nature genetics

ARTICLES

https://doi.org/10.1038/s41588-020-00710-0

Cancer therapy shapes the fitness landscape of clonal hematopoiesis

Kelly L. Bolton[1], Ryan N. Ptashkin[2,34], Teng Gao[3,34], Lior Braunstein[4], Sean M. Devlin[5], Daniel Kelly[6], Minal Patel[7], Antonin Berthon[3], Aijazuddin Syed[2], Mariko Yabe[8], Catherine C. Coombs[9], Nicole M. Caltabellotta[7], Mike Walsh[10], Kenneth Offit[10], Zsofia Stadler[11], Diana Mandelker[2], Jessica Schulman[7], Akshar Patel[7], John Philip[12], Elsa Bernard[3], Gunes Gundem[3], Juan E. Arango Ossa[7], Max Levine[13], Juan S. Medina Martinez[13], Noushin Farnoud[7], Dominik Glodzik[3], Sonya Li[10], Mark E. Robson[10], Choonsik Lee[14], Paul D. P. Pharoah[15,16], Konrad H. Stopsack[10], Barbara Spitzer[13], Simon Mantha[17], James Fagin[10,18], Laura Boucai[19], Christopher J. Gibson[20], Benjamin L. Ebert[20], Andrew L. Young[21], Todd Druley[22], Koichi Takahashi[23], Nancy Gillis[24,25], Markus Ball[25,26], Eric Padron[25], David M. Hyman[10,27], Jose Baselga[28], Larry Norton[10,27], Stuart Gardos[10,27], Virginia M. Klimek[10,27], Howard Scher[10,27], Dean Bajorin[10,27], Eder Paraiso[19,29], Ryma Benayed[2], Maria E. Arcila[2], Marc Ladanyi[2], David B. Solit[10,19,30], Michael F. Berger[2,19,30], Martin Tallman[1], Montserrat Garcia-Closas[14], Nilanjan Chatterjee[31], Luis A. Diaz Jr[10,32,33], Ross L. Levine[1], Lindsay M. Morton[14], Ahmet Zehir[2,34,35] ✉ and Elli Papaemmanuil[3,34,35] ✉

Acquired mutations are pervasive across normal tissues. However, understanding of the processes that drive transformation of certain clones to cancer is limited. Here we study this phenomenon in the context of clonal hematopoiesis (CH) and the development of therapy-related myeloid neoplasms (tMNs). We find that mutations are selected differentially based on exposures. Mutations in *ASXL1* are enriched in current or former smokers, whereas cancer therapy with radiation, platinum and topoisomerase II inhibitors preferentially selects for mutations in DNA damage response genes (*TP53*, *PPM1D*, *CHEK2*). Sequential sampling provides definitive evidence that DNA damage response clones outcompete other clones when exposed to certain therapies. Among cases in which CH was previously detected, the CH mutation was present at tMN diagnosis. We identify the molecular characteristics of CH that increase risk of tMN. The increasing implementation of clinical sequencing at diagnosis provides an opportunity to identify patients at risk of tMN for prevention strategies.

The multistage model of carcinogenesis suggests that the successive acquisition of somatic mutations predates cancer development[1]. Each mutation contributes to a clone's fitness advantage, resulting in clonal expansions that culminate in malignant transformation, in a process that parallels Darwinian evolution[2]. This evolutionary process results from a complex interplay between the mechanisms that drive mutagenesis, the genetic targets of selection and the contexts in which these mutations contribute to differential clonal fitness.

Systematic cancer sequencing studies have delivered a detailed understanding of the processes that lead to mutations, the resulting mutation signature and the genetic drivers of malignant disease[3,4]. However, our understanding of the evolutionary trajectories that underlie cancer development is primarily based on retrospective modeling of clonal structures observed at diagnosis[5] or disease progression[6]. Such approaches do not allow characterization of the genetic and clonal dynamics of early oncogenesis. Recent sequencing studies of normal tissues show that acquisition of somatic mutations is pervasive with aging[7–16]. Our understanding of the environmental factors that drive a subset of these mutated clones towards malignant transformation is limited and largely based on in vitro and animal studies[17–19]. Progress in this regard has been challenged by the paucity of longitudinal genetic and clonal studies with detailed annotation of intervening exposures.

Studies of CH present a unique opportunity to study the evolutionary process underlying malignant transformation in blood. Noninvasive sampling enables acquisition of statistically powered cohorts and longitudinal samples that permit assessment of the transition from normal to transformed disease. Population studies show that individuals with CH are at increased risk of transformation to myeloid neoplasms[20,21]. However, only a small proportion of people with CH progress to myeloid neoplasms. Patients with cancer are at heightened risk of subsequent tMNs such as acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS)[22,23]. tMN was traditionally thought to develop from the mutagenic effects of cancer therapy[23]. However, recent studies show that tMN-initiating mutations can predate cancer therapy[19], consistent with CH[24]. Here, we sought to characterize the relationships between CH

A full list of author affiliations appears at the end of the paper.

and environmental exposures and determine how cancer therapy shapes patterns of selection that contribute towards progression to overt leukemia.

Molecular characteristics and clinical determinants of CH

Utilizing prospective targeted sequencing data (Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT)) from 24,146 patients with cancer representing a wide range of primary tumor types ($n=56$) and ages (Extended Data Table 1), we established a stringent variant calling and filtration workflow to detect CH variants in blood, with a minimum variant allele fraction (VAF) of 2% (Methods and Supplementary Note). We identified 11,076 unique CH mutations in 7,216 individuals, representing 30% of patients in our cohort. The median VAF of CH mutations was 5.0% (range, 2–78%). Among individuals with CH, 69% ($n=4{,}952$) had one mutation and 31% (2,264) had two or more. The spectrum of CH mutations followed expected patterns of positive selection for truncating variants and missense mutations in tumor suppressors and oncogenes, respectively (Supplementary Fig. 1). As the design of our panel limits interrogation to bona fide cancer genes, we annotated each mutation on the basis of its putative role in cancer pathogenesis using OncoKB[25] and recurrence in an in-house dataset of myeloid neoplasms[26–28] (Methods). Over half of the CH mutations that we detected were classified as putative cancer-driver mutations (CH-PD, 52%, $n=5{,}810$). Almost all CH-PD variants (91%, $n=5{,}301$) were recurrent mutations in myeloid neoplasms (CH-myeloid-PD) (Supplementary Fig. 2).

Overall, mutations in myeloid driver genes (median = 0.047) and CH-PD (0.050) showed higher VAFs than nonmyeloid (0.038) and non-PD (0.038) mutations, respectively (Supplementary Fig. 3a,b and Extended Data Table 2). Similarly, hotspot mutations at R882 within *DNMT3A* had higher VAFs compared with nonhotspot mutations, even after accounting for total number of mutations (Supplementary Fig. 4). The VAFs of mutations within individuals who harbored multiple mutations were higher compared with individuals with one mutation (Extended Data Table 2 and Supplementary Fig. 3c). Consistent with earlier literature[13,14,24], CH mutations were most frequently identified in *DNMT3A*, *TET2* and *ASXL1*. Overall, 48% of CH mutations identified were in myeloid driver genes, while only 20% of genes on the MSK-IMPACT panel are myeloid driver genes. The strong enrichment of myeloid variants highlights the strength of the fitness advantage imparted on hematopoietic stem and progenitor cells (HSPCs) by mutations in genes implicated in myeloid pathogenesis as compared with bona fide oncogenic mutations in other cancer-driver genes (Supplementary Fig. 2).

To assess the role of cancer therapy alongside other factors in driving selection of CH clones, we extracted and curated detailed clinical data for 10,138 patients who had received all of their cancer care at Memorial Sloan Kettering (MSK) (Supplementary Note). These patients' demographic characteristics and solid tumor primary site did not differ from those who received treatment outside of MSK or whose treatment information was unavailable ($n=14{,}008$) (Supplementary Table 1). As previously reported[24], older age strongly correlated with the presence of CH clones in patients with cancer (odds ratio (OR) = 1.9, $P<10^{-6}$) (Extended Data Table 3). CH was less common in patients of Asian ancestry relative to white ancestry (OR = 0.7, $P=1\times10^{-3}$) (Extended Data Table 3), consistent with recent reports[29].

Overall, a total of 5,978 patients (59%) were exposed to cancer therapy (including cytotoxic therapy, radiation therapy, targeted therapy and immunotherapy) before blood draw (Extended Data Fig. 1), whereas 4,160 (41%) were treatment-naive. Patients who had received previous cancer treatment were more likely to have CH compared with treatment-naive patients at the time of testing (OR = 1.3, $P=1\times10^{-6}$). The same was true for current and former smokers (OR = 1.1, $P=5\times10^{-3}$), and effect sizes were similar between current ($n=729$, OR = 1.2, $P=0.10$) and former smokers ($n=4{,}260$, OR = 1.1, $P=8\times10^{-3}$). The number of CH mutations in each patient was positively associated with cancer therapy and smoking, and clone size was also positively associated with smoking (Extended Data Tables 2 and 4). The association among age, therapy and CH was stronger for CH-PD compared with mutations not known to be putative cancer drivers (Extended Data Table 2). All subsequent analyses were limited to CH-PD.

The odds of having CH among patients with cancer differed by primary tumor type even after adjustment for age (Extended Data Fig. 2). The overall mutational spectrum of CH was similar across cancer types, with the exception of DNA damage response (DDR) gene mutations being more frequent in patients with ovarian and endometrial cancers. This enrichment was most striking for mutations in *PPM1D*, which were found in 13% of patients with ovarian cancer and 7% of patients with endometrial cancer as compared with <5% in other cancer subgroups (Extended Data Fig. 3). However, among patients who received no cancer therapy before blood draw, 8% of women with ovarian cancer and 0% of women with endometrial cancer had CH in *PPM1D*, suggesting that differences in the spectrum of CH mutations across tumor type could be explained by interactions between mutations in specific genes and specific classes of cancer therapy.

Clinical parameters shape the fitness landscape of CH

We next sought to determine how specific external exposures might influence the fitness landscape of CH mutations and found that age, treatment and smoking correlated with specific molecular subtypes of CH (Fig. 1a,b and Supplementary Fig. 5). For example, mutations in the spliceosome genes *SRSF2* and *SF3B1* were less common in our cohort relative to other CH mutations, but showed the strongest association with age ($\mathrm{OR}_{SRSF2}=3.6$, Q (false discovery rate (FDR)-corrected P value) $=7\times10^{-6}$; $\mathrm{OR}_{SF3B1}=5.0$, $Q\leq10^{-6}$) (Fig. 1b,c). Overall, in tests of heterogeneity, *DNMT3A* showed significantly weaker associations with age than other mutations, including spliceosome genes (Supplementary Fig. 5). CH mutations in the DDR genes *TP53*, *PPM1D* and *CHEK2* were most strongly associated with previous exposure to cancer therapy ($\mathrm{OR}_{TP53}=2.8$, $Q=2\times10^{-4}$; $\mathrm{OR}_{PPM1D}=4.3$, $Q\leq10^{-6}$; $\mathrm{OR}_{CHEK2}=4.5$, $Q=6\times10^{-6}$; Fig. 1c). Besides differences in the frequency of DDR mutations, CH mutational features were otherwise similar between treated and untreated individuals (Supplementary Fig. 6). Mutations in *ASXL1* were significantly associated with smoking history (OR = 2.5, $Q=1\times10^{-4}$; Fig. 1c). Current smokers had a stronger association with CH in *ASXL1* (OR = 3.1, $P=1\times10^{-3}$) compared with former smokers (OR = 2.4, $P=1\times10^{-4}$) although the OR did not significantly differ ($P=0.4$). While CH was more frequent overall among patients who received cancer-specific therapy, CH defined by mutations in epigenetic modifiers (*DNMT3A*, *TET2*) or splicing regulators (*SRSF2*, *SF3B1*, *U2AF1*) was not strongly affected by exposure to therapy (Fig. 1b,c). Together, these observations provide evidence that the relative fitness of acquired mutations in HSPCs is modulated by environmental factors such as cancer treatment, smoking or the aging microenvironment in a gene-dependent manner.

Given the variety of cancer therapies, different therapeutic classes may impart distinct effects on CH. In our study, patients were exposed to 490 different agents (Supplementary Note and Supplementary Table 2). To this point, we found evidence of heterogeneity in the strength of association between class agent and CH gene mutations. For example, of all treatment modalities, external beam radiation therapy (OR = 1.4, $P<10^{-6}$), cytotoxic chemotherapy (OR = 1.2, $P=2\times10^{-3}$) and radionuclide therapy (OR = 1.6, $P=0.01$) were most strongly associated with CH-PD (global test of heterogeneity, $P_{\mathrm{het}}=0.03$). With respect to subclasses of cytotoxic therapy, CH-PD was most strongly associated with previous

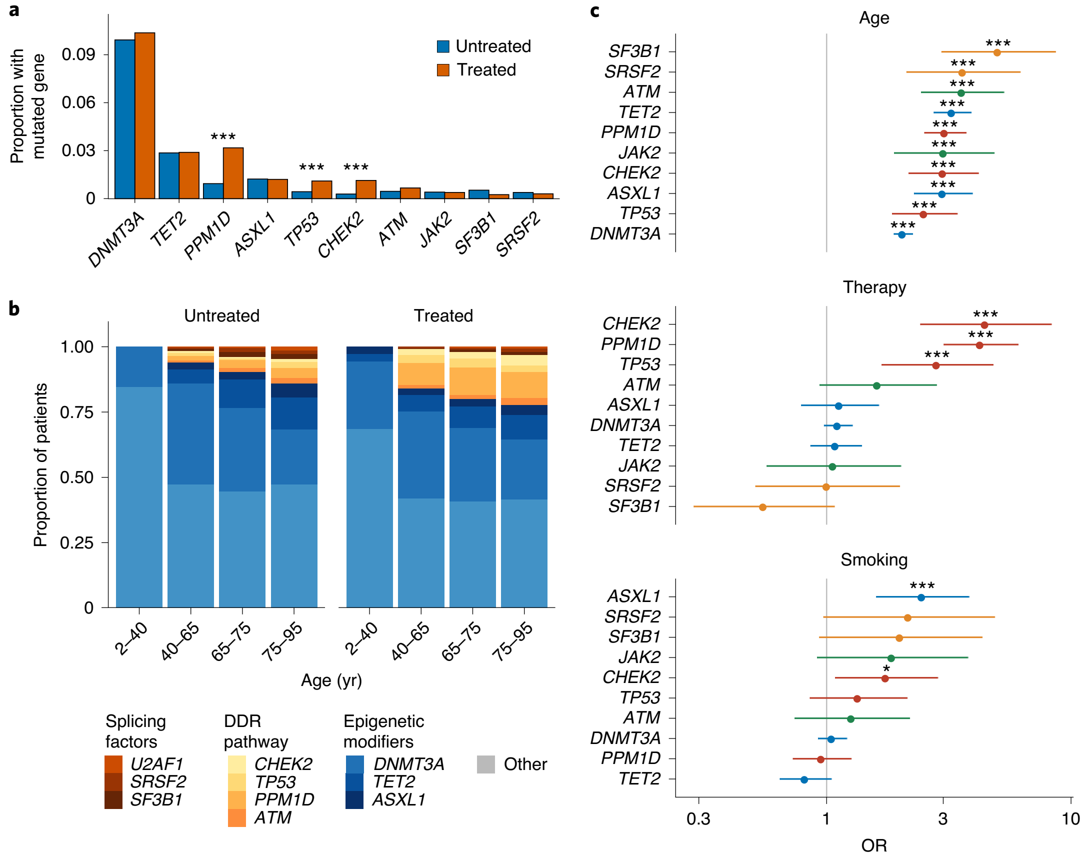


Fig. 1 | Specific molecular subtypes of CH-PD correlate with age, previous therapy exposure and smoking history. a, Proportion of patients with CH-PD mutations in specific genes among treated and untreated patients. Multivariable logistic regression was used to test whether the odds of having a specific gene mutated differed between treated ($n = 5{,}978$) and untreated ($n = 4{,}160$) patients after adjustment for age, sex, smoking and ancestry. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$. **b**, Among patients with CH-PD, the proportion with mutations in specific genes, by age group and treatment status. **c**, OR with 95% confidence interval for CH-PD mutation in the ten most commonly mutated genes, with top, increasing age ($n = 10{,}138$); middle, for patients previously exposed to cancer therapy ($n = 5{,}978$) compared with those with no exposure ($n = 4{,}160$); bottom, for current/former smokers ($n = 4{,}989$) compared with nonsmokers ($n = 5{,}145$), in multivariable logistic regression models adjusted for therapy, smoking, ancestry, age, sex and time from diagnosis to blood draw. *Q value (FDR-corrected P value) < 0.05, **$Q < 0.01$, ***$Q < 0.001$. Age is expressed as the mean centered values.

exposure to topoisomerase II inhibitors (OR = 1.3, $P = 0.01$) and platinum agents (OR = 1.2, $P = 0.02$), and, of the platinum agents, carboplatin (OR = 1.4, $P = 0.001$) was associated with CH, unlike cisplatin (OR = 1.1, $P = 0.10$) and oxaliplatin (OR = 0.98, $P = 0.88$) (Fig. 2a). Targeted therapies and immunotherapeutic agent exposure were not significantly associated with CH (Fig. 2a).

Associations with therapy exposure also varied by gene. Mutations in *PPM1D* were most strongly associated with previous exposure to platinum (OR = 3.2, $Q < 10^{-6}$) or radionuclide therapy (OR = 6.2, $Q = 7 \times 10^{-6}$) and also showed associations with topoisomerase II inhibitors (OR = 2.0, $Q = 0.002$), taxanes (OR = 1.8, $Q = 0.003$), topoisomerase I inhibitors (OR = 1.7, $Q = 0.002$) and external beam radiation therapy (OR = 1.8, $Q = 0.04$) (Fig. 2b). Mutations in *TP53* were associated with previous platinum (OR = 2.1, $Q = 0.03$), radiation therapy (OR = 1.8, $Q = 0.04$) and taxane (OR = 1.9, $Q = 0.05$) exposure, whereas *CHEK2* was associated with platinum (OR = 2.4, $Q = 0.02$) and topoisomerase II inhibitors (OR = 2.2, $Q = 0.02$) (Fig. 2b). The strength of the association between DDR CH and cytotoxic therapy differed by cytotoxic therapy subclass ($P = 4 \times 10^{-6}$) and platinum subclass ($P = 0.03$).

To evaluate whether treatment dose modulated these relationships, we calculated each patient's relative cumulative exposure to specific therapy classes (Supplementary Note and Supplementary Fig. 7). Increasing exposure to platinum chemotherapy was associated with CH-PD (P-trend = 0.04). Among platinum agents, CH-PD was associated with higher cumulative doses of carboplatin (P-trend = 3×10^{-5}) and cisplatin (P-trend = 0.04) (Fig. 2c). Evidence of dose–response further supports a possible causal relationship between the associated exposures and CH.

Clonal dynamics of CH in response to cancer therapy

Our retrospective analysis suggests that exposure to cancer therapy results in a higher likelihood of CH, particularly in patients with mutations in DDR genes, following exposure to specific therapies. To definitively characterize how treatment affects mutational presentation and clonal dominance of CH across time, we collected sequential blood samples from 525 patients with solid tumors (median sampling interval time = 23 months, range: 6–53 months), of whom 61% received cytotoxic therapy or external beam radiation therapy and 39% received either targeted therapy or immunotherapy or were untreated (Methods and Supplementary Fig. 8). None of these patients developed secondary hematologic malignancies during follow-up. Of these patients, 389 (74%) had CH, defined as a mutation present at a VAF of ≥2%, at the time of first sampling. The majority of CH mutations were present at both time points ($n = 590$ of 620, 95%), allowing us to examine how clones evolved in the presence or absence of therapy and whether the clone-defining mutations influenced these trajectories.

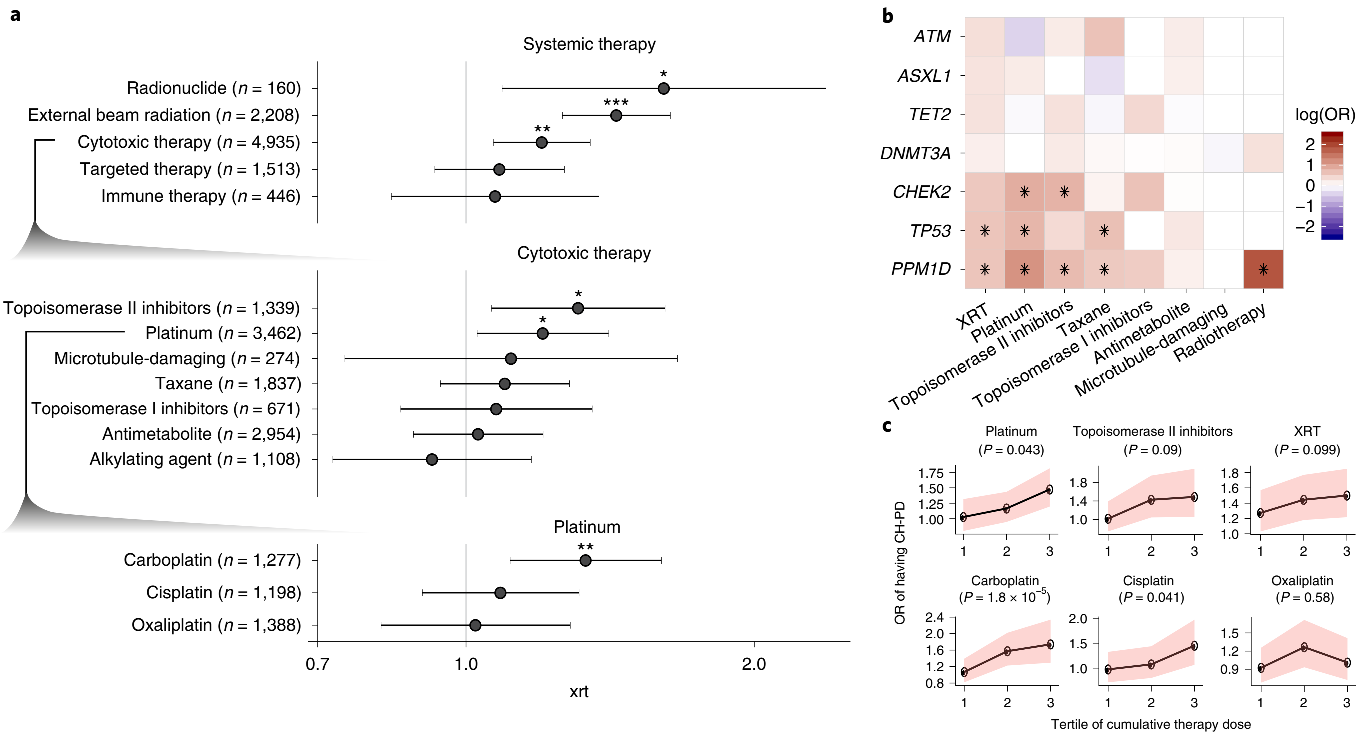


Fig. 2 | Association between CH-PD and previous exposure to cancer therapy. **a**, ORs and 95% confidence intervals for CH-PD and specific classes of cancer therapy in multivariable logistic regression adjusted for each other, smoking, ancestry, sex and time from diagnosis to blood draw. Top, OR for broad classes of cancer therapy; middle, OR between CH-PD and previous exposure to subclasses of cytotoxic therapy; bottom, OR between CH-PD and exposure to specific platinum-based drugs. **b**, OR between previous receipt of cancer therapy and CH-PD stratified by tertile of cumulative exposure for the agent. Multivariable logistic regression was used adjusted as in **a** but with cumulative weight-adjusted dose of systemic therapy classes and cumulative radiation dose (as expressed in equivalent dose in 2-Gy fractions, EQD_2). The *P*-trend was calculated to test for association between CH and increasing tertiles of cumulative cancer therapy exposure among those who received the therapy in the multivariable model. Shaded bands indicate 95% confidence intervals. XRT, external beam radiation. **c**, Heatmap showing the log(OR) between CH-PD in specific genes and previous exposure to the major classes of cytotoxic therapy and radiation therapy in logistic regression models adjusted for therapy subclass, smoking, ancestry, sex and time from diagnosis to blood draw. *$Q < 0.05$, **$Q < 0.01$, ***$Q < 0.001$.

We found evidence of both positive and negative changes in clone size across treatment modalities (Fig. 3a). Among mutations detected at both time points, the majority (62% ($n = 367$) of CH mutations remained stable, 28% ($n = 164$) had evidence of growth and 10% ($n = 59$) decreased in clonal size. Among patients receiving external beam radiation therapy or cytotoxic therapy, growth was most pronounced for CH with mutations in DDR genes *TP53*, *CHEK2* and *PPM1D* (Fig. 3b,c). Similar to our retrospective series, increasing cumulative exposure to these therapies resulted in faster clone growth in patients whose CH was defined by DDR mutations (Fig. 3d). We did not see evidence of a significant association between change in VAF and time from end of cytotoxic therapy to the second blood sampling. Future studies with sequential sampling before, during and after therapy will be needed to characterize the kinetics of CH. Patients with multiple mutations exhibited faster CH growth[20] as compared with those with one mutation ($P = 0.03$) irrespective of mutation type and treatment status (Supplementary Fig. 9). This likely reflects the greater competitive advantage of a subset of clones harboring multiple mutations, although this cannot be determined with certainty in the absence of single-cell sequencing. The proportion of patients with newly detected mutations among those who received interval cytotoxic/radiation therapy (4%, $n = 13$) was nonsignificantly higher as compared with those who did not (1%, $n = 2$, $P = 0.06$) (Supplementary Fig. 10). Thus, in addition to therapy selecting for CH, therapy may have mutagenic effects on HSPCs.

Many parameters likely influence evolutionary trajectories of emerging CH clones. To study competing clonal dynamics, we identified 34 patients in our prospective serial sampling series with one mutation in a DDR gene and one in a non-DDR gene (Fig. 3e). The presence of these distinct classes of gene mutations within the same patient controls for any confounding parameters. In patients receiving interval cytotoxic therapy or radiation therapy, CH clones with DDR mutations grew faster compared with clones with other CH mutations in the same patient. However, the reverse was true in untreated patients: clones with mutations in non-DDR CH genes (for example, *DNMT3A*) outcompeted clones with DDR mutations (Fig. 3e). In summary, our serial sampling data provide direct evidence in patients that cancer therapy selects for clones with mutations in the DDR genes *TP53*, *PPM1D* and *CHEK2* and that these clones have lower competitive fitness relative to non-DDR gene mutations in the absence of cytotoxic or radiation therapy.

Genetic and clonal evolution to tMN

Recent studies have shown that tMN-initiating mutations can predate cancer therapy[19], challenging the traditional hypothesis that tMN develops from the mutagenic effects of cancer therapy[23] and suggesting a relationship with CH. We hypothesized that tMN development is at least in part mediated by therapeutic selection of mutant clones in a gene-dependent manner.

To study the molecular events defining progression of CH to tMN, we analyzed 35 cases for which paired samples were available at the time of molecular profiling for primary cancer and at the time of leukemic transformation for tMN (median inter-sampling time of 24 months, range: 5–90 months) (Supplementary Table 3).

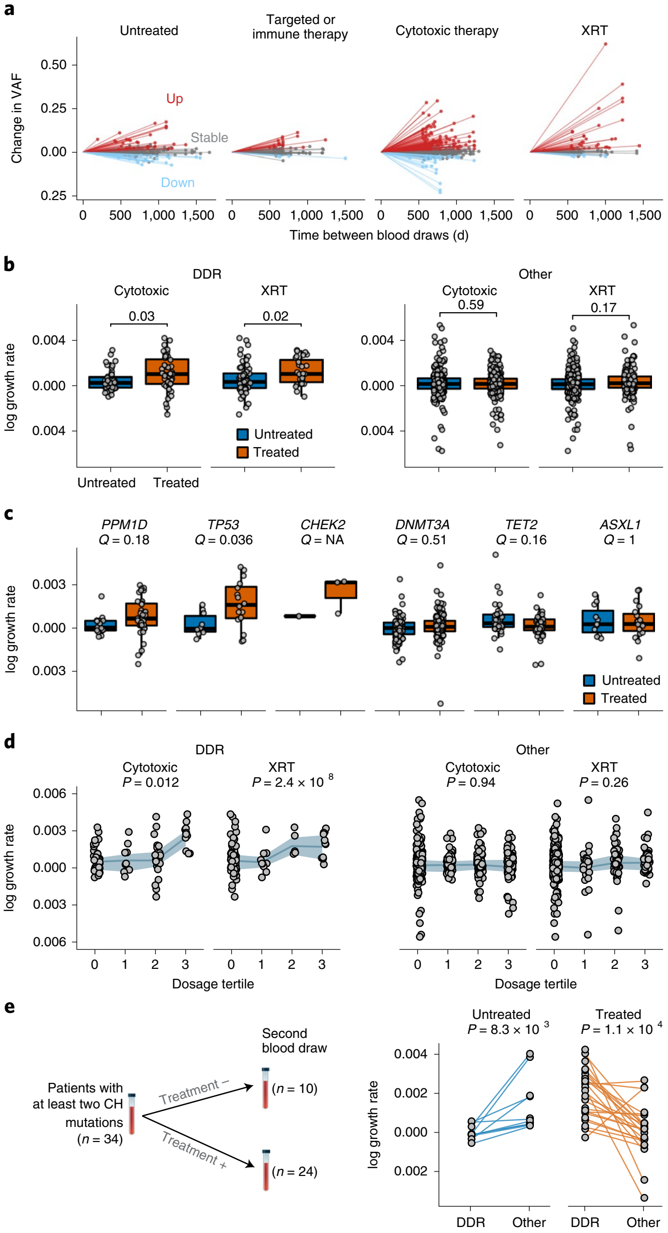


Fig. 3 | Clonal evolution of CH mutations under the selective pressure of cancer therapy. a, Change in VAF for CH mutations from initial to follow-up sequencing for patients stratified by type of therapy received during the follow-up period. **b**, Change in growth rate for DDR and non-DDR CH mutations among those who received XRT ($n = 167$) or cytotoxic therapy ($n = 285$) during the follow-up period. Shown are the P values generated from t-tests comparing the growth rate of CH mutations among patients exposed to either of these therapies compared with untreated patients. **c**, Change in growth rate for specific CH mutations stratified by whether patients received cytotoxic or radiation therapy ($n = 268$) or no therapy ($n = 177$) during the follow-up period. Shown are the FDR-corrected P values (Q values) from a t-test comparing the growth rates of mutations in treated and untreated patients. **d**, Change in growth rate for DDR and non-DDR CH mutations stratified by tertile of cumulative exposure to cytotoxic therapy and XRT. Shown are the P values for a trend test for increasing growth rate of CH with increasing tertile of therapy exposure using generalized linear regression adjusted for age, sex and smoking. Shaded bands indicate interquartile ranges. **e**, Intra-patient competition between DDR and non-DDR CH mutations. Connecting lines show the difference in growth rate between DDR versus other genes in patients who received XRT or cytotoxic therapy versus those who did not receive such therapy during the follow-up period. A paired t-test was used to test for significance in the difference between growth rates of DDR and non-DDR CH mutations within individuals. All P values are two-sided.

We called mutations present at a VAF of ≥2% in at least one time point. We detected disease-defining events at the time of tMN in 34 patients. Strikingly at least one of these mutations was present at the time of CH (with at least one supporting read) in 19 patients (59%), with 13 (41%) harboring two or more. In all of these cases, the CH mutation was present at the time of tMN diagnosis (Extended Data Fig. 4). However, these mutations are unlikely sufficient for leukemic transformation. In 91% of cases, transformation was associated with acquisition of additional somatic mutations, including chromosomal aneuploidies or mutations in genes (for example, *FLT3*, *KRAS*, *NRAS*) known to drive late progression to myeloid disease[27,30–32] (Supplementary Fig. 11).

Nearly half ($n = 14$, 40%) of the tMN patients had mutations in *TP53*. Overall, 10 of 14 *TP53* mutations were detectable at the time

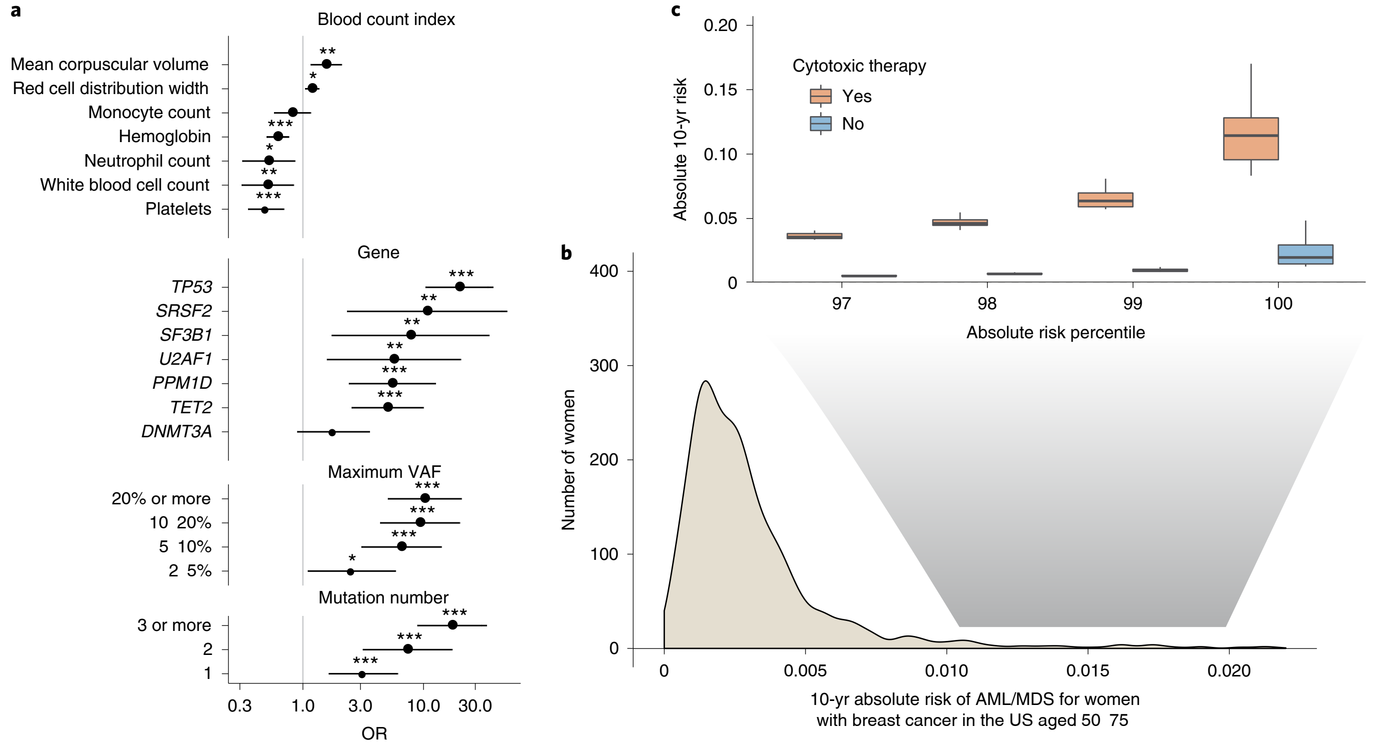


Fig. 4 | Risk of AML or MDS by clinical and CH-PD mutational characteristics in patients with solid tumors. a, HRs and 95% confidence intervals from Cox regression for blood count indexes, and CH-PD mutational characteristics for tMNs (AML or MDS, $n = 75$). All models were adjusted for age and sex and stratified by study center. Blood counts are expressed as the mean centered score (the OR is per 1 s.d. of the blood count). $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. **b**, Projected distribution of absolute 10-yr risk of AML or MDS after a breast cancer diagnosis for women in the United States aged 50–75 at presentation based on our synthetic model. **c**, Comparison of distribution of absolute 10-yr risk of AML or MDS among women at the top percentiles of risk between those who go on to receive adjuvant cytotoxic chemotherapy and those who receive surgery only. $n = 9{,}437$.

of CH testing. Of these, four cases had a concomitant *TP53* mutation and another non-DDR mutation at the time of CH. In agreement with prospective serial sequencing, in the presence of therapy, the *TP53* clone had consistently attained dominance by the time of tMN (Extended Data Fig. 4). At transformation, in 12 of 13 (92%) cases with available karyotype, *TP53* mutations co-occurred with isolated chromosomal aneuploidies or complex karyotype. This provides a direct mechanistic link, whereby cells carrying mutations in *TP53* are positively selected when exposed to oncologic therapy and attain clonal dominance with further genetic diversification, such as the acquisition of chromosomal aneuploidies.

Clinical implications of CH in patients with cancer

Based on the direct evidence that CH mutations led to tMN transformation in our paired sample data, we sought to identify risk factors associated with tMN. By combining patient data from our cohort with detailed clinical histories and three previously published studies[33–35], we created a cohort of 9,437 patients with cancer exposed to cancer therapy, of whom 75 developed tMN (Supplementary Table 2 and Supplementary Note). Cause-specific Cox proportional hazards analysis (Supplementary Table 2) showed that CH present at a VAF of >2% was associated with an increased tMN risk (hazard ratio (HR) = 6.9, $P < 10^{-6}$), and increased with the total number of mutations and clone size (Fig. 4a). The strongest associations were observed for mutations in *TP53*, further validating the relevance of *TP53* in tMN, and for mutations in spliceosome genes (*SRSF2*, *U2AF1* and *SF3B1*). Future studies using error-corrected sequencing methods will clarify the relationship between CH and tMN at VAFs of <2%. Comparison of HRs for tMN and AML risk showed similar effect sizes (Supplementary Fig. 12) in our cohort as in recent studies of healthy individuals[20,36]. These data suggest that the relative risk of myeloid neoplasms associated with CH and related parameters (gene, VAF and mutation number) is similar between healthy individuals and patients with cancer.

We next sought to evaluate how CH, in combination with clinical parameters such as age and peripheral blood counts, might help stratify tMN risk for patients with cancer. For example, in patients with solid tumors undergoing surgical resection, adjuvant cancer therapy can improve overall survival by reducing cancer recurrence. However, in some situations, the absolute survival benefit of adjuvant therapy is modest and is countered, at least in part, by the risk for subsequent tMN, which is almost universally fatal, with a 5-yr survival of 10% (ref. [37]). In the absence of prospective clinical studies, we performed an exploratory analysis using a synthetic model to quantify the absolute risk of AML/MDS following a breast cancer diagnosis. Using previously established methodology[38,39], we combined estimates of HR parameters obtained from our multivariable analysis with the distribution of CH mutational features and blood count parameters from untreated patients at MSK and external sources to model the 10-yr cumulative absolute AML/MDS risk distribution for women with breast cancer aged 50–75 yr in the United States. This risk model assumes a multiplicative effect of CH mutational features and cancer therapy on risk of tMN, based on the similarity between risk estimates for CH mutational features in AML that develops in individuals never exposed to therapy and tMN (Supplementary Fig. 12). We determined how the risk distribution would change with receipt of adjuvant therapy by shifting the population between receiving and not receiving adjuvant chemotherapy.

In our model, the majority (96%) of patients with breast cancer have a low 10-yr absolute risk (<1%) for myeloid neoplasm

(Fig. 4b), and, for these patients, deferment of adjuvant chemotherapy would not affect their absolute myeloid neoplasm risk (Fig. 4c). However, for women at the highest risk of myeloid neoplasm based on CH and blood count parameters in our synthetic model (top 1%), adjuvant chemotherapy increased the absolute risk of myeloid neoplasm by approximately 9%. This would exceed the predicted absolute benefit in overall survival of chemotherapy in many women with early-stage breast cancer[40]. While not appropriate for clinical implementation, our findings may inform the design of and provide a rationale for future studies to formally estimate the benefits of risk-adapted treatment decisions in patients with cancer with CH.

Discussion

Longitudinal studies of CH present a unique opportunity to study the patterns of early mutagenesis and the dynamics of clonal selection in the progression towards malignant transformation. Here, by combining epidemiologic and genetic approaches, we provide insights into the mechanisms that drive the transition of a normal HSPC to a cell with a considerably stronger proliferation advantage, and study how the ensuing trajectories are shaped by host and environmental exposures including age, ancestry, smoking and cancer therapy. We provide evidence that the fate of CH mutations is dictated by a complex interplay between the inherent fitness advantage of the mutation(s) in HSPCs and parameters that preferentially select for specific mutations, that is, aging for spliceosome mutations, smoking for mutations in *ASXL1* and cancer therapy for specific genes involved in DDR (Extended Data Fig. 5). These relationships provide insight into disease biology and may inform early detection and prevention strategies in cancer. We refine the relevance of CH as a predictor and precursor of tMN in patients with cancer and show that CH mutations detected before tMN diagnosis were consistently part of the dominant clone at transformation. We demonstrate that cancer therapy directly favors growth of clones with mutations in genes such as *TP53*, which is associated with chemo-resistant disease and is strongly enriched in tMN. This provides a direct mechanistic link among genetic subtypes of CH, receipt of subsequent cancer therapy and how these modulate the transition from CH to attainment of clonal dominance and, for a subset of cases, development of tMN.

Previous murine and in vitro modeling studies have provided evidence supporting an association between cancer therapy and increased fitness of DDR clones in CH. However, these observations have not been verified in humans, nor do they define how therapy enables the transition of CH to myeloid neoplasm. Here we show that clones with DDR mutations are positively selected in the presence of cancer therapy but not in its absence. We also show that beyond clonal dominance the transition to tMN is most parsimoniously associated with the acquisition of further genetic lesions. Our detailed treatment information including agent class, dose and mechanism of action allowed us to refine the specificity and strength of the association between cancer therapy and CH and characterize distinct gene–treatment effects. We show that radiation therapy and cytotoxic therapy are significantly associated with CH, with regimens containing platinum and topoisomerase II inhibitors most strongly correlating with CH in specific DDR pathway genes including *TP53*, *PPM1D* and *CHEK2*. Serial sampling before and after therapy provided clear, definitive evidence that therapy induces gene-specific clonal expansion, whereby clones with mutations in DDR genes outcompete other clones in the setting of cancer therapy, but not in its absence. Last, the dose–response relationships observed in both our cross-sectional arm and longitudinal study further support a causal relationship between platinum and CH and the cumulative effect of therapy on selection.

The specificity of the associations at a genetic and exposure level (that is, therapeutic subclasses and agents such as carboplatin) sets a framework for future correlative and mechanistic studies in early oncogenesis for blood disorders. The specific mechanisms and pathways through which chemotherapeutic agents induce hematopoietic stem cell injury may be agent specific[41,42]. Further work will be needed to elucidate the mechanisms responsible for the differential fitness effects of cancer therapy and other environmental exposures such as smoking on CH both during and after exposure, and how these relate to tMN risk. Beyond the most frequent cancer genes surveyed here, comprehensive genome studies such as deep whole-exome or whole-genome analyses in cohorts linked to detailed registries of environmental exposures are warranted to uncover the full repertoire of selection in CH.

We find overlap in the types of cancer therapy associated with selection of DDR CH and those linked to tMN risk (carboplatin, topoisomerase II inhibitors and radiation). Selection of *TP53* is only one mechanism driving tMN and may be distinct from the processes driving initiation and selection for other tMN-associated alterations including chromosomal aneuploidies and genomic rearrangement (that is, *MLL* fusion genes). Our work adds to early evidence[43,44] that external stressors are critical in shaping gene-dependent selection of clonal mosaicism. Characterization of the complex interplay among genotype, fitness challenges and environmental factors will be key to understanding age-associated clonal mosaicism and the associated exposures that result in malignant transformation. These insights would provide the premise for risk stratification and prevention strategies.

Our observations provide a rationale for clinical therapeutic intervention, including the development of therapies aimed to target high-risk CH clones and modulation of the use of adjuvant systemic cancer therapy in patients at highest risk of subsequent myeloid neoplasm. The latter could entail deferring adjuvant cytotoxic therapy or substituting therapies shown to promote high-risk CH with alternative agents when clinically appropriate. We showcase this with a prototype synthetic model; however, development and validation of risk prediction models for specific clinical scenarios are needed before implementation. The realization of precision medicine is reliant upon the development of evidence-based guidelines that consider molecular biomarkers alongside standard clinical criteria to inform clinical care. The decreasing cost of prospective clinical sequencing assays and the high frequency of CH in patients with cancer suggest that screening for CH before initiation of cancer therapy may be feasible, and may enable molecularly based early detection and interception.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-00710-0.

Received: 5 April 2020; Accepted: 2 September 2020;
Published online: 26 October 2020

References

1. Armitage, P. & Doll, R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer* **8**, 1–12 (1954).
2. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
3. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
4. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
5. Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nat. Rev. Genet.* **13**, 795–806 (2012).
6. Ding, L. et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).

7. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
8. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
9. Martincorena, I., Jones, P. H. & Campbell, P. J. Constrained positive selection on cancer mutations in normal skin. *Proc. Natl Acad. Sci. USA* **113**, E1128–E1129 (2016).
10. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
11. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
12. Suda, K. et al. Clonal expansion and diversification of cancer-associated mutations in endometriosis and normal endometrium. *Cell Rep.* **24**, 1777–1789 (2018).
13. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
14. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
15. McKerrell, T. et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).
16. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
17. Fernandez-Antoran, D. et al. Outcompeting p53-mutant cells in the normal esophagus by redox manipulation. *Cell Stem Cell* **25**, e6 (2019).
18. Hsu, J. I. et al. PPM1D mutations drive clonal hematopoiesis in response to cytotoxic chemotherapy. *Cell Stem Cell* **23**, 700–713.e6 (2018).
19. Wong, T. N. et al. Role of TP53 mutations in the origin and evolution of therapy-related acute myeloid leukaemia. *Nature* **518**, 552–555 (2015).
20. Abelson, S. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
21. Desai, P. et al. Somatic mutations predict acute myeloid leukemia years before diagnosis. *Nat. Med.* **24**, 1015–1023 (2018).
22. Morton, L. M. et al. Evolving risk of therapy-related acute myeloid leukemia following cancer chemotherapy among adults in the United States, 1975–2008. *Blood* **121**, 2996–3004 (2013).
23. McNerney, M. E., Godley, L. A. & Le Beau, M. M. Therapy-related myeloid neoplasms: when genetics and environment collide. *Nat. Rev. Cancer* **17**, 513–527 (2017).
24. Coombs, C. C. et al. Therapy-related clonal hematopoiesis in patients with non-hematologic cancers is common and associated with adverse clinical outcomes. *Cell Stem Cell* **21**, 374–382.e4 (2017).
25. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* **2017**, PO.17.00011 (2017).
26. Papaemmanuil, E. et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N. Engl. J. Med.* **365**, 1384–1395 (2011).
27. Papaemmanuil, E. et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
28. Grinfeld, J. et al. Classification and personalized prognosis in myeloproliferative neoplasms. *N. Engl. J. Med.* **379**, 1416–1430 (2018).
29. Bick, A. G. et al. Inherited causes of clonal hematopoiesis of indeterminate potential in TOPMed whole genomes. Preprint at *bioRxiv* https://doi.org/10.1101/782748 (2019).
30. Lindsley, R. C. et al. Acute myeloid leukemia ontogeny is defined by distinct somatic mutations. *Blood* **125**, 1367–1376 (2015).
31. Welch, J. S. et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
32. Cancer Genome Atlas Research Network et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
33. Gillis, N. K. et al. Clonal haemopoiesis and therapy-related myeloid malignancies in elderly patients: a proof-of-concept, case-control study. *Lancet Oncol.* **18**, 112–121 (2017).
34. Takahashi, K. Germline polymorphisms and the risk of therapy-related myeloid neoplasms. *Best Pract. Res. Clin. Haematol.* **32**, 24–30 (2019).
35. Gibson, C. J. et al. Clonal hematopoiesis associated with adverse outcomes after autologous stem-cell transplantation for lymphoma. *J. Clin. Oncol.* **35**, 1598–1605 (2017).
36. Young, A. L., Tong, R. S., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis and risk of acute myeloid leukemia. *Haematologica* **104**, 2410–2417 (2019).
37. Fianchi, L. et al. Characteristics and outcome of therapy-related myeloid neoplasms: report from the Italian network on secondary leukemias. *Am. J. Hematol.* **90**, E80–E85 (2015).
38. Choudhury, P. P. et al. iCARE: an R package to build, validate and apply absolute risk models. *PLoS ONE* 15, e0228198 (2020).
39. Maas, P. et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncol.* **2**, 1295–1302 (2016).
40. Candido Dos Reis, F. J. et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Res.* **19**, 58 (2017).
41. Meng, A., Wang, Y., Van Zant, G. & Zhou, D. Ionizing radiation and busulfan induce premature senescence in murine bone marrow hematopoietic cells. *Cancer Res.* **63**, 5414–5419 (2003).
42. Hu, W. et al. Mechanistic investigation of bone marrow suppression associated with palbociclib and its differentiation from cytotoxic chemotherapies. *Clin. Cancer Res.* **22**, 2000–2008 (2016).
43. Meisel, M. et al. Microbial signals drive pre-leukaemic myeloproliferation in a Tet2-deficient host. *Nature* **557**, 580–584 (2018).
44. Zhu, M. et al. Somatic mutations increase hepatic clonal fitness and regeneration in chronic liver disease. *Cell* **177**, 608–621.e12 (2019).

[1]Department of Medicine, Leukemia Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [3]Computational Oncology Service, Department of Epidemiology & Biostatistics, Center for Computational Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [4]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [5]Department of Epidemiology & Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [6]Department of Information Systems, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [7]Center for Hematologic Malignancies, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [8]Department of Pathology, Hematopathology Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [9]Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [10]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [11]Department of Medicine, Clinical Genetics Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [12]Department of Health Informatics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [13]Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [14]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [15]Department of Oncology, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. [16]Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Cambridge, UK. [17]Department of Medicine, Hematology Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [18]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [19]Department of Medicine, Endocrinology Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [20]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [21]Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA. [22]Department of Pediatrics, Washington University School of Medicine, St. Louis, MO, USA. [23]Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [24]Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL, USA. [25]Department of Malignant Hematology, Moffitt Cancer Center, Tampa, FL, USA. [26]Institute of Pathology, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. [27]Weill Cornell Medical College, New York, NY, USA. [28]Research & Development, AstraZeneca, Milton, Cambridge, UK. [29]Center for Strategy & Innovation, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [30]Marie-Josée and Henry R. Kravis Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [31]Department of Biostatistics, Bloomberg School of Public Health Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA. [32]Program in Precision Interception and Prevention, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [33]Department of Medicine, Solid Tumor Division, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [34]These authors contributed equally: Ryan N. Ptashkin, Teng Gao. [35]These authors jointly supervised this work: Ahmet Zehir, Elli Papaemmanuil. ✉e-mail: zehira@mskcc.org; papaemme@mskcc.org

Methods

MSK-IMPACT cohort. The study population included patients with nonhematologic cancers at Memorial Sloan Kettering Cancer Center (MSKCC) who underwent matched tumor and blood sequencing using the MSK-IMPACT panel on an institutional prospective tumor sequencing protocol (ClinicalTrials.gov number, NCT01775072) before 1 July 2018; all patients enrolled on this protocol provided informed consent. This study was approved by the MSKCC Institutional Review Board (IRB). A subset of patients who underwent tumor-genomic profiling as standard of care did not directly consent, in which case an IRB waiver was obtained to allow for inclusion into this study.

We extracted data on ancestry, smoking, date of birth and cancer history through the MSK cancer registry. Patients who had a hematologic malignancy diagnosed within 3 yr before blood collection for MSK-IMPACT testing or who had an active hematologic malignancy at the time of blood draw were excluded. Patients who were diagnosed with a hematologic malignancy less than 3 months following MSK-IMPACT were considered to have an active hematologic malignancy at the time of MSK-IMPACT and were also excluded. When unavailable through the cancer registry, we extracted data on ancestry and smoking through structured fields in clinician medical notes, if available. Patients for whom age was not available were excluded. Blood indices were taken from clinical laboratory results closest to the date of blood collection for MSK-IMPACT, within 1 yr before or after blood collection (median 0 d). The 8,810 individuals included in the previous MSK-IMPACT publication studying CH[24] are included in the current manuscript. A major difference between the two studies, in addition to an expanded sample size, is the comprehensiveness of the clinical data, including therapeutic exposure data, that were obtained as detailed in the Supplementary Note.

Serial sampling cohort. To study the growth rate of CH mutations over time, we collected additional blood samples on patients sequenced using MSK-IMPACT for repeat CH mutation testing. These came from three sources: first, from 372 patients with CH in whom we obtained a second blood sample at least 18 months after initial MSK-IMPACT blood collection; second, from 21 samples from patients with CH on MSK-IMPACT who had a blood sample banked at least 12 months before MSK-IMPACT testing; and third, from 132 samples that were taken for repeat MSK-IMPACT testing for clinical purposes at least 6 months after the first MSK-IMPACT testing irrespective of CH status (Supplementary Fig. 8). For all patients who had sequential sampling data, we manually reviewed their medical records to capture receipt of cancer therapy at outside institutions during the follow-up period. If patients received therapy outside of MSK during the follow-up period, we excluded them from analyses of dose–response relationships since cumulative dose of therapy could not be consistently collected from outside records. This study was approved by the MSKCC IRB.

Targeted capture-based sequencing. Participants had a tumor and blood sample (as a matched normal) sequenced using MSK-IMPACT, a Food and Drug Administration-authorized hybridization capture-based next-generation sequencing assay encompassing all protein-coding exons from the canonical transcript of 341, 410 or 468 cancer-associated genes (Supplementary Table 4). MSK-IMPACT is validated and approved for clinical use by the New York State Department of Health Clinical Laboratory Evaluation Program and is used to sequence patients with cancer at MSK. DNA was extracted from formalin-fixed paraffin-embedded tumor tissue and patient-matched blood samples and sheared, and DNA fragments were captured using custom probes[45]. MSK-IMPACT contains most of the commonly reported CH genes with few exceptions. Earlier versions of the panel did not contain *PPM1D* or *SRSF2*. Additionally, three genes commonly reported to be observed in patients with malignancies, *SRCAP*, *BRCC3* and *ZNF318*, were not included, the first two belonging to the DDR pathway.

The blood samples in the serial sampling cohort that were obtained for repeat CH testing were sequenced using a comparable capture-based custom panel using 163 genes implicated in myeloid pathogenesis, which included the most commonly mutated genes in our MSK-IMPACT study, with the exception of *ATM*. The median sequencing depth was 665× (range = 111–1,987×), which was comparable to that obtained in the blood by using MSK-IMPACT. For all subsequent analyses using the serial sampling cohort, we only considered mutations that were present in both the initial and follow-up panels.

Variant calling. Pooled libraries were sequenced on an Illumina HiSeq 2500 with 2 × 100-base-pair paired-end reads. Sequencing reads were aligned to the human genome (hg19) using BWA (0.7.5a). Reads were re-aligned around indels using ABRA (0.92), followed by base quality score recalibration with the Genome Analysis Toolkit (3.3-0). Median coverage in the blood samples was 497×, and median coverage in the tumors was 790×. Variant calling for each blood sample was performed unmatched, using a pooled control sample of DNA from ten unrelated individuals as a comparator. Single-nucleotide variants were called using Mutect and VarDict. Insertions and deletions were called using Somatic Indel Detector and VarDict. Variants that were called by two callers were retained. Dinucleotide substitution variants were detected by VarDict and retained if any base overlapped a single-nucleotide variant called by Mutect. All called mutations were genotyped in the patient-matched tumor sample. Mutations were annotated with VEP (v.86) and OncoKb.

Postprocessing filters for CH calling. We applied a series of postprocessing filters to further remove false-positive variants caused by sequencing artifacts and putative germline polymorphisms. We removed variants that were found (with a VAF of >2% at least once) in a panel of sequencing data from 300 blood samples obtained from persons under 20 yr of age and without evidence of CH. We further filtered single-nucleotide deletions within a homopolymer stretch (≥3 base repetition) of the same deleted base pair, single-nucleotide substitutions completing a stretch of a ≥5-base-pair-long homopolymer (for example, GGCGG → GGGGG), in-frame deletions or insertions in a highly repetitive region (DUST[46] algorithm score of ≥5) and variants with unequal proportions of forward/reverse direction supporting reads based on a Fisher test. We performed manual review in Integrative Genomics Viewer of recurrent mutations not previously reported in public databases. We required a VAF of at least 2% and at least ten supporting reads. All genotypes were calculated using sequencing reads and bases with a quality value of at least 20. Because somatic mutations in the blood would be expected to be detected in the blood but not in other tissue compartments, we compared the VAF of mutations in the blood compared with the matched tumor. Variant calls that were present in the blood with a VAF of at least twice that in the tumor, or 1.5 times the VAF if the tumor biopsy site was a lymph node, were considered somatic. This ratio was chosen based on maximizing the sensitivity and specificity of CH calls through simulations of leukocyte contamination in the tumor (Supplementary Note and Supplementary Figs. 11 and 12). To further filter putative germline polymorphisms that passed the blood/tumor solid tissue ratio due to allelic imbalance in the tumor specimen, we removed any variant reported in any population in the gnomAD database at a frequency greater than 0.005.

Validation of calls. To test the reproducibility of our CH mutation calling, we compared the mutational calling results from 1,173 samples, where the same DNA library for a blood sample was sequenced and analyzed twice using MSK-IMPACT. We detected 91% of variants in both samples using our calling criteria, with a correlation coefficient of 0.98 for the VAF between the two calls indicating that the reproducibility of our calls was high. In ten cases with CH, we obtained a second blood sample and re-sequenced using a custom capture-based panel with unique molecular identifiers and found that this independent method confirmed all 18 of our CH calls using MSK-IMPACT.

Variant annotation. Variants were annotated according to evidence for functional relevance in cancer (putative driver or CH-PD) and for relevance to myeloid neoplasms specifically (CH-myeloid-PD). We annotated variants as oncogenic in myeloid disease (CH-myeloid-PD) if they were in a gene hypothesized to drive myeloid/hematologic malignancies (Supplementary Table 5) and if they fulfilled any of the following criteria: (1) truncating variants in *NF1*, *DNMT3A*, *TET2*, *IKZF1*, *RAD21*, *WT1*, *KMT2D*, *SH2B3*, *TP53*, *CEBPA*, *ASXL1*, *RUNX1*, *BCOR*, *KDM6A*, *STAG2*, *PHF6*, *KMT2C*, *PPM1D*, *ATM*, *ARID1A*, *ARID2*, *ASXL2*, *CHEK2*, *CREBBP*, *ETV6*, *EZH2*, *FBXW7*, *MGA*, *MPL*, *RB1*, *SETD2*, *SUZ12* or *ZRSR2* or in *CALR* exon 9; (2) translation start site mutations in *SH2B3*; (3) *TERT* promoter mutations; (4) *FLT3* internal tandem duplications; (5) in-frame indels in *CALR*, *CEBPA*, *CHEK2*, *ETV6* or *EZH2*; (6) any variant occurring in the COSMIC ‘hematopoietic and lymphoid’ category ≥10 times; and (7) any variant noted as potentially oncogenic in an in-house dataset of 7,000 individuals with myeloid neoplasm ≥5 times. We annotated variants as oncogenic (CH-PD) if they fulfilled any of the following criteria: (1) any variant noted as oncogenic or likely oncogenic in OncoKB[25]; (2) any truncating mutations (nonsense, essential splice site or frameshift indel) in known tumor suppressor genes as per the Cancer Gene Census, OncoKB or the scientific literature; (3) any variant reported as somatic at least 20 times in COSMIC[47]; and (4) any variant meeting criteria for CH-Myeloid-PD as above. All missense variants not meeting the above criteria were individually reviewed for potential oncogenicity as previously described[48].

Calculation of dN/dS ratios. We used the dNdScv (https://github.com/im3sanger/dndscv) package to quantify the ratios of the number of nonsynonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site (dN/dS) for missense and truncating mutations at the gene level as well as on the panel level. Due to the differences in the gene panel between different MSK-IMPACT panel versions, we excluded all MSK-IMPACT-341 samples and included only genes that were present on both MSK-IMPACT-410 and MSK-IMPACT-468 panels in the analysis. Finally, to generate the overall dN/dS landscape in CH, we presented only genes that reached a significance level of $Q < 0.1$ after multiple testing correction and contained more than 25 variants.

Modeling the association between CH and previous exposure to cancer therapy. We used multivariable logistic regression to evaluate for an association between CH (including gene- and variant-specific factors) and therapy, age, sex and smoking history. In addition to these variables, we adjusted for time from cancer diagnosis to blood draw for MSK-IMPACT testing because trends in preferred oncologic agents vary over time and CH is known to associate with survival.

We did not adjust for primary tumor type since we hypothesized that most of the difference in CH-PD frequencies across tumor types reflected differences in treatment regimens. Indeed, among untreated patients, a global Wald test for differences in CH-PD prevalence by tumor type was not significant ($P=0.98$). Analyses stratified by the time since start and by completion of external beam radiation and chemotherapy showed no clear evidence of a time-dependence/latency between CH-PD and cumulative exposure to therapy. Thus, the time from start or stop of therapy was not adjusted for. While considering exploratory analyses, we performed multiple hypothesis correction using the FDR Q values for gene-specific analyses to control for inflation of type I error. We did not perform multiple hypothesis correction for analyses testing an association between subclasses of cancer therapy and CH because the association between cancer therapy and CH is known and our goal was to define the relative strength of these associations with subtypes of therapy rather than hypothesis testing. Heterogeneity P values to test for differences in the strength of the association between subclasses of CH and clinical variables were calculated through logistic regression models limited to CH-positive individuals testing for a difference in the odds of having CH with the mutational feature of interest (for example, CH-PD) versus having CH without the mutational feature (for example, non-CH-PD). Generalized estimating equations were used to test for an association between CH VAF and selected clinical and mutational features among CH-positive individuals, accounting for correlation between the VAF of mutations in the same person. Ordinal logistic regression among CH-positive individuals was used to test for an association between clinical characteristics and increasing CH mutation number. A test for trend between increasing cumulative exposure to cancer therapy and the odds of CH-PD was performed using multivariable logistic regression limited to individuals exposed to the therapy of interest.

Modeling the effect of cancer therapy on mutation growth rate. For each mutation in each individual with sequential sequencing data available, we modeled the growth rate of the mutation between the two time points according to the following formula:

$$\alpha = \log(V/V_0)/(T - T_0)$$

where T and T_0 indicate the age of the individual (in days) at the two measurement time points and V and V_0 correspond to the VAF at T and T_0, respectively. We also classified mutations as having increased, decreased or remained constant during the follow-up period based on a binomial test comparing the two VAFs. Generalized estimating equations were used to test for an association between exposure to cytotoxic therapy and external beam radiation therapy and CH growth rate, adjusting for age, sex and smoking status, accounting for correlation between the growth rate of mutations in the same person. Among patients with at least one mutation in a DDR CH gene and another non-DDR CH gene, we calculated the difference in the growth rate between mutations. When patients had more than two mutations in the same gene category, we used the highest growth rate for that category. A paired t-test was used to test for significance in the difference between growth rates of DDR mutations compared with non-DDR mutations within individuals who received cytotoxic therapy and/or external beam radiation therapy and within those who were untreated during the follow-up period.

Combined analysis for AML/MDS risk. We combined data from MSK and three previously published studies, Gillis et al., abbreviated MOF ($n=68$); Takahashi et al., abbreviated MDA ($n=67$); and Gibson et al., abbreviated DFC ($n=401$), studying the effect of CH on tMN risk in patients with cancer. We defined tMN as an MDS or AML diagnosed following exposure to therapeutic radiation or cytotoxic therapy as per the World Health Organization criteria[49]. For all samples, uniform postprocessing filters were applied to ensure retention of variants in accordance with the quality control standards of the MSK cohort, including a universal 2% minimum VAF cutoff. We only included mutations within genes that are present on the panel from all centers and on all panel versions from each center (Supplementary Table 6). The only exceptions were *SRSF2*, which the MSK-IMPACT-341 sequencing panel did not cover, and *PPM1D*, which was not sequenced in MSK-IMPACT-341, MDA or MOF. We performed mean imputation of missing clinical data for blood counts. Only mutations that we classified as CH-PD were included in analyses. We performed univariate cause-specific Cox proportional hazards regression for the effects of maximum VAF, total number of CH mutations, CH in specific genes and blood count parameters adjusted for age and sex and stratified by study site. Interaction terms between study and CH were used to test for heterogeneity between studies on the effect of CH on tMN risk. The proportional hazards assumption was tested through visual inspection of residual plots and through the inclusion of time-varying covariates. We performed a multivariable analysis including age, sex and all variables that were significant in the univariate analysis, with the exception of the genes not included in all studies to prevent reduction of sample size, *PPM1D* and *SRSF2*. Because our sample set was limited to individuals who received cancer therapy, we were unable to study gene–treatment interactions in the risk of myeloid neoplasm. Thus, in our combined model, CH and cancer therapy are modeled as having multiplicative effects, that is, no multiplicative interaction on myeloid neoplasm risk. We think this is a reasonable assumption for an exploratory analysis such as the one presented in our study. Much larger studies (including patients with solid tumors who did and did not receive any cancer therapy besides surgery) would be needed to define the magnitudes of CH–treatment interactions.

We also combined data from two studies investigating the effect of CH on AML risk in healthy individuals, Abelson et al., abbreviated PMC ($n=969$), and Young et al., abbreviated WSU ($n=103$), with data from MSK and applied uniform processing to mutation data from different centers. As in the solid tumor combined analysis, the same postprocessing filters used in the main MSK cohort, including a universal 2% minimum VAF cutoff, were applied to these studies and only mutations that we classified as CH-PD were included in analyses. We performed a multivariable Cox regression adjusted for age and sex including the variables used in the multivariable tMN risk analysis in patients with solid tumors.

Modeling absolute risk of AML/MDS. We used the iCARE R package[38,39] to build a model for absolute risk of AML/MDS in women with breast cancer aged 50–75 yr in the United States, by combining (1) the multivariate HR estimates from our study that were significant in the univariate model, including maximum VAF of CH, gene-specific effects and peripheral blood count indexes (red cell distribution width, hemoglobin); (2) age-specific AML/MDS rates in breast cancer using data provided by the National Comprehensive Cancer Network (NCCN)[50]; (3) competing hazards for mortality in women with breast cancer in the United States aged 50–75 yr as reported in SEER[51]; (4) previously published HR estimates for chemotherapy on the risk of tMN in women with breast cancer from the NCCN[50]; (5) the distribution of CH VAF, number of mutations, CH gene and peripheral blood count indexes using our cohort of MSK solid tumor cancer patients aged 50–75 yr who were untreated before blood draw; and (6) the proportion of women who receive adjuvant chemotherapy for breast cancer in the United States from SEER[51]. While our IMPACT cohort is not representative of the general breast cancer population in the United States, since the distribution of CH mutational features is largely driven by age and since we do not see major differences in rates of CH between sexes or untreated tumor types, we believe that the distribution of CH mutational features in untreated patients with solid tumors sequenced on IMPACT reasonably approximates an age-matched untreated breast cancer population. While blood count indexes are known to differ by sex, we chose to use the distribution of blood counts from the entire treatment-naive IMPACT population (both male and female) to capture the inter-relationship between blood count indexes and CH mutational features. Sensitivity analyses using the distribution of blood count parameters from only female IMPACT patients produced similar results. This risk model assumes an additive association on the log scale of CH mutational features and cancer therapy for risk of tMN. This assumption is supported by the similarity between risk estimates for CH mutational features between AML in healthy individuals never exposed to therapy and tMN (Supplementary Fig. 10).

All of the statistical analyses were performed using the R statistical package (www.r-project.org). The code used in statistical analyses is provided in the Supplementary Note.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The minimal clinical and mutational data necessary to replicate the findings in the article, except those shown in Extended Data Fig. 5 and Supplementary Fig. 12, are publicly available on GitHub: https://github.com/papaemmelab/bolton_NG_CH. Data for the excepted figures (individual drug names and start and stop dates, and combinations of mutations at tMN diagnosis, respectively) cannot be made public to preserve patient anonymity. Raw sequencing data cannot be publicly deposited for legal and privacy reasons, as sequencing was performed for clinical purposes. Mutation calls are available on cBioPortal: http://www.cbioportal.org/study/summary?id=msk_ch_2020

Code availability

The codes to replicate the findings in the article, except those shown in Extended Data Fig. 5 and Supplementary Fig. 12, are publicly available on GitHub: https://github.com/papaemmelab/bolton_NG_CH. The codes used to generate the excepted figures are not included because the data cannot be shared (see above).

References

45. Cheng, D. T. et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagn.* **17**, 251–264 (2015).
46. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
47. Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

48. Papaemmanuil, E. et al. Identification of novel somatic mutations in SF3B1, a gene encoding a core component of RNA splicing machinery, in myelodysplasia with ring sideroblasts and other common cancers. *Eur. J. Cancer* **47**, 7 (2011).
49. Campo, E. et al. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues* 4th edn, Vol. 2 (IARC, 2017).
50. Wolff, A. C. et al. Risk of marrow neoplasms after adjuvant breast cancer therapy: the National Comprehensive Cancer Network experience. *J. Clin. Oncol.* **33**, 340–348 (2015).
51. *Surveillance, Epidemiology, and End Results (SEER) Program Populations (1969–2017)* (National Cancer Institute, DCCPS, Surveillance Research Program, 2018); www.seer.cancer.gov/popdata

Acknowledgements

This work was supported by the National Institutes of Health (grant no. K08 CA241318 to K.L.B., grant no. K12 CA120780 to C.C.C., grant no. P50 CA172012 to L.B., grant no. P50 CA172012 to J.F., grant no. UG1 HL069315 to V.M.K.), the American Society of Hematology (K.L.B. and E. Papaemmanuil), the EvansMDS Foundation (K.L.B.), the European Hematology Association (E. Papaemmanuil), Gabrielle's Angels Foundation (E. Papaemmanuil), the V Foundation (E. Papaemmanuil), the Geoffrey Beene Foundation (E. Papaemmanuil), the UNC Oncology Clinical Translational Research Training Program (C.C.C.), Cycle for Survival (V.M.K.), the Starr Cancer Consortium (to R.L.L., A.Z., M.F.B., R.N.P.) and the Cancer Colorectal Cancer Dream Team Translational Research Grant (grant no. SU2C-AACR-DT22-17 to L.A.D.). E. Papaemmanuil is a Josie Robertson Investigator. C.C.C. is a recipient of the Conquer Cancer Foundation Young Investigator Award and the Prostate Cancer Foundation Young Investigator Award. K.H.S. is a recipient of the Defense Early Investigator Research Award (grant no. W81XWH-18-1-0330), the Prostate Cancer Foundation Young Investigator Award and the Prostate Cancer Foundation Challenge Award. C.L., M.G.-C. and L.M.M. are supported by funds from the Intramural Research Program of the National Cancer Institute, National Institutes of Health. Work performed at Memorial Sloan Kettering Cancer Center was supported in part by the Cancer Center Support Grant (grant no. P30 CA008748). N.G.'s work was supported in part by the Tissue Core and Genomic Core Facilities at the H. Lee Moffitt Cancer Center & Research Institute, an NCI-designated Comprehensive Cancer Center (grant no. P30 CA076292). The University of Cambridge has received salary support in respect of P.D.P.P. from the NHS in the East of England through the Clinical Academic Reserve.

Author contributions

K.L.B., R.L.L., A.Z. and E. Papaemmanuil conceived and designed the study. K.L.B., D.K., M.P., A.P., L.B. and N.C. collected clinical data. R.N.P., A.S., R.B., M.E.A., M. Ladanyi, M.F.B. and A.Z. led the generation of IMPACT sequencing data. K.L.B., M.P., A.P., N.C., D.M.H., M.S.T. and R.L.L. collected sequential samples. R.N.P., T.G. and K.L.B. called variants and performed postprocessing of sequencing data. K.L.B., T.G., S.M.D., A.B., M.G.-C., N.C., L.M.M., A.Z. and E. Papaemmanuil performed statistical analyses and/or participated in data interpretation. K.L.B., R.N.P., T.G., L.B., S.M.D., D.K., M.P., A.B., A.S., M.Y., C.C.C., N.M.C., M.W., K.O., Z.S., D.M., J.S., A.P., J.P., E.B., G.G., J.E.A.O., M. Levine, J.S.M.M., N.F., D.G., S.L., M.E.R., C.L., P.D.P.P., K.H.S., B.S., S.M., J.F., L.B., C.J.G., B.L.E., A.L.Y., T.D., K.T., N.G., M.B., E. Padron, D.M.H., J.B., L.N., S.G., V.M.K., H.S., D.B., E. Paraiso, R.B., M.E.A., M. Ladanyi, D.B.S., M.F.B., M.S.T., M.G.-C., N.C., L.A.D., R.L.L., L.M.M., A.Z. and E. Papaemmanuil contributed to the writing of the manuscript and approved it for submission.

Competing interests

The authors declare the following competing interests: K.L.B. has received research funding from GRAIL. C.C.C. has received honoraria from AbbVie, Loxo, H3 Biomedicine, Medscape, Octapharma and Pharmacyclics; has served as a consultant for AbbVie, Covance, Cowen & Co. and Dedham Group; and has received institutional research funding from AROG, Gilead, Loxo, H3 Biomedicine and Incyte. Z.S. has an immediate family member who holds consulting/advisory roles within the field of ophthalmology with Allergan, Adverum Biotechnologies, Alimera Sciences, Biomarin, Fortress Biotech, Genentech, Novartis, Optos, Regeneron, Regenxbio and Spark Therapeutics. E.B. receives research funding from Celgene. D.G. is a consultant of MNM Diagnostics and has received honoraria for speaking and scientific advisory engagements with Celgene, Prime Oncology, Novartis, Illumina and Kyowa Hakko Kirin. S.L. is an employee of GRAIL. M.E.R. holds an uncompensated advisory role with AstraZeneca, Daiichi-Sankyo, Merck and Pfizer and receives institutional research funding from AstraZeneca, AbbVie, Medivation and Pfizer. B.L.E. has received research funding from Celgene and Deerfield. T.D. is the Chief Medical Officer, ArcherDX, Inc. and receives salary from and holds an ownership stake in the company. K.T. receives consultancy fees from Symbio Pharmaceuticals. D.M.H. has consulted for Fount, Chugai, Boehringer Ingelheim, AstraZeneca, Pfizer, Bayer and Genentech/Roche; has equity in Fount; and has received research grants from Loxo, Bayer, Puma and AstraZeneca. J.B. is an employee of AstraZeneca; is on the Board of Directors of Foghorn and is a past board member of Varian Medical Systems, Bristol-Myers Squibb, Grail, Aura Biosciences and Infinity Pharmaceuticals; has performed consulting and/or advisory work for Grail, PMV Pharma, ApoGen, Juno, Eli Lilly, Seragon, Novartis and Northern Biologics; has stock or other ownership interests in PMV Pharma, Grail, Juno, Varian, Foghorn, Aura, Infinity Pharmaceuticals, ApoGen and Northern Biologics, as well as Tango and Venthera, for which he is a co-founder; and has previously received honoraria or travel expenses from Roche, Novartis and Eli Lilly. M. Ladanyi serves on the advisory boards for AstraZeneca, Bristol Myers Squibb, Takeda, Bayer and Merck, and has received research support from Loxo Oncology and Helsinn Therapeutics. D.B.S. has served as a consultant for or received honoraria from Pfizer, Loxo Oncology, Lilly Oncology, Illumina and Vivideon Therapeutics. M.F.B. is on the advisory board for Roche and receives research support from Illumina. M.S.T. receives research funding from AbbVie, Cellerant, Orsenix, ADC Therapeutics and Biosight; serves on the advisory boards of Daiichi-Sankyo, KAHR, Rigel, Nohla, Delta Fly Pharma, Tetraphase, Oncolyze and Jazz Pharma; has received royalties from UpToDate; and has received research funding from Incyte, Kura Oncology and Celgene. L.A.D. is a member of the board of directors of Personal Genome Diagnostics (PGDx) and Jounce Therapeutics; is a paid consultant to PGDx and Neophore; is an uncompensated consultant for Merck (with the exception of travel and research support for clinical trials); is an inventor of multiple licensed patents related to technology for circulating tumor DNA analyses and mismatch repair deficiency for diagnosis and therapy from Johns Hopkins University, some of which are associated with equity or royalty payments directly to Johns Hopkins and L.A.D.; and holds equity in PGDx, Jounce Therapeutics, Thrive Earlier Detection and Neophore; his wife holds equity in Amgen. The terms of all of these arrangements are being managed by Johns Hopkins and Memorial Sloan Kettering in accordance with their conflict of interest policies. R.L.L. is on the supervisory board of Qiagen and is a scientific advisor to Loxo, Imago, C4 Therapeutics and Isoplexis, which include equity interest; receives research support from and has consulted for Celgene and Roche and has consulted for Lilly, Janssen, Astellas, Morphosys and Novartis; and has received honoraria from Roche, Lilly and Amgen for invited lectures and from Gilead for grant reviews. A.Z. received honoraria from Illumina. E. Papaemmanuil receives research funding from Celgene and is a cofounder in Isabl Technologies, a software analytics company for high-throughput clinical whole-genome and RNA-sequencing analyses. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-00710-0.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-00710-0.

Correspondence and requests for materials should be addressed to A.Z. or E. Papaemmanuil.

Reprints and permissions information is available at www.nature.com/reprints.

Extended Data Table 1 | Clinical characteristics of solid tumor patients assessed for CH

	CH–	CH+
Total	**16930 (70%)**	**7216 (30%)**
Smoking status		
Non-smoker	8979 (74%)	3086 (26%)
Current/former	7255 (65%)	3894 (35%)
Missing	696 (75%)	236 (25%)
Gender		
Male	7710 (70%)	3315 (30%)
Female	9220 (70%)	3901 (30%)
Age		
0-10	324 (96%)	13 (3.9%)
10-20	284 (96%)	13 (4.4%)
20-30	672 (95%)	36 (5.1%)
30-40	1398 (92%)	121 (8%)
40-50	2757 (87%)	420 (13%)
50-60	4490 (78%)	1298 (22%)
60-70	4499 (64%)	2575 (36%)
70-80	2127 (50%)	2092 (50%)
80-90	379 (37%)	648 (63%)
Ethnicity		
White	12628 (69%)	5802 (31%)
Asian	1274 (78%)	356 (22%)
Black	1081 (73%)	410 (27%)
Other	1175 (77%)	355 (23%)
Unknown	772 (72%)	293 (28%)
Therapy		
Treated	4193 (70%)	1785 (30%)
Untreated	3027 (73%)	1133 (27%)
Unknown	9710 (69%)	4298 (31%)
Primary tumor subtype		
Ampullary carcinoma	47 (76%)	15 (24%)
Anal cancer	38 (67%)	19 (33%)
Appendiceal cancer	128 (79%)	34 (21%)
Biliary cancer	351 (69%)	157 (31%)
Bladder cancer	445 (62%)	267 (38%)
Breast carcinoma	2610 (74%)	930 (26%)
Cancer of unknown primary	484 (67%)	239 (33%)
Cervical cancer	91 (77%)	27 (23%)
Chondroblastoma	1 (100%)	0 (0%)
Chondrosarcoma	42 (78%)	12 (22%)
Chordoma	27 (75%)	9 (25%)
Choroid plexus tumor	3 (100%)	0 (0%)
Colorectal cancer	1625 (75%)	528 (25%)
Embryonal tumor	153 (89%)	18 (11%)
Endometrial cancer	510 (61%)	321 (39%)
Ependymomal tumor	26 (90%)	3 (10%)
Esophagogastric carcinoma	464 (70%)	196 (30%)
Ewing sarcoma	66 (89%)	8 (11%)
Gastrointestinal neuroendocrine tumor	73 (68%)	34 (32%)

Continued

Extended Data Table 1 | Clinical characteristics of solid tumor patients assessed for CH (continued)

	CH−	CH+
Total	**16930 (70%)**	**7216 (30%)**
Gastrointestinal stromal tumor	200 (70%)	84 (30%)
Germ cell tumor	352 (91%)	35 (9%)
Gestational trophoblastic disease	10 (77%)	3 (23%)
Glioma	834 (76%)	260 (24%)
Head and neck carcinoma	252 (69%)	111 (31%)
Hepatocellular carcinoma	134 (71%)	55 (29%)
Melanoma	612 (69%)	269 (31%)
Meningothelial tumor	52 (79%)	14 (21%)
Mesothelioma	146 (65%)	78 (35%)
Miscellaneous brain tumor	22 (85%)	4 (15%)
Miscellaneous neuroepithelial tumor	11 (65%)	6 (35%)
Nerve sheath tumor	43 (88%)	6 (12%)
Non-small cell lung cancer	2235 (63%)	1324 (37%)
Osteosarcoma	98 (90%)	11 (10%)
Ovarian cancer	411 (62%)	254 (38%)
Pancreatic cancer	964 (68%)	452 (32%)
Penile cancer	7 (78%)	2 (22%)
Pheochromocytoma	6 (86%)	1 (14%)
Pineal tumor	1 (25%)	3 (75%)
Prostate cancer	971 (65%)	523 (35%)
Renal cell carcinoma	445 (78%)	128 (22%)
Retinoblastoma	38 (95%)	2 (5%)
Salivary carcinoma	161 (76%)	52 (24%)
Sellar tumor	53 (88%)	7 (12%)
Sex cord stromal tumor	29 (81%)	7 (19%)
Skin cancer, non-melanoma	137 (60%)	91 (40%)
Small bowel cancer	66 (77%)	20 (23%)
Small cell lung cancer	128 (60%)	84 (40%)
Soft tissue sarcoma	751 (76%)	233 (24%)
Thymic tumor	35 (70%)	15 (30%)
Thyroid cancer	267 (62%)	165 (38%)
Uterine sarcoma	124 (73%)	46 (27%)
Vaginal cancer	10 (67%)	5 (33%)
Wilms tumor	23 (96%)	1 (4.2%)
Unknown	75 (69%)	34 (31%)

Extended Data Table 2 | Association between variant allele fraction (VAF) of CH mutations and clinical characteristics

Variable (reference)		OR	95% CI	p
Age	–	1	1–1.1	0.0011
Ethnicity (white)	Asian	1	0.94–1.2	0.42
	Black	0.9	0.82–1	0.053
	Other	0.93	0.83–1	0.24
	Unknown	0.92	0.8–1.1	0.22
Smoking status (non-smoker)	Smoker	1.1	1.1–1.2	0.000023
Therapy (untreated)	Treated	1	0.96–1.1	0.8
PD status (Non-PD non-myeloid)	Myeloid PD	1.3	1.3–1.4	$<1\times10^{-6}$
	Non-myeloid PD	1.3	1.2–1.5	0.000052
	Non-PD myeloid	0.99	0.92–1.1	0.8
Number of mutations (1)	≥ 2	1.1	1.1–1.2	0.0000038

Generalized estimating equations were used to test for association between VAF of CH mutations (among those with a mutation) and selected clinical and mutational features, accounting for correlation between the VAF of mutations in the same person. Age expressed in decile.

Extended Data Table 3 | Association among clinical characteristics and CH mutational characteristics

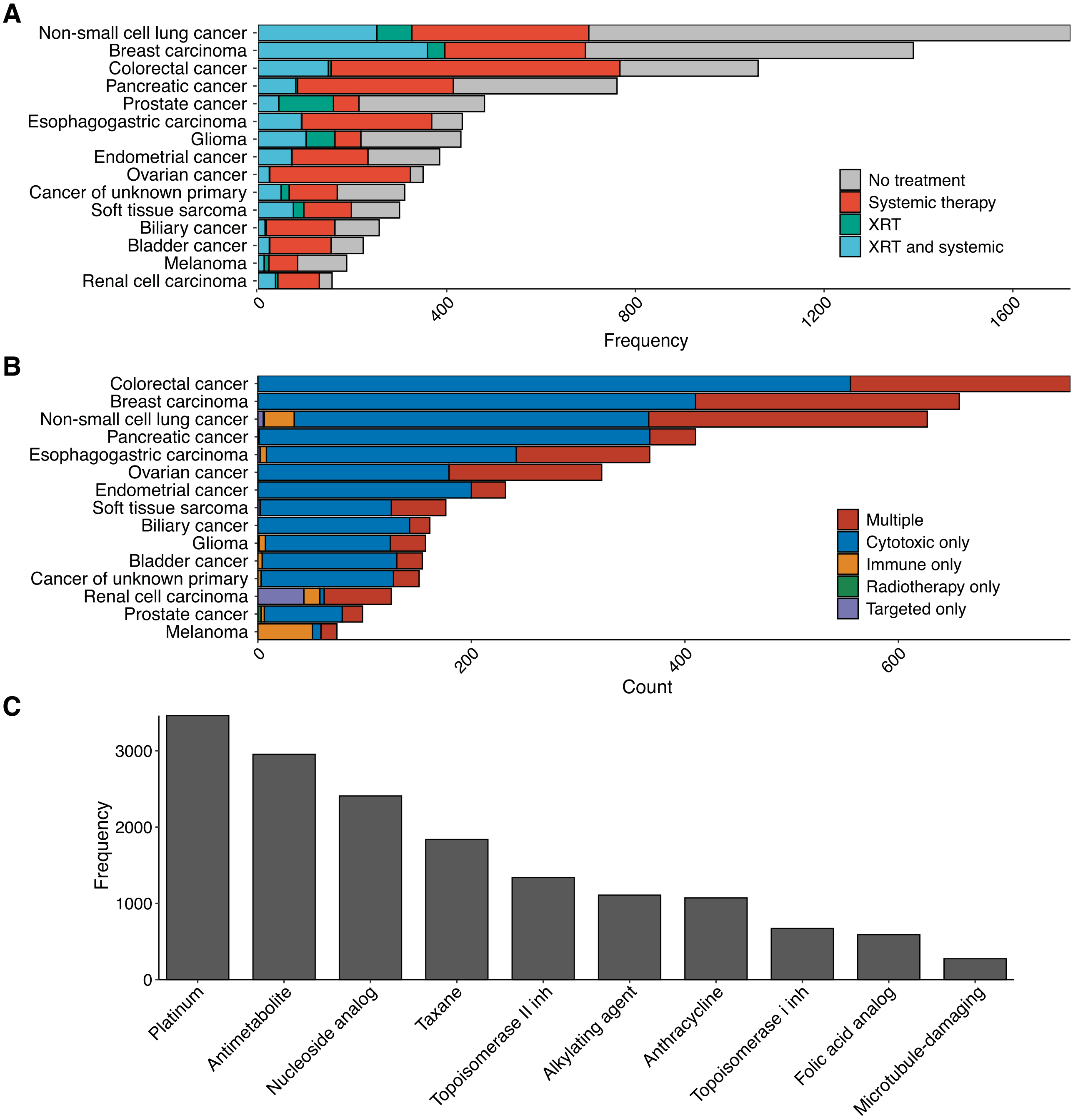
Variable (reference)		OR	95% CI	p
Age	–	1	1–1.1	0.0011
Ethnicity (white)	Asian	1	0.94–1.2	0.42
	Black	0.9	0.82–1	0.053
	Other	0.93	0.83–1	0.24
	Unknown	0.92	0.8–1.1	0.22
Smoke (non-smoker)	Smoker	1.1	1.1–1.2	0.000023
Therapy (untreated)	Treated	1	0.96–1.1	0.8
PD status (non-PD non-myeloid)	Myeloid PD	1.3	1.3–1.4	$< 1\times10^{-6}$
	Non-myeloid PD	1.3	1.2–1.5	0.000052
	Non-PD myeloid	0.99	0.92–1.1	0.8
Number of mutations (1)	≥ 2	1.1	1.1–1.2	0.0000038

Myeloid PD, genes mutated in myeloid neoplasms; non-myeloid, genes not linked to myeloid neoplasms; myeloid PD, variants known to be myeloid drivers or putative somatic driver mutations in myeloid neoplasms; myeloid non-PD, mutations within genes linked to myeloid neoplasms but that are not putative drivers; non-myeloid PD, mutations that are putative somatic driver mutations of cancer in genes not linked to myeloid neoplasms; non-myeloid non-PD, mutations within genes not linked to myeloid neoplasms that are not putative drivers of cancer. Associations were evaluated using multivariable logistic regression models to generate heterogeneity p-values. Sensitivity analyses restricted to individuals with only one mutation yielded similar results. Age expressed in decile.

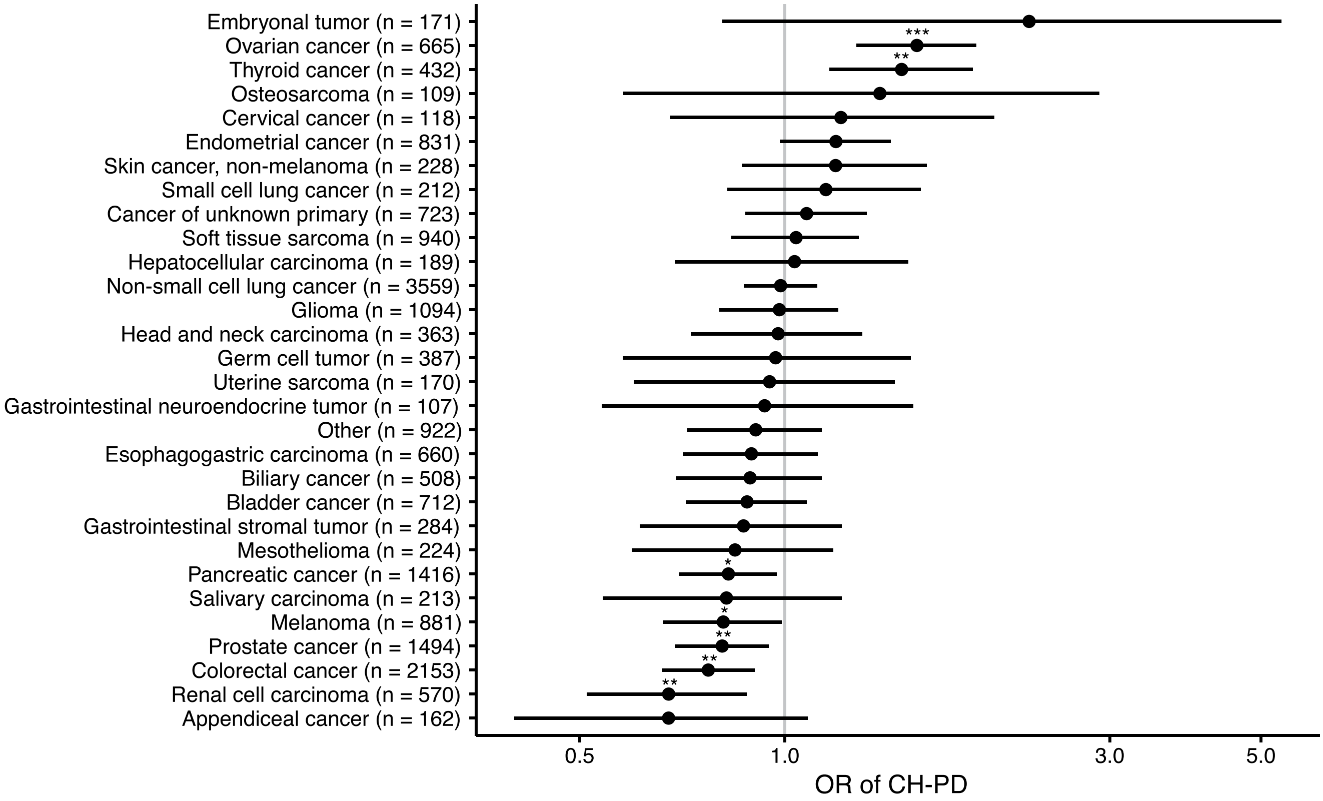
Extended Data Table 4 | Association between CH mutation number and clinical characteristics

Variable (reference)		OR	95% CI	p
Age (0–10)	> 10	2.3	2–2.6	$< 1 \times 10^{-6}$
Gender (male)	Female	1.1	0.94–1.3	0.2
Ancestry (white)	Non-white	0.83	0.67–1	0.087
Smoke (non-smoker)	Smoker	1.2	1–1.4	0.027
Therapy (untreated)	Treated	1.2	1.1–1.5	0.011

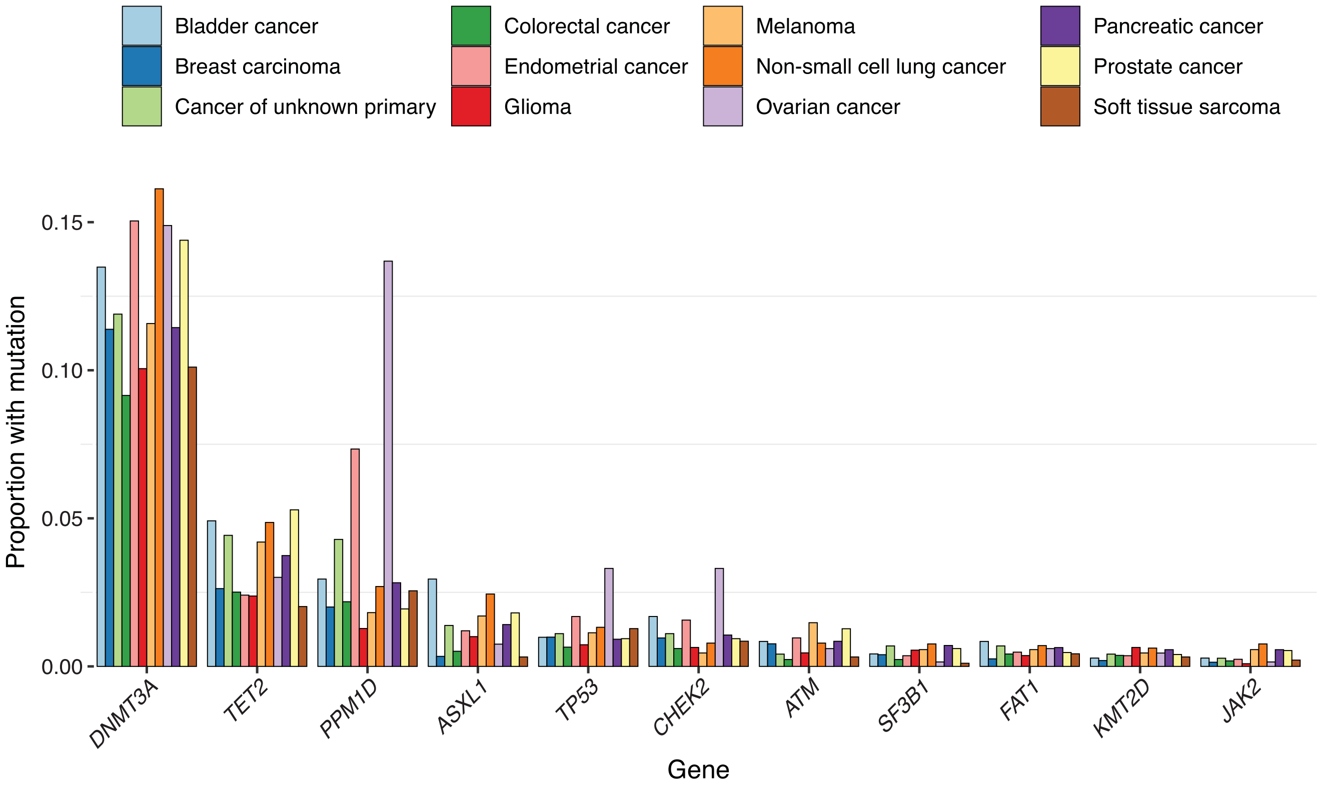
Ordinal logistic regression was used to test for association between clinical characteristics and mutation number in patients with clonal hematopoiesis in a multivariable model. Age expressed in decile.



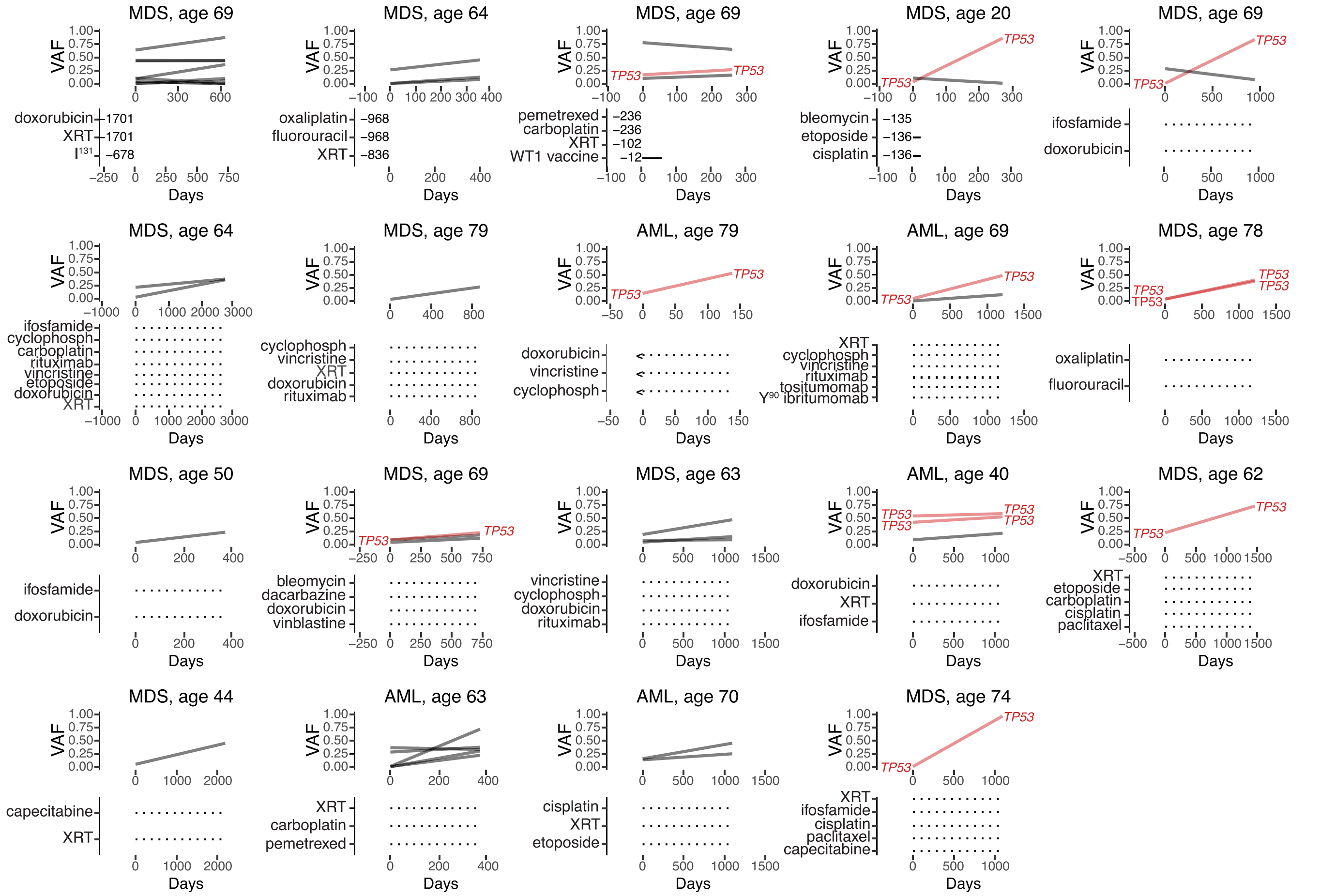
Extended Data Fig. 1 | Distribution of cancer therapy received prior to blood collection for sequencing. a, Frequency of patients receiving systemic therapy or external beam radiation therapy by primary tumor type. **b**, Frequency of patients receiving specific classes of systemic therapy by primary tumor type. **c**, Frequency of patients receiving top ten subclasses of cytotoxic therapy. Most patients (91%) who received at least one of these cytotoxic therapy classes received multiple classes.



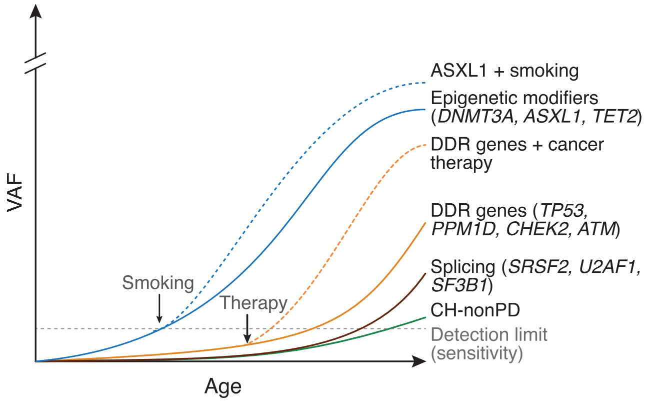
Extended Data Fig. 2 | Association between primary tumor site and CH-PD. Odds ratios (circle) and 95% confidence intervals for CH-PD in selected primary tumor types with at least 100 subjects compared to breast cancer (n = 3540) in a logistic regression model adjusted for age. * p < 0.05, ** p < 0.01, *** p < 0.001.



Extended Data Fig. 3 | Proportion of patients with common CH-PD mutations by primary tumor sites. Genes mutated in at least 75 individuals and the top 12 primary tumor sites are shown.



Extended Data Fig. 4 | Variant frequencies (VAF) at time of pre-tMN testing and tMN diagnosis. Plots show changes in mutational frequencies in relation to cancer therapy exposure in 19 CH cases. Below each graph are listed treatments received prior to pre-tMN testing and the number of days between the end of treatment and the pre-tMN sample.



Extended Data Fig. 5 | Differences in the fitness effect of CH mutations and the environment shape clonal dominance over an individual's lifetime. Conceptual graph illustrating how associations between specific exposures and CH mutations may shape clonal dominance over an individual's lifetime. AML, acute myeloid leukemia; cyclophosph, cyclophosphamide; MDS, myelodysplastic syndrome.

nature research

Corresponding author(s): Elli Papaemmanuil, Ahmet Zehir

Last updated by author(s): Jun 17, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: Clinical data was generated from the EMR. De-identified and binned clinical data are available on GitHub.

Data analysis: R version 4.0.1 was used to analyze the majority of the data in this study; R code used is available on GitHub. Sequencing data were aligned using BWA (0.7.5a), reads were re-aligned around indels using ABRA (0.92), and base quality scores were recalibrated using the Genome Analysis Toolkit (GATK) (3.3-0). Variants were called using Mutect, VarDict, and Somatic Indel Detector (SID).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The following data availability statement appears following the methods: "The minimal clinical and mutational data necessary to replicate the findings in the article, except those shown in Extended Data Figure 5 and Supplementary Figure 12, are publicly available on Github: https://github.com/papaemmelab/bolton_NG_CH. Data for the excepted figures (individual drug names and start and stop dates, and combinations of mutations at tMN diagnosis, respectively) cannot be made public to preserve patient anonymity."

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was determined based on the number of individuals with blood sequencing data using the MSK-IMPACT assay as of July 2019. This dataset provided adequate (80%) power to detect an association between subclasses of cancer therapy with a frequency of at least 10% and CH with an OR of at least 1.2.
Data exclusions	Individuals were excluded only if they had an active hematologic malignancy at the time of blood sequencing or if sequencing failed quality control (metrics outlined in methods).
Replication	We assessed the reproducibility of our mutational calling strategy as detailed in the methods. Major findings from the retrospective analysis were supported from prospectively collected data as detailed in the manuscript. All data were collected from real-world patients; no experiments were performed. Because of the unique nature of the data required to study how oncologic therapy relates to CH prevalence and clonal evolution (large numbers of cancer patients with prospectively sequenced blood and complete clinical histories), replication would not be possible. However, the findings from two separate cohorts, our retrospective data and our prospective collection provide parallel lines of evidence supporting our main conclusions.
Randomization	As subjects were not randomized between treatment groups, we adjusted for possible confounders using multi-variable regression. In our prospective study, we assessed for mutational competition within the same individual to investigate the effect of therapy independent of individual-specific factors that may differ between those who receive therapy and those who did not during the follow-up period.
Blinding	Data collection took place independently of mutational analysis. Investigators were not blinded, but separate investigators assembled the clinical data (K.B) and the mutational data (R.P). Clinical and mutational data frames were processed and analyzed separately before combining.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
□	☒ Human research participants
□	☒ Clinical data

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Extensive patient characteristics were ascertained from the MSK electronic medical record system. Please see methods and supplementary notes for details on how this data was ascertained and from what sources.
Recruitment	Participants were included who had blood and matched tumor sequencing data available on MSK-IMPACT. Patients sequenced on MSK-IMPACT are more likely to have advanced disease compared to the general solid tumor population in the U.S. However, since our study was focused on the association between oncologic therapy and clonal hematopoiesis and since this contained a mixture of treated and untreated patients, we do not anticipate the study population would bias our association results.
Ethics oversight	This study was approved by the Memorial Sloan Kettering IRB under protocol 12-245 part C and protocol 18-288. As stated in the manuscript, all patients enrolled on the MSK-IMPACT clinical protocol provided informed consent. A subset of patients that underwent tumor-genomic profiling as standard of care were not directly consented, in which case an IRB waiver was obtained to allow for inclusion into this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	NCT01775072
Study protocol	Details available at ClinicalTrials.gov NCT01775072 or upon request.
Data collection	Data was collected between December 1, 2014 through July 1, 2018 at Memorial Sloan Kettering Cancer Center, a tertiary / referral cancer center in New York, NY.
Outcomes	For the development of therapy-related myeloid neoplasms, the clinical endpoint was the time from blood draw to the development of myelodysplastic syndrome or acute myeloid leukemia, or censoring in cases lost to follow-up or who died. All outcome-associated P values and estimates of hazard ratios were generated from multivariable Cox proportional hazards models.

nature research

https://doi.org/10.1038/s41588-020-00710-0

Supplementary information

Cancer therapy shapes the fitness landscape of clonal hematopoiesis

In the format provided by the authors and unedited

Supplement to "*Cancer therapy shapes the fitness landscape of clonal hematopoiesis*"

Supplementary Notes for "*Cancer therapy shapes the fitness landscape of clonal hematopoiesis*"

Measurement of cumulative therapy dose

To define cumulative exposure to a therapeutic subclass of cytotoxic therapy, a simple sum of the mg/kg of drugs within a class cannot be used because drugs even within the same drug class are delivered on different dosing scales. To derive metrics for cumulative exposure to cytotoxic therapy subclasses, we applied the approach used by the Late Effects Study Group[1]. For each drug the total dose per kg received prior to blood draw was summed for each patient. The dose distribution for each agent was divided into tertiles and the patient's dose was assigned a score based on tertile of total exposure. An individual patient's scores for each drug in a specific drug class were summed. The distribution of the resulting sum across all patients was used to derive tertiles of total exposure to the drug class in the entire cohort (Supplementary Figure 7).

Given the variety of radiotherapy fractionation schemes and prescribed tumor doses, we calculated the cumulative radiation dose received by each patient prior to blood draw in 2-Gy per fraction equivalents (EQD_2) using an α/β of 3 Gy, considering CH to be a late-responding tissue effect[2]. We calculated tertiles of dose based on the distribution of cumulative EQD_2 received over the entire cohort and assigned each individual a score based on their tertile of exposure (*e.g.* a patient who did not receive external beam radiation received a score of zero for that particular agent. If the patient's cumulative radiation dose, as expressed in EQD_2, was within the first tertile, a score of one was assigned, and so forth).

Clinical characteristics of previously published studies included in combined analyses

To study the relationships between CH and tMN we aggregated data from MSK with 5 previously published studies including Gillis et al, (MOF) Takahshi et al. (MDA), Gibson et al. (DFC), Young et al., (WSU) and Abelson et al. (EPI).

MOF

Gillis et al.[3] performed a nested case-control study for tMN risk using subjects from an internal biorepository of 123,357 cancer patients who consented to participate in the Total Cancer Care biobanking protocol at Moffitt Cancer Center (Tampa, FL, USA) between Jan 1, 2006, and June 1, 2016. Cases were individuals diagnosed with a primary malignancy, treated with chemotherapy who subsequently developed a therapy-related myeloid neoplasm, and were 70 years or older at either diagnosis. Controls were individuals who were diagnosed with a primary malignancy at age 70 years or older and were treated with chemotherapy but did not develop therapy-related myeloid neoplasms. Controls were matched to cases in at least a 4:1 ratio on the basis of sex, primary tumour type, age at diagnosis, smoking status, chemotherapy drug class, and duration of follow-up. DNA was isolated from peripheral blood collected before therapy-related myeloid neoplasm diagnosis and subjected to Droplet-partitioned, targeted, amplicon-based, next-generation sequencing was used in accordance with the manufacturer's instructions (RainDance Technologies, Billerica, MA, USA) to identify somatic mutations in 49 myeloid-driver genes (ThunderBolts Myeloid Panel, RainDance, Billerica, MA, USA).

MDA

Takahashi et al.[4] performed a case-control study for cancer patients who developed therapy-related myeloid neoplasms (cases) and lymphoma patients who did not develop therapy-related myeloid neoplasms (controls).Cases were identified using a clinical database at the Department of Leukemia of The University of Texas MD Anderson Cancer Center (Houston, TX, USA) including 40 000 patients who have consented for their data to be used in research. Inclusion criteria were that patients had to have been treated for a primary cancer from June 11, 1997, and subsequently had diagnoses of therapy-related myeloid neoplasms between Jan 1, 2003, and Dec 31, 2015, and had paired samples of diagnostic bone marrow at the time of therapy-related myeloid neoplasm diagnosis and peripheral blood samples obtained at the time of primary cancer diagnosis. An aged-matched control group (using a 3:1 control to case ratio) was identified using a clinical database of patients treated for lymphoma from 2008 to 2015. Eligible patients were those who had a pre-treatment blood sample available, had received a combination chemotherapy regimen including an alkylating agent, had at least 5 years of follow-up with no clinical evidence of therapy-related myeloid neoplasm development, and had no evidence of bone marrow metastasis of lymphoma in a bilateral bone marrow biopsy. Targeted sequencing of

32 myeloid genes was performed using an amplicon-based targeted deep sequencing method, including unique molecular barcodes.

DFC

Gibson et al.[5] performed a cohort study among 401 adult patients who underwent ASCT for non-Hodgkin lymphoma between 2003 and 2010 (Dana-Farber Cancer Institute, Boston, MA; targeted sequencing cohort) with mobilized stem-cell products available at the time of ASCT. All subjects had been exposed to cancer therapy prior to stem cell collection. During the follow-up period 18 patients developed tMN. Targeted sequencing was performed using mobilized stem-cell products at the time of ASCT and on bone marrow aspirates obtained at the time of TMN diagnosis for all patients who had an available specimen (N=9). Targeted deep sequencing was performed using 86 known myeloid genes genes using the Custom SureSelect hybrid capture system (Agilent Technologies, Santa Clara, CA).

WSU

Young et al.[6] utilized a nested case-control design for AML using data from two large cohort studies, the Nurses Health Study (NHS) and the Health Professionals Follow-Up Study. Subjects were drawn from the "blood sub-cohorts" of these two studies which included 32,826 women (NHS) with a blood sample from 1989-90 as well as 18,018 men (HPFS) who provided a whole blood sample from 1993-95. The case definition included all blood subcohort participants with confirmed diagnoses of AML (ICD-8=205.0) occurring after blood draw. Two matched controls were selected per case on cohort (sex), race, birthdate (± 1 year), and blood draw details (date ± 1 year, time ± 4 hours, fasting status <8, 8+ hours). In total this included 35 cases (16 NHS, 19 HPFS) and 70 controls (32 NHS, 38 HPFS). Samples were sequenced using the Illumina TruSight Myeloid Sequencing Panel for targeted capture from 54 leukemia-associated genes Technical replicate libraries were sequenced on different machine runs. Error corrected sequencing analysis of raw sequencing results was performed as described elsewhere[7].

EPI

Abelson et al.[8] performed a case-control study for AML using samples from EPIC[9]. We used data from both the discovery and validation sets. The discovery set included 509 DNA samples from individuals who enrolled on the EPIC study between 1993 and 1998 across 17 different centres. This included 95 individuals who developed AML and 414 age- and gender-matched controls who did not develop any haematological disorders during the follow-up period. Subjects for the validation cohort were obtained from individuals enrolled in the EPIC-Norfolk longitudinal cohort study between 1994 and 2010. Samples were available from 37 patients with AML (of which 8 were already included in the discovery cohort) and 262 age- and gender-matched controls without a history of cancer or any haematological conditions. Targeted deep sequencing in the discovery cohort was performed using error-corrected, custom capture-based sequencing using the xGen AML Cancer Panel. Targeted sequencing in the validation set was performed using a custom complementary RNA bait set (SureSelect, Agilent, ELID 0537771) designed complementary to all coding exons of 111 myeloid driver-genes.

Eliminating germline events and technical artifacts using tumor comparator

Using a synthetic dataset, we profiled the error rates of several methods that use the matched tumor as a comparator to eliminate germline events and false positive calls (artifacts).
We simulated pairs of observed variant allele fractions in the blood and the tumor as follows:

Let f_b be the true variant allele fraction in blood, f_t be the true variant allele fraction in blood and c be the level of blood contamination in the tumor and $r \in \{0,1\}$ be an indicator for whether the variant is real (r=1) or artifact ($r = 0$). Let v_b be the observed VAF in the blood, v_t be the observed VAF in tumor and d be the sequencing depth in both blood and tumor. For convenience, d is fixed to be 500 as per the typical coverage for IMPACT sequencing panels.

A called mutation m can be either be true CH (a somatic mutation in the blood), an artifact, or a germline mutation. If m is a real CH mutation, then we would expect the tumor VAF to be a product of the amount of blood contamination in the tumor and the true VAF in the blood, $f_\mathrm{t} = cf_\mathrm{b}$. If m is an artifact or a germline mutation, we would expect the tumor VAF to be same as the blood VAF, namely $f_\mathrm{t} = f_\mathrm{b}$. v_t is expected to follow a binomial distribution based on the sequencing depth d and true VAF in the blood and tumor respectively. Thus, the observed blood VAF can be modeled by $v_b \sim Bin(d,f_b)$ while the observed tumor VAF can be modeled by $v_t \sim Bin(d,c,f_b)$. We simulated the observed blood and tumor VAFs for real and artifactual mutations under a range of blood contamination levels ($c = \{0.05, 0.1, \ldots, 0.5\}$) and true blood VAFs ($f_b$.= $\{0.02, 0.04, \ldots, 0.2\}$). Using this synthetic dataset, we evaluated two methods (with different threshold parameters) that aim to differentiate real CH variants from non-CH variants using the observed VAFs:

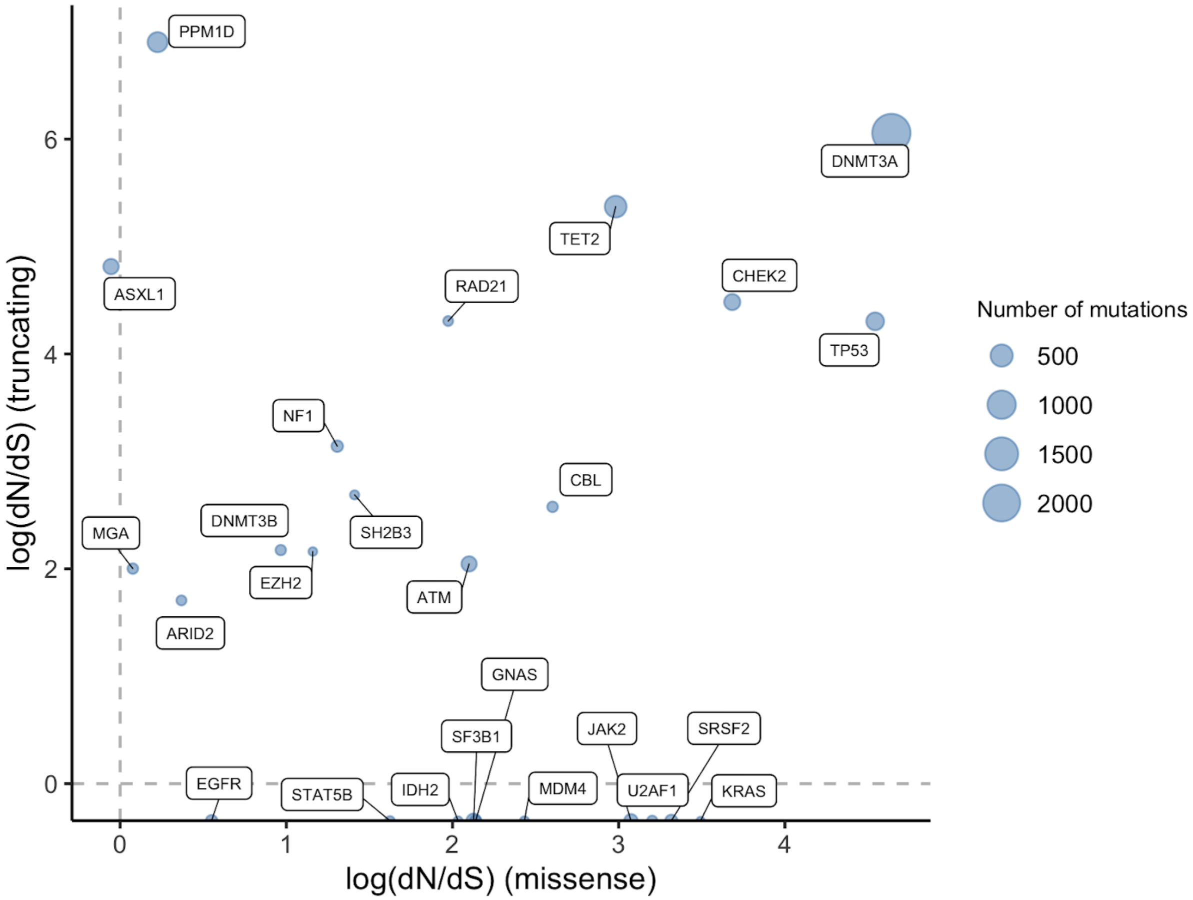
1. Blood-tumor Ratio: Predict mutation is real if $v_b / v_t >=$ C otherwise consider it an artifact. We evaluated a range of cutoffs for C {1, 1.5, 2, 3, 4}.
2. Binomial test: Predict mutation is real if $p<0.05$ from a binomial test with the null hypothesis of $v_b = v_t$

The predictions by each method were evaluated against the true mutation types that gave rise to the data points, and were classified as true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The overall precision of various methods/cutoffs were calculated as TP/TP+FP and its recall as TP/TP+FN (Supplementary Figures 13 and 14).

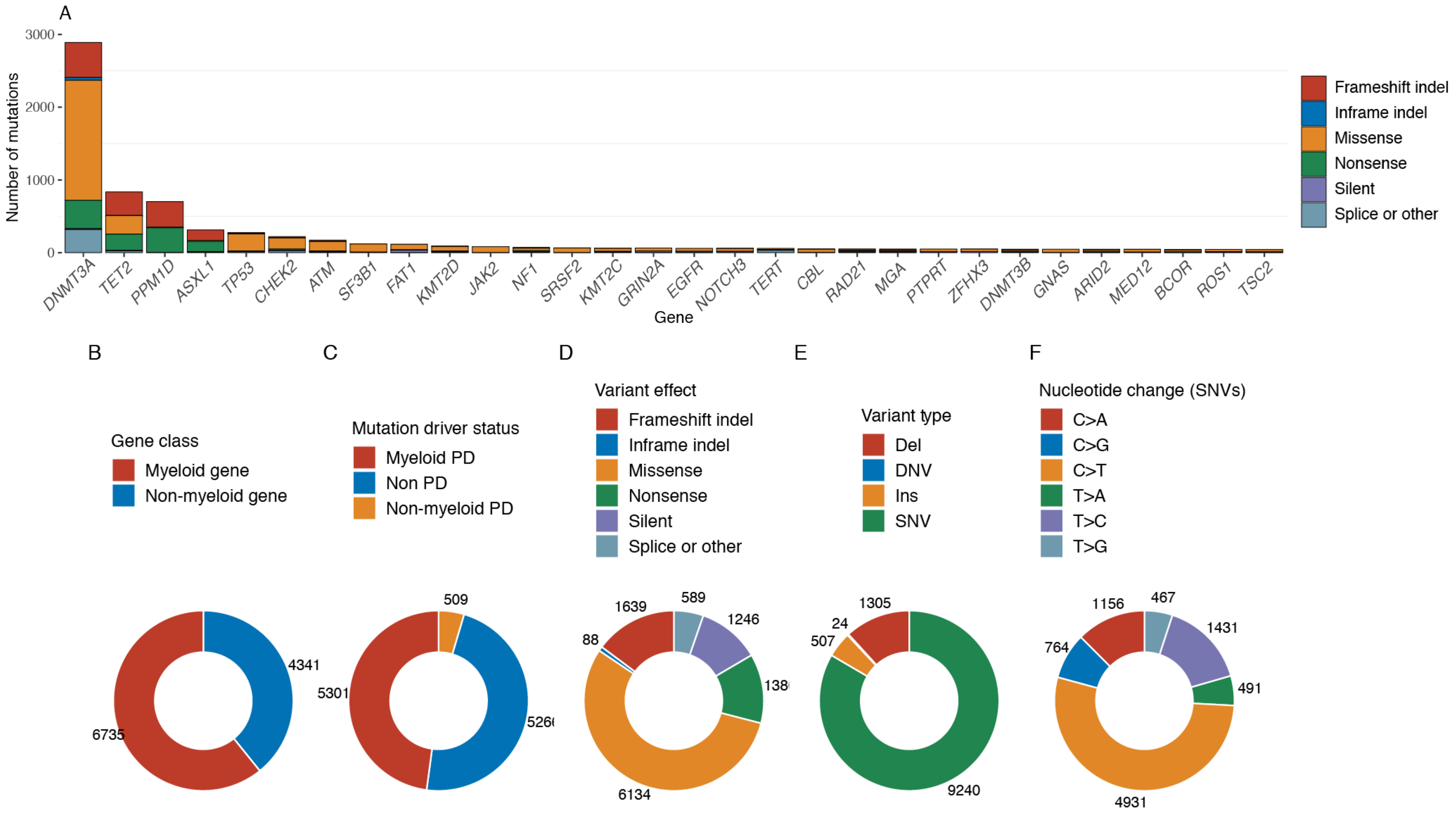
Since we expect most CH mutations to have a true variant allele frequency of less than 10% and since we expect most solid tumors to a range of contamination levels with leukocytes (but generally less than 20%), based on our simulations we used v_b / v_t cutoff of two. However, in the situation where a lymph node with metastatic disease was chosen as the source for tumor material, a high level (greater than 30%) of leukocyte contamination could be present in some cases. Thus when the biopsy site for the tumor was a lymph node, we used v_b / v_t cutoff of 1.5.

REFERENCES

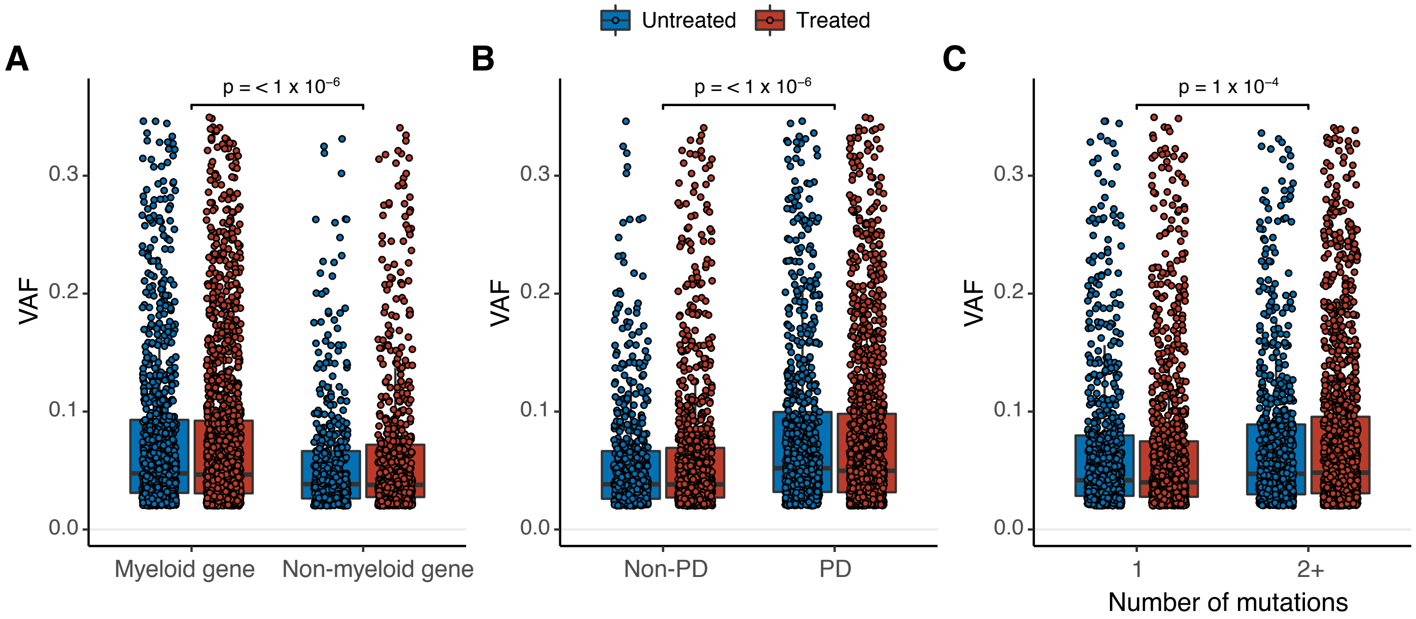
1. Tucker, M. A. *et al.* Leukemia After Therapy With Alkylating Agents for Childhood Cancer2. *JNCI: Journal of the National Cancer Institute* vol. 78 459–464 (1987).
2. Chan, T. A., Hristov, B. & Powell, S. N. Radiation Biology for Radiation Oncologists. *Clinical Radiation Oncology* 15–61 (2017) doi:10.1002/9781119341154.ch2.
3. Gillis, N. K. *et al.* Clonal haemopoiesis and therapy-related myeloid malignancies in elderly patients: a proof-of-concept, case-control study. *Lancet Oncol.* **18**, 112–121 (2017).
4. Takahashi, K. *et al.* Preleukaemic clonal haemopoiesis and risk of therapy-related myeloid neoplasms: a case-control study. *Lancet Oncol.* **18**, 100–111 (2017).
5. Gibson, C. J. *et al.* Clonal Hematopoiesis Associated With Adverse Outcomes After Autologous Stem-Cell Transplantation for Lymphoma. *J. Clin. Oncol.* **35**, 1598–1605 (2017).
6. Young, A. L., Tong, R. S., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis and risk of acute myeloid leukemia. *Haematologica* (2019) doi:10.3324/haematol.2018.215269.
7. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat. Commun.* **7**, 12484 (2016).
8. Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
9. Riboli, E. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *International Journal of Epidemiology* vol. 26 6S–14 (1997).



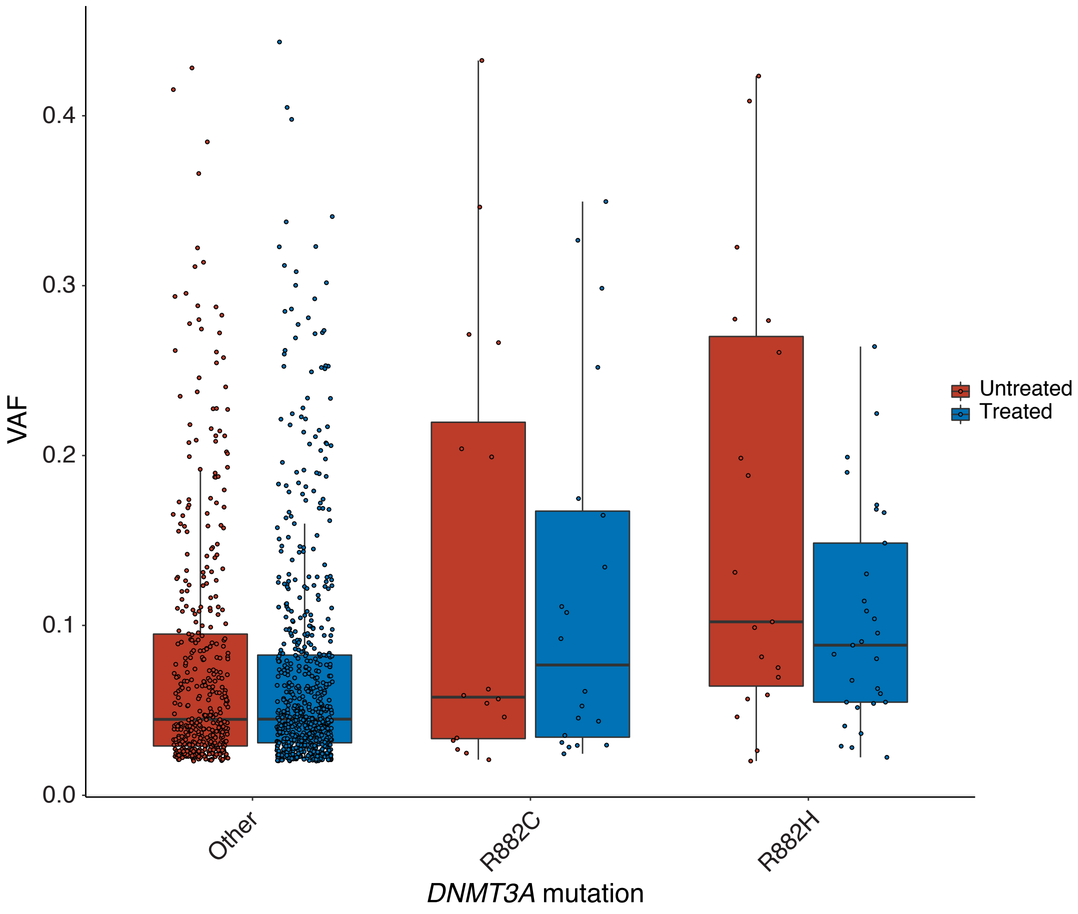
Supplementary Figure 1. Quantification of the extent of natural selection by gene in clonal hematopoiesis using a dN/dS method. Using the dNdScv method (see Methods), we quantified the dN/dS ratios for missense and nonsense and essential splice mutations (truncating) at the level of individual genes. This includes CH mutations from 24,146 solid tumor patients. Shown are the dN/dS ratios for genes mutated at least 25 times showing evidence of significant selection. log(dN/dS)<0 provides evidence of negative selection and log(dN/dS)>0 provides evidence of positive selection.



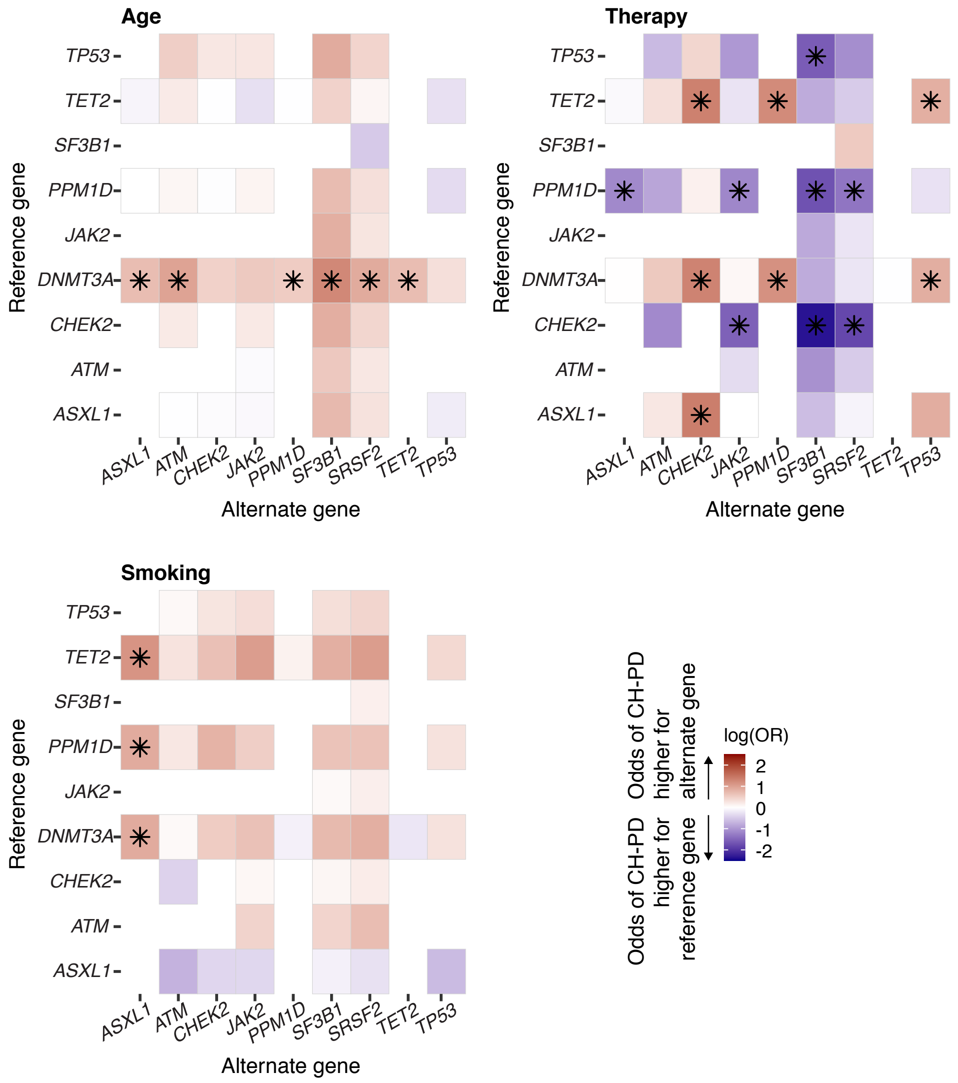
Supplementary Figure 2. Mutational characteristics of CH in 24,146 **solid tumor patients.** (A) Number of mutations observed in the 30 most common genes. (B) Proportion of mutations in a myeloid gene and those not in a myeloid gene. (C) Proportion of mutations considered to be a possible driver of myeloid neoplasms (myeloid PD), a driver of non-myeloid neoplasms (non-myeloid PD) and those not considered to be a possible cancer driver. (D) Proportion of mutations by functional effect. (E). Proportion of deletions (DEL), insertions (INS) or SNVs. (F). Proportion of SNVs with specific nucleotide.



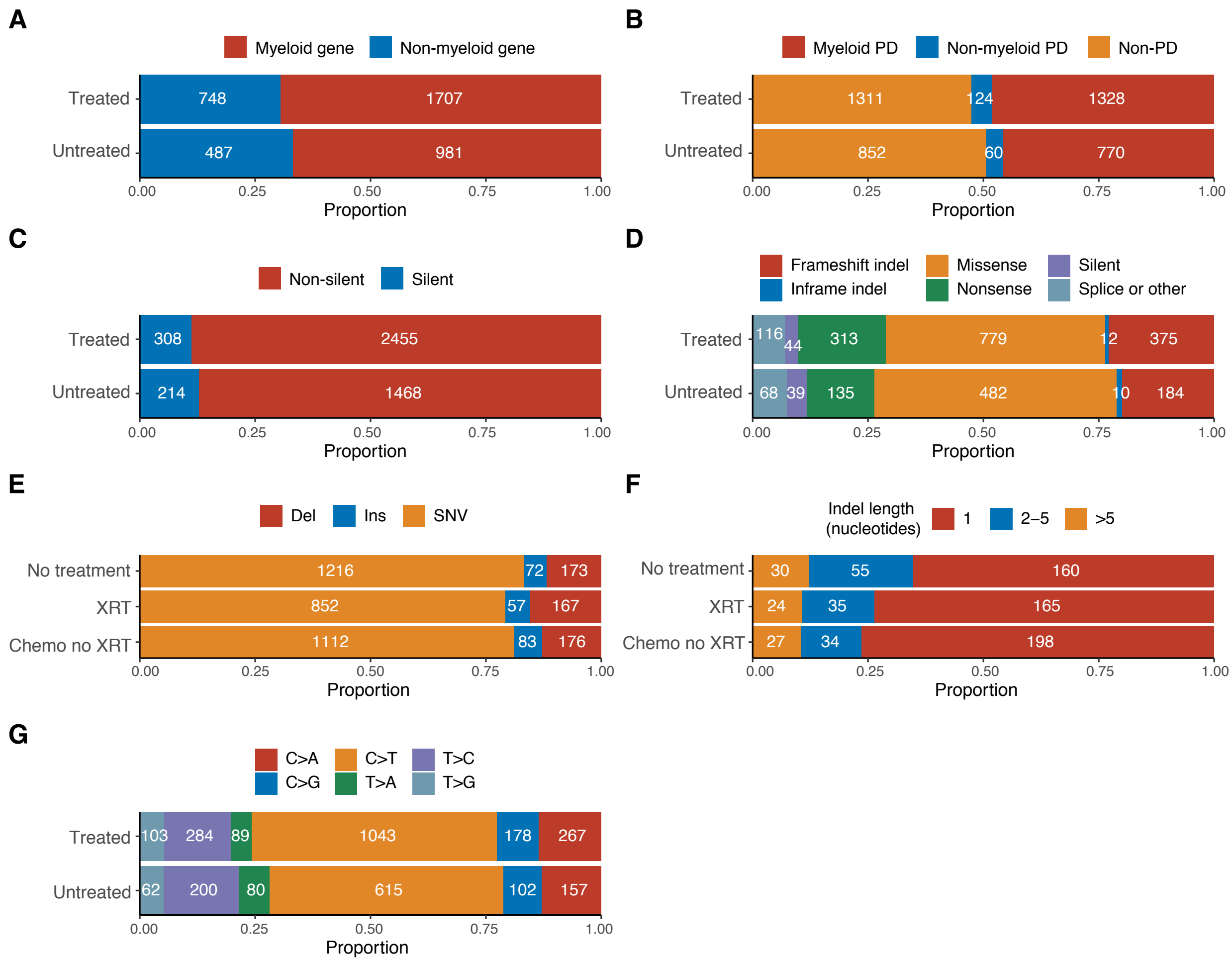
Supplementary Figure 3. Relationship between variant allele fraction (VAF) and CH mutational features. Differences in VAF between: (A) genes recurrently mutated in myeloid neoplasms (myeloid gene) vs. those not implicated in myeloid disease (non-myeloid gene), (B) variants thought to be putative cancer drivers (PD) vs. variants not known to be cancer drivers (non-PD) and (C) individuals with 1 vs. 2+ mutations. P-values were calculated from generalized estimating equations testing for associations between VAF and mutational features adjusted for age, sex, race, smoking history and exposure to oncologic therapy accounting for the within-subject correlation in VAF. n=10,138. Boxplots display interquartile ranges.



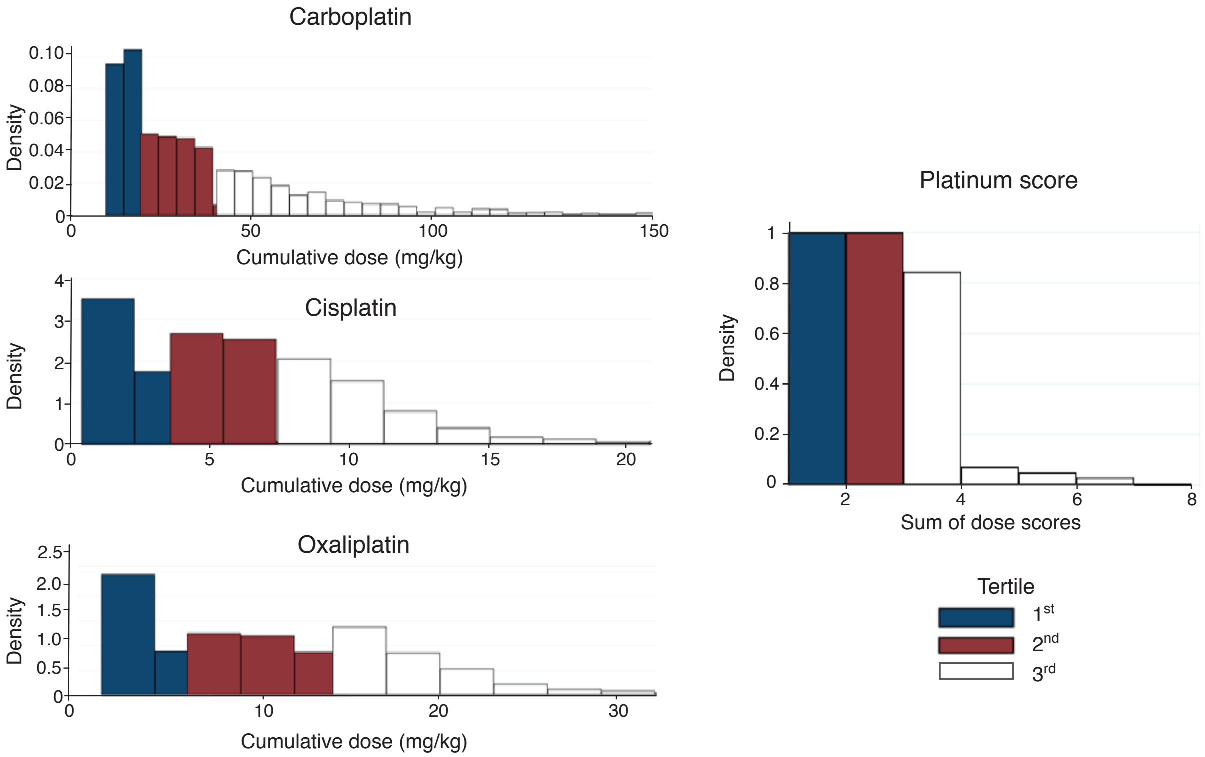
Supplementary Figure 4. Relationship between variant allele fraction (VAF) and *DNMT3A* mutations. p-values were calculated using generalized estimating equations adjusted for age, sex, race, smoking history and exposure to oncologic therapy accounting for the within-subject correlation in VAF. **, p-value 0.01; ***, p<1x10^{-6}, respectively). n=10,138. Boxplots display interquartile ranges.



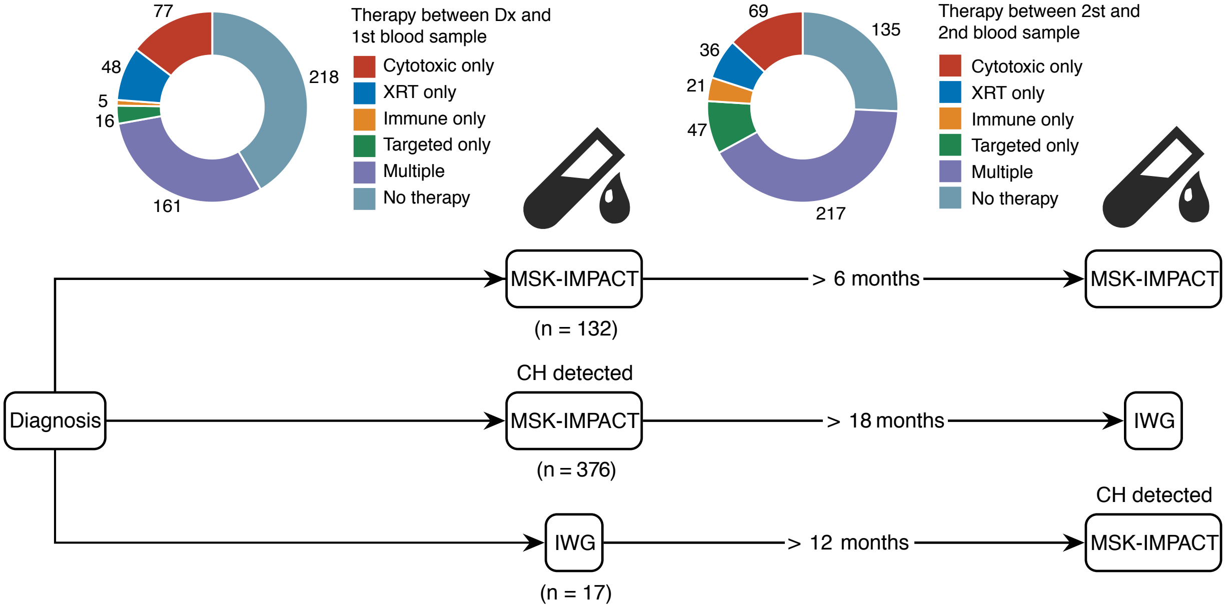
Supplementary Figure 5. Relative odds of CH in specific genes by clinical characteristics. Heatmap of the natural log of the odds ratio for CH-PD in the alternate gene compared to the reference gene from multivariable logistic regression. All combinations (N=45) of the nine most common genes were included. The more common gene from the combination used as the reference. All models were adjusted for age, race, smoking, gender and therapy exposure prior to blood draw. * q (FDR-corrected p-values) <0.05.



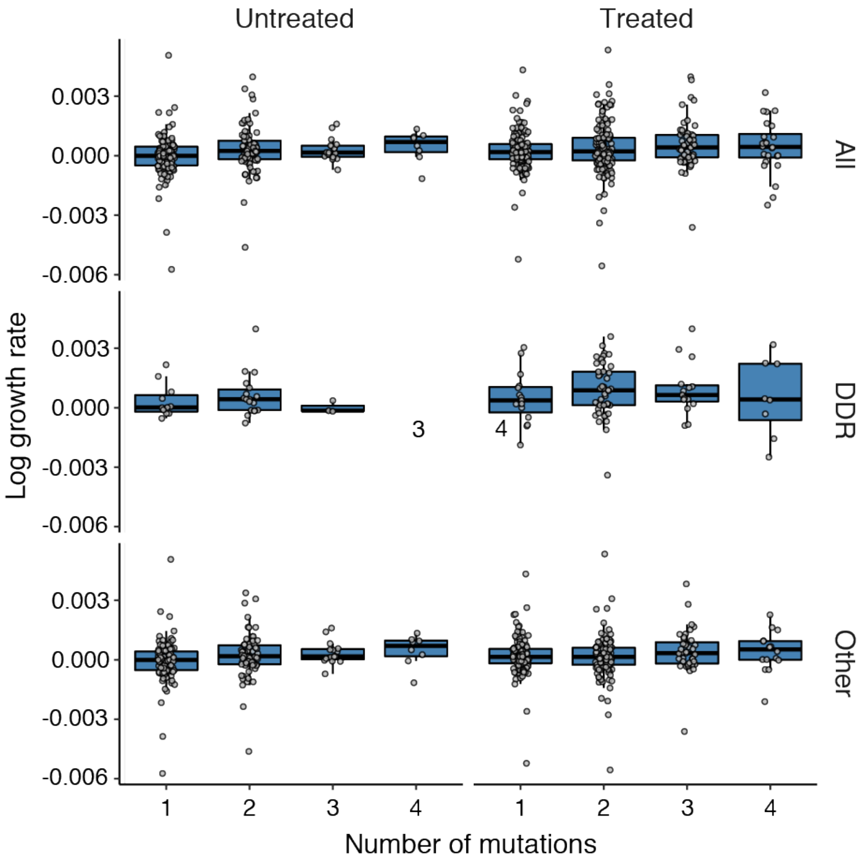
Supplementary Figure 6. Mutational characteristics of CH and prior exposure to cancer therapy in patients with solid tumors. Among patients who received any cancer treatment prior to blood draw for mutational testing (treated) and those who did not receive therapy prior to blood draw (untreated) we compared proportions of: (A) Mutations in a hypothesized myeloid neoplasm driver gene (myeloid gene) vs. those in a gene not known to be a driver of myeloid neoplasms (non-myeloid gene). (B) Mutations considered to be a possible driver of myeloid neoplasms (myeloid PD), a driver of non-myeloid neoplasms (non-myeloid PD) vs. those not considered to be a possible cancer driver (non-PD). (C) Proportion of non-synonymous (non-silent) vs. synonymous (silent) mutations. (D). Proportion of mutations within major functional effect categories. (E) Proportion of deletions (Del), insertions (Ins) or single nucleotide variants (SNV). (F) Proportion of insertions or deletions by the nucleotide length of the alteration. (G) Proportion of SNVs with specific nucleotide changes. n = 10,138.



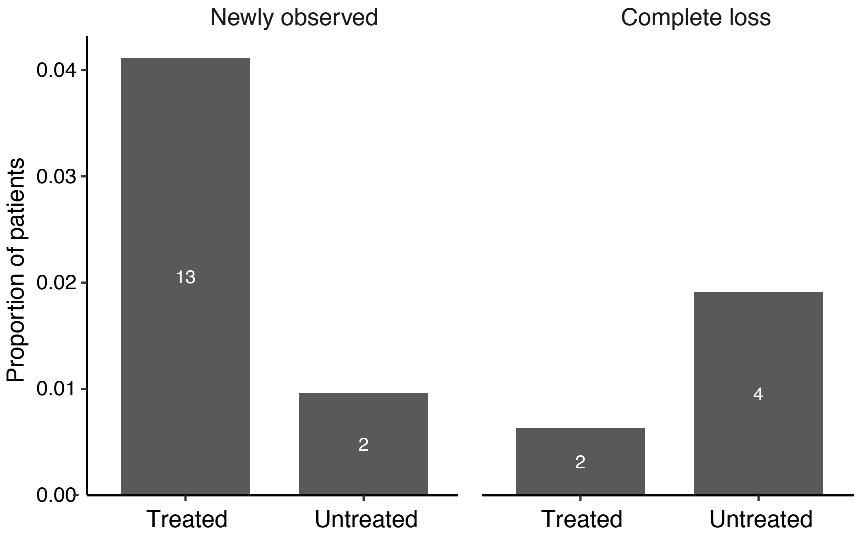
Supplementary Figure 7. Calculation of cumulative exposure for therapy subclasses. For each drug and for all platinum agents, the total dose per kilogram of body weight received prior to blood draw was summed for each patient. The dose distribution for each agent was divided into tertiles and the patient's dose was assigned a score based on tertile.



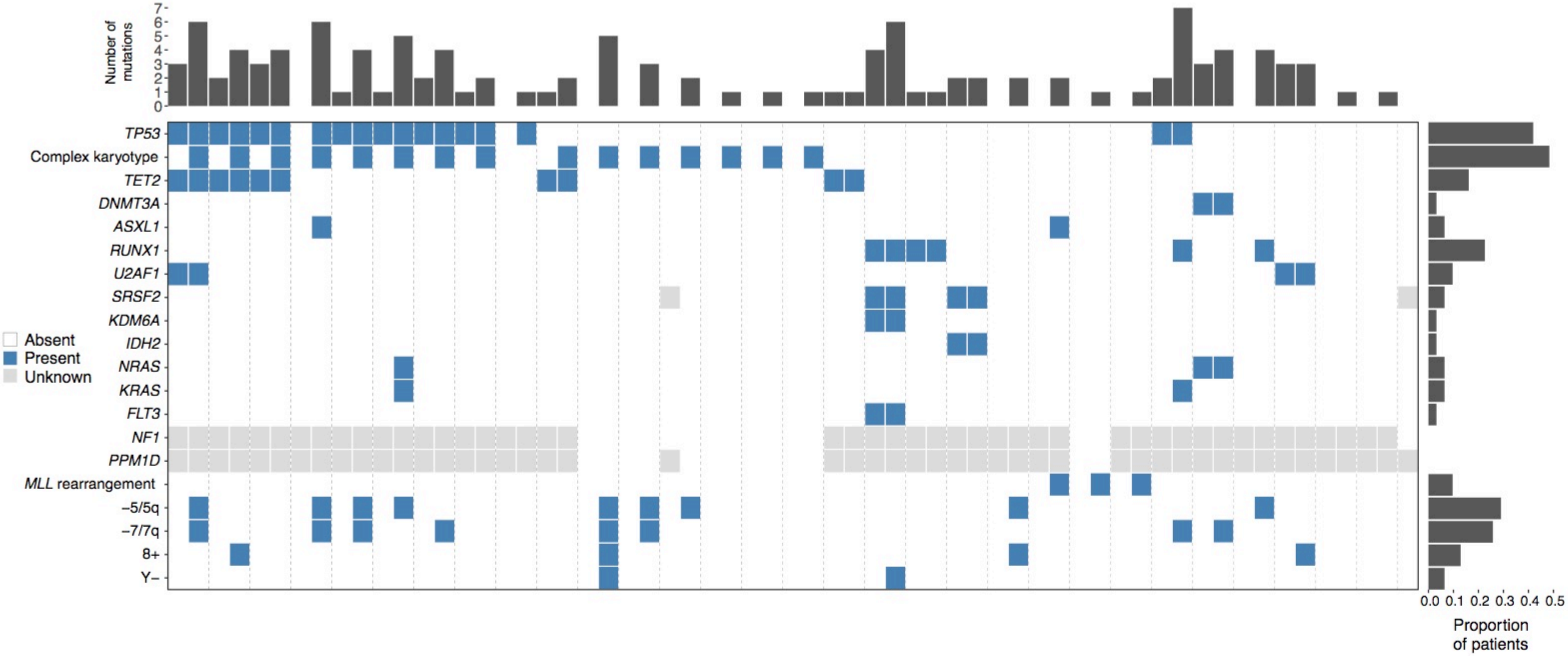
Supplementary Figure 8. Overview of serial samples included in the study relative to MSK-IMPACT testing. Serial samples included in the study, of which at least one was analyzed by MSK-IMPACT testing. Top, treatment received; bottom, timing and type of genomic analysis. n=525.



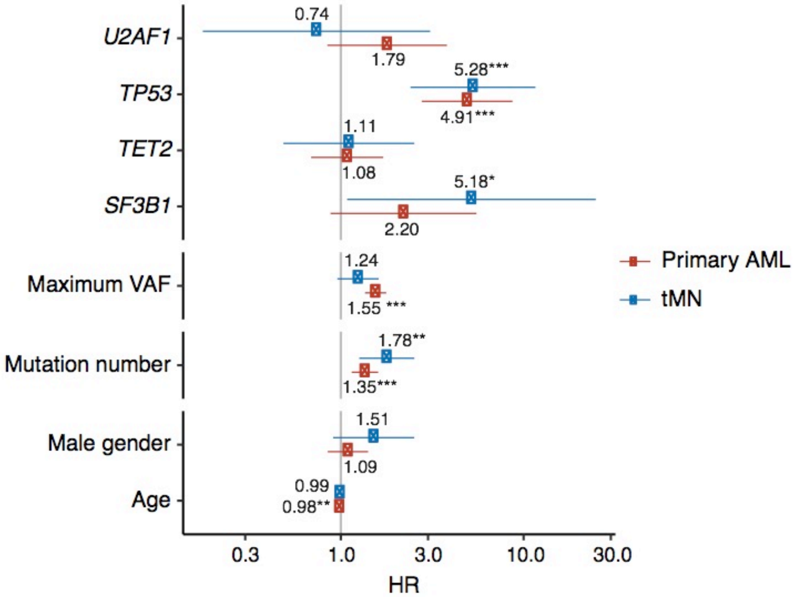
Supplementary Figure 9. Growth rate of CH mutations vs. mutation number. Growth rates for each mutation during follow-up according to the total number of mutations in that individual stratified by receipt of therapy in 525 individuals. Regression using generalized estimating equations was used to test for trend toward increasing growth rate with mutation number among subjects with clonal hematopoiesis adjusted for age, gender, treatment and smoking accounting for correlation between the VAF of mutations in the same person, yielding a p-trend=0.03. Boxplots display interquartile ranges.



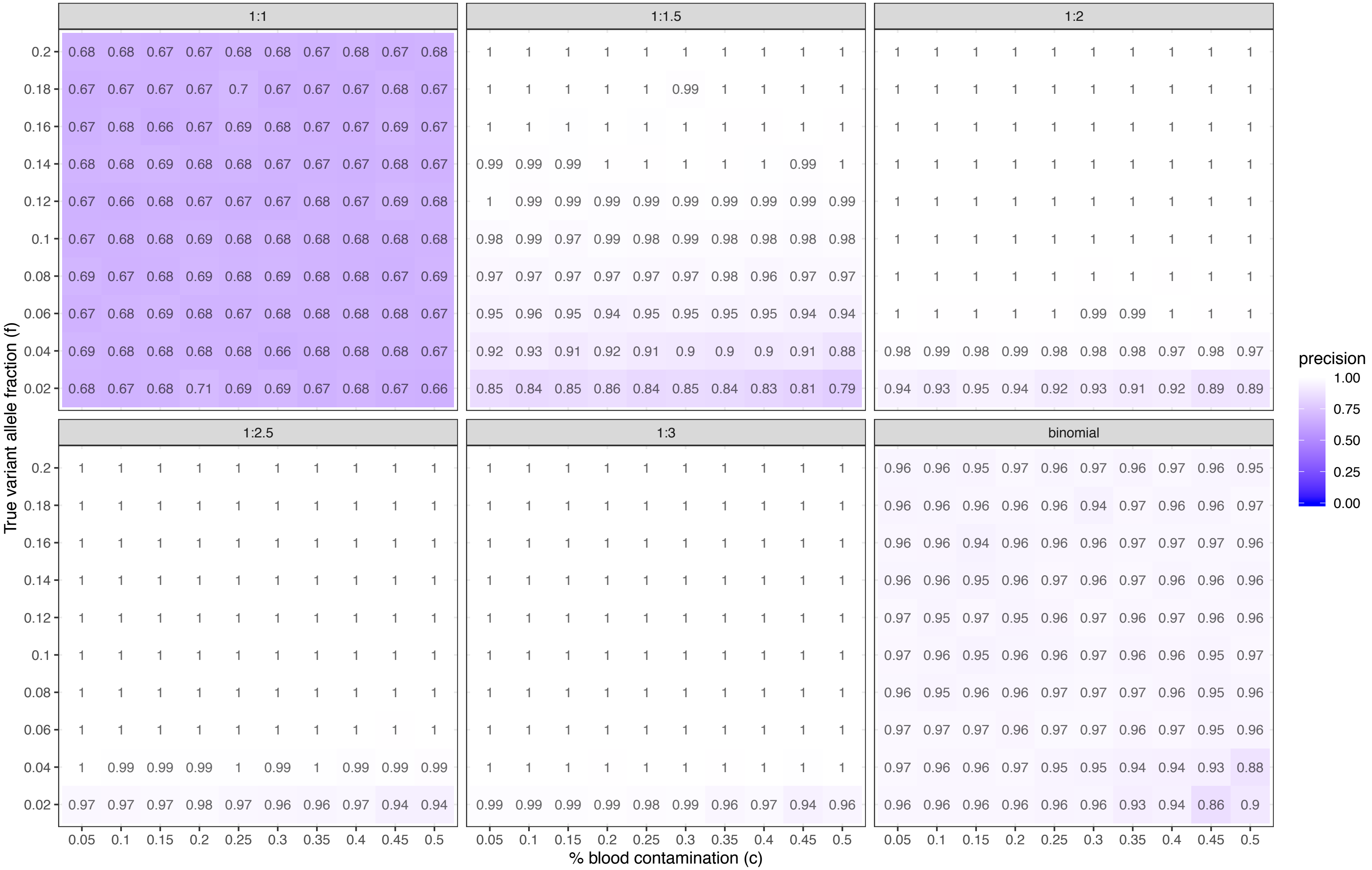
Supplementary Figure 10. Proportion of patients who gained or lost a mutation during follow-up stratified by receipt of therapy. Newly observed, no initial read, VAF ≥ 2% at follow-up; complete loss, initial VAF ≥ 2%, no read at follow-up.



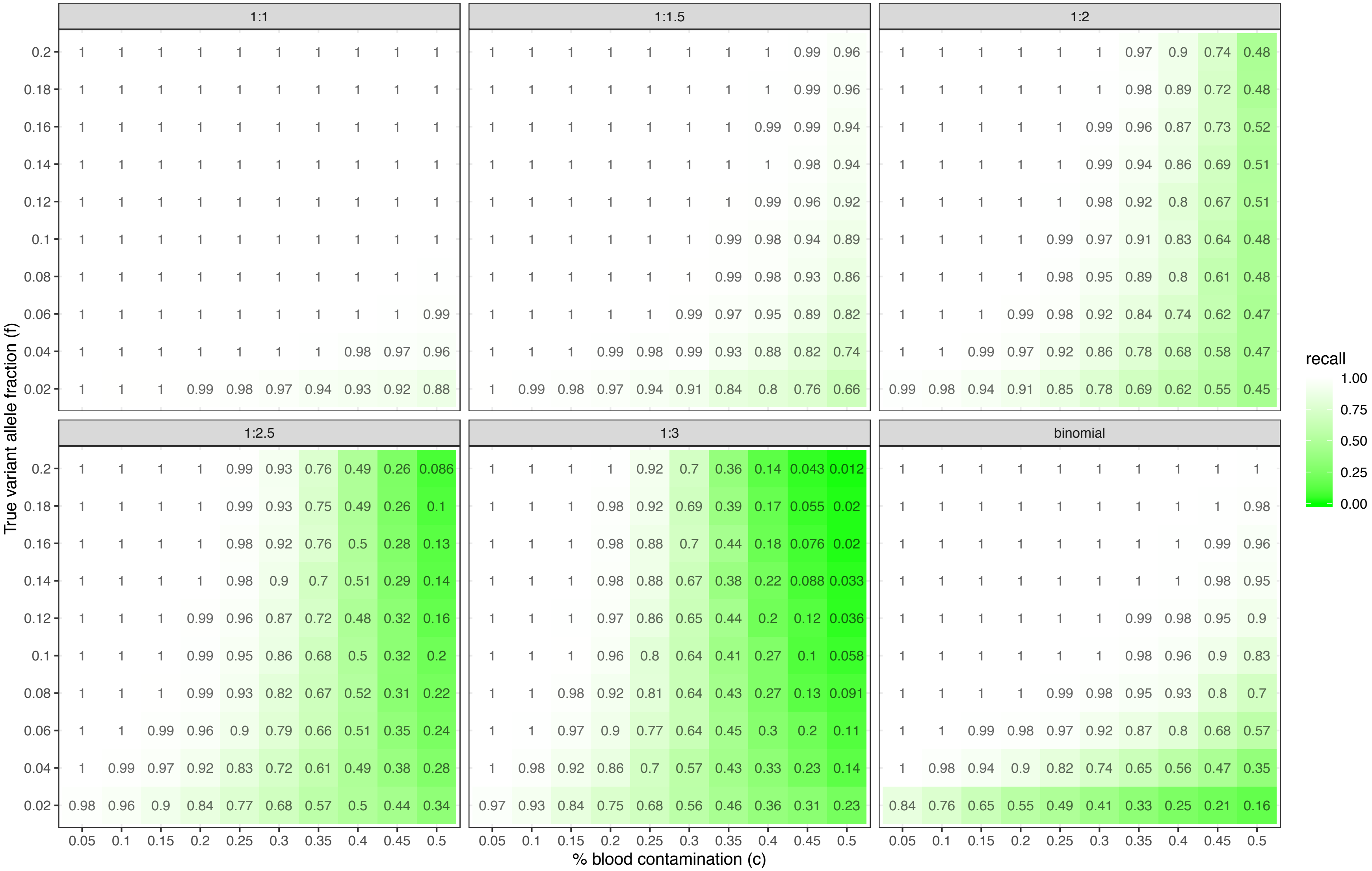
Supplementary Figure 11. Mutation landscape in tMN cases with at least one genetic alteration present at the time of tMN diagnosis. Out of 35 samples with paired pre- and tMN samples, 34 had at least one genetic alteration and are displayed here. N01, pre-tMN sample; -tMN, sample attained at time of tMN diagnosis. Chromosomal abnormalities were not evaluated at the time of pre-tMN testing. Genes included in this analysis are listed in Supplemental Table 5.



Supplementary Figure 12. Risk of myeloid malignancy by clinical and mutational characteristics comparing studies for tMN in solid tumor patients and AML risk in healthy individuals. Hazard ratios and 95% confidence intervals from multivariable Cox regression for including CH mutational characteristics. All models were adjusted for age and gender and stratified by study center. Patients with solid tumors, n=9,437; healthy individuals, n=1,072.



Supplementary Figure 13. Precision of CH calling by simulation. Precision for discrimination of true CH calls from artifacts using a range of cutoffs (1:1, 1:1.5, 1:2, 1:2.5, 1:3) for the ratio of VAF in blood to VAF in tumor and a binomial test of the null hypothesis for an equal VAF in blood and tumor.



1:1	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.2	1	1	1	1	1	1	1	1	1	1
0.18	1	1	1	1	1	1	1	1	1	1
0.16	1	1	1	1	1	1	1	1	1	1
0.14	1	1	1	1	1	1	1	1	1	1
0.12	1	1	1	1	1	1	1	1	1	1
0.1	1	1	1	1	1	1	1	1	1	1
0.08	1	1	1	1	1	1	1	1	1	1
0.06	1	1	1	1	1	1	1	1	1	0.99
0.04	1	1	1	1	1	1	1	0.98	0.97	0.96
0.02	1	1	1	0.99	0.98	0.97	0.94	0.93	0.92	0.88

1:1.5	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.2	1	1	1	1	1	1	1	1	0.99	0.96
0.18	1	1	1	1	1	1	1	1	0.99	0.96
0.16	1	1	1	1	1	1	1	0.99	0.99	0.94
0.14	1	1	1	1	1	1	1	1	0.98	0.94
0.12	1	1	1	1	1	1	1	0.99	0.96	0.92
0.1	1	1	1	1	1	1	0.99	0.98	0.94	0.89
0.08	1	1	1	1	1	1	0.99	0.98	0.93	0.86
0.06	1	1	1	1	1	0.99	0.97	0.95	0.89	0.82
0.04	1	1	1	0.99	0.98	0.99	0.93	0.88	0.82	0.74
0.02	1	0.99	0.98	0.97	0.94	0.91	0.84	0.8	0.76	0.66

1:2	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.2	1	1	1	1	1	1	0.97	0.9	0.74	0.48
0.18	1	1	1	1	1	1	0.98	0.89	0.72	0.48
0.16	1	1	1	1	1	0.99	0.96	0.87	0.73	0.52
0.14	1	1	1	1	1	0.99	0.94	0.86	0.69	0.51
0.12	1	1	1	1	1	0.98	0.92	0.8	0.67	0.51
0.1	1	1	1	1	0.99	0.97	0.91	0.83	0.64	0.48
0.08	1	1	1	1	0.98	0.95	0.89	0.8	0.61	0.48
0.06	1	1	1	0.99	0.98	0.92	0.84	0.74	0.62	0.47
0.04	1	1	0.99	0.97	0.92	0.86	0.78	0.68	0.58	0.47
0.02	0.99	0.98	0.94	0.91	0.85	0.78	0.69	0.62	0.55	0.45

1:2.5	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.2	1	1	1	1	0.99	0.93	0.76	0.49	0.26	0.086
0.18	1	1	1	1	0.99	0.93	0.75	0.49	0.26	0.1
0.16	1	1	1	1	0.98	0.92	0.76	0.5	0.28	0.13
0.14	1	1	1	1	0.98	0.9	0.7	0.51	0.29	0.14
0.12	1	1	1	0.99	0.96	0.87	0.72	0.48	0.32	0.16
0.1	1	1	1	0.99	0.95	0.86	0.68	0.5	0.32	0.2
0.08	1	1	1	0.99	0.93	0.82	0.67	0.52	0.31	0.22
0.06	1	1	0.99	0.96	0.9	0.79	0.66	0.51	0.35	0.24
0.04	1	0.99	0.97	0.92	0.83	0.72	0.61	0.49	0.38	0.28
0.02	0.98	0.96	0.9	0.84	0.77	0.68	0.57	0.5	0.44	0.34

1:3	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.2	1	1	1	1	0.92	0.7	0.36	0.14	0.043	0.012
0.18	1	1	1	0.98	0.92	0.69	0.39	0.17	0.055	0.02
0.16	1	1	1	0.98	0.88	0.7	0.44	0.18	0.076	0.02
0.14	1	1	1	0.98	0.88	0.67	0.38	0.22	0.088	0.033
0.12	1	1	1	0.97	0.86	0.65	0.44	0.2	0.12	0.036
0.1	1	1	1	0.96	0.8	0.64	0.41	0.27	0.1	0.058
0.08	1	1	0.98	0.92	0.81	0.64	0.43	0.27	0.13	0.091
0.06	1	1	0.97	0.9	0.77	0.64	0.45	0.3	0.2	0.11
0.04	1	0.98	0.92	0.86	0.7	0.57	0.43	0.33	0.23	0.14
0.02	0.97	0.93	0.84	0.75	0.68	0.56	0.46	0.36	0.31	0.23

binomial	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.2	1	1	1	1	1	1	1	1	1	1
0.18	1	1	1	1	1	1	1	1	1	0.98
0.16	1	1	1	1	1	1	1	1	0.99	0.96
0.14	1	1	1	1	1	1	1	1	0.98	0.95
0.12	1	1	1	1	1	1	0.99	0.98	0.95	0.9
0.1	1	1	1	1	1	1	0.98	0.96	0.9	0.83
0.08	1	1	1	1	0.99	0.98	0.95	0.93	0.8	0.7
0.06	1	1	0.99	0.98	0.97	0.92	0.87	0.8	0.68	0.57
0.04	1	0.98	0.94	0.9	0.82	0.74	0.65	0.56	0.47	0.35
0.02	0.84	0.76	0.65	0.55	0.49	0.41	0.33	0.25	0.21	0.16

Supplementary Figure 14. Recall of CH calling by simulation. Recall for discrimination of true CH calls from artifacts using a range of cutoffs (1:1, 1:1.5, 1:2, 1:2.5, 1:3) for the ratio of VAF in blood to VAF in tumor and a binomial test of the null hypothesis for an equal VAF in blood and tumor.