Epo: erythropoietin; Hct: hematocrit; SD: standard deviation; SaO_2: saturation of oxygen; P50: partial pressure of oxygen for which 50% of Hb is saturated with oxygen; 2,3-BPG: 2,3-bisphosphoglycerate; O_2: oxygen; ECC: erythroid colony culture; MRI: magnetic resonance imaging).

Supplementary Tables:

Table S 1: Primers used for Sanger sequencing validation

Sequence (5'->3')	F/R	Ta	Size amplicon (bp)	Chr	Start	End	Gene	GC rich solution added
TGCCTTACGATGACAGAAATGG	F	58	342	9	5022000	5022341	*JAK2*	no
AAACACACACGCCAGCCATA	R							
GGTTTTAGTGGCGGCATGAT	F	55	251	9	5050678	5050928	*JAK2*	no
ACAAAATCAAAAGGCATGGGTAA	R							
AGGAAGCGAATAAGGTACAGATT	F	55	239	9	5054624	5054862	*JAK2*	no
TCTTTATGTTTCCCTCTTGACCA	R							
CCTTTGCTCAGAGGGACCTG	F	55	181	12	26275882	26276062	*BHLHE41*	no
ATGCGTTCCACTCGGGATTT	R							

GTGCCATATGCACAGTGTATGC	F	58	225	12	58109416	58109640	*OS9*	no
CGAGGAGTGCGAATGGTCAG	R							
GAACGGCAGAGAGAGATGGA	F	55	267	12	58112016	58112282	*OS9*	no
TTCTTTTAGCCCGTCAGCCT	R							
CTTCGTCAGAGCCCTTGGAG	F	58	184	19	41306536	41306719	*EGLN2*	no
CCAAAACCGTCCCGAAGAGG	R							
GTCTCCAGGTACGCCATCAC	F	58	425	19	41313381	41313805	*EGLN2*	no
TTTAAGCTGTCTGCCATGCG	R							
GAACAGCTCCGCGCAAATAG	F	55	273	X	44921898	44922170	*KDM6A*	no
TGCTTTTGTCTAAAAATCTCCGTCA	R							
AGAGTGCCTAGCGTCTCTCA	F	55	232	X	44922784	44923015	*KDM6A*	no
TGGTCAGGTTTGTGCGGTTA	R							
ACTTACTTTCGTCCGGCCAT	F	58	285	1	231556738	231557022	*EGLN1*	no
CACGGCATCTGTGTGGTG	R							
GCCAGATCTCGGCGAAGTAA	F	55	243	14	62187140	62187382	*HIF1A*	no
TCAAAACATTGCGACCACCT	R							
TGCCTATCAGTTAACTTGGGAGG	F	58	262	14	62199049	62199310	*HIF1A*	no
GCCAAACTGTACAGAGGTTGC	R							
GAGCAGGGGAATGAGGATGG	F	58	248	19	46811401	46811648	*HIF3A*	no
ACAGGAGGTGGGGATATGCT	R							
GCTCTGGACATATGAGGGCC	F	58	276	19	46812418	46812693	*HIF3A*	no
AGAAATGCGGGAGTGTGGAG	R							
CAGGGCAGTATCGCTTCCTG	F	58	261	19	46815787	46816047	*HIF3A*	no
CGTGCACACTCCCTCACATA	R							
CACCTCCCTTTCTGCCTTGT	F	58	291	19	46823673	46823963	*HIF3A*	no
GCTGTGTGTTTTGGAGGCTG	R							
CCTGGCATTTGATCCCCACT	F	60	199	19	46828688	46828886	*HIF3A*	no
TCTAAATCTGTCTCCACTGCC	R							
ACCCCTCTGCGCAAAAGTAA	F	55	300	19	46842623	46842922	*HIF3A*	no
TCCTTTCTGGGGGAGGAGAA	R							
CAAAGCAGGTTGTGTGTGGC	F	58	310	2	46573851	46574160	*EPAS1*	no
GTCGCATGATGGAGGCCTT	R							
CAACCCTGTTCCCTTCCTCC	F	58	210	12	111884719	111884928	*SH2B3*	no
CTGCTGGAGAAGAGGCTGAG	R							
CTGCCAGAAGACGGACCATT	F	58	357	12	111885165	111885521	*SH2B3*	no
GAGGGAAAGTGGAGGTGCTG	R							
TCTCTAGTCTCACGAGGGGT	F	55	284	14	62162373	62162656	*HIF1A*	yes
CCCAATCCCATTAACGCCG	R							
GAGCCTGCCTGCCTTCAC	F	58	252	2	46605139	46605390	*EPAS1*	no
AGAAAACAGCTCTGATACCTGGT	R							
CGTTTGAGCAGCACTGTGAA	F	58	364	2	46607245	46607608	*EPAS1*	no
GGGCTCTGTCTTCTTGCTCT	R							
AGACACCACTGAAGGAGCA	F	55	305	2	46611521	46611825	*EPAS1*	no
GGTGCTGCCCAGGTAGAA	R							
GCGGAGAACTGGGACGAG	F	58	383	3	10183544	10183926	*VHL*	no
GCTTCAGACCGTGCTATCGT	R							
AGCCTCTTGTTCGTTCCTTGT	F	55	432	3	10191403	10191834	*VHL*	no
TGTTTGCCCCTAAACATCACA	R							

AGAGACGTGGGGATGAAGGA	F	58	263	7	100319096	100319358	*EPO*	no
TTCTGTGGAATGTGCTGGGG	R							
GCAGGAGGGAGAGGGTGA	F	60	255	7	100320232	100320486	*EPO*	no
AAGTGTCCGCTCCTACTCAC	R							
AAAGCCAGCAGATCCTACGG	F	58	174	7	100319504	100319677	*EPO*	no
ACTCACCTCCATCCTCTTCCA	R							
CCACCCAACCATGTCTTCCA	F	58	288	9	5029845	5030132	*JAK2*	no
GCATCCCACCAAAACAGCTT	R							
GGAAGCTTTGTCTTTCGTGTCAT	F	58	132	9	5064906	5065037	*JAK2*	no
ATTGGGCCATGACAGTTGCT	R							
TCTTGTTCCTACTTCGTTCTCCA	F	58	289	9	5072430	5072718	*JAK2*	no
TGAAAAGCTGCACACATGAGT	R							
TCAGGGGATTTGTGTTGAGTTTA	F	58	267	9	5126105	5126371	*JAK2*	no
CTGTCCTTGTTTGTCATTGCCA	R							
TGACATGTGCCCTGTATTGAA	F	55	202	9	5126590	5126791	*JAK2*	no
TCATCCAGCCATGTTATCCCT	R							
CAGCCTGGTGTCCTAAGAGC	F	58	198	19	11488648	11488845	*EPOR*	no
ACCCCACCCCACCTAAAGTA	R							
CTCCGACCACTCCTCCATTC	F	58	202	19	11492573	11492774	*EPOR*	no
CCATGGAAGCTGTGTCGC	R							
CGCTTCCTCCCAGAAACACA	F	58	118	19	11493823	11493940	*EPOR*	no
CCCCATGCCCTTCCTTTGTC	R							
AGCGAGTTCTGTGAGTTGCA	F	58	201	12	111856028	111856228	*SH2B3*	yes
GCCTGTGTCCCGGTAGTC	R							
CGGAGAGGCTGCTGAGAC	F	58	236	12	111856447	111856682	*SH2B3*	yes
CCTTGGGTGGGTCGAAGA	R							
AGTTGGACTTAGGGAACAAAGGA	F	58	242	11	5246770	5247011	*HBB*	no
TCCAAGCTAGGCCCTTTTGC	R							
TCCCATAGACTCACCCTGAAGT	F	58	437	11	5247793	5248229	*HBB*	no
AGAAGTCTGCCGTTACTGCC	R							
TCATGAGCAGCCCAATGGTT	F	58	499	1	231556815	231557313	*EGLN1*	yes
GCCAAGGGAAAAGTAAAGGCC	R							
CCTCAAACAGCAGGGGACAT	F	58	363	19	11493676	11494038	*EPOR*	no
CTTGGGAAGATGGCAGGGC	R							
GAAGCTCGGAACTGTGGGAA	F	58	356	7	134346368	134346723	*BPGM*	no
GACCGTGGCAGTTGATCCAA	R							
TTTACTTTTCCCTTGGCCGC	F	58	448	1	231557297	231557744	*EGLN1*	yes
CAGTAACGGCCCCTATCTCT	R							
GTGAGTAGGAGCGGACACTT	F	58	300	7	100320467	100320766	*EPO*	no
ACACACCTGGTCATCTGTCC	R							
GGAATCCCTCAGTACCTGCA	F	58	384	7	134363477	134363860	*BPGM*	no
GAGAGAGAGTTCAGCACCCA	R							
TCTGATGTACCAACCTCACCA	F	58	168	9	5069963	5070130	*JAK2*	no
TCACATGAATGTAAATCAAGAAAACA	R							
CCAGTACCACACACCTGCTA	F	58	398	12	58111835	58112232	*OS9*	no
AAGCAACAGGAGGAAGAGCT	R							
GGTCAGAGTTCACATTCGGC	F	58	363	X	44928836	44929198	*KDM6A*	no
ATCTCCATGGCTAGGACTGC	R							

CATGAGACATCTGGACCCCA	F	58	459	3	44670797	44671255	*ZNF197*	yes
TCTCAGTTGGACCCGAAGAC	R							
TCTCTCTCGCAGCTCATCTC	F	58	442	10	102295595	102296036	*HIF1AN*	yes
TCCGTATCTCCTCGCCTTTC	R							
TCCAGCTTCATCCTCTTGGG	F	58	424	12	26275600	26276023	*BHLHE41*	yes
TCTTGCAATACCTCTCCCGG	R							
TGGGGAATACGTGCTCACTT	F	58	470	12	111885306	111885775	*SH2B3*	no
GACCAAGAGAGACCACACCA	R							
CGATGGACTTGGTTGTGTGT	F	58	495	14	62207107	62207601	*HIF1A*	no
TTGAGGACTTGCGCTTTCAG	R							
AGGGACCTTAGCACCAAGTC	F	58	438	19	11494456	11494893	*EPOR*	no
GGGCTGTATCATGGACCACC	R							
AGCTCAGACTGTTGACCACA	F	58	475	19	46832351	46832825	*HIF3A*	no
AAATGGTGAGGGATGAGGCT	R							
GAAGAAGACGGCGGGGAG	F	58	400	3	10183607	10184006	*VHL*	no
AGCAGCGTCACCCTGGAT	R							

Official gene symbols according to HUGO Gene Nomenclature Committee are given here. Other gene symbols used frequently in the literature are: *HIF2A* (*EPAS1*), *PHD2* (*EGLN1*), *PHD1* (*EGLN2*), *PHD3* (*EGLN3*), *FIH* (*HIF1AN*), *LNK* (*SH2B3*), *DEC2* (BHLHE41***)***.
F indicates forward primer; R, reverse primer; Ta, annealing temperature; bp, base pairs; and Chr, chromosome.

Table S 2: Variants identified by whole genome sequencing (WGS500 project)

Chr	Position	Ref	Alt	Gene	Transcript ID	cDNA Change	Protein Change	Genotype	No of cases	dbSNP142 (allelic freq)	Sample ID
7	100318468	G	A	*EPO*	NM_000799	c.-136G>A	NA	Het	4	Not found	PAR07, PAR09, PAR15, PAR16
7	134346528	G	A	*BPGM*	NM_001724	c.G269A	p.R90H	Het	1	Not found	PAR03
9	135863848	G	T	*GFI1B*	NM_004188	c.G503T	p.C168F	Hom	1	rs527297896 (0.001)	PAR02
12	26273317	C	T	*BHLHE41*	NM_030762	c.*1682G>A	NA	Hom	2	rs76268917 (0.038)	PAR04, PAR12
12	26274410	T	C	*BHLHE41*	NM_030762	c.*589A>G	NA	Hom	2	rs76306214 (0.036)	PAR04, PAR12
X	44920641	T	C	*KDM6A*	NM_021140	c.T1402C	p.C468R	X-linked (Male)	1	rs138723332 (0.00132)	PAR11

Sporadic Cases: PAR02, PAR03, PAR11; Family M: PAR15, PAR16 (affected siblings); Family S: PAR07, PAR09 (affected mother and daughter); Family T: PAR04, PAR12 (distant relatives, both affected). Official gene symbols according to HUGO Gene Nomenclature Committee are given here. Other gene symbols used frequently in the literature are: *DEC2* (*BHLHE41*).
Chr indicates chromosome; Ref, reference allele; Alt, alternate allele; NA, non-applicable; Het, heterozygous; and Hom, homozygous.

Table S 3: Amplicons generated by the erythrocytosis gene panel with poor coverage

Amplicon ID	Gene	Gene region	Average No reads	Max N reads	Chr	Start	End	Length	Description
AMPL3630659372	*BHLHE41*	CDS	0.26	3	12	26275506	26275647	141	Failed in all samples
AMPL3774175241	*KDM6A*	3'UTR	0.31	4	X	44971547	44971692	145	Failed in all samples
AMPL3774508291	*EGLN1*	5'UTR	0.34	3	1	231558051	231558231	180	Failed in all samples
AMPL3774242165	*EGLN2*	5'UTR	0.55	4	19	41305339	41305449	110	Failed in all samples
AMPL3630748538	*EGLN2*	5'UTR	0.69	15	19	41304979	41305160	181	Failed in all samples
AMPL3630468797	*HIF3A*	5'UTR	1.56	12	19	46806842	46806957	115	Failed in all samples

AMPL706138063	*VHL*	CDS	1.82	7	3	10183733	10183902	169	Failed in all samples
AMPL844175990	*EGLN1*	CDS	2.22	8	1	231557578	231557757	179	Failed in all samples
AMPL3773690924	*HIF3A*	CDS	2.82	8	19	46838138	46838313	175	Failed in all samples
AMPL3774166343	*KDM6A*	CDS	4.75	18	X	44732714	44732847	133	Failed in all samples
AMPL3774152291	*EPAS1*	CDS	9.37	29	2	46607762	46607856	94	Average coverage lower than 20X
AMPL3630701854	*BHLHE41*	CDS	9.82	44	12	26275365	26275467	102	Average coverage lower than 20X
AMPL3774508232	*EGLN1*	5'UTR	12.41	47	1	231557881	231558041	160	Average coverage lower than 20X
AMPL2721812404	*SH2B3*	5'UTR	14.60	56	12	111843727	111843905	178	Average coverage lower than 20X
AMPL3630680367	*GFI1B*	CDS	15.82	47	9	135864528	135864712	184	Average coverage lower than 20X
AMPL3630701704	*BHLHE41*	CDS	16.12	63	12	26275093	26275281	188	Average coverage lower than 20X
AMPL844172391	*EGLN1*	5'UTR	18.88	64	1	231558201	231558343	142	Average coverage lower than 20X

These amplicons are encompassing CDS regions from *BHLHE41*, *EPAS1*, *GFI1B*, *VHL*, *HIF3A* and *KDM6A* genes (1,365bp in total), as well as UTR regions from *EGLN1*, *HIF3A*, *EGLN2*, *SH2B3* and *KDM6A* genes (1,211bp in total). Official gene symbols according to HUGO Gene Nomenclature Committee are given here. Other gene symbols used frequently in the literature are: *HIF2A* (*EPAS1*), *PHD2* (*EGLN1*), *PHD1* (*EGLN2*), *LNK* (*SH2B3*), *DEC2* (BHLHE41***)***.

Chr indicates chromosome; CDS, coding DNA sequence; and UTR, untranslated region.

Table S 4: Variants in positive controls successfully identified by the erythrocytosis gene panel

Sample ID	Chr	Position	Ref	Alt	Genotype	Gene region	Gene	Transcript ID	cDNA change	Protein change	Previous detection
C_001	19	11488877	C	T	Het	exonic	*EPOR*	NM_000121	c.G1310A	p.R437H	Sanger
C_002	1	231509737	A	G	Het	exonic	*EGLN1*	NM_022051	c.T1000C	p.W334R	Sanger
C_003	3	10191593	A	G	Het	exonic	*VHL*	NM_000551	c.A586G	p.K196E	Sanger
C_004	2	46607719	T	C	Het	exonic	*EPAS1*	NM_001430	c.T1908C	p.N636N	Sanger
PAR02	9	135863848	G	T	Hom	exonic	*GFI1B*	NM_004188	c.G503T	p.C168F	WGS
PAR03	7	134346528	G	A	Het	exonic	*BPGM*	NM_001724	c.G269A	p.R90H	WGS
PAR11	X	44920641	T	C	Hom	exonic	*KDM6A*	NM_021140	c.T1402C	p.C468R	WGS
PAR12	12	26273317	C	T	Hom	3'UTR	*BHLHE41*	NM_030762	c.*1682G>A	NA	WGS
	12	26274410	T	C	Hom	3'UTR	*BHLHE41*	NM_030762	c.*589A>G	NA	WGS
PAR04	12	26273317	C	T	Hom	3'UTR	*BHLHE41*	NM_030762	c.*1682G>A	NA	WGS
	12	26274410	T	C	Hom	3'UTR	*BHLHE41*	NM_030762	c.*589A>G	NA	WGS
PAR07	7	100318468	G	A	Het	5'UTR	*EPO*	NM_000799	c.-136G>A	NA	WGS

Official gene symbols according to HUGO Gene Nomenclature Committee are given here. Other gene symbols used frequently in the literature are: *HIF2A* (*EPAS1*), *PHD2* (*EGLN1*), *DEC2* (*BHLHE41*). Chr indicates chromosome; Ref, reference allele; Alt, alternate allele; Het, heterozygous; Hom, homozygous; NA, non-applicable; and WGS, whole genome sequencing.

Table S 5: All 51 variants detected by the erythrocytosis gene panel across 57 out of 125 patients (and validated by Sanger sequencing)

chr	Position	Ref	Alt	Gene	Transcript ID	cDNA change	Protein change	Total No of Cases	No of heteroz.	No of homoz.
1	231556799	A	G	*EGLN1*	NM_022051	c.T836C	p.L279P	1	1	0
1	231557164	C	G	*EGLN1*	NM_022051	c.G471C	p.Q157H	13	12	1
2	46574031	AAGG	A	*EPAS1*	NM_001430	c.47delAGG	p.del17E	1	1	0
2	46607405	T	C	*EPAS1*	NM_001430	c.T1594C	p.Y532H	2	2	0
2	46607420	G	A	*EPAS1*	NM_001430	c.G1609A	p.G537R	1	1	0
2	46611651	T	C	*EPAS1*	NM_001430	c.T2465C	p.M822T	1	1	0
3	10183605	C	T	*VHL*	NM_000551	c.C74T	p.P25L	2	2	0
3	10183685	G	T	*VHL*	NM_000551	c.G154T	p.E52X	1	1	0
3	10191578	C	G	*VHL*	NM_000551	c.C571G	p.H191D	1	0	1
3	10191605	C	T	*VHL*	NM_000551	c.C598T	p.R200W	4	4	0
7	100319185	TC	T	*EPO*	NM_000799	c.19delC	p.P7fs	1	1	0
7	100319633	G	A	*EPO*	NM_000799	c.G208A	p.D70N	1	1	0
7	100320290	G	C	*EPO*	NM_000799	c.G250C	p.G84R	2	2	0
7	100320336	A	G	*EPO*	NM_000799	c.A296G	p.E99G	1	1	0
7	100320381	C	T	*EPO*	NM_000799	c.C341T	p.P114L	1	2	0
7	100320614	C	G	*EPO*	NM_000799	c.C440G	p.S147C	1	1	0
7	134346563	C	A	*BPGM*	NM_001724	c.C304A	p.Q102K	1	1	0
9	5022168	G	A	*JAK2*	NM_004972	c.G181A	p.E61K	1	1	0
9	5029893	C	G	*JAK2*	NM_004972	c.C337G	p.L113V	1	1	0
9	5050747	A	T	*JAK2*	NM_004972	c.A530T	p.E177V	1	1	0
9	5054775	G	C	*JAK2*	NM_004972	c.G827C	p.G276A	1	1	0
9	5065003	C	G	*JAK2*	NM_004972	c.C1177G	p.L393V	3	3	0
9	5070026	AA	TT	*JAK2*	NM_004972	c.1615_1616invAA	p.K539L	1	1	0
9	5072561	G	A	*JAK2*	NM_004972	c.G1711A	p.G571S	1	1	0
9	5126343	G	A	*JAK2*	NM_004972	c.G3188A	p.R1063H	1	1	0
9	5126715	A	G	*JAK2*	NM_004972	c.A3323G	p.N1108S	2	2	0
11	5246832	T	G	*HBB*	NM_000518	c.A440C	p.H147P	1	1	0
11	5246840	G	C	*HBB*	NM_000518	c.C432G	p.H144Q	1	1	0
11	5246944	C	T	*HBB*	NM_000518	c.G328A	p.V110M	1	1	0
11	5247816	C	G	*HBB*	NM_000518	c.G306C	p.E102D	1	1	0
12	26276001	A	C	*BHLHE41*	NM_030762	c.T447G	p.F149L	1	1	0
12	58109559	G	A	*OS9*	NM_001261421	c.G497A	p.G166D	1	1	0
12	58112155	C	T	*OS9*	NM_001261421	c.C1265T	p.S422L	1	1	0
12	111856181	G	A	*SH2B3*	NM_005475	c.G232A	p.E78K	1	1	0
12	111856506	G	T	*SH2B3*	NM_005475	c.G557T	p.S186I	2	2	0
12	111856571	G	C	*SH2B3*	NM_005475	c.G622C	p.E208Q	1	1	0
12	111884812	G	A	*SH2B3*	NM_005475	c.G901A	p.E301K	1	1	0
12	111885310	G	A	*SH2B3*	NM_005475	c.G1198A	p.E400K	1	1	0
12	111885466	C	T	*SH2B3*	NM_005475	c.C1243T	p.R415C	1	1	0
14	62187212	G	C	*HIF1A*	NM_001530	c.G148C	p.V50L	1	1	0

19	11488727	T	C	*EPOR*	NM_000121	c.A1460G	p.N487S	2	2	0
19	11492737	G	A	*EPOR*	NM_000121	c.C296T	p.A99V	2	2	0
19	11493887	C	T	*EPOR*	NM_000121	c.G137A	p.G46E	3	3	0
19	41306650	C	T	*EGLN2*	NM_053046	c.C173T	p.S58L	4	4	0
19	41313427	G	T	*EGLN2*	NM_053046	c.G1139T	p.R380L	1	1	0
19	41313759	C	T	*EGLN2*	NM_053046	c.C1214T	p.T405M	1	1	0
19	46811511	A	C	*HIF3A*	NM_022462	c.A190C	p.I64L	1	1	0
19	46823777	C	A	*HIF3A*	NM_022462	c.C896A	p.A299D	1	1	0
19	46823803	C	T	*HIF3A*	NM_022462	c.C922T	p.P308S	1	1	0
19	46828843	T	C	*HIF3A*	NM_022462	c.T1180C	p.F394L	1	1	0
X	44922890	C	T	*KDM6A*	NM_021140	c.C1751T	p.T584M	1	1	0

Official gene symbols according to HUGO Gene Nomenclature Committee are given here. Other gene symbols used frequently in the literature are: *HIF2A* (*EPAS1*), *PHD2* (*EGLN1*), *PHD1* (*EGLN2*), *LNK* (*SH2B3*), *DEC2* (*BHLHE41*). Chr indicates chromosome; Ref, reference allele and Alt, alternate allele. These variants were subsequently classified in 3 groups: known causal variants related to erythrocytosis (Table 2 in main manuscript), novel variants not found before in erythrocytosis (Table 3 in main manuscript), and likely non-causative polymorphisms (Table S6).

Table S 6: Non disease-causing variants detected by the erythrocytosis gene panel, also found in the *in silico* control cohort.

Chr	Position	Ref	Alt	Gene	Transcript ID	cDNA change	Protein change	Genotype	No of cases	ERY freq	Control freq	Adj Fisher pval
1	231557164	C	G	*EGLN1*	NM_022051	c.G471C	p.Q157H	Het/ Hom	12 Het/ 1 Hom	0.056	0.053	0.8713
7	100319633	G	A	*EPO*	NM_000799	c.G208A	p.D70N	Het	1	0.004	0.004	1.0000
7	100320381	C	T	*EPO*	NM_000799	c.C341T	p.P114L	Het	1	0.004	0.002	0.6703
7	100320614	C	G	*EPO*	NM_000799	c.C440G	p.S147C	Het	1	0.004	0.0005	0.4621
9	5029893	C	G	*JAK2*	NM_004972	c.C337G	p.L113V	Het	1	0.004	0.0005	0.4621
9	5065003	C	G	*JAK2*	NM_004972	c.C1177G	p.L393V	Het	3	0.012	0.012	1.0000
9	5126343	G	A	*JAK2*	NM_004972	c.G3188A	p.R1063H	Het	1	0.004	0.003	0.7561
9	5126715	A	G	*JAK2*	NM_004972	c.A3323G	p.N1108S	Het	2	0.008	0.001	0.3758
12	58112155	C	T	*OS9*	NM_001261421	c.C1265T	p.S422L	Het	1	0.004	0.015	0.4621
12	111856506	G	T	*SH2B3*	NM_005475	c.G557T	p.S186I	Het	1	0.008	0.188	2.83E-17
14	62187212	G	C	*HIF1A*	NM_001530	c.G148C	p.V50L	Het	1	0.004	0.002	0.6329
19	11488727	T	C	*EPOR*	NM_000121	c.A1460G	p.N487S	Het	2	0.008	0.003	0.4621
19	11492737	G	A	*EPOR*	NM_000121	c.C296T	p.A99V	Het	2	0.008	0.004	0.4621
19	11493887	C	T	*EPOR*	NM_000121	c.G137A	p.G46E	Het	3	0.012	0.003	0.3929
19	41306650	C	T	*EGLN2*	NM_053046	c.C173T	p.S58L	Het	3	0.016	0.014	0.8713
19	41313759	C	T	*EGLN2*	NM_053046	c.C1214T	p.T405M	Het	1	0.004	0.001	0.4621
19	46823803	C	T	*HIF3A*	NM_022462	c.C922T	p.P308S	Het	1	0.004	0.014	0.4621
19	46828843	T	C	*HIF3A*	NM_022462	c.T1180C	p.F394L	Het	1	0.004	0.029	0.1730
X	44922890	C	T	*KDM6A*	NM_021140	c.C1751T	p.T584M	Het	1*	0.004	0.001	0.4621

Variant calling files from 1000 Genomes project, generated by integration of exome and low coverage data across 1041 individuals, were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/consensus_call_sets/snps. Vcf tools were used to extract the variants identified within the coordinates of the amplicons generated by Ampliseq gene panel. The variants were annotated with ANNOVAR and filtered following the same criteria described previously for erythrocytosis cohort and Ion Torrent sequencing data. Common variants between the erythrocytosis and *in silico* control cohorts were identified and differences in their allelic frequencies were assessed with Fisher exact test followed by Benjamini and Hochberg false discovery correction (all analysis were performed using RStudio)[18].

Official gene symbols according to HUGO Gene Nomenclature Committee are given here. Other gene symbols used frequently in the literature are: *PHD2* (*EGLN1*), *PHD1* (*EGLN2*), *LNK* (*SH2B3*). Chr indicates chromosome; Ref, reference allele; Alt, alternate allele; Het, heterozygous and Hom, homozygous.

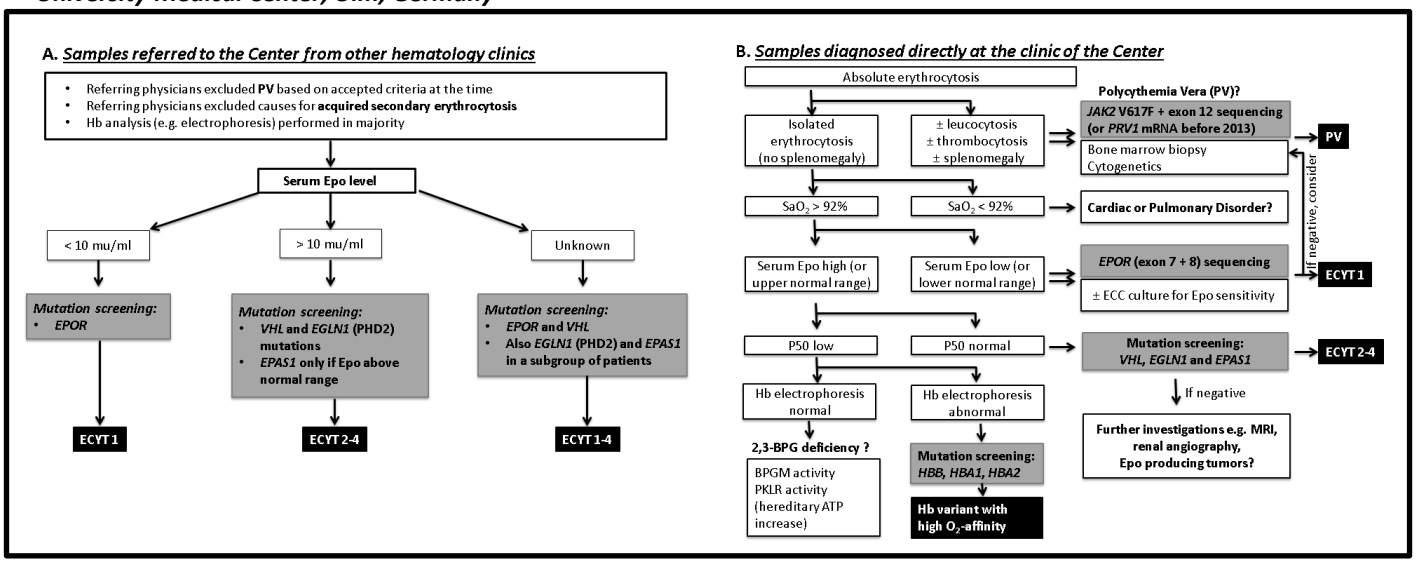
*this patient is a female

References:

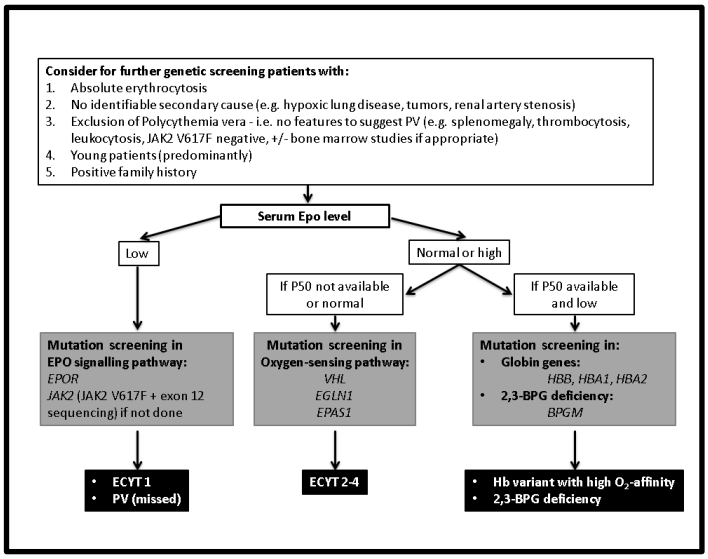
1. Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet. 2015;47(7):717-726.
2. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011;21(6):936-939.
3. Rimmer A MI, Lunter G and McVean G.(2012) Platypus: An Integrated Variant Caller. 2012.
4. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
5. Petousi N, Copley RR, Lappin TR, et al. Erythrocytosis associated with a novel missense mutation in the BPGM gene. Haematologica. 2014;99(10):e201-204.
6. Lemarchandel V, Joulin V, Valentin C, et al. Compound heterozygosity in a complete erythrocyte bisphosphoglycerate mutase deficiency. Blood. 1992;80(10):2643-2649.
7. Lorenzo V FR, Rebecca M, Sabina S, Kimberly H, Karl V, Josef P. A Novel EPO Gene Mutation In a Family With Autosomal Dominant Polycythemia. 55th ASH Annual Meeting and Exposition; New Orleans, LA; 2013.
8. Saleque S, Kim J, Rooke HM, Orkin SH. Epigenetic regulation of hematopoietic differentiation by Gfi-1 and Gfi-1b is mediated by the cofactors CoREST and LSD1. Mol Cell. 2007;27(4):562-572.
9. Saleque S, Cameron S, Orkin SH. The zinc-finger proto-oncogene Gfi-1b is essential for development of the erythroid and megakaryocytic lineages. Genes Dev. 2002;16(3):301-306.
10. Xia X, Lemieux ME, Li W, et al. Integrative analysis of HIF binding and transactivation reveals its role in maintaining histone methylation homeostasis. Proc Natl Acad Sci U S A. 2009;106(11):4260-4265.
11. van Haaften G, Dalgliesh GL, Davies H, et al. Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. Nat Genet. 2009;41(5):521-523.
12. Dalgliesh GL, Furge K, Greenman C, et al. Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. Nature. 2010;463(7279):360-363.
13. Miyake N, Mizuno S, Okamoto N, et al. KDM6A Point Mutations Cause Kabuki Syndrome. Hum Mutat. 2012;
14. Lederer D, Grisart B, Digilio MC, et al. Deletion of KDM6A, a histone demethylase interacting with MLL2, in three patients with Kabuki syndrome. Am J Hum Genet. 2012;90(1):119-124.
15. Montagner M, Enzo E, Forcato M, et al. SHARP1 suppresses breast cancer metastasis by promoting degradation of hypoxia-inducible factors. Nature. 2012;487(7407):380-384.
16. Jessop L BP, Machiela M, Myers T, Sikdar N, Colli L, and Chanock S. Abstract 5061: Post-GWAS functional characterization of the 12p11.23 renal cancer susceptibility locus implicates BHLHE41. Cancer Research; 2014.
17. Huerta-Sanchez E, Degiorgio M, Pagani L, et al. Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. Mol Biol Evol. 2013;30(8):1877-1888.
18. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological. 1995;57(1):289-300.

Figure S1

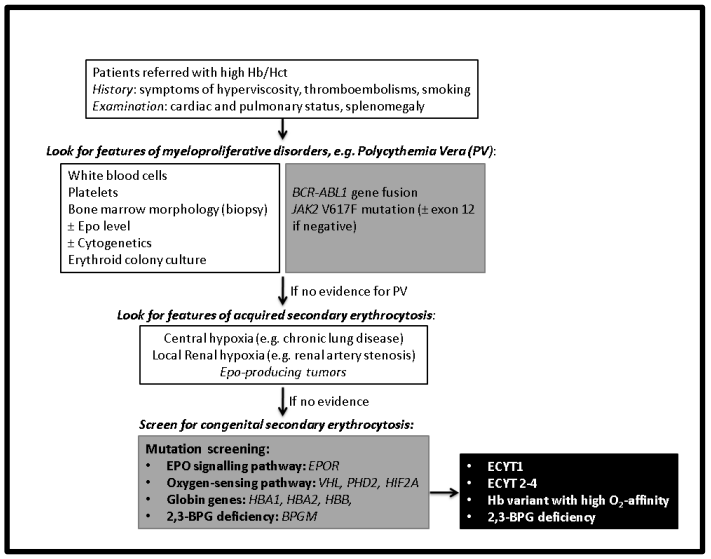
University Medical Center, Ulm, Germany



Queen's University, Belfast, United Kingdom



University Medical Center, Utrecht, The Netherlands



Centro Hospitalar e Universitário de Coimbra, Portugal