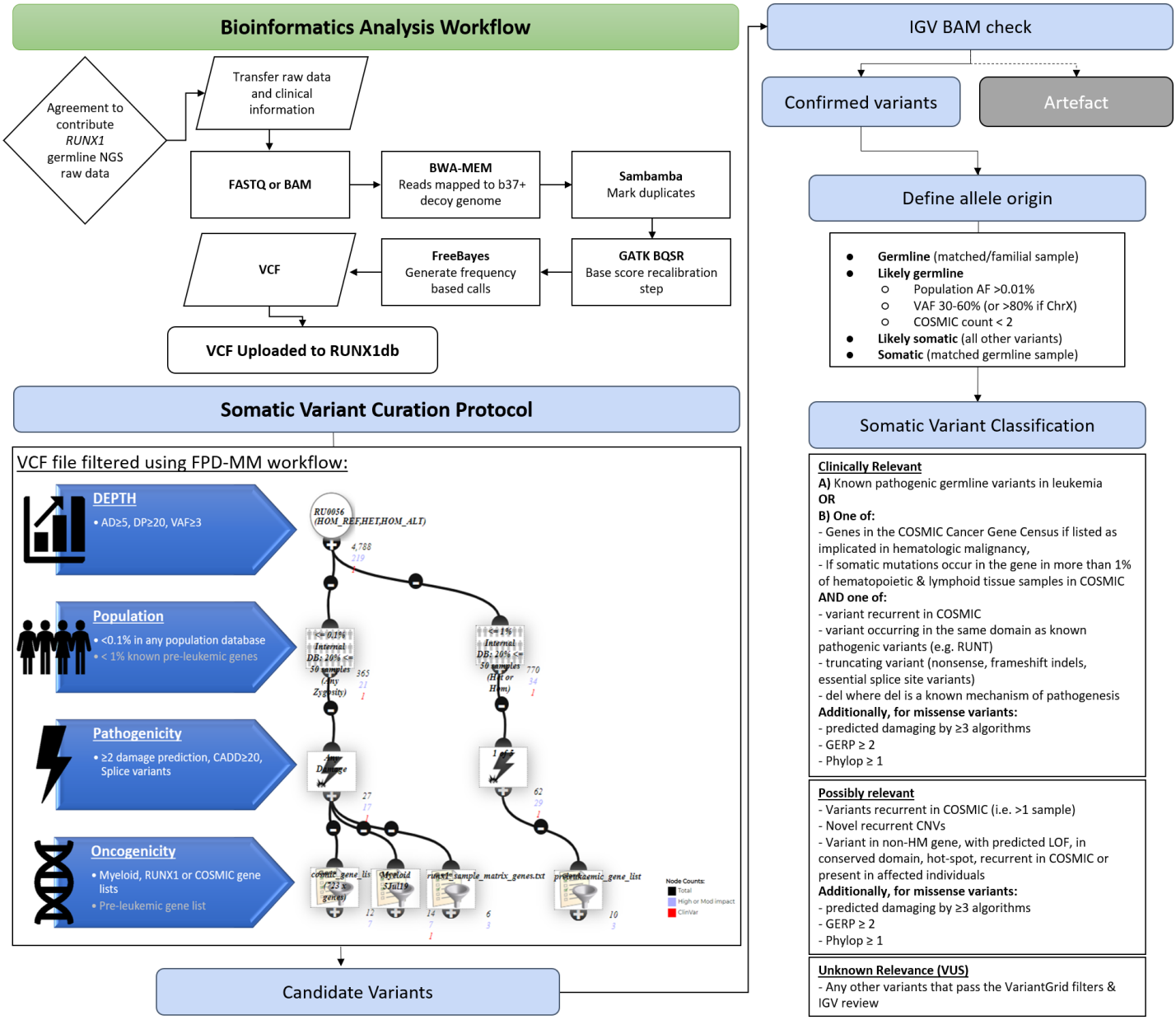
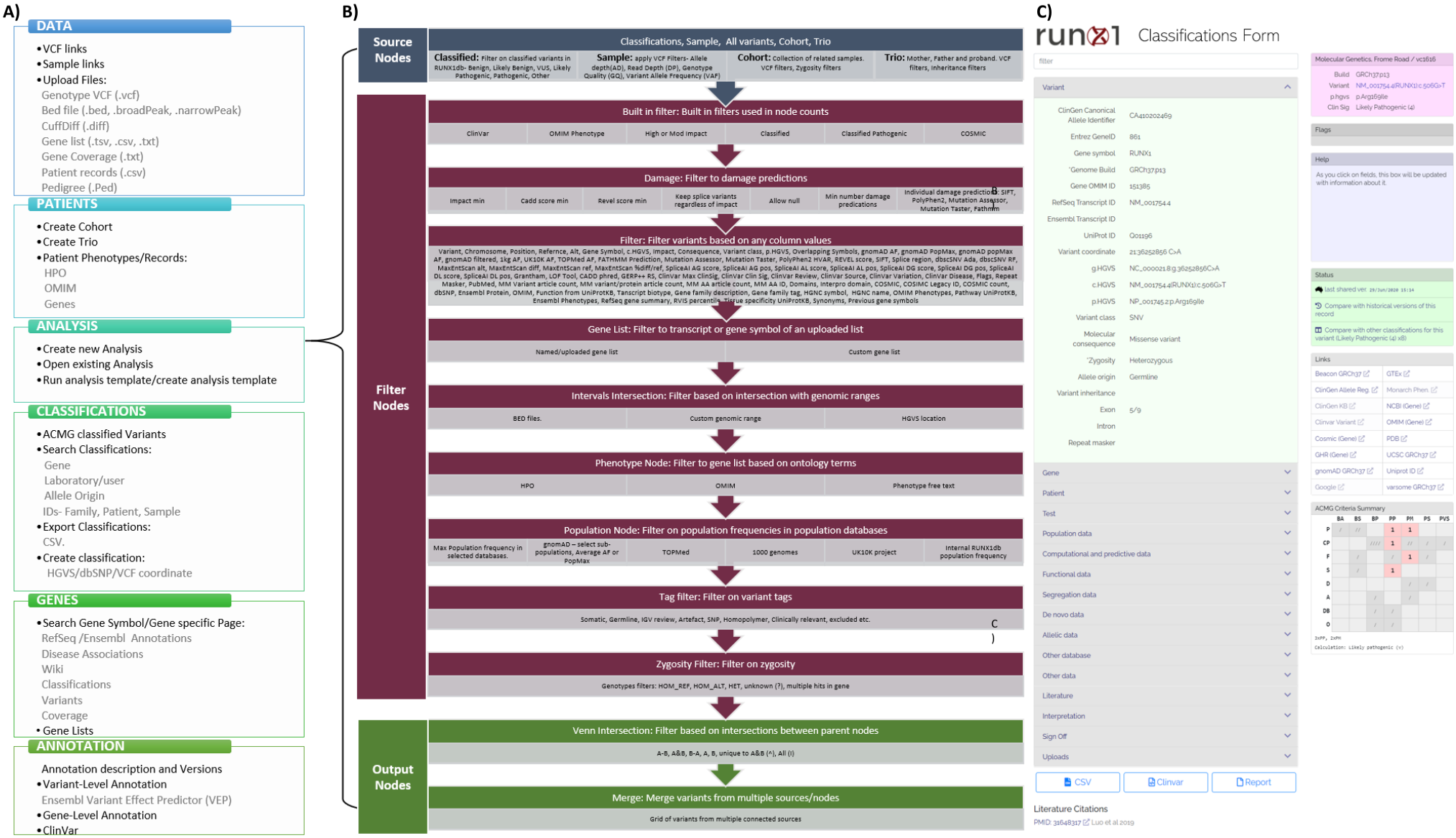


Supplementary Figure 1: Flow diagram outlining the RUNX1db genomics cohort bioinformatics analysis and somatic curation pipelines: Current NGS platforms represented within the database include: Whole exome Sequencing (Illumina), TruSight Myeloid Sequencing Panel (Illumina), Custom amplicon panels, custom 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[1]. VCF output was uploaded onto the RUNX1 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

1. Singhal D, Wee LYA, Kutyna MM, et al. The mutational burden of therapy-related myeloid neoplasms is similar to primary myelodysplastic syndrome but has a distinctive distribution. Leukemia. 2019;33(12):2842-2853.