The RUNX1 Database (RUNX1db): establishment of an expert curated RUNX1 registry and genomics database as a public resource for familial platelet disorder with myeloid malignancy


Haematologica 2021 [Epub ahead of print]


Publisher’s Disclaimer.
E-publishing ahead of print is increasingly important for the rapid dissemination of science. Haematologica is, therefore, E-publishing PDF files of an early version of manuscripts that have completed a regular peer review and have been accepted for publication. E-publishing of this PDF file has been approved by the authors. After having E-published Ahead of Print, manuscripts will then undergo technical and English editing, typesetting, proof correction and be presented for the authors’ final approval; the final version of the manuscript will then appear in print on a regular issue of the journal. All legal disclaimers that apply to the journal also pertain to this production process.
The RUNX1 Database (RUNX1db): establishment of an expert curated RUNX1 registry and genomics database as a public resource for familial platelet disorder with myeloid malignancy.

Claire C. Homan1,2, Sarah L. King-Smith1,2, David M. Lawrence1,2,3, Peer Arts1,2, Jinghua Feng2,3, James Andrews1,2, Mark Armstrong1,2, Thuong Ha1,2, Julia Dobbins1,2, Michael W. Draisey, Kai Yu, Csaba Bódor4, Alan Cantor4, Maria Cazzola4,5, Erin Degelman1,2, Courtney D. DiNardo1,2,3, Nicolas Duployez1,2,12, Remi Favre1,2, Stefan Fröhling15,16, Jude FitzGibbon17, Jeffery M. Kloor18, Alwin Krämer19, Mineo Kurokawa20, Joanne Lee21, Luca Malcovati18,20, Neil V. Morgan22, Georges Natsoulis23, Carolyn Owen20, Keyur P. Patel1,2, Claude Preudhomme12,13, Hana Raslova24, Hugh Rennhoff23, Tim Ripperger25, Rachael Schulte26, Kiran Tawadry27, Elvira Velloso28,29, Benedict Yan3, Paul Liu4, Lucy A. Godley4, Andreas W. Schreiber,3,30, Christopher N. Hahn31,2,31, Hamish S. Scott1,2,3,31, Anna L. Brown1,2,3,31. On behalf of the RUNX1 international data-sharing consortium.

1Department of Genetics and Molecular Pathology, SA Pathology, Adelaide, SA, Australia; 2Centre for Cancer Biology, SA Pathology and University of South Australia, Adelaide, SA, Australia; 3Australian Cancer Research Foundation (ACRF) Cancer Genomics Facility, Centre for Cancer Biology, SA Pathology, Adelaide, SA, Australia; 4Section of Hematology/Oncology, Departments of Medicine and Human Genetics, Center for Clinical Cancer Genetics, and The University of Chicago Comprehensive Cancer Center, The University of Chicago, Chicago, IL; 5National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892; 6HCEMM-SE Molecular Oncohematology Research Group, 1st Department of Pathology and Experimental Cancer Research, Semmelweis University, Budapest, Hungary; 7Division of Hematology/Oncology, Boston Children’s Hospital and Dana Farber Cancer Institute, Harvard Medical School, Boston, MA 02115; 8Department of Molecular Medicine, University of Pavia, Pavia, Italy; 9Department of Hematology Oncology, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy; 10Division of Hematology and Hematological Malignancies, Foothills Medical Centre, Calgary, AB, Canada; 11Department of Leukemia, University of Texas MD Anderson Cancer Center, Houston, TX; 12Laboratory of Hematology, Biology and Pathology Center, Centre Hospitalier Regional Universitaire de Lille, Lille, France; 13Jean-Pierre Aubert Research Center, INSERM, Universitaire de Lille, Lille, France; 14Assistance Publique-Hôpitaux de Paris, Armand Trousseau children’s Hospital, Paris, France; 15Department of Translational Medical Oncology, National Center for Tumor Diseases (NCT) and German Cancer Research Center (DKFZ), Heidelberg, Germany; 16German Cancer Consortium (DKTK), Heidelberg, Germany; 17Centre for Haematology-Oncology, Bar’s Cancer Institute, Queen Mary University of London, London, UK; 18St Jude Children’s Research Hospital, Memphis, Tennessee, United States; 19Clinical Cooperation Unit Molecular Hematology/Oncology, German Cancer Research Center (DKFZ) and Dept. of Internal Medicine V, University of Heidelberg, Heidelberg, Germany; 20Department of Hematology & Oncology, Graduate School of Medicine, The University of Tokyo, Japan; 21Department of Haematology-Oncology, National University Cancer Institute, National University Health System, Singapore; 22Institute of Cardiovascular Sciences, College of Medical and Dental Sciences, University of Birmingham, UK; 23Imago Biosciences, Inc., San Francisco, CA, USA; 24Institut Gustave Roussy, Université Paris Sud, Equipe Labellisée par la Ligue Nationale Contre le Cancer, Villejuif, France; 25Department of Human Genetics, Hannover Medical School, Hannover, Germany; 26Department of Pediatrics, Division of Pediatric Hematology and Oncology, Monroe Carell Jr. Children’s Hospital, Vanderbilt University Medical Center, Nashville, TN, USA; 27Department of Haematology, Addenbrooke’s Hospital, Cambridge, CB2 0QQ, 28Service of Hematology, Transfusion and Cell Therapy and Laboratory of Medical Investigation in Pathogenesis and Directed Therapy in Onco-Immuo-Hematology (LIM-31) HCFMUSP, University of Sao Paulo Medical School, Sao Paulo, Brazil; 29Genetics Laboratory, Hospital Israelita Albert Einstein, Sao Paulo, Brazil; 30School of Biological Sciences, University of Adelaide, Adelaide, SA, Australia; 31School of Medicine, University of Adelaide, Adelaide, SA, Australia.

° Clinical Genome Resources Variation Myeloid Malignancy Expert Panel committee members
*Corresponding author: anna.brown@sa.gov.au

Competing Interests: The authors declare no competing financial interests

Acknowledgments: This work is supported by a grant from the RUNX1 Research Program. The authors would also like to thank the RUNX1 Research Program for their support in helping to facilitate the development of the database and fostering collaborations. We also thank the patients and their family members for their willingness to participate in this research and the RUNX1 international data-sharing consortium for their valuable contributions. This project is also proudly supported by funding from the Leukaemia Foundation of Australia, and project grants APP1145278 and APP1164601 from the National Health and Medical Research Council of Australia. This work was produced with the financial and other support of Cancer Council SA’s Beat Cancer Project on behalf of its donors and the State Government of South Australia through the Department of Health (PRF
Dear Editor,

Familial platelet disorder with associated myeloid malignancy (FPD-MM, OMIM: 601399)\textsuperscript{1,2} is a rare cancer predisposition syndrome caused by pathogenic germline variants in \textit{RUNX1}\textsuperscript{3}. Despite over two decades of research, many challenges remain to improve outcomes for individuals with FPD-MM\textsuperscript{4}. Firstly, the syndrome may go unrecognized due to poor recognition of family history and/or access to appropriate genetic testing. Secondly, intentional screening or incidental detection (e.g. tumour-sequencing) of \textit{RUNX1} variants requires access to expert interpretation. Thirdly, after diagnosis, the relative rarity of the disorder inhibits collation of sizeable local cohorts, making identify commonalities in disease course and/or outcome highly challenging. To help overcome these significant challenges we have developed an interactive public web-based international collaborative database for \textit{RUNX1}, RUNX1db (https://runx1db.runx1-fpd.org/). RUNX1db is a

---

Fellowship to H.S.S.). P.A. is supported by a fellowship from The Hospital Research Foundation. Part of this project was undertaken whilst P.A. was holding a Royal Adelaide Hospital Mary Overton Early Career Fellowship. L.M is supported by Associazione Italiana per la Ricerca sul Cancro (AIRC) (Accelerator Award Project 22796; 5x1000 Project 21267; Investigator Grant 2017 Project 20125). L.A.G is supported by the Cancer Research Foundation. KY and PL are supported by the Division of Intramural Research, National Human Genome Research Institute, NIH. T.R. is supported by a grant of the European Hematology Association (EHA) and BMBF MyPred (01GM1911B). CB is supported by EU’s Horizon 2020 Research and Innovation Program under grant agreement No. 739593.

The \textit{RUNX1} international data-sharing consortium includes all co-authors and others, including Michael Doubek (Masaryk University, Czechia), Stephen Langabeer (St. James’s Hospital, Ireland), Koneti Rao (Sol Sherry Thrombosis Research Center, USA), Josée Hébert (Université de Montréal, Canada), Lauren M. Bear (Massachusetts General Hospital, USA), Timothy A. Graubert (Massachusetts General Hospital, USA), Akiko Shimamura (Harvard Medical School, USA), Peter Ganly (Canterbury District Health Board, NZ), Marc H.G.P. Raaijmakers (Erasmus Medical Center Cancer Institute, Netherlands), Peter J.M. Valk (Erasmus Medical Center Cancer Institute, Netherlands), Paula Heller (Instituto de Investigaciones Médicas (IDIM) Alfredo Lanari, Argentina).

\textbf{Contributions}: CCH and ALB designed the research, wrote the manuscript, collected NGS and clinical data, curated NGS data, performed ACMG \textit{RUNX1} variant classification and analyzed the data. SLK collected and curated NGS data, and performed ACMG \textit{RUNX1} variant classification. PA, JD collected and curated NGS data. PA designed Figure 1. MA, TH, JF, AWS designed and performed bioinformatics analysis. DML, JA designed the database and VariantGrid software. MWD, KY, CB, AC, MC, ED, ND, RF, SF, JF, JMK, AK, MK, JL, NVM, GN, CO, KPP, CP, HR, TR, RS, KT, EV, BY, PL, CDD, LAG, LM, ALB contributed NGS data, clinical patient information and scientific insight. CDD, LAG, LM, ALB as members of the MM-VCEP advised on \textit{RUNX1} variant classification. ALB, CNH, HSS conceived and designed the study. All authors critically reviewed and approved the manuscript.
centralized repository for germline \textit{RUNX1} variant information, associated next-generation sequencing (NGS) data, and expert-curated variant information (both germline and somatic).

We recently identified, from publications, 140 different families with germline \textit{RUNX1} variants\(^6\). While a rich resource, historically reported variants are largely not classified according to the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) guidelines, only established in 2015\(^5\). Additionally, the Clinical Genome Resources Variation Myeloid Malignancy Expert Panel (ClinGen MM-VCEP) recently created guidelines specific for classification of germline \textit{RUNX1} variants\(^6\). Gene-specific guidelines, while important, add additional complexity to curation of identified variants. Making available expert knowledge to accurately classify these germline variants, prevents both missing pathogenic variants or misattribution of benign variants as causative in families\(^7\). Additionally, variants identified through clinical services and research studies don't always make it into the public domain, due to constraints associated with reporting of variants through publication or variant repositories. To address some of these challenges, we updated curated variants from publications and undertook an international survey of colleagues identifying unpublished variants. This study identified an additional 119 families (259 total), with 164 unique variants, including 10 new variants not previously described (Figure 1 and Table S1). Using this data, we created the first comprehensive \textit{RUNX1} germline registry and performed expert curation of all variants according to the \textit{RUNX1}-specific ACMG classification rules (ALB, CNH, LAG, LM, CDD MM-VCEP members). The registry represents the largest collection of curated and clinically classified \textit{RUNX1} germline variants to date, providing a unique clinical resource for researchers, clinical genomics laboratories, and haematologists (Figure 1, Table S1). Utilizing this resource, we have identified 97 pathogenic/likely pathogenic \textit{RUNX1} variants, with 54 located within the \textit{RUNT} domain (RHD)(75% of RHD variants) of which 24 are missense mutations. Only one pathogenic missense variant is observed outside of the RHD, suggesting the RHD is highly intolerant to genetic-variation. Most commonly observed pathogenic germline \textit{RUNX1} variations are whole-gene deletions (21 probands), deletion of exons 1-2 (9 probands), and mutation of amino acid
p.Arg201 within the RHD (8 probands) (Table S1). Accessibility and update-ability of this information is available through a live-web portal which hosts the registry (https://runx1db.runx1-fpd.org/classification/classifications). Each curated variant has links to patient-phenotypic information and the current clinical classification, including the evidence for each ACMG code assessed and links to external clinical databases including ClinVar and associated publications. Importantly, expert crowdsourcing allows the real-time updating of the database through user profile accounts. Newly identified variants can be easily added to the database and are automatically annotated with over 137 parameters required for accurate classification (e.g. population frequency, pathogenicity predictions). These parameters populate a classification-tool that guides users stepwise through ACMG classification of new variants (or updating current classifications with new information). Once curated and classified, collated information can be exported as an automated classification report summary, flagged for expert-review, shared with other users, and uploaded to ClinVar.

In addition to a germline RUNX1 variant registry, RUNX1db has the capacity to house NGS datasets, creating the first international genomics cohort of this rare disease. This initiative intends to enable researchers to answer questions about FPD-MM beyond germline variant detection. For example, family members, heterozygous for RUNX1 mutations, can have varying clinical presentations indicating variable penetrance and expressivity. In almost all cases, germline RUNX1 carriers present with thrombocytopenia and qualitative platelet defects, and progression to haematologic malignancies (HM) is incompletely penetrant with variable age of onset reaching from early childhood to late adulthood. Patients develop myeloid malignancies most frequently, and T-cell and more rarely B-cell acute lymphoblastic leukaemia (ALL). Currently, there is no way to predict which individuals will progress to myelodysplastic syndrome (MDS), acute myeloid leukaemia (AML), or other HM. Accumulation of somatic mutations and additional germline modifier variants are mechanisms proposed to contribute to this heterogeneity. NGS technology is widely used for surveillance and diagnosis of HM, accumulating large amounts of data often not utilized beyond
RUNX1 variant detection. Individual laboratories often only have small numbers of patients with deleterious RUNX1 germline variants, which makes asking larger questions about commonalities of genotype-phenotype, disease progression, monitoring, treatment and outcome, difficult⁹. To accumulate the data required to make evidence-based clinical decisions in FPD-MM, a dedicated resource utilizing the collective wealth of NGS data generated from research and diagnostic laboratories internationally is ideal to standardize and collate disease-specific clinical and genomics data. The database has also been designed for accumulation, sharing and curation of genomics data acquired from individuals with germline RUNX1 mutations both pre- and post-malignancy progression. We have collated 179 NGS datasets, both whole-exome sequencing (WES) and HM gene panel data, from 19 distinct research centres worldwide, including NGS from 60 FPD-MM families and 120 individuals, making it the largest FPD-MM NGS dataset (Figure 2). This includes from individuals ranging in age from 1-76 years old, malignancy phenotypes of AML, MDS, myelodysplastic syndrome/myeloproliferative neoplasm (MDS/MPN), ALL, and pre-Leukemic phenotypes including thrombocytopenia and asymptomatic carriers (Table 1). Detailed clinical information for each patient and associated samples, are stored on the database and can be updated, enabling specific phenotypic-genotypic cohort studies to be performed on the clinical spectrum of FPD-MM. Additionally, the database can be updated easily with new NGS data as available, including longitudinal datasets from serial testing of individual patients. The database allows for a comprehensive, unbiased and customizable review of all RUNX1 germline datasets with all raw sequencing data being analyzed through a standardized bioinformatics pipeline, designed to identify both somatic and germline variants, and is available on the database as variant level data (VCF, Figure S1). Using the integrated VariantGrid (https://github.com/SACGF/variantgrid) genomics analysis software, we have curated a panel of somatic variants for each dataset (including all malignancy and pre-leukemic samples), prioritizing the identification of potentially pathogenic variants in HM (2,643 variants, 167 samples). Standard filtering criteria were adapted for identifying somatic variants (Figure S1). Variants that passed all filtering criteria were subsequently manually curated. Variants classified as having no clinical significance (benign/likely benign) according to
ACMG/AMP guidelines were excluded. Remaining variants were either classified as 1) Clinically relevant, 2) Possibly relevant, or 3) Unknown relevance (Figure S1). Curated somatic variant data is available through the interactive-oncoplot on the database homepage or variant page. Shared in real-time with the scientific community, this curated dataset has already allowed the selection of secondary mutations to model FPD-MM disease and therapy in vitro and in animals. Importantly, investigators can interrogate the data to answer additional research questions as the software provides a fully automated annotation of variants and allows non-bioinformaticians to filter, sort, analyze, and curate genetic variants stored in the database via a graphical interface (Figure S2).

This project serves as a model for data accumulation for rare cancer predisposition syndromes. Adoption of a single database that serves as a repository for patient demographic and clinical data, a mutational germline registry, and patient genomics data, which can be interrogated as a large cohort are essential components for diagnosis and treatment of patients with a rare-disorder such as FPD-MM. This resource is especially useful in FPD-MM, where the genetic cause is well established, but variability in clinical presentation and disease development render diagnosis challenging. Aggregation of multiple families, individuals, and disease stages into a centralized database where all data undergo rigorous quality control using a single bioinformatics analysis strategy will aid in the exploration and discovery of the molecular progression of the disorder. Harmonized interpretation of genomic variants is imperative to understanding the mutational profile of a malignancy, which is achieved through a curated list of variants displayed for each sample. Institutional, national, and international ethics and data sharing guidelines may initially limit contributions to initiatives like this that are supported by patient advocates but need to be overcome given the importance of the work. We envision that information from this database will guide precision-based approaches to patient care plans with reasonable surveillance and adequate counselling, and eventually application of new targeted therapies and interventions prior to malignancy development for germline RUNX1 carriers.

With continued accumulation of data and clinical information, this type of gene-specific database can provide the basis to develop evidence-based clinical decisions such as when to watch and wait,
and when to apply more aggressive therapies such as stem cell transplant. Finally, we hope that this database will serve as a model from which similar efforts will emerge for other HMs for the benefit of all our patients and families.

References:

<table>
<thead>
<tr>
<th>Parameters</th>
<th>RUNX1 Database Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Individuals/Unique Mutations</td>
<td>120/47</td>
</tr>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>All samples: Median (range), years</td>
<td>40 (1-76)</td>
</tr>
<tr>
<td>Malignancy Samples: Median (range), years</td>
<td>43 (3-69)</td>
</tr>
<tr>
<td>Pre-leukemic: Median (range), years</td>
<td>34 (1-76)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Males: Total/Malignancy</td>
<td>51/23</td>
</tr>
<tr>
<td>Females: Total/Malignancy</td>
<td>67/29</td>
</tr>
<tr>
<td>Malignancy Subtype</td>
<td></td>
</tr>
<tr>
<td>Total Samples/Individuals</td>
<td>62/48</td>
</tr>
<tr>
<td>AML: #Individuals</td>
<td>29</td>
</tr>
<tr>
<td>MDS: #Individuals</td>
<td>14</td>
</tr>
<tr>
<td>MDS/MPN: #Individuals</td>
<td>2</td>
</tr>
<tr>
<td>MPN : #Individuals</td>
<td>2</td>
</tr>
<tr>
<td>ALL: #Individuals</td>
<td>4</td>
</tr>
<tr>
<td>AL : #Individuals</td>
<td>1</td>
</tr>
<tr>
<td>Pre-Leukemic</td>
<td></td>
</tr>
<tr>
<td>(#Samples/#Individuals)</td>
<td>65/56</td>
</tr>
</tbody>
</table>

Table 1: RUNX1db genomics cohort demographics. Acute myeloid Leukemia (AML), myelodysplastic syndrome (MDS), myelodysplastic syndrome/myeloproliferative neoplasm (MDS/MPN), myeloproliferative neoplasm (MPN), acute lymphoblastic leukemia (ALL), acute leukemia (AL).
Figure Legends:

**Figure 1: Registry of germline RUNX1 mutations.** Germline RUNX1 variants currently included in the RUNX1db registry are visualised using the ProteinPaint web application (https://pecan.stjude.cloud/home). Variants (displayed as protein changes where possible) are colour coded according to pathogenicity classification as determined by the MM-VCEP RUNX1-specific recommendations. The number of probands for each variant is indicated within the circle where the number is greater than one. All variants are annotated to RUNX1c; NM_001754.4; LRG_482.

**Figure 2: RUNX1 database genomics cohort demographic.** A) Breakdown of the number and types of NGS samples currently stored in the RUNX1db. Pre-Leukemic: thrombocytopenia, asymptomatic Other: includes post-transplant/post-treatment and saliva samples. Both WES and panel data is analysed and stored in the database. B) Scatter plot displaying the age of the individual when each sample was collected. Major RUNX1db cohorts (malignancy and pre-leukemic samples) are displayed. The median age for each cohort is represented by the vertical line. Clinical demographics of the malignancy cohort is shown with the number of individuals with different types of FPD-MM malignancy presentation and the C) gender and D) age distribution; Adult≥40years, AYA=15-39years, Children ≤14years. Acute myeloid Leukaemia (AML); Myelodysplastic syndromes (MDS); Myelodysplastic Syndrome/Myeloproliferative Neoplasm overlap (MDS/MPN); Myeloproliferative Neoplasm (MPN); Acute lymphoblastic leukaemia (ALL); Acute undifferentiated leukaemia (AL).
Current NGS platforms represented within the database include: Whole exome Sequencing (Illumina), Custom amplicon panels, custom capture panels and AmpliSeq panels (Ion torrent). Datasets for inclusion were preferentially obtained as raw data in the FASTQ format. Sequence reads were aligned to the GRCh37 (hs37d5) human reference genome with BWA-MEM (ver 0.7.12). Sambamba (ver 0.6.5) was used for marking PCR duplicates and GATK (ver 3.8-1) for recalibrating base-quality scores. Freebayes (ver 1.2) was used to call single nucleotide variants (SNVs) and insertions/deletions (INDELs). To increase sensitivity and permit the joint analysis of many samples, Freebayes was run in two passes, as previously described. VCF output was uploaded onto the RUNX1db database. Variant, gene and protein level annotation were performed using an in-house pipeline (https://github.com/SACGF/variantgrid). VCFs were subsequently filtered (VariantGrid analysis software) and curated according to the outlined procedure to identify somatic variants of relevance. Grey writing in the FPD-MM filtering workflow indicates additional filtering applied to pre-leukemic samples only. Somatic variant filtering: Utilising the VariantGrid analysis software a somatic variant curation pipeline was developed. Sample Filter: AD=5, DP=20, VAF=3%. Population Filter: Max population frequency of 0.1% in gnomAD (selected populations: African/African American, East Asian, Latino/Mixed Amerindian, non-Finnish European, South Asian), 1.0% for pre-leukemic samples. Damage Filter: Impact minimum=moderate, CADD score ≥20, Minimum 2 damage predictions, allow null (frameshift considered damaging) and keep splice variants. Oncogenicity Filters (https://runx1db.runx1-fpd.org/genes/gene_lists): Variants which passed all filtering criteria were subsequently manually curated.
Supplementary Figure 2: Capabilities and functionality of the RUNX1 database for genomics analysis: A) Highlighting the six top-level menus of the RUNX1db and the functions and/or data which can be utilised/accessed within each menu. B) Description of the nodes available and filters which can be applied for users to create their own analysis workflow utilising the sequencing data stored in the RUNX1db. Coloured boxes indicate the filtering node and description of filter use. Grey boxes indicate the sub-filters which can be applied within the node. C) Image depicts an example of the classification form which has been used to classify a RUNX1 likely pathogenic germline variant. A number of fields in the classification form are auto populated from annotation data. Based on the input to the classification form the ACMG Criteria summary will output the variant pathogenicity prediction. This form is available for users to classify novel variants in the database.
Supplementary Table 1: RUNX1 germline variant registry. All variants are classified according to MM-VCEP RUNX1-specific recommendations and links to the MM-VCEP variant interpretation page provided where available. Variants are annotated to RUNX1c; NM_001754.4; LRG_482.

Supplementary data reference list