ELSEVIER

Blood 142 (2023) 1318–1320

The 65th ASH Annual Meeting Abstracts

POSTER ABSTRACTS

503.CLONAL HEMATOPOIESIS, AGING AND INFLAMMATION

Machine Learning for Identification of High-Risk Clonal Haematopoiesis Using Blood Count Data

William G Dunn, MBChB (Hons) MSc MRCP[1], *Isabella Withnell*[2], *Muxin Gu*[1], *Pedro Quiros*[1], *Sruthi Cheloor Kovilakam*[1], *Ludovica Marando*[1], *Margarete Fabre*[3,4], *Irina Mohorianu*[1], *Dragana Vuckovic*[2], *George Vassiliou, FRCPath, MRCP, PhD MBBS*[5]

[1]Wellcome-MRC Cambridge Stem Cell Institute, Cambridge, United Kingdom
[2]Imperial College London, London, United Kingdom
[3]Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, United Kingdom
[4]Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom
[5]Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, United Kingdom

Introduction:
Clonal haematopoiesis (CH) is a common, age-related phenomenon associated with an increased risk of myeloid malignancies (MM), raising the prospect of screening to identify individuals at high risk of progression from CH to MM. The nature of the driver mutation(s) is the major determinant of the risk of progression to MM. Presently, the identification of these mutations relies on next generation sequencing (NGS) of blood DNA, making it impractical to identify those at risk at scale. We previously reported that the presence of CH is associated with changes in complete blood count (CBC) parameters such as the red cell distribution width (RDW), mean corpuscular volume (MCV) and others. However, individual parameters have limited discriminative power to distinguish those with CH versus those without. Here, we present a machine learning (ML) pipeline applied to a combined CBC dataset, which can be harnessed to improve our ability to identify individuals with high likelihood of carrying CH, who can subsequently be prioritised for genetic testing.
Methods:
Whole exome sequencing (WES) data from 454,052 UK Biobank (UKB) participants were analysed using Mutect2 to detect CH-driver mutations across the exons of 43 CH-associated genes. Germline variants were filtered using a panel of ormal and a one-sided exact binomial test. A final set of driver mutations was derived by filtering variants based on a pre-defined list of recurrent CH mutations or recurrence of >= 7 times in haematological cancers in the catalogue of somatic mutations in cancer. Fields pertaining to blood count variables were extracted and filtered to remove individuals with missing data, extreme outliers, highly correlated parameters and variables with near zero-variance resulting in 434,057 participants with complete datasets. Individuals were annotated as "CH" or "no-CH"; a Random Forest (RF) binary classifier was optimised on age, sex and 13 blood count variables as features. To quantify the robustness and stability of the model, the RF optimisation was iterated ten times, on a customised cross validation approach by using a different random sample from the majority (control) class each time. All ML models were constructed using the Caret package (R version 3.6.3).
Results:
We first found that RF classifiers to identify the presence of CH overall performed poorly (median AUC: 0.640/0.671 for any/large clone VAF >=10% CH respectively). However, gene-specific models showed that the presence of CH associated with mutations in certain high risk genes (*SF3B1*, *SRSF2*, *JAK2*, *CALR*) could be predicted from blood counts more confidently (median AUC 0.776, 0.834, 0.933, 0.849 respectively). By contrast, mutations in the more common CH driver genes (*DNMT3A* and *TET2*) could not be accurately predicted from blood count data (median AUC 0.618 and 0.666 respectively) (Figure 1). In light of these findings, we proceeded to build a single RF model to predict the presence of high-risk CH associated with mutations in any of *SF3B1*, *SRSF2*, *JAK2* or *CALR*. This performed well, with median AUC of 0.816/0.888 for any/large clone high-risk CH respectively (Figure 2). Our RF models were also able to give insights into novel blood count association with specific types of CH; in particular we found that the basophil count was an important predictor of the presence of CH driven by mutation in codon K57 of the *GNB1* gene. To investigate this further, we analysed basophil counts in the UKB and confirmed that basophilia (basophils >0.2 x 10^9/L) was more common amongst carriers of *GNB1*-K57 mutations (14/234 = 5.9%) than controls (2597/433,863 = 0.5%).
Conclusions:

Data-driven ML models applied to blood count data can identify individuals with high-risk CH driven by genes associated with progression to myelodysplastic syndrome and myeloproliferative neoplasms, who should be prioritised for genetic testing for somatic mutations. The future integration of raw blood analyser and longitudinal blood count data into predictive models would be an important next step towards improving our ability to identify individuals that may harbour CH and help facilitate the introduction of large-scale screening for the phenomenon.

Disclosures Fabre: *AstraZeneca:* Current Employment. **Vassiliou:** *AstraZeneca:* Other: Educational Grant; *STRM.BIO:* Consultancy.

Figure 1

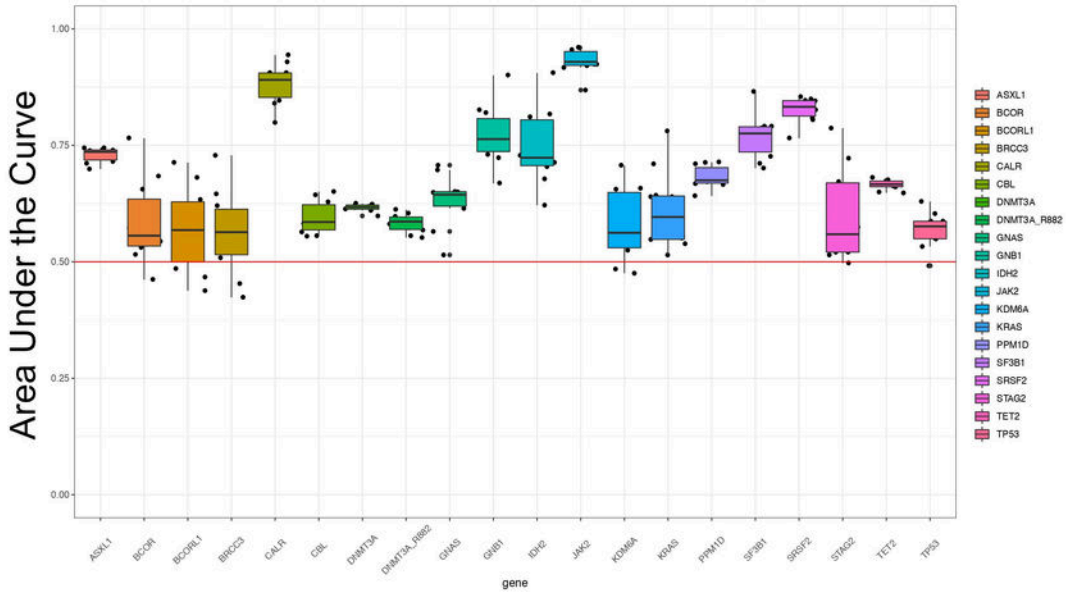

Figure 2

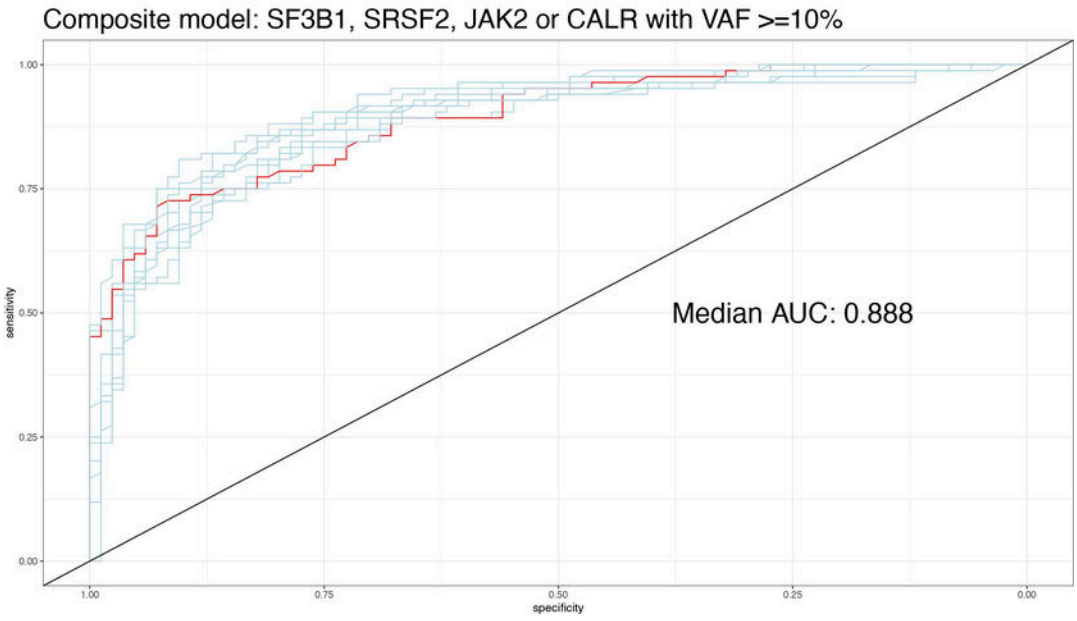

Figure 1

https://doi.org/10.1182/blood-2023-184721