

Identification of Clonal Hematopoiesis Driver Mutations through In Silico Saturation Mutagenesis

Santiago Demajo^{1,2,^}, Joan Enric Ramis-Zaldivar^{1,^}, Ferran Muiños^{1,2,^}, Miguel L Grau¹, Maria Andrianova¹, Núria López-Bigas^{1,2,3,4,*}, Abel González-Pérez^{1,2,4*}

[^]Equal contribution

*Correspondence should be addressed to: abel.gonzalez@irbbarcelona.org, nuria.lopez@irbbarcelona.org

Affiliations

1. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain.
2. Centro de Investigación Biomédica en Red en Cáncer (CIBERONC), Instituto de Salud Carlos III, Madrid, Spain.
3. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
4. University Pompeu Fabra, Barcelona, Spain

Running title

Clonal hematopoiesis driver mutations

Corresponding authors and contact information

Abel Gonzalez-Perez and Nuria Lopez-Bigas

Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028, Barcelona, Spain.

Phone: +34 93 40 39912

Email: abel.gonzalez@irbbarcelona.org, nuria.lopez@irbbarcelona.org

Authors' disclosures

The authors declare that although no patent application is envisioned giving the type of technology (software), IRB Barcelona retains any and all commercial rights over it. In addition, IRB innovation department is evaluating the registration of the source code.

Abstract

Clonal hematopoiesis (CH) is a phenomenon of clonal expansion of hematopoietic stem cells driven by somatic mutations affecting certain genes. Recently, CH has been linked to the development of hematologic malignancies, cardiovascular diseases, and other conditions. Although the most frequently mutated CH driver genes have been identified, a systematic landscape of the mutations capable of initiating this phenomenon is still lacking. Here, we trained machine-learning models for 12 of the most recurrent CH genes to identify their driver mutations. These models outperform expert-curated rules based on prior knowledge of the function of these genes. Moreover, their application to identify CH driver mutations across almost half a million donors of the UK Biobank reproduces known associations between CH driver mutations and age, and the prevalence of several diseases and conditions. We thus propose that these models support the accurate identification of CH across healthy individuals.

Significance

We developed and validated gene-specific machine learning models to identify CH driver mutations, showing their advantage with respect to expert-curated rules. These models can support the identification and clinical interpretation of CH mutations in newly sequenced individuals.

Introduction

In healthy hematopoiesis, a pool of hematopoietic stem cells (HSCs) contributes to all blood-related lineages. During aging, this process frequently gives place to clonal hematopoiesis (CH), a state in which one stem cell-derived population occupies a large fraction of the blood cells and platelets (1–6). This phenomenon of clonal expansion is driven by somatic mutations acquired by the HSCs at some point during life and is highly prevalent in elderly populations (1–3,5,7–13). These mutations, that affect an array of genes collectively named CH drivers, confer the HSCs bearing them growth advantages with respect to their neighbors, and are thus under positive selection in hematopoiesis. In the past decade, a wide array of studies has linked the presence of CH to an increased risk of development of hematological cancers, cardiovascular diseases, all-cause mortality, and more recently, solid tumors and infectious diseases (2,7,14–19,19,20).

While intensive research in recent years has led to the identification of some 60 CH driver genes (1,12,13,21), we have only a fragmented picture of which of their mutations can drive clonal expansions. Across large cohorts of blood donors from the general population (sequenced at ~40X), such as the UK Biobank (22) or TOPMed (23), the ability to detect true CH driver mutations is hindered by the contamination with germline variants and sequencing artifacts (24). However, uncovering the repertoire of CH driver mutations is key to understanding the molecular mechanisms underpinning CH, accurately identifying individuals with this condition (powering epidemiological studies of its relationship with more serious health conditions), and monitoring the potential impact of CH driver mutations on their health.

Faced with this reality, several research groups and health-related international institutions have summarized the knowledge accumulated on several CH genes in a series of expert-curated rules to select the mutations most likely to drive CH (25–27). These are normally applied together with stringent filters to variants identified in the blood samples of healthy individuals to identify CH cases (24). Despite their practical usefulness, these expert-curated rules have a series of caveats. They cannot be learned –or systematically updated– directly from information on CH mutations; rather this information must be first sedimented into shared knowledge, and their coverage of genes is heterogeneous, due to the difference in the amount of knowledge available for each gene or even across different domains of one gene.

We reasoned that these hurdles could be overcome by applying a machine learning-based approach that produces explainable models, trained on high-quality available CH mutations, as we recently demonstrated for cancer driver mutations (28). As in the case of cancer, these models would not suffer from any biases from sedimented knowledge, could reveal complex patterns in CH mutations that have not been apparent so far, and would be easy to scale as more datasets of CH mutations become available. We have thus now repurposed this approach to build explainable models of 12 CH driver genes (collectively referred to as boostDM-CH) using *bona fide* CH mutations identified across known CH driver genes (12) and sets of neutral hematopoiesis mutations. When tested on CH mutations not included in the training, the models

show a performance that is in general terms above that of expert-curated rules. Models also perform on par with deep saturation and experimental base-editing approaches. When applied to mutations identified in the blood of 470,000 donors from the general population (UK Biobank), the mutations identified by boostDM-CH models as CH drivers show a very significant association with age and the development of hematological malignancies and other diseases, while mutations identified as CH non-drivers show no meaningful associations.

Results

BoostDM-CH models accurately identify CH driver mutations

Training machine learning models aimed at distinguishing CH driver mutations requires a high-quality dataset of blood somatic mutations identified across individuals. We have shown before that repurposing tumor-blood sequencing data from large cancer cohorts using a reverse mutation calling eliminates germline contamination and provides a high quality set of blood somatic mutations (Supp. Note) (12). Using these blood somatic mutations across more than 33,000 cancer patients from three cancer genomics cohorts, we identified 64 CH driver genes (12) through the detection of signals of positive selection in their pattern of mutations (29) (Fig. 1a).

To accurately identify the CH driver mutations in those genes we aimed to build gene-specific machine learning models that captured combinations of features that define them in each gene. These features include, for example, the significant clustering of mutations in specific regions of the linear sequence or the three-dimensional folded structure of the protein, the enrichment of mutations in certain domains, the consequence type of each mutation, and some additional features such as the conservation of the residue and post-translational modifications (28).

The most important and challenging step to train these models is to start with a good quality set of driver/positive and non-driver/negative blood mutations. Using a set of known CH driver mutations as the positive set would produce models that learn our current knowledge biases. To solve this problem we reasoned –as we did previously in the case of cancer– that somatic mutations detected in human blood, an unbiased set highly enriched for driver mutations of each CH gene, constitute the best positive training set. The ideal negative set contains mutations that could have occurred by neutral mutagenesis processes in HSCs but do not provide a selective advantage (non-drivers). We thus generated synthetic mutations by simulating neutral mutagenesis in HSCs, following the probabilities of tri-nucleotide changes observed across blood samples as the negative training set (28). Although these sets are highly enriched in driver and non-driver mutations respectively, they are imperfect, as the group of observed CH mutations may contain non-drivers, whereas the synthetic set of mutations may contain drivers. These imperfections need to be taken into account in the strategy used to train the models (Methods and Supp. Methods).

For example, to build a model for *DNMT3A*, we used as the positive set the 2,650 blood somatic mutations identified across all cancer patients from the three discovery cohorts via the reverse calling, and created fifty negative sets of 2,650 synthetic mutations (Fig. 1b; Methods and Supp. Methods). We then trained a robust model able to classify any mutation in the gene as CH driver or non-driver and provide an explanation of the features used for this classification (boostDM-CH). For example, the R882H mutation of *DNMT3A* (Fig. 1c) is predicted as a CH driver primarily from a combination of three salient features: its location within a linear (30) or a three-dimensional (31) cluster and its conservation across vertebrate species (32).

We followed the same strategy for the 25 CH driver genes with sufficient number of mutations to carry out the training (Methods). The boostDM-CH models of 12 CH driver genes (*ASXL1*, *CHEK2*, *DNMT3A*, *GNAS*, *IDH2*, *MDM4*, *PPM1D*, *SF3B1*, *SRSF2*, *TET2*, *TP53*, and *U2AF1*) had good performance (F_{50} above 0.80) and were deemed sufficiently representative of all their potential driver mutations (high discovery index, see Methods and Supp. Methods) (genes in blue in Fig. 1d; Supplementary Fig. S1). As a trend, genes with more observed CH mutations (or higher discovery index) yielded models with higher F_{50} (Fig. 1d and Supplementary Fig. S2a). Nevertheless, some CH driver genes with relatively low number of mutations found in the three discovery cohorts, with most of them concentrated within significant clusters (e.g., *U2AF1* and *IDH2*), also yielded high-quality models. The area under the cross-validation precision-recall curves (AUC) for these models ranged between 0.86 and 0.99 (Fig. 1e; Supplementary Fig. S2b).

To benchmark the performance of the boostDM-CH model of *DNMT3A*, the most recurrent CH driver gene, we compared it with a recently published experimental base editing assay that quantified the reporter methylation activity of mutants of several specific residues (33) (Supp. Methods). The *DNMT3A* model showed a better performance than this experimental assay in the separation between observed and neutral CH mutations (Fig. 1f; Supplementary Fig. S3a; Supp. Note). Moreover, when applied to an independent set of CH mutations identified across the general population (Japanese Biobank) (34), the *DNMT3A* boostDM-CH model also showed higher F_{50} than the experimental assay (Fig. 1f; Supplementary Fig. S3a). Furthermore, the *TP53* boostDM-CH model exhibited comparable performance to four experimental deep mutagenesis studies of this gene (35–38) on the task of identifying CH driver mutations observed across cancer patient cohorts and the Japanese Biobank (Supplementary Fig. S3b).

We then compared the collective performance of boostDM-CH models to that of three sets of expert-curated rules (referred to as Niroula (25), Bick (26), and WHO (27), Supplementary Table S1) on three different sets of CH mutations. We first evaluated their performance on the classification of cross-validation CH observed and synthetic neutral mutations across cancer patients in 10 CH genes (*MDM4* and *CHEK2* are not covered by the three rule sets). Next, we compared the performance of boostDM-CH models and the expert-curated rules in the classification of a set of uncommon CH mutations in 6 genes (10), included neither in the training nor in the cross-validation of models (Methods). Finally, we compared boostDM-CH and

the three sets of expert-curated rules in the identification of CH mutations across a fully independent cohort, the Japanese Biobank. Interestingly, the performance of boostDM-CH models is systematically greater than that of the three sets of expert-curated rules, even when analyzing mutations not employed in their training (Fig. 1g left panel; Supplementary Fig. S3c). This performance comparison reflects a variability across the 10 CH driver genes (Fig. 1g right panel). These results demonstrate the potential of machine-learning models trained from observed mutations –and thus, unbiased by prior knowledge– to capture the genetic mechanisms underlying CH.

The repertoire of CH driver mutations reveals mechanisms underlying CH

We next used boostDM-CH models to classify all possible mutations in each CH driver gene (*in silico* saturation mutagenesis; Fig. 2a,b; Supplementary Fig. S4). Each mutation is classified on the basis of the boostDM-CH score provided by each model (Methods): mutations attaining a score equal to or above 0.5 are classified as CH drivers, while those with a score lower than 0.5 are deemed CH non-drivers. The boostDM-CH score can be used to further tier the mutations into high-confidence (score equal to or greater than 0.9) and other (between 0.5 and 0.9) drivers.

Missense driver mutations (and specifically high-confidence ones) identified in *DNMT3A* cluster at specific regions of the protein, such as the PWWP and DNA methylase domains (Fig. 2a). These clusters likely capture the interference with different aspects of the DNMT3A methylation activity, such as hindrance of the tetramerization (R882) (5), or impairment of the recognition of histone modifications (mutations within the PWWP domain) (33,39). Driver nonsense mutations, expected to trigger nonsense-mediated decay, are distributed along its entire sequence. Similarly, in *TET2*, the identified missense CH driver mutations (some of which act through haploinsufficiency) (40) appear enriched at the beginning and the end of the Tet-JBP domain, and may interfere with the removal of methyl groups from CpGs (41,42) (Fig. 2b). A few missense mutations that appear in the N-terminal part of the protein are likely false positives of the model, but, interestingly, they are not included in the high-confidence set of driver mutations (Supp. Note). In contrast, driver nonsense mutations appear distributed along the entire protein sequence and all of them are tiered as high-confidence (Fig. 2b).

The analysis of the distribution of CH driver mutations within the remaining 10 CH driver genes with boostDM-CH models reveals different landscapes (Supplementary Figs. S4 and S5). A group of CH drivers (such as *ASXL1* and *TP53*) appears to act in a loss-of-function manner, showing a broad distribution of missense and nonsense CH driver mutations along long tracts of the protein sequence. This contrasts with another group of genes (*SF3B1*, *PPM1D*, *U2AF1*, and *IDH2*) known to act in a gain-of-function manner, where CH driver mutations appear confined to specific regions of the protein, and in extreme cases, to a single mutational hotspot. These clusters of CH driver mutations are related to underlying alterations in the biological function of these genes.

A detailed explanation of the contribution of different features to the classification of four illustrative CH driver mutations across the same number of genes is presented through the radial plots in Figure 2c. For example, in the case of *PPM1D*, all nonsense high-confidence CH driver mutations concentrate within the C-terminal portion of the protein. These mutations escape nonsense-mediated decay and result in a form of the protein lacking a degron sequence and, thus, with increased stability (43). This, in turn, causes altered phosphorylation of proteins involved in the response to DNA damage, such as *TP53*, providing mutant cells with an advantage when exposed to certain cytotoxic therapies (11,43).

Whereas there is a high degree of overlap between mutations identified as drivers by boostDM-CH and by the expert-curated rule sets, clear differences are also apparent between them (Supplementary Fig. S6). There are also similarities and differences in the configuration of CH and cancer driver mutations in 3 of these CH driver genes for which we have also been able to build boostDM myeloid cancer models (28) (Fig. 3a-c and Supplementary Fig. S7). One of the differences corresponds to a mutational hotspot of *IDH2* observed in myeloid malignancies, but absent in CH mutations (Fig. 3c). This hotspot is considered a driver CH mutation by the expert-curated rules, which are based in part on data from hematological malignancies. However, its absence from the driver output of the *IDH2* boostDM-CH model suggests it may be an example of incorrect definition of the rules (see below). This is an example of the benefits provided by models trained from observed data over expert-curated rules established from sedimented knowledge.

In summary, boostDM-CH models support the exploration of underlying CH mechanisms across genes. To facilitate this process, the results of the *in silico* saturation mutagenesis provided by boostDM-CH models of the 12 CH driver genes included in this section are available to the research community at www.intogen.org/ch/boostdm.

BoostDM-CH models identify CH driver mutations in a large general population cohort

One of the main hurdles to exploiting large cohorts of donors (such as the UK Biobank, UKB) for population-wide CH epidemiological studies is the difficulty to accurately identify blood somatic mutations (in a relatively shallow sequencing) without a reference germline sample from the same donor (12,24). Potential somatic mutations identified across these blood samples may contain a non-negligible fraction of germline variants and sequencing artifacts, as well as passenger mutations. We reasoned that boostDM-CH models –trained on high-quality blood somatic mutations obtained from the reverse mutation calling across tumor cohorts– could be employed to accurately identify CH driver mutations in this setting. We could then use these CH driver mutations to analyze the relationship between CH and several phenotypes across the population.

Thus, we next identified 201,857 potential somatic variants in the 12 genes across 467,202 individuals in the UKB (Methods). The application of BoostDM-CH models classified 41,306 of them as CH driver mutations (28,508, or 69.0%, missense; 12,098, or 29.3%, nonsense; and

700, 1.7%, splice site affecting) (Fig. 4a,b and Supplementary Fig. S8a). CH driver mutations appeared in 8.2% (38,124) of the donors of the UKB cohort (92.5% of them bearing a single driver mutation). Differences in the tri-nucleotide mutational profiles of variants identified as CH drivers or non-drivers by boostDM-CH suggest that the latter contain a mixture of sequencing artifacts, germline variants and somatic passenger mutations (Supplementary Fig. S8b). Most (72%) driver mutations were identified with high-confidence (boostDM-Ch score equal to or greater than 0.9), as were most (90%) non-driver mutations (boostDM-score below 0.1) (Supplementary Fig. S9a,b).

The proportion of individuals with CH driver mutations identified by boostDM-CH models (relative to those in the youngest age bracket, 38-45 years old) grows with age, as expected of true CH mutations (Fig. 4c). In contrast, the proportion of donors with non-driver CH mutations remains constant. While all potential somatic SNVs in the 12 CH driver genes show a significant association with age (Fig. 4c, black dot), the significance is virtually entirely due to CH driver mutations (top red dot), since non-driver mutations (gray dot) show almost no association. The effect size is larger for individuals with CH mutations with VAF > 10% (driver large) than for those with CH mutations with VAF ≤ 10% (driver small), and it also appears larger for donors bearing several co-occurring CH driver mutations (multiple drivers). The association with age is also clearer when high-confidence CH driver and non-driver mutations are compared (Supplementary Fig. S9c-e). CH driver mutations in almost all individual genes show higher association with age than non-driver mutations (Fig. 4d). The significant association with age is also verified across mutations never observed (or observed only once) in the three cancer genomics training cohorts, suggesting that the boostDM-CH models are capable of identifying new CH driver mutations (Supplementary Fig. S10a,b; Supp Note), and it is also apparent across donors with different ethnic backgrounds in UKB (Supplementary Fig. S10c). Among 85 recurrent variants identified in 12 CH genes in UKB, those identified as drivers by boostDM-CH tend to confer high fitness (estimated as described by (44)), and show a significant association with age. In contrast, recurrent non-driver variants lack these two features (Supplementary Methods; Supplementary Fig. S11a-d; Supplementary Note).

We next verified known associations of CH (defined by the presence of CH driver mutations identified by boostDM-CH) with the exposure to external CH promoters across UKB individuals. We corroborated that individuals with a history of smoking have significantly higher likelihood of carrying CH driver mutations than non-smokers, as have donors who suffered a solid malignancy (as a proxy of exposure to cytotoxic drugs) (1,12,13) prior to their enrollment in UKB (Fig. 5a, Supplementary Fig. S12a,b). We also tested the association of CH driver mutations identified by boostDM-CH in UKB with subsequent conditions known to be linked to CH (Supplementary Table S2). First, we corroborated the known association of the presence of CH driver mutations with the development of hematological malignancies, in particular of myeloid origin (Fig. 5b; Supplementary Fig. S13a-e). Both the significance of the association and the increase in risk appear higher for donors bearing large CH clones, clones with multiple mutations (Fig. 5b), or high-confidence CH drivers (Supplementary Fig. S13f). Conversely, the

risk of donors bearing CH non-driver mutations to subsequently develop a hematological malignancy is comparable to that of the general UKB population. We also verified that the known associations between CH driver mutations and increased risk of subsequent heart failure, development of a solid malignancy, risk of any type of infection, and all-cause mortality are reproduced (Fig. 5c, Supplementary Fig. S14, S15 and S16). Finally, we compared driver mutations identified by boostDM-CH with those identified by the expert-curated rules (Supplementary Fig. S17) on the basis of the aforementioned associations. In all cases, the association with boostDM-CH drivers showed equal or higher significance than that obtained using the established expert-curated rules (Supplementary Fig. S18a-c). This suggests that boostDM-CH drivers are more specific than those identified through expert-curated rules.

Newly identified CH driver mutations

We next focused on mutations identified across the UKB cohort which are classified as drivers by boostDM-CH models but not by the expert-curated rule sets (Supplementary Table S3). To this end, we first selected the mutations that were classified as drivers by the consensus of the three sets of expert-curated rules. These expert-curated rule consensus CH drivers were compared with those identified by the 10 corresponding boostDM-CH models (i.e., excluding *MDM4* and *CHEK2* not covered by all rule sets), yielding three groups of mutations: rules-boostDM-overlapping (N=28,901), boostDM-exclusive (N=10,079), and rules-exclusive (N=7,436) (Fig. 6a and Supplementary Fig. S19a,b). We observed a strong association with the age of individuals carrying either rules-boostDM-overlapping or boostDM-exclusive mutations, while a very weak association was observed for individuals carrying rules-exclusive mutations (Fig. 6b). A stronger association with age can also be observed for mutations identified uniquely by boostDM-CH than for those identified by each of the three sets of rules (Supplementary Fig. S19c,d). Furthermore, while both rules-boostDM-overlapping and boostDM-exclusive mutations showed significant association with all-cause mortality and development of myeloid malignancies, rules-exclusive mutations were not significantly associated with either condition (Fig. 6c). We also analyzed 52 recurrent single nucleotide variants identified as drivers by the Bick rule set (26) in the 12 genes included in this study (Supplementary Table S4), which were deemed non-drivers by a recent analysis that demonstrated they are not associated with age across UKB donors (24). Out of these, 28 (54%) are correctly classified by boostDM-CH as non-drivers, a number which rises to 34 (65%) if only high-confidence boostDM-CH drivers are considered.

We then focused on mutations identified as drivers by boostDM-CH and/or the expert-curated rule sets in specific CH genes across the UKB. In the exemplary cases of *DNMT3A* and *TP53*, 4,605 and 501 (respectively) mutations missed by the consensus of the three sets of expert-curated rules, but identified by the corresponding boostDM-CH models, show a significant association with age and are thus likely drivers of CH (Fig. 6d). Accordingly, *DNMT3A* driver mutations identified exclusively by boostDM-CH show a very similar distribution of methylation capability as that observed for rules-boostDM-overlapping mutations in the base editing assay mentioned above (Supplementary Fig. S19e) (33). We found that 134 boostDM-exclusive

mutations in *ASXL1*, all nonsense variants located upstream exon 11 (see Supplementary Fig. S17) are not significantly associated with age or prior smoking (in agreement with a recent experimental report (45)), suggesting they may be false positives of the boostDM-CH model (Supplementary Fig. S20a,b). In the case of *TET2* the lack of association with age of 271 boostDM-exclusive missense mutations identified outside the two segments of the protein coincident with the three sets of expert-curated rules also implies that there are, among them, false positives of the model (Supplementary Fig. S20c,d). In comparison, 3,380 mutations deemed drivers by the three expert-curated rule sets but not by boostDM-CH appear to be false positives (Supplementary Fig. S20e). Thus, mutations identified as CH drivers in *TET2* by expert-curated rule sets have a much higher rate of false positives than those identified by boostDM-CH. Moreover, the boostDM-CH score of the 271 suspected false positive boostDM-exclusive missense mutations is lower than 0.9, and they are thus not included in the high-confidence driver set (Supplementary Fig. S20f). Next, we explored the intriguing case of two *IDH2* hotspots, one of which (R140) is rules-boostDM-overlapping, while the other (R172) is rules-exclusive (see Fig. 3c, above). We verified that while the mutations in the R140 hotspot are significantly associated with age, those in position R172 are not, suggesting that only mutations in R140 are CH drivers (Fig. 6e; Supp. Note). Finally, we found that *CHEK2* and *MDM4* driver mutations also show a significant association with age after removing potential contaminating germline variants (Supplementary Fig. S21a-d; Supp. Note).

The results of all these analyses, taken together, indicate that CH driver mutations identified exclusively by boostDM-CH across UKB are, for the most part, true CH drivers. BoostDM-CH models outperform the expert-curated rule sets in the detection of CH driver mutations in a cohort representing the general population.

Discussion

Here we built machine learning-based models to identify all CH driver mutations in 12 well-known CH genes. We were primarily motivated by the lack of unbiased methods to identify the mutations responsible for the development of CH, even within well-established CH driver genes (24). One key requirement to train such machine learning models is the availability of a big enough set of high-quality blood somatic mutations. To fulfill this, we resorted to cohorts of cancer patients where blood somatic mutations can be reliably identified through a comparison of blood and tumor samples (12). The models were then trained on features that distinguish these blood somatic mutations observed in CH genes across donors from those expected to arise under neutral mutagenesis (28).

We demonstrate that, in general, these models outperform expert-curated rules designed by CH researchers on the basis of years of accumulated knowledge (25–27) in identifying CH driver mutations across cancer genomics cohorts and other cohorts representing the general population. Machine learning models trained unbiasedly on observed blood somatic mutations are thus able to recapitulate sedimented knowledge on CH development. Of note, the similarity

in the blueprints of CH driver mutations produced by the models and by expert-curated rules and the applicability of the models to identify CH driver mutations in the UKB indicate that CH across cancer patients is largely driven by age, which suggests that it follows a similar evolutionary path as in the general population (Supp. Note).

BoostDM-CH models can be employed to gain new insights into the mutational mechanisms underlying CH in different genes. Here, we provide examples with *DNMT3A*, *TET2*, and the models of other genes. The 12 boostDM-CH models trained in this work and their associated data are available to the CH research community at www.intogen.org/ch/boostdm. In the future, as more cohorts for which blood and a second sample (such as in cancer patients cohorts) become available in the public domain, growing sets of reliable blood somatic mutations will be identified, and good-quality models for more CH genes will be within reach.

Moreover, we have shown a new path for the identification of CH driver mutations across donors in a large cohort, from whom only a blood sample is available, using boostDM-CH models. Applying this rationale to the UKB, we are able to recapitulate known associations of CH with age and the development of several phenotypes. Driver mutations newly identified by boostDM-CH across CH driver genes have higher specificity than those identified uniquely by expert-curated rule sets. In summary, BoostDM-CH models can distinguish CH driver mutations from a combination of sequencing artifacts, germline contamination, and passenger somatic mutations across half a million donors in the UKB (22) with better accuracy than expert-curated rules. This work constitutes a proof-of-principle for a wider use of boostDM-CH models in large retrospective or prospective clinical studies aimed at discovering such associations. Furthermore, it also illustrates how the models could assist in detecting CH across large populations, to subsequently monitor individuals at risk of developing different conditions.

Methods

Blood somatic mutations in three cancer genomics cohorts

Blood somatic mutations identified across two of our discovery cohorts –TCGA (whole-exome sequencing) (46), HMF (whole-genome sequencing) (47)– through reverse calling were obtained from our previous study Pich *et al.*, 2022 (12). Briefly, aligned sequencing reads from normal (blood) and tumor BAM files from patients with solid tumors in TCGA and HMF cohorts were compared (using Strelka2) to call blood somatic mutations, which were subjected to a strict filtering post-process described in detail in the aforementioned study. We thus retrieved all mutations identified in each of these two discovery cohorts. Blood somatic mutations in a third discovery cohort –MSK-IMPACT (in a gene sequencing panel) (48) called by Bolton *et al.*, 2020 (1)– were obtained from cBioPortal (<https://www.cbioportal.org/>) (49).

Compendium of mutational CH genes

CH driver genes were obtained from our previous work (12). We successfully built boostDM-CH models for 25 of them: *ASXL1*, *ATM*, *CBL*, *CHEK2*, *CTCF*, *DNMT3A*, *EZH2*, *GNAS*, *IDH2*, *JAK2*, *KDM5C*, *KRAS*, *MDM4*, *NF1*, *PPM1D*, *RAD21*, *STAT5B*, *SF3B1*, *SH2B3*, *SRSF2*, *STAG2*, *TET2*, *TP53*, *U2AF1*, *ZRSR2*. High-quality boostDM-CH models (with the criteria defined in the main text and Supplementary Methods) were built for 12 of these genes (*ASXL1*, *CHEK2*, *DNMT3A*, *GNAS*, *IDH2*, *MDM4*, *PPM1D*, *SF3B1*, *SRSF2*, *TET2*, *TP53*, and *U2AF1*). See details in Supplementary Methods.

BoostDM-CH implementation

BoostDM-CH is a supervised learning strategy based on mutations observed across blood samples and synthetic mutations that follow neutral mutagenesis. The boostDM-CH pipeline has been implemented in Python and Nextflow (50). The base classifiers were trained with XGboost (51) v.0.90 and we used SHAP (52) v.0.28.5 to compute the local explanations for the predictions of the base classifiers. The pipeline is available as a GitHub repository: <https://github.com/bbglab/boostdm-pipeline/tree/ch>. In silico saturation mutagenesis is a term that indicates the assessment of all possible changes in a gene or protein with a computational approach (28). To assess the driver potential of all possible mutations in CH driver genes, we used boostDM-CH models. The training of the models, the derivation of local explanations for each mutation, and their validation are described in Supplementary Methods.

Expert-curated rules to identify CH driver mutations

Three sets of expert-curated rules designed by CH researchers (referred to as Niroula (25); Bick (26), and WHO (27)) were used for boostDM-CH performance comparison (details in Supplementary Table S1). For most of the benchmarking we only consider those mutations in genes included by all the three sets of expert-curated rules (10 genes).

Comparison with myeloid boostDM models

The myeloid boostDM models were trained with data from 11 tumor-type specific cohorts from IntOGen (Release v2023.05.31) (28,29) comprising Acute Myeloid Leukemias (9 cohorts), Myelodysplastic Syndromes (1 cohort) and Chronic Myelogenous Leukemias (1 cohort).

Application of boostDM-CH models to a population-based cohort (UK Biobank)

All analyses were performed with the UK Biobank Research Analysis Platform (22). The cohort used in the study comprises 467,202 individuals (54% females) with no history of hematologic malignancy prior to enrollment in UKB, with age range 37–73 and median age 58 years old. Briefly, mutations were called in the 12 genes with selected boostDM-CH models from UKB CRAM files using a nextflow (50) (v20.10.0) pipeline implementing Mutect2 (GATK, v.4.2.2.0) in tumor only mode. Variants identified were filtered to eliminate potential germline contamination using a Panel of Normals, the 1000 Genomes Project and the Genome Aggregation Database (53), as well as *ad hoc* filters to exclude low-quality calls. Find details in Supplementary Methods.

Clinical data from the UKB were downloaded in November 2022 and individual traits were pulled out from the whole phenotype file classified in data-fields. Basic information from the individuals used for the analyses was age of recruitment (data-field: 21022), sex (data-field: 31), genetic principal components (data-field: 22009), death status (data-field 40007), and body mass index (BMI, data-field: 21001). Smoking status was defined as never smoker or ever smoker using smoking status information (data-field: 20116). Presence of cancer was defined from reported occurrences of cancer (data-field: 40009). Years to first cancer was assessed using the age of recruitment and the age of first cancer (data-field: 22008). Specific cancer type, cardiovascular diseases (CVD) traits, infectious disease and other conditions such as hypertension or diabetes mellitus type II were generated combining information from different data-fields (Supplementary Table S2) including ICD-10 diagnosis (data-fields: 40006, 41202, 41270), ICD-9 diagnosis (data-fields: 40013, 41203, 41271), self-reported cancer (data-field: 20001), self-reported non-cancer illness (data-field: 20002), underlying cause of death (data-field: 40001), contributory cause of death (data-field: 40002), operation (data-field: 20004), and OPSC4 (data-field: 41272, 41200), similarly definitions outlined by refs. (54,55). For each definition, the first diagnosis event that occurred was selected. Years to first occurrence of cancer, CVD and infectious diseases was calculated also using the difference between date of recruitment and specific diagnosis dates (data-fields: 40005, 40000, 41260, 41262, 41263, 41280, 41281, 41282) and diagnosis age (data-fields: 20007, 20009) based on disease definitions. Regarding some of the covariates used for the association, diabetes mellitus type II was defined as its diagnosis or treatment with insulin or oral hypoglycemic medication (data-field: 6177); dyslipidemia was defined as cholesterol ≥ 240 mg/dL (data-field: 30690), LDL-direct ≥ 160 mg/dL (data-field: 30780), HDL-cholesterol < 40 mg/dL (data-field: 30760), or use of lipid-lowering drugs (data-

field: 6177); hypertension was defined by its diagnosis or by having a systolic blood pressure ≥ 140 mmHg (data-field: 4080), diastolic blood pressure ≥ 90 mmHg (data-field: 4079), or use of antihypertensive medication (data-field: 6177).

Unless otherwise specified, all regression models included age, sex, and the first ten ancestry principal components as covariates. For cancer regressions, we also included smoking status as covariate, while for CVD we included smoking status, BMI, diabetes mellitus type II, dyslipidemia, and hypertension status. Infectious diseases included smoking status and hematological cancer. For age, cancer, smoking status, CVD, and death association with CH we performed logistic regression analysis using logit (Python statsmodel package v.0.14.0). To analyze the risk of having hematological malignancies and myeloid neoplasms in CH carriers, we performed a Cox proportional hazards model CoxPHFitter (Python lifelines package v.0.27.8). We count as an event any reported diagnosis of hematological or myeloid cancer, respectively, after enrollment in the UKB. Individuals without the event who died before the end of the follow-up were censored at the time of death, while the rest were censored at the last follow-up reported (2021-06-25, from data-field 40005). The maximum number of years to an event was restricted to the 97th percentile of the UKB population. Kaplan-Meier curves were built using KaplanMeierFitter and logrank_test functions (Python lifelines package v.0.27.8).

Data and software availability

Blood somatic mutation data required to train boostDM-CH models is available through Hartwig Medical Foundation (HMF) and dbGaP following the same procedure to access the original datasets used in the reverse calling approach (12). HMF blood somatic mutations are available as part of the data access request to HMF (<https://www.hartwigmedicalfoundation.nl>). TCGA blood somatic mutations are available through dbGaP (phs002867) to researchers who have obtained permission to access protected TCGA data. Panel-sequenced data from the IMPACT targeted cohort is available through cBioPortal (https://www.cbioportal.org/study/summary?id=msk_ch_2020). The compendium of CH driver genes is available through IntOGen (intogen.org/ch). The *in silico* saturation mutagenesis of CH drivers described in this study is available through the boostDM-CH (intogen.org/ch/boostdm). Data required to reproduce UKB and Japanese Biobank analyses is available upon access request to both entities (details in Supplementary Methods). Software required to train boostDM-CH models and carry out analyses described in the article will be publicly available upon publication.

Acknowledgements

The authors want to acknowledge Martina Gasull and Paula Gomis for support in data access and Federica Brando for key support in the development of the boostDM-CH website. N.L.-B. acknowledges funding from the European Research Council (consolidator grant 682398). S. D. was supported by a Juan de la Cierva fellowship from Spanish Ministerio de Ciencia e Innovación (IJC2020-044728-I). J. E. R-Z. was supported by a Postdoctoral AECC 2023 fellowship from fundación científica asociación Española Contra el Cáncer (AECC) (POSTD234814RAMI). This project was supported by the CHEMOHEALTH project, funded by the Spanish Ministry of Science (MCIN), AEI /10.13039/501100011033/ and by FEDER, the AtheroClonal project, also funded by the Spanish Ministry of Science (MCIN), AEI and the European Union “NextGenerationEU”/PRTR”, and the MyoClonal project, funded by la Caixa, HR22-00732. This work was delivered as part of the PROMINENT team supported by the Cancer Grand Challenges partnership funded by Cancer Research UK (CGCATF-2021/100008), the National Cancer Institute (OT2CA278668) and the Scientific Foundation of the Spanish Association Against Cancer, AECC. It has also been supported by the “CGI-Clinics” project, funded by the European Union's Horizon Europe programme under grant agreement 101057509. IRB Barcelona is a recipient of a Severo Ochoa Centre of Excellence Award from the Spanish Ministry of Economy and Competitiveness (MINECO; Government of Spain) and an Excellence Institutional grant by the Asociación Española contra el Cáncer, and is supported by CERCA (Generalitat de Catalunya). This publication and the underlying research are partly facilitated by Hartwig Medical Foundation and the Center for Personalized Cancer Treatment (CPCT) which have generated, analyzed and made available data for this research. This publication and the underlying research are also partly facilitated by data collected and made public by The Cancer Genome Atlas network. This research has been conducted using the UK Biobank Resource under Application Number 69794.

Author contributions

SD and JER-Z, carried out most analyses of the results of the application of the boostDM-CH models to the mutations in the discovery cohorts, validation cohorts and UKB. FM carried out the training and validation of boostDM-CH models and contributed to downstream analyses. SD, JER-Z, and FM generated the figures in the manuscript. MLG carried out the identification of somatic mutations in 12 CH driver genes across UKB donors. MA carried out the analysis of fitness of recurrent CH mutations across UKB donors. SD, FM, AG-P, and NL-B conceived the project. AG-P and NL-B supervised the project and its analyses and prepared the first draft of the manuscript. All authors contributed to the discussion of results and the completion of the manuscript.

References

1. Bolton KL, Ptashkin RN, Gao T, Braunstein L, Devlin SM, Kelly D, et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nat Genet. Nature Publishing Group*; 2020;52:1219–26.
2. Bowman RL, Busque L, Levine RL. Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. *Cell Stem Cell. Elsevier*; 2018;22:157–70.
3. Busque L, Patel JP, Figueroa ME, Vasanthakumar A, Provost S, Hamilou Z, et al. Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat Genet. 2012*;44:1179–81.
4. Busque L, Buscarlet M, Mollica L, Levine RL. Concise Review: Age-Related Clonal Hematopoiesis: Stem Cells Tempting the Devil. *STEM CELLS. 2018*;36:1287–94.
5. Challen GA, Goodell MA. Clonal hematopoiesis: mechanisms driving dominance of stem cell clones. *Blood. American Society of Hematology*; 2020;136:1590–8.
6. Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. *Science. 2019*;366(6465):4673.
7. Chen S, Wang Q, Yu H, Capitano ML, Vemula S, Nabinger SC, et al. Mutant p53 drives clonal hematopoiesis through modulating epigenetic pathway. *Nat Commun. Nature Publishing Group*; 2019;10:5649.
8. Coombs CC, Zehir A, Devlin SM, Kishtagari A, Syed A, Jonsson P, et al. Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes. *Cell Stem Cell. 2017*;21:374-382.e4.
9. Fuster José J., Walsh Kenneth. Somatic Mutations and Clonal Hematopoiesis. *Circ Res. American Heart Association*; 2018;122:523–32.
10. Gao T, Ptashkin R, Bolton KL, Sirenko M, Fong C, Spitzer B, et al. Interplay between chromosomal alterations and gene mutations shapes the evolutionary trajectory of clonal hematopoiesis. *Nat Commun. 2021*;12:338.
11. Hsu JI, Dayaram T, Tovy A, De Braekeleer E, Jeong M, Wang F, et al. PPM1D Mutations Drive Clonal Hematopoiesis in Response to Cytotoxic Chemotherapy. *Cell Stem Cell. 2018*;23:700-713.e6.
12. Pich O, Reyes-Salazar I, Gonzalez-Perez A, Lopez-Bigas N. Discovering the drivers of clonal hematopoiesis. *Nat Commun. Nature Publishing Group*; 2022;13:4267.
13. Hagiwara K, Natarajan S, Wang Z, Zubair H, Mulder HL, Dong L, et al. Dynamics of Age-versus Therapy-Related Clonal Hematopoiesis in Long-term Survivors of Pediatric Cancer. *Cancer Discov. 2023*;13:844–57.
14. Marnell CS, Bick A, Natarajan P. Clonal hematopoiesis of indeterminate potential (CHIP): Linking somatic mutations, hematopoiesis, chronic inflammation and cardiovascular disease. *J Mol Cell Cardiol. Elsevier*; 2021;161:98–105.
15. Calvillo-Argüelles O, Jaiswal S, Shlush LI, Moslehi JJ, Schimmer A, Barac A, et al. Connections Between Clonal Hematopoiesis, Cardiovascular Disease, and Cancer: A Review. *JAMA Cardiol. American Medical Association*; 2019;4:380–7.
16. Arends CM, Liman TG, Strzelecka PM, Kufner A, Löwe P, Huo S, et al. Associations of clonal hematopoiesis with recurrent vascular events and death in patients with incident ischemic stroke. *Blood. 2023*;141:787–99.
17. Miller PG, Qiao D, Rojas-Quintero J, Honigberg MC, Sperling AS, Gibson CJ, et al. Association of clonal hematopoiesis with chronic obstructive pulmonary disease. *Blood. 2022*;139:357–68.
18. Yu B, Roberts MB, Raffield LM, Zekavat SM, Nguyen NQH, Biggs ML, et al. Association of Clonal Hematopoiesis With Incident Heart Failure. *J Am Coll Cardiol. 2021*;78:42–52.

19. Mitchell SR, Gopakumar J, Jaiswal S. Insights into clonal hematopoiesis and its relation to cancer risk. *Curr Opin Genet Dev.* 2021;66:63–9.
20. Bolton KL, Koh Y, Foote MB, Im H, Jee J, Sun CH, et al. Clonal hematopoiesis is associated with risk of severe Covid-19. *Nat Commun.* Nature Publishing Group; 2021;12:5975.
21. Wang Y, Sano S, Ogawa H, Horitani K, Evans MA, Yura Y, et al. Murine models of clonal haematopoiesis to assess mechanisms of cardiovascular disease. *Cardiovasc Res.* 2022;118:1413–32.
22. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* Public Library of Science; 2015;12:e1001779.
23. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature.* Nature Publishing Group; 2021;590:290–9.
24. Vlasschaert C, Mack T, Heimlich JB, Niroula A, Uddin MM, Weinstock J, et al. A practical approach to curate clonal hematopoiesis of indeterminate potential in human genetic data sets. *Blood.* 2023;141:2214–23.
25. Niroula A, Sekar A, Murakami MA, Trinder M, Agrawal M, Wong WJ, et al. Distinction of lymphoid and myeloid clonal hematopoiesis. *Nat Med.* 2021;27:1921–7.
26. Bick AG, Weinstock JS, Nandakumar SK, Fulco CP, Bao EL, Zekavat SM, et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature.* 2020;586:763–8.
27. Swerdlow S, Campo E, Harris E, Pileri S, Stein H, Thiele J. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. 4th ed. IARC; 2017.
28. Muiños F, Martínez-Jiménez F, Pich O, Gonzalez-Perez A, Lopez-Bigas N. In silico saturation mutagenesis of cancer genes. *Nature.* 2021;596:428–32.
29. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. *Nat Rev Cancer.* Nature Publishing Group; 2020;20:555–72.
30. Arnedo-Pac C, Mularoni L, Muiños F, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics.* Oxford Academic; 2019;35:4788–90.
31. Tokheim C, Bhattacharya R, Niknafs N, Gygi DM, Kim R, Ryan M, et al. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res.* 2016;76:3719–31.
32. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 2010;20:110–21.
33. Lue NZ, Garcia EM, Ngan KC, Lee C, Doench JG, Liao BB. Base editor scanning charts the DNMT3A activity landscape. *Nat Chem Biol.* 2023;19:176–86.
34. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, et al. Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol.* 2017;27:S2–8.
35. Kato S, Han S-Y, Liu W, Otsuka K, Shibata H, Kanamaru R, et al. Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc Natl Acad Sci U S A.* 2003;100:8424–9.
36. Giacomelli AO, Yang X, Lintner RE, McFarland JM, Duby M, Kim J, et al. Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat Genet.* Nature Publishing Group; 2018;50:1381–7.
37. Ursu O, Neal JT, Shea E, Thakore PI, Jerby-Arnon L, Nguyen L, et al. Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nat Biotechnol.* Nature Publishing Group; 2022;40:896–905.

38. Kotler E, Shani O, Goldfeld G, Lotan-Pompan M, Tarcic O, Gershoni A, et al. A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. *Mol Cell*. Elsevier; 2018;71:178-190.e8.
39. Sendžikaitė G, Hanna CW, Stewart-Morgan KR, Ivanova E, Kelsey G. A DNMT3A PWWP mutation leads to methylation of bivalent chromatin and growth retardation in mice. *Nat Commun*. Nature Publishing Group; 2019;10:1884.
40. Kaasinen E, Kuusmin O, Rajamäki K, Ristolainen H, Aavikko M, Kondelin J, et al. Impact of constitutional TET2 haploinsufficiency on molecular and clinical phenotype in humans. *Nat Commun*. 2019;10:1252.
41. Waheed SO, Chaturvedi SS, Karabenchewa-Christova TG, Christov CZ. Catalytic Mechanism of Human Ten-Eleven Translocation-2 (TET2) Enzyme: Effects of Conformational Changes, Electric Field, and Mutations. *ACS Catal*. American Chemical Society; 2021;11:3877–90.
42. Tulstrup M, Soerensen M, Hansen JW, Gillberg L, Needhamsen M, Kaastrup K, et al. TET2 mutations are associated with hypermethylation at key regulatory enhancers in normal and malignant hematopoiesis. *Nat Commun*. 2021;12:6061.
43. Kahn JD, Miller PG, Silver AJ, Sellar RS, Bhatt S, Gibson C, et al. PPM1D-truncating mutations confer resistance to chemotherapy and sensitivity to PPM1D inhibition in hematopoietic cells. *Blood*. 2018;132:1095–105.
44. Watson CJ, Papula AL, Poon GYP, Wong WH, Young AL, Druley TE, et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science*. American Association for the Advancement of Science; 2020;367:1449–54.
45. Kohnke T, Nuno KA, Alder CC, Gars EJ, Phan P, Fan AC, et al. Human ASXL1-mutant hematopoiesis is driven by a truncated protein associated with aberrant de-ubiquitination of H2AK119. *Blood Cancer Discov*. 2024;5:1–22.
46. Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell*. Elsevier; 2018;173:283–5.
47. Martínez-Jiménez F, Movasati A, Brunner SR, Nguyen L, Priestley P, Cuppen E, et al. Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature*. Nature Publishing Group; 2023;618:333–41.
48. Cheng DT, Mitchell T, Zehir A, Shah RH, Benayed R, Syed A, et al. MSK-IMPACT: A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J Mol Diagn JMD*. 2015;17:251–64.
49. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov*. 2012;2:401–4.
50. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol*. Nature Publishing Group; 2017;35:316–9.
51. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min*. San Francisco, California, USA: Association for Computing Machinery; 2016. page 785–94.
52. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Proc 31st Int Conf Neural Inf Process Syst*. Red Hook, NY, USA: Curran Associates Inc.; 2017. page 4768–77.
53. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. Nature Publishing Group; 2020;581:434–43.
54. Trinder M, Walley KR, Boyd JH, Brunham LR. Causal Inference for Genetically Determined

Levels of High-Density Lipoprotein Cholesterol and Risk of Infectious Disease. *Arterioscler Thromb Vasc Biol.* 2020;40:267–78.

55. Kar SP, Quiros PM, Gu M, Jiang T, Mitchell J, Langdon R, et al. Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. *Nat Genet.* Nature Publishing Group; 2022;54:1155–66.

Figure legends

Figure 1. Building and evaluating BoostDM-CH models.

- Blood somatic mutations used to train the boostDM-CH models were identified across three cancer genomics cohorts through reverse calling (blood vs tumor sample); then, CH driver genes were identified by uncovering signals of positive selection using the IntOGen pipeline; finally, models to identify CH driver mutations were built.
 - Training and cross-validation of machine learning-based boostDM-CH models, exemplified through *DNMT3A*. In the outcome of the model, mutations with a score equal to or greater than 0.5 are deemed CH drivers.
 - Explanation of the classification of the *DNMT3A*-R882H mutation as a driver based on the contribution of the features employed in training the model. The numbers in the radar plot correspond to the SHAP values of each feature. Features with positive SHAP value (i.e., positive contribution to the classification of a mutation as driver) appear above the '0' line in the radar plot, and the main contributing features are shown in bold.
 - Performance (median $\pm$ interquartile range (IQR) F_{50}) of the cross-validation of 25 CH models as a function of their number of observed mutations. Blue dots represent the genes with models having F_{50} above 0.8 and sufficient discovery index to deem the set of mutations across cancer genomics cohorts representative of their CH driver mutations (high-quality models).
 - Precision-Recall curves (with the value of the area indicated) of the 12 high-quality boostDM-CH models.
 - Performance (median $\pm$ IQR F_{50}) in the classification of *DNMT3A* blood somatic mutations of *DNMT3A* boostDM-CH model and a *DNMT3A* experimental base editing assay.
 - Performance (median $\pm$ IQR F_{50}) in the classification of blood somatic mutations of boostDM-CH models and three sets of expert-curated rules. Left, overall performance in three CH datasets; right, gene-specific performance in one of the datasets.
- PTM: post-translational modifications; NMD: nonsense-mediated decay.

Figure 2. *In Silico* Saturation Mutagenesis of CH genes.

Blueprints of CH driver mutations in *DNMT3A* (a) and *TET2* (b). The plots represent the distribution of driver (red) and non-driver (gray) mutations. From top to bottom, the first plot contains observed exonic mutations (i.e., excluding intronic splice mutations, which are included in the training) across cancer genomics cohorts classified as drivers or non-drivers by boostDM-CH, with the height of each needle denoting the recurrence of the mutation. The plot below the observed mutations presents the boostDM-CH scores of all possible SNVs along the protein sequence (*in silico* saturation mutagenesis). Vertical ticks below this plot represent all positions of the CH drivers predicted by boostDM-CH (in red). Vertical ticks immediately below (in gold) show either missense or nonsense drivers classified in high-confidence (boostDM-CH score ≥ 0.9), or other (boostDM-CH score ≥ 0.5 and < 0.9). The values of mutational features used to train the models are shown linearly along the protein sequence in the plot at the bottom of the figure.

The histograms to the right show the distribution of boostDM-CH scores of all mutations identified in the three cancer genomics discovery cohorts. The dashed line indicates thresholds at 0.5 and 0.9 boostDM-CH score, which define high confidence drivers and other drivers. The bars below show the number (and proportion) of non-drivers and drivers (divided into high-confidence and others) observed in these

cohorts. The two bars represent the partition of all (top) or unique (that is, considering each different mutation only once, bottom) observed mutations.

c) Radar plots representing SHAP values (decomposition of the boostDM-CH score of mutations to the additive contribution of features; see Supp. Methods) for several illustrative driver mutations. In each mutation, the main features contributing to the driver prediction are highlighted in bold.

Figure 3. Comparison of boostDM-CH models and myeloid boostDM models in three genes.

For each gene (a-c), the needle plots represent the distribution of driver (red) and non-driver (gray) observed mutations in CH (top) and myeloid cancer (bottom) cohorts. The internal plots represent the distribution of driver mutation along the sequence of each gene using *in silico* saturation mutagenesis by boostDM-CH (red) and boostDM-myeloid cancer (blue).

Figure 4. Application of boostDM-CH to identify CH driver mutations across 467,202 donors.

a) Identification of CH driver mutations across UKB donors using boostDM-CH models.

b) Distribution of VAF of CH driver and non-driver mutations; inner barplot, number of driver mutations identified in each of the 12 CH driver genes studied.

c) Fold increase in the proportion of cases with CH driver and non-driver mutations across age groups (left). Association of CH with age measured via logistic regression (right).

In this and subsequent regression plots in Figure 5, the results of several sets of donors in the UKB are shown: “driver”, donors bearing at least one CH driver mutation (according to boostDM-CH); “multiple driver”, donors bearing more than one CH driver mutation; “driver large”, donors bearing at least one CH driver mutation with VAF $\geq 10\%$; “driver small”, donors bearing a CH driver mutation(s) with VAF $< 10\%$; “non-driver”, donors bearing a potential mutation in a CH driver gene classified as non-driver by boostDM-CH; “potential mutation”, donors bearing any potential mutation (driver or non-driver) in a CH driver gene (see Methods).

d) Association between carriers of CH mutations in the different CH genes and age. Mutations predicted as drivers by boostDM-CH increase significantly with age in all genes, except *CHEK2* and *MDM4* (Supp. Note).

Figure 5. BoostDM-CH driver mutations are significantly associated with known phenotypes.

a) Association of CH with known CH promoting factors, including smoking and solid tumor history prior to enrollment in UKB (as a proxy of previous chemotherapy treatment), measured via logistic regression.

b) Association of CH with the risk of developing myeloid malignancies, measured via Kaplan-Maier analysis (upper) and Cox regression analysis (bottom).

c) Association of CH with all-cause death, heart failure, the occurrence of a solid tumor, and any infection (obtained via a composite variable), measured by logistic regression.

Asterisks represent significant associations (FDR <0.05); ns: non-significant.

In logistic and Cox regression analyses presented in this and subsequent figures, age, sex, and 10 ancestry principal components are included as covariables (depending on the dependent variable in each case). Tumor-related regression analyses in this and other figures also include smoking history as covariate. Heart failure regression analyses presented in this and other figures also include smoking history, dyslipidemia, body mass index, hypertension, and diabetes type II status as covariables. Infectious disease regression analyses also include smoking history and occurrence of hematological neoplasms as covariables (Supp. Methods).

Figure 6. CH driver mutations identified *de novo* by boostDM-CH.

- a) Number of CH driver mutations from the UKB cohort identified by boostDM-CH and three expert-curated rule sets in the 10 CH genes in common. The comparison of the consensus between the three sets of rules and boostDM-CH defines three groups of mutations: boostDM-exclusive, rules-boostDM-overlapping, and rules-exclusive. The subsequent plots present analyses with these three groups of mutations.
- b) Fold increase in the proportion of cases bearing CH mutations in each of the three groups of mutations across age bins (left). Association of the presence of CH mutations of each of the three groups and age in the 10 CH genes, measured by logistic regression (right). On the right-hand of the plot, the number of donors in the UKB cohort bearing mutations of each of the three groups is shown.
- c) Association of the presence of the different groups of CH mutations and all-cause mortality (top) and development of myeloid cancer (bottom) in the 10 CH genes, measured by logistic regression.
- d) Association with age of the presence of CH mutations of each of the three groups in either *DNMT3A* or *TP53*.
- e) Association with age of hotspot mutations (at positions R140 and R172), as well as all mutations identified as non-drivers and all potential mutations identified in *IDH2*. The bottom panel presents the distribution of effect size (odds-ratio) resulting from randomly sampling 5 disjoint sets of 9 mutations (the number of observed *IDH2* R172 mutations) from two driver (top) and two non-driver (bottom) hotspots (according to boostDM-CH). See Supplementary Note for details.
- Asterisks represent significant associations (FDR<0.05); ns: non-significant.

Figure 1

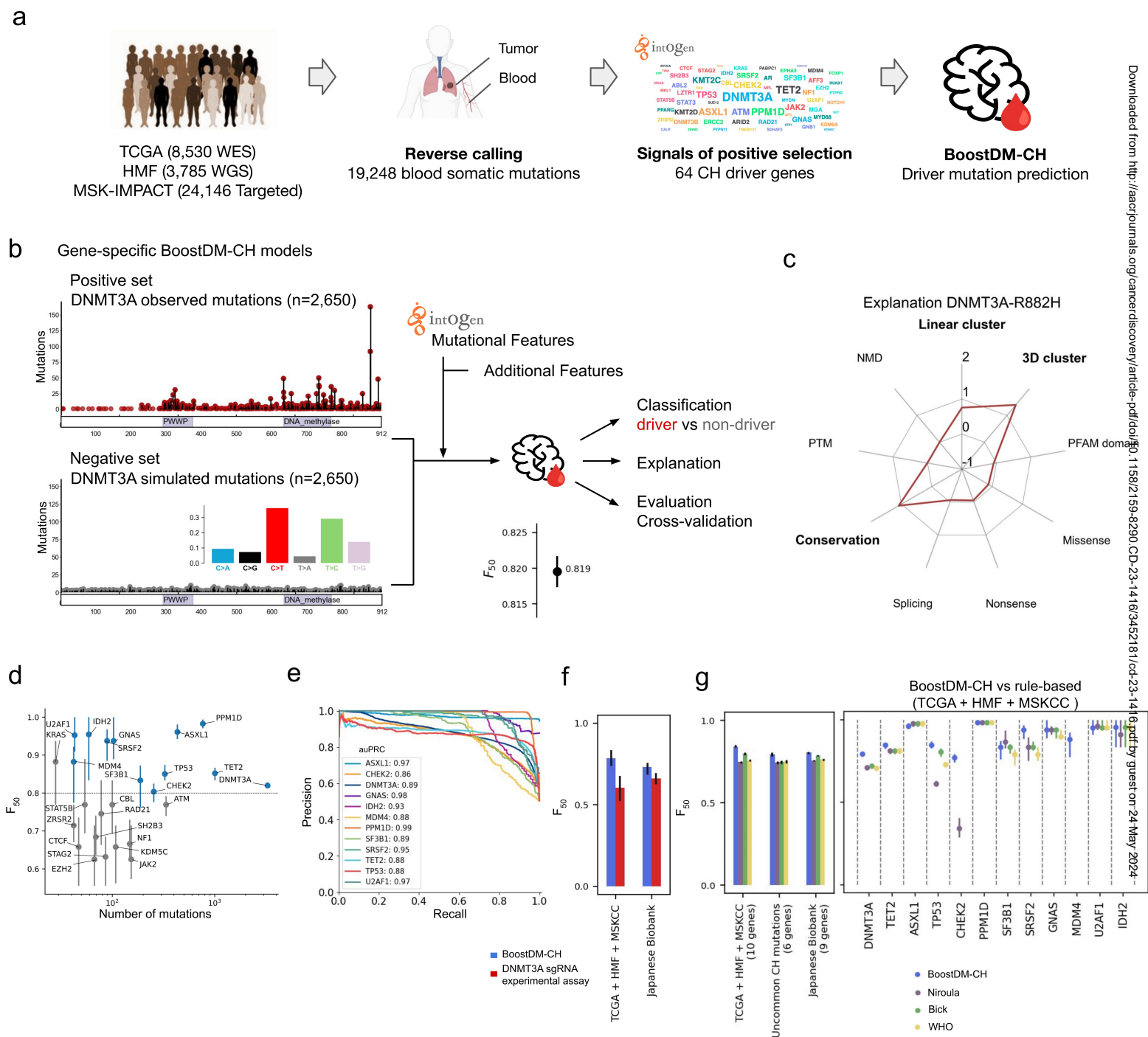


Figure 2

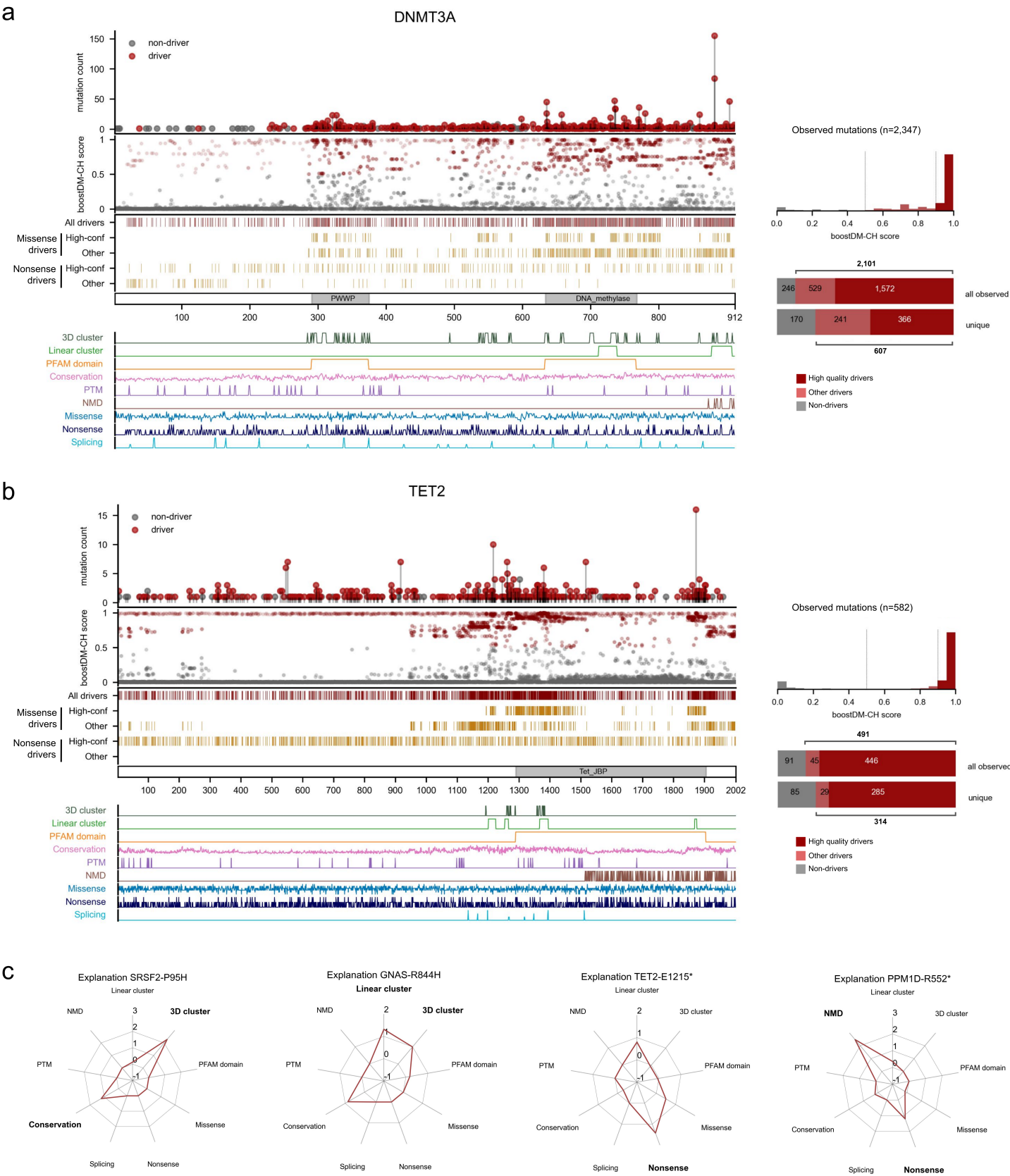


Figure 3

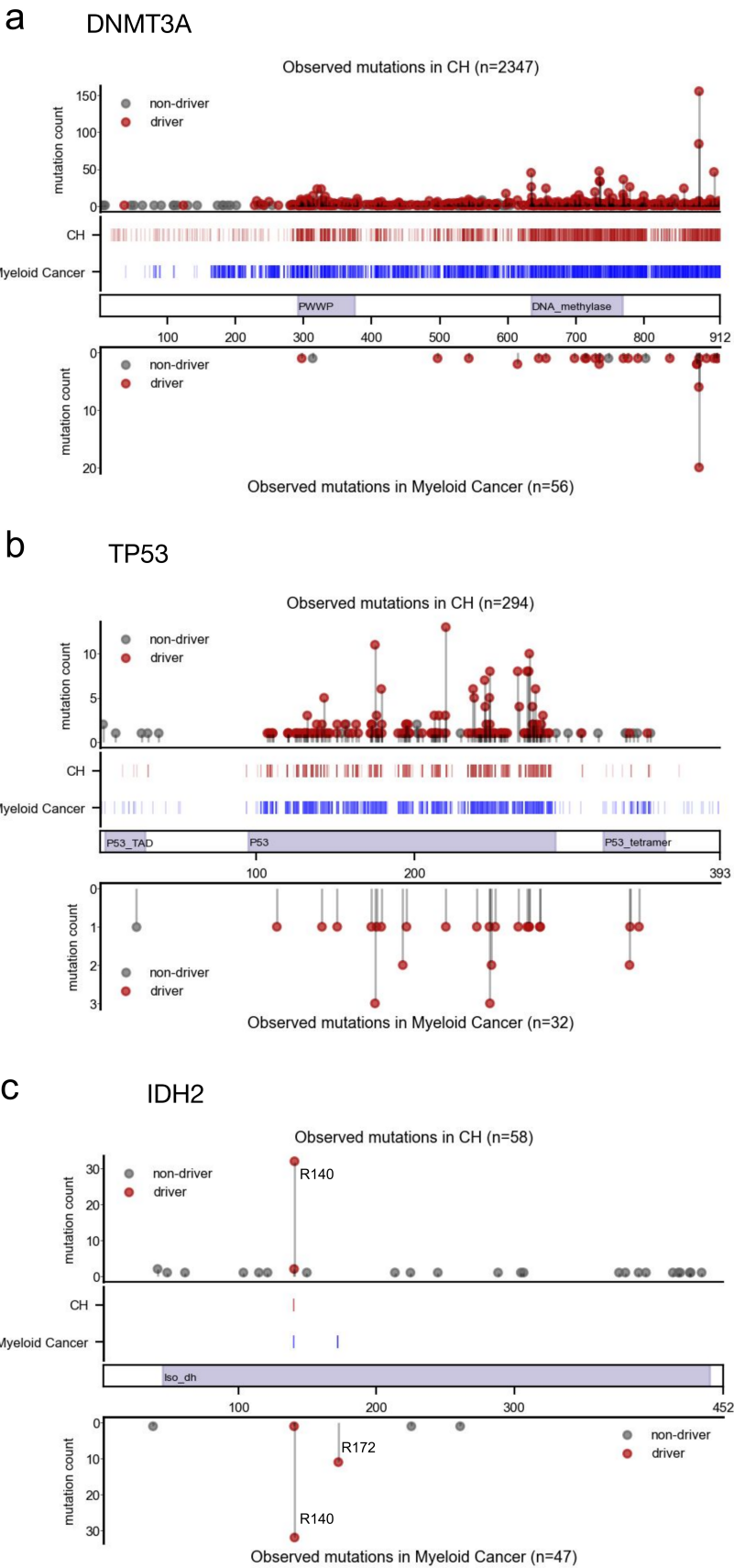


Figure 4

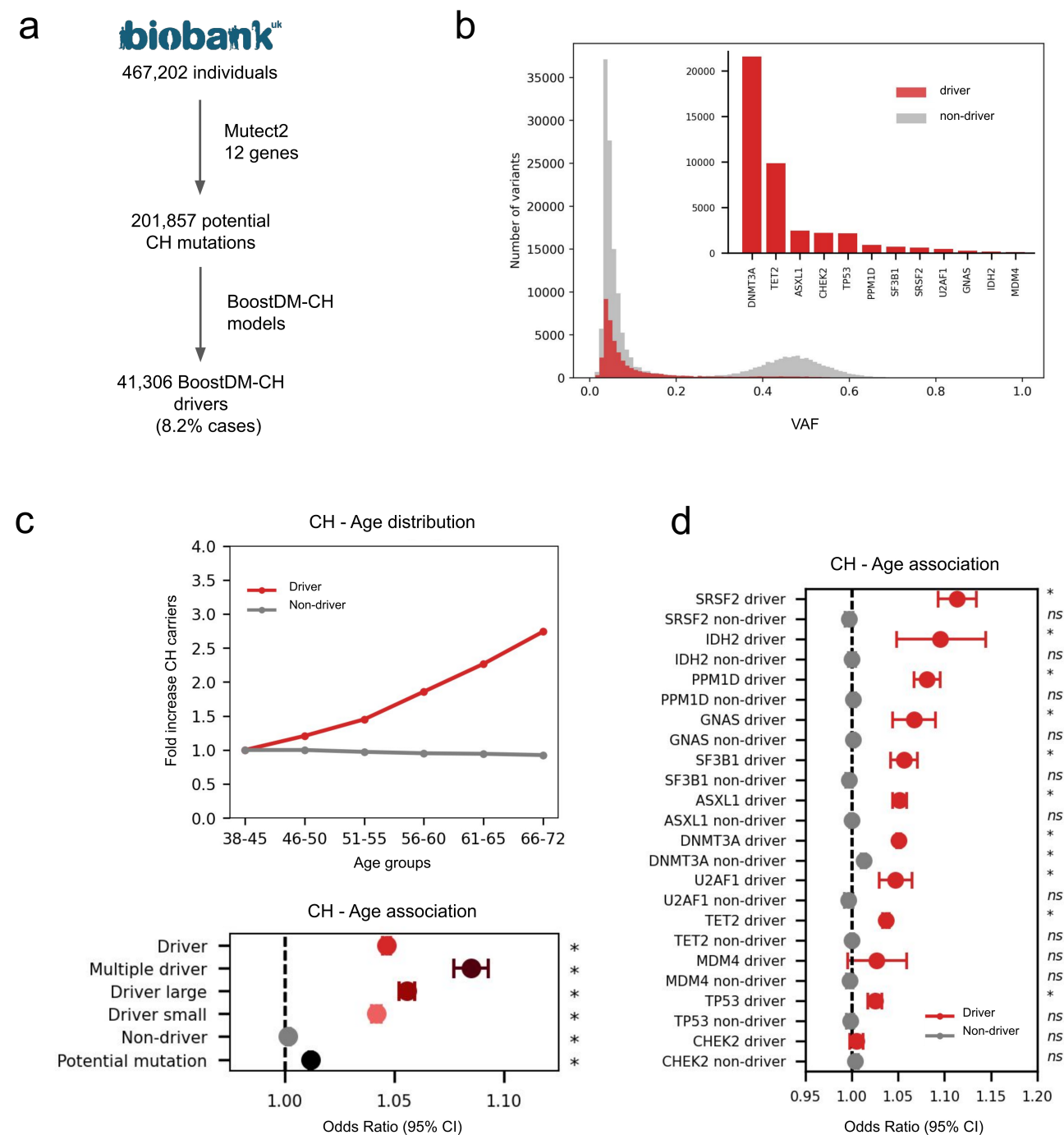


Figure 5

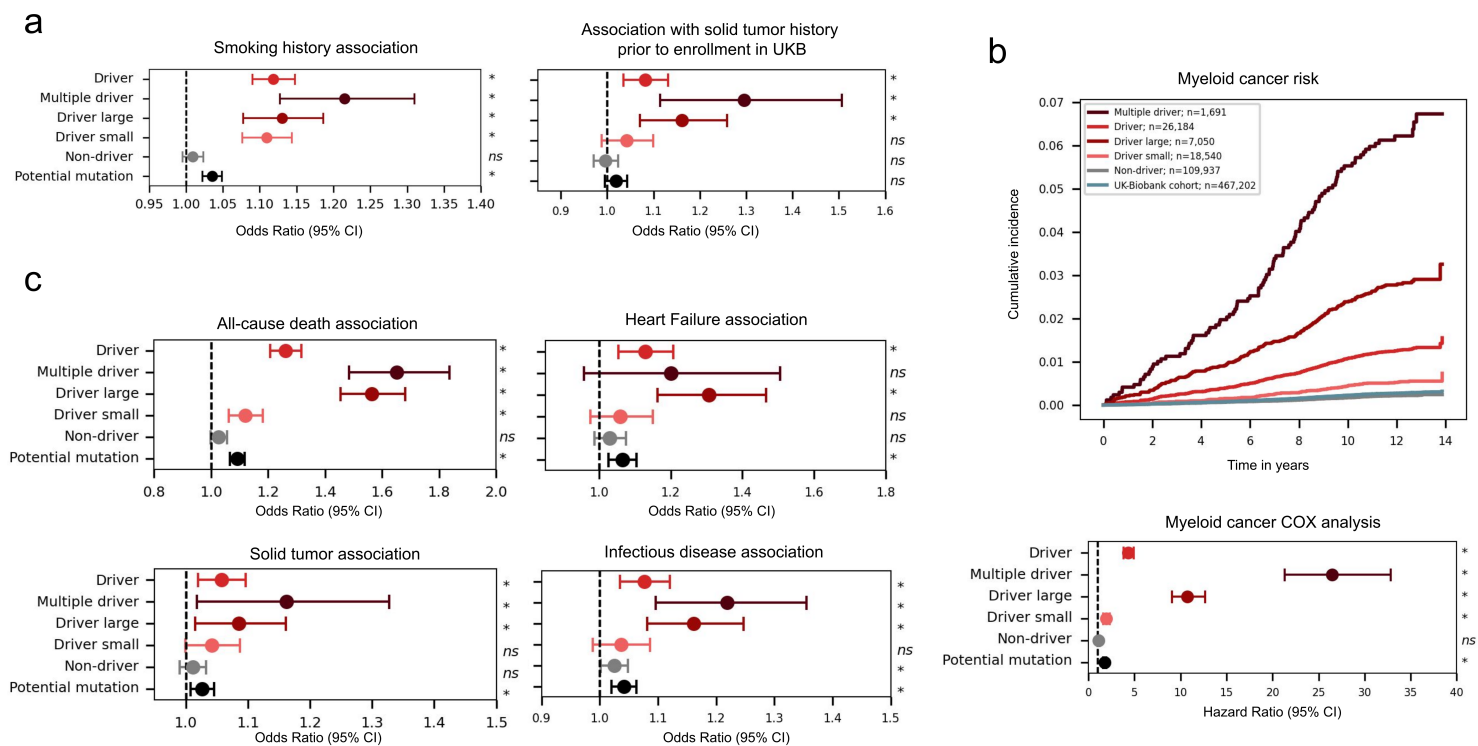
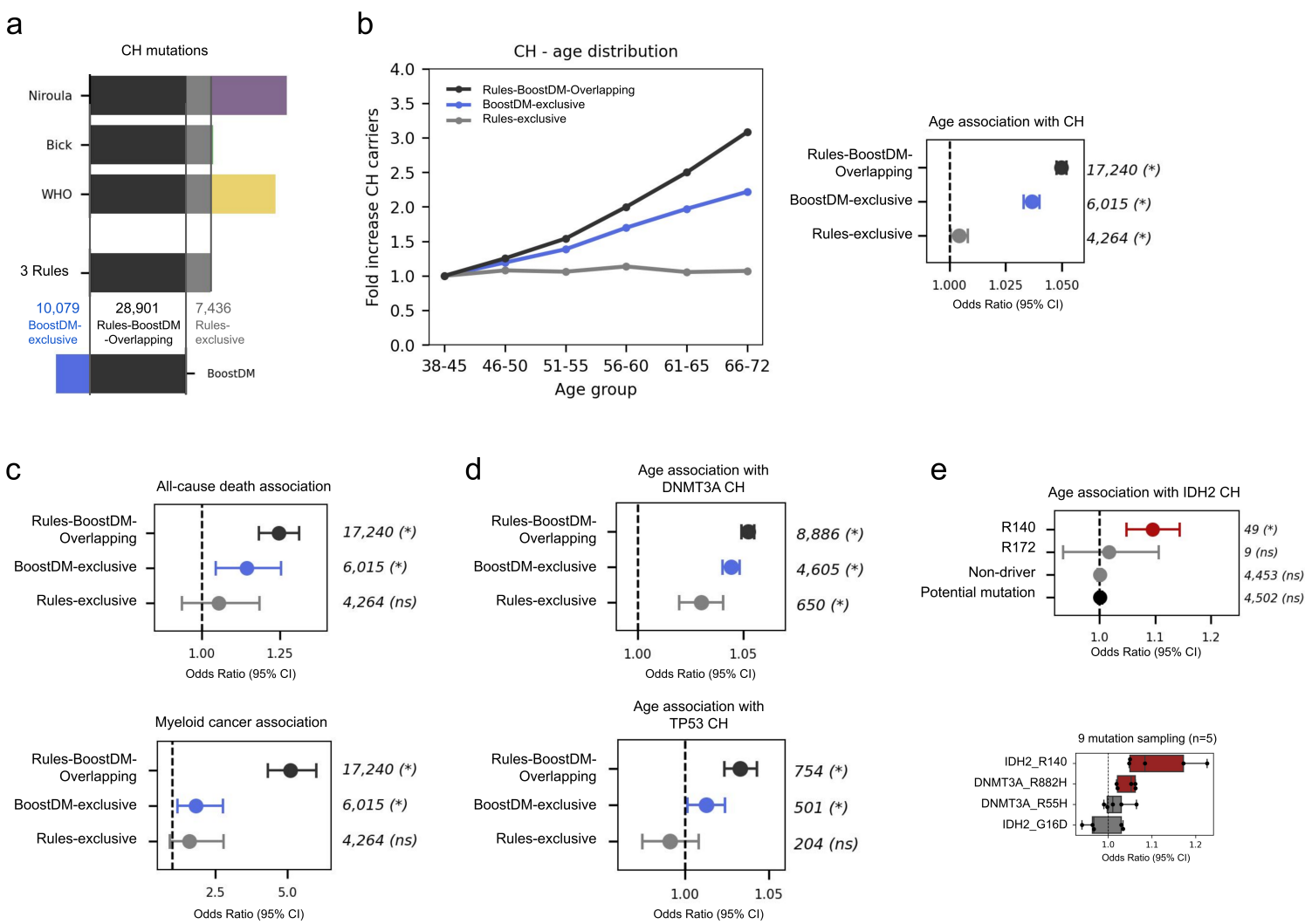


Figure 6



Parsed Citations

- 1. Bolton KL, Ptashkin RN, Gao T, Braunstein L, Devlin SM, Kelly D, et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. Nat Genet. Nature Publishing Group; 2020;52:1219-26.**
Pubmed: [Author and Title](#)
Google Scholar: [Google Scholar Search](#)
- 2. Bowman RL, Busque L, Levine RL. Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. Cell Stem Cell. Elsevier; 2018;22:157-70.**
Pubmed: [Author and Title](#)
Google Scholar: [Google Scholar Search](#)
- 3. Busque L, Patel JP, Figueroa ME, Vasanthakumar A, Provost S, Hamilou Z, et al. Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. Nat Genet. 2012;44:1179-81.**
Pubmed: [Author and Title](#)
Google Scholar: [Google Scholar Search](#)
- 4. Busque L, Buscarlet M, Mollica L, Levine RL. Concise Review: Age-Related Clonal Hematopoiesis: Stem Cells Tempting the Devil. STEM CELLS. 2018;36:1287-94.**
Pubmed: [Author and Title](#)
Google Scholar: [Google Scholar Search](#)
- 5. Challen GA, Goodell MA. Clonal hematopoiesis: mechanisms driving dominance of stem cell clones. Blood. American Society of Hematology; 2020;136:1590-8.**
Pubmed: [Author and Title](#)
Google Scholar: [Google Scholar Search](#)
- 6. Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. Science. 2019;366(6465):4673.**
Pubmed: [Author and Title](#)
Google Scholar: [Google Scholar Search](#)
- 7. Chen S, Wang Q, Yu H, Capitano ML, Vemula S, Nabinger SC, et al. Mutant p53 drives clonal hematopoiesis through modulating epigenetic pathway. Nat Commun. Nature Publishing Group; 2019;10:5649.**
Pubmed: [Author and Title](#)
Google Scholar: [Google Scholar Search](#)
- 8. Coombs CC, Zehir A, Devlin SM, Kishtagari A, Syed A, Jonsson P, et al. Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes. Cell Stem Cell. 2017;21:374-382.e4.**
Pubmed: [Author and Title](#)
Google Scholar: [Google Scholar Search](#)
- 9. Fuster Jose J., Walsh Kenneth. Somatic Mutations and Clonal Hematopoiesis. Circ Res. American Heart Association; 2018;122:523-32.**
Pubmed: [Author and Title](#)
Google Scholar: [Google Scholar Search](#)
- 10. Gao T, Ptashkin R, Bolton KL, Sirenko M, Fong C, Spitzer B, et al. Interplay between chromosomal alterations and gene mutations shapes the evolutionary trajectory of clonal hematopoiesis. Nat Commun. 2021;12:338.**
Pubmed: [Author and Title](#)
Google Scholar: [Google Scholar Search](#)
- 11. Hsu JI, Dayaram T, Tovy A, De Braekeleer E, Jeong M, Wang F, et al. PPM1D Mutations Drive Clonal Hematopoiesis in Response to Cytotoxic Chemotherapy. Cell Stem Cell. 2018;23:700-713.e6.**
Pubmed: [Author and Title](#)
Google Scholar: [Google Scholar Search](#)

12. Pich O, Reyes-Salazar I, Gonzalez-Perez A, Lopez-Bigas N. Discovering the drivers of clonal hematopoiesis. Nat Commun. Nature Publishing Group; 2022;13:4267.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

13. Hagiwara K, Natarajan S, Wang Z, Zubair H, Mulder HL, Dong L, et al. Dynamics of Age- versus Therapy-Related Clonal Hematopoiesis in Long-term Survivors of Pediatric Cancer. Cancer Discov. 2023;13:844-57.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

14. Marnell CS, Bick A, Natarajan P. Clonal hematopoiesis of indeterminate potential (CHIP): Linking somatic mutations, hematopoiesis, chronic inflammation and cardiovascular disease. J Mol Cell Cardiol. Elsevier; 2021;161:98-105.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

15. Calvillo-Arguelles O, Jaiswal S, Shlush LI, Moslehi JJ, Schimmer A, Barac A, et al. Connections Between Clonal Hematopoiesis, Cardiovascular Disease, and Cancer: A Review. JAMA Cardiol. American Medical Association; 2019;4:380-7.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

16. Arends CM, Liman TG, Strzelecka PM, Kufner A, Lowe P, Huo S, et al. Associations of clonal hematopoiesis with recurrent vascular events and death in patients with incident ischemic stroke. Blood. 2023;141:787-99.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

17. Miller PG, Qiao D, Rojas-Quintero J, Honigberg MC, Sperling AS, Gibson CJ, et al. Association of clonal hematopoiesis with chronic obstructive pulmonary disease. Blood. 2022;139:357-68.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

18. Yu B, Roberts MB, Raffield LM, Zekavat SM, Nguyen NQH, Biggs ML, et al. Association of Clonal Hematopoiesis With Incident Heart Failure. J Am Coll Cardiol. 2021;78:42-52.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

19. Mitchell SR, Gopakumar J, Jaiswal S. Insights into clonal hematopoiesis and its relation to cancer risk. Curr Opin Genet Dev. 2021;66:63-9.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

20. Bolton KL, Koh Y, Foote MB, Im H, Jee J, Sun CH, et al. Clonal hematopoiesis is associated with risk of severe Covid-19. Nat Commun. Nature Publishing Group; 2021;12:5975.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

21. Wang Y, Sano S, Ogawa H, Horitani K, Evans MA, Yura Y, et al. Murine models of clonal haematopoiesis to assess mechanisms of cardiovascular disease. Cardiovasc Res. 2022;118:1413-32.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

22. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med. Public Library of Science; 2015;12:e1001779.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

23. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. Nature Publishing Group; 2021;590:290-9.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

24. Vlasschaert C, Mack T, Heimlich JB, Niroula A, Uddin MM, Weinstock J, et al. A practical approach to curate clonal hematopoiesis of indeterminate potential in human genetic data sets. Blood. 2023;141:2214-23.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

25. Niroula A, Sekar A, Murakami MA, Trinder M, Agrawal M, Wong WJ, et al. Distinction of lymphoid and myeloid clonal hematopoiesis. Nat Med. 2021;27:1921-7.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

26. Bick AG, Weinstock JS, Nandakumar SK, Fulco CP, Bao EL, Zekavat SM, et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature. 2020;586:763-8.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

27. Swerdlow S, Campo E, Harris E, Pileri S, Stein H, Thiele J. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. 4th ed. IARC; 2017.

28. Muinos F, Martinez-Jimenez F, Pich O, Gonzalez-Perez A, Lopez-Bigas N. In silico saturation mutagenesis of cancer genes. Nature. 2021;596:428-32.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

29. Martinez-Jimenez F, Muinos F, Sentis I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. Nature Publishing Group; 2020;20:555-72.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

30. Arnedo-Pac C, Mularoni L, Muinos F, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. Bioinformatics. Oxford Academic; 2019;35:4788-90.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

31. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. Cancer Res. 2016;76:3719-31.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

32. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20:110-21.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

33. Lue NZ, Garcia EM, Ngan KC, Lee C, Doench JG, Liao BB. Base editor scanning charts the DNMT3A activity landscape. Nat Chem Biol. 2023;19:176-86.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

34. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, et al. Overview of the BioBank Japan Project: Study design and profile. J Epidemiol. 2017;27:S2-8.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

35. Kato S, Han S-Y, Liu W, Otsuka K, Shibata H, Kanamaru R, et al. Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. Proc Natl Acad Sci U S A. 2003;100:8424-9.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

36. Giacomelli AO, Yang X, Lintner RE, McFarland JM, Duby M, Kim J, et al. Mutational processes shape the landscape of TP53 mutations in human cancer. Nat Genet. Nature Publishing Group; 2018;50:1381-7.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

37. Ursu O, Neal JT, Shea E, Thakore PI, Jerby-Arnon L, Nguyen L, et al. Massively parallel phenotyping of coding variants in cancer with Perturb-seq. Nat Biotechnol. Nature Publishing Group; 2022;40:896-905.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

38. Kotler E, Shani O, Goldfeld G, Lotan-Pompan M, Tarcic O, Gershoni A, et al. A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. Mol Cell. Elsevier; 2018;71:178-190.e8.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

39. Sendzikaite G, Hanna CW, Stewart-Morgan KR, Ivanova E, Kelsey G. A DNMT3A PWWP mutation leads to methylation of bivalent chromatin and growth retardation in mice. Nat Commun. Nature Publishing Group; 2019;10:1884.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

40. Kaasinen E, Kuismin O, Rajamaki K, Ristolainen H, Aavikko M, Kondelin J, et al. Impact of constitutional TET2 haploinsufficiency on molecular and clinical phenotype in humans. Nat Commun. 2019;10:1252.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

41. Waheed SO, Chaturvedi SS, Karabenchewa-Christova TG, Christov CZ. Catalytic Mechanism of Human Ten-Eleven Translocation-2 (TET2) Enzyme: Effects of Conformational Changes, Electric Field, and Mutations. ACS Catal. American Chemical Society; 2021;11:3877-90.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

42. Tulstrup M, Soerensen M, Hansen JW, Gillberg L, Needhamsen M, Kaastrup K, et al. TET2 mutations are associated with hypermethylation at key regulatory enhancers in normal and malignant hematopoiesis. Nat Commun. 2021;12:6061.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

43. Kahn JD, Miller PG, Silver AJ, Sellar RS, Bhatt S, Gibson C, et al. PPM1D-truncating mutations confer resistance to chemotherapy and sensitivity to PPM1D inhibition in hematopoietic cells. Blood. 2018;132:1095-105.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

44. Watson CJ, Papula AL, Poon GYP, Wong WH, Young AL, Druley TE, et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Science. American Association for the Advancement of Science; 2020;367:1449-54.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

45. Kohnke T, Nuno KA, Alder CC, Gars EJ, Phan P, Fan AC, et al. Human ASXL1-mutant hematopoiesis is driven by a truncated protein associated with aberrant de-ubiquitination of H2AK119. Blood Cancer Discov. 2024;5:1-22.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

46. Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell. Elsevier; 2018;173:283-5.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

47. Martinez-Jimenez F, Movasati A, Brunner SR, Nguyen L, Priestley P, Cuppen E, et al. Pan-cancer whole-genome comparison of primary and metastatic solid tumours. Nature. Nature Publishing Group; 2023;618:333-41.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

48. Cheng DT, Mitchell T, Zehir A, Shah RH, Benayed R, Syed A, et al. MSK-IMPACT: A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. J Mol Diagn JMD. 2015;17:251-64.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

49. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discov. 2012;2:401-4.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

50. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. Nature Publishing Group; 2017;35:316-9.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

51. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. San Francisco, California, USA: Association for Computing Machinery; 2016. page 785-94.

52. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proc 31st Int Conf Neural Inf Process Syst. Red Hook, NY, USA: Curran Associates Inc.; 2017. page 4768-77.

53. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. Nature Publishing Group; 2020;581:434-43.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

54. Trinder M, Walley KR, Boyd JH, Brunham LR. Causal Inference for Genetically Determined Levels of High-Density Lipoprotein Cholesterol and Risk of Infectious Disease. Arterioscler Thromb Vasc Biol. 2020;40:267-78.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)

55. Kar SP, Quiros PM, Gu M, Jiang T, Mitchell J, Langdon R, et al. Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. Nat Genet. Nature Publishing Group; 2022;54:1155-66.

Pubmed: [Author and Title](#)

Google Scholar: [Google Scholar Search](#)