SUPPLEMENTARY APPENDIX

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 van Wijk,[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, UK; [3]Nuffield Department of Medicine, University of Oxford, UK; [4]Hematology Department, Centro Hospitalar e Universitário de 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, the Netherlands; [8]A list of members and affiliations is provided in the Online Supplementary Information; and [9]Department of Physiology, Anatomy and Genetics, University of Oxford, UK

**CC and NP contributed equally to this work*

***PJR, PAR and JCT jointly supervised this work*

 doi:10.3324/haematol.2016.144063

Received: February 9, 2016.

Accepted: July 26, 2016.

Pre-published: September 20, 2016.

Correspondence: nayia.petousi@ndm.ox.ac.uk

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.

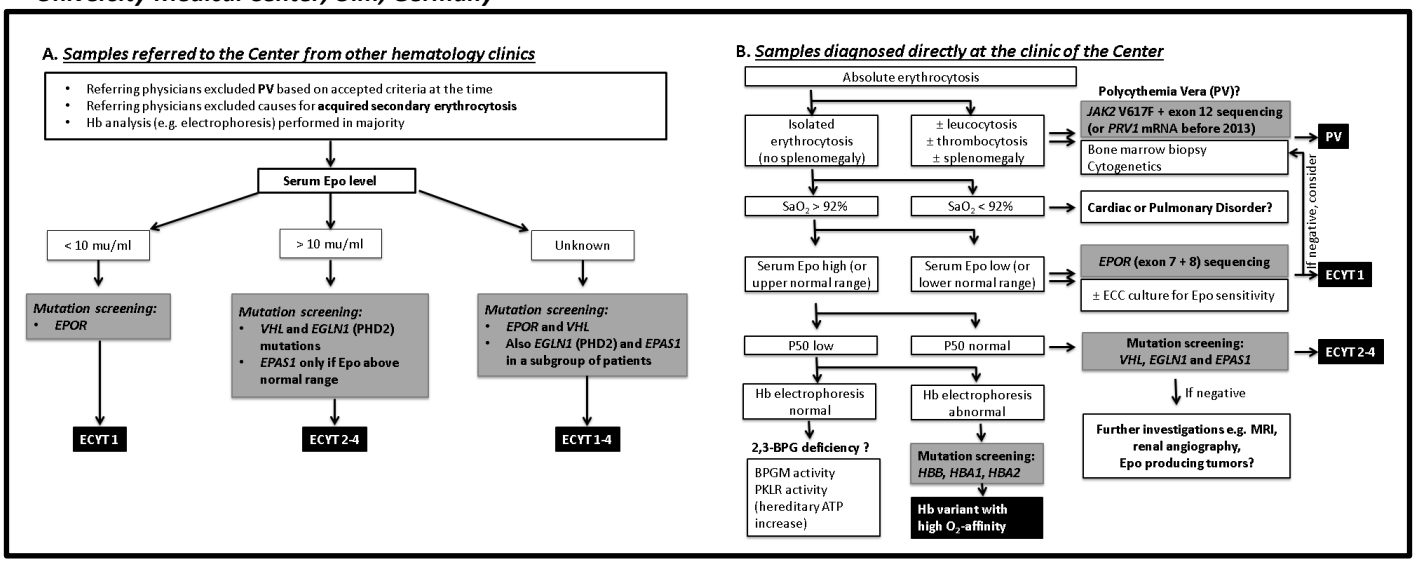
*this patient is a female

References:

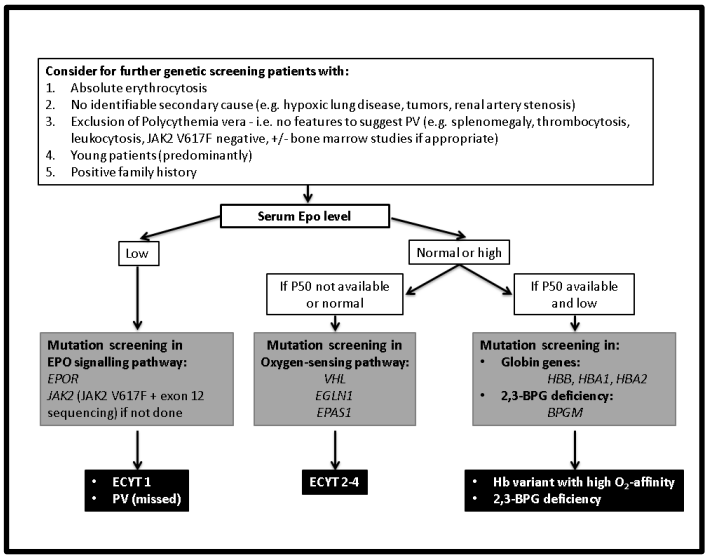
1. Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet. 2015;47(7):717-726.
2. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011;21(6):936-939.
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

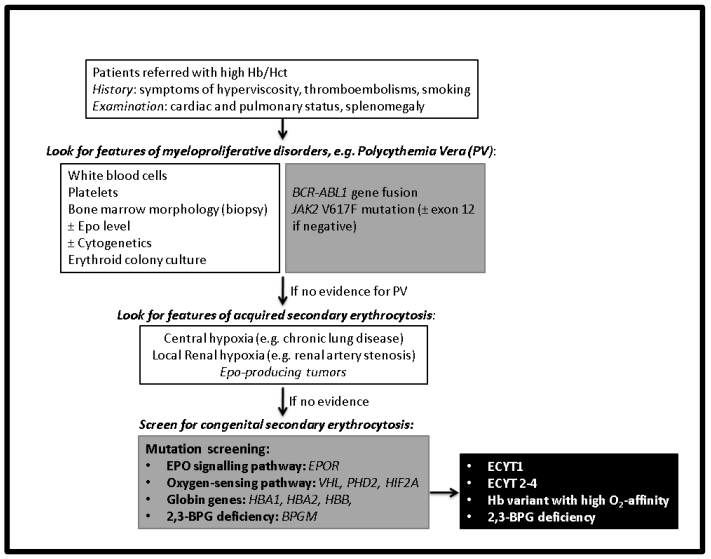
University Medical Center, Ulm, Germany



Queen's University, Belfast, United Kingdom



University Medical Center, Utrecht, The Netherlands



Centro Hospitalar e Universitário de Coimbra, Portugal