Cancer therapy shapes the fitness landscape of clonal hematopoiesis

Kelly L. Bolton1, Ryan N. Ptashkin2,4, Teng Gao3,4, Lior Braunstein4, Sean M. Devlin5, Daniel Kelly6, Minal Patel7, Antonin Berthon3, Aijazuddin Syed2, Mariko Yabe8, Catherine C. Coombs9, Nicole M. Caltabellotta2, Mike Walsh10, Kenneth Offit10, Zsófia Stadler11, Diana Mandelker2, Jessica Schulman7, Akshar Patel7, John Philip12, Elsa Bernard3, Guine Gundem3, Juan E. Arango Ossa4, Max Levine13, Juan S. Medina Martinez12, Noushin Farnoud7, Dominik Glodzik3, Sonya Li10, Mark E. Robson10, Choonsik Lee14, Paul D. P. Pharoah15,16, Konrad H. Stopsack10, Barbara Spitzer13, Simon Mantha17, James Fagin10,18, Laura Boucai19, Christopher J. Gibson20, Benjamin L. Ebert20, Andrew L. Young21, Todd Druley22, Koichi Takahashi23, Nancy Gillis24,25, Markus Ball25,26, Eric Padron25, David M. Hyman10,27, Jose Baselga28, Larry Norton10,27, Stuart Gardos10,27, Virginia M. Klimke10,27, Howard Scher10,27, Dean Bajorin10,27, Eder Paraiso19,29, Ryma Benayed2, Maria E. Arcila2, Marc Ladanyi2, David B. Solit20,29,30, Michael F. Berger20,29,30, Martin Tallman1, Montserrat Garcia-Closas14, Nilanjan Chatterjee31, Luis A. Diaz Jr10,32,33, Ross L. Levine10, Lindsay M. Morton14, Ahmet Zehir2,3,35 and Elli Papamannui3,34,35

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, PPMID, 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 development1. Each mutation contributes to a clone’s fitness advantage, resulting in clonal expansions that culminate in malignant transformation, in a process that parallels Darwinian evolution2. 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 disease4–11. However, our understanding of the evolutionary trajectories that underlie cancer development is primarily based on retrospective modeling of clonal structures observed at diagnosis1 or disease progression4. 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 aging12–14. 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 studies15–19. 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 neoplasms10,21. 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)12,23. tMN was traditionally thought to develop from the mutagenic effects of cancer therapy19. However, recent studies show that tMN-initiating mutations can predate cancer therapy19, consistent with CH14. Here, we sought to characterize the relationships between CH

NATURE GENETICS | VOL 52 | NOVEMBER 2020 | 1219–1226 | www.nature.com/naturegenetics
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 and recurrence in an in-house dataset of myeloid neoplasms (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, 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, older age strongly correlated with the presence of CH clones in patients with cancer (odds ratio (OR) = 1.9, P < 10^{-10}) (Extended Data Table 3). CH was less common in patients of Asian ancestry relative to white ancestry (OR = 0.7, P = 1 × 10^{-3}) (Extended Data Table 3), consistent with recent reports.

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 × 10^{-4}). The same was true for current and former smokers (OR = 1.1, P = 5 × 10^{-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 × 10^{-4}). 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 (OR_{het} = 3.6, Q (false discovery rate (FDR)-corrected) P value) = 7 × 10^{-6}; OR_{het} = 5.0, Q < 10^{-4}) (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 (OR = 2.8, Q = 2 × 10^{-4}; OR_{het} = 4.3, Q < 10^{-4}; OR_{het} = 4.5, Q = 6 × 10^{-4}; 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 × 10^{-4}; Fig. 1c). Current smokers had a stronger association with CH in ASXL1 (OR = 3.1, P = 1 × 10^{-3}) compared with former smokers (OR = 2.4, P = 1 × 10^{-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, all of treatment modalities, external beam radiation therapy (OR = 1.4, P < 10^{-4}), cytotoxic chemotherapy (OR = 1.2, P = 2 × 10^{-3}) and radionuclide therapy (OR = 1.6, P = 0.01) were most strongly associated with CH-PD (global test of heterogeneity, P_{het} = 0.03). With respect to subclasses of cytotoxic therapy, CH-PD was most strongly associated with previous
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 \( \text{PPM1D} \) were most strongly associated with previous exposure to platinum (OR = 3.2, \( Q < 10^{-4} \)) or radionuclide therapy (OR = 6.2, \( Q = 7 \times 10^{-4} \)) 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 \( \text{TP53} \) 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 \( \text{CHEK2} \) 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 (\( P = 4 \times 10^{-4} \)) 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 = \( 3 \times 10^{-9} \)) 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 \( \geq 2\% \), 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.
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. 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, EQD2). 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. *P < 0.05, **P < 0.01, ***P < 0.001.

Genetic and clonal evolution to tMN
Recent studies have shown that tMN-initiating mutations can predate cancer therapy21, challenging the traditional hypothesis that tMN develops from the mutagenic effects of cancer therapy22 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).
We called mutations present at a VAF of $\geq 2\%$ 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 disease27,30–32 (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 of CH.
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, 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, \( P < 10^{-3} \), 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. 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, 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 and other myeloid neoplasms.

**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). \( * P < 0.05, ** P < 0.01, *** P < 0.001 \), \( **** P < 0.0001 \).

**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. \( n = 9,437 \).
(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\(^\text{45}\). 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 splicesome 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\(^\text{46,47}\). 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\(^\text{48,49}\) 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](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* 6, 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. Martinezcorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science 362*, 911–917 (2018).
9. Martinezcorena, 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. Martinezcorena, 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 hematopoesis 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 clonal manipulation. *Cell Stem Cell 25*, e6 (2019).
18. Hsu, J. L. 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 leukemia. *Nature 518*, 552–555 (2015).
20. Abelson, S. et al. Prediction of acute myeloid leukemia 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. Lindsey, R. C. et al. Acute myeloid leukemia ontology 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 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 haemopoiesis 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 a network on secondary leukaemias. *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. Israe1, M. et al. Microbial signals drive pre-leukaemia 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).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020
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 in 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 required to have had 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 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 sampling; second, from 21 samples for repeat CH mutation testing. These came from three sources: first, from collected additional blood samples on patients sequenced using MSK-IMPACT is the comprehensiveness of the clinical data, including therapeutic exposure data, that were obtained as detailed in the Supplementary Note. 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 NFI, DNMT3A, TET2, IKZF1, RAD21, WT1, KMT2D, SH2B3, TP53, CEBPA, ASXL1, RUNX1, BCR, KDM6A, STAG2, PHH2, KMT2C, PPM1D, ATM, ARID1A, ARID2, ASXL2, CBL, CREBBP, ETV6, EZH2, FBXW7, MGA, MPL, RBL, SETD2, SUZ12 or ZRS2 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, CLEC2, 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 7,880 individuals with hematologic malignancies ≥2 times. We annotated variants as CH-PD if they fulfilled any of the following criteria: (1) any variant noted as oncogenic or likely oncogenic in OncoKB; (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; 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.

**Calculation of dN/dS ratios.** We used the dNdScv (https://github.com/im3sagner/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 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 multivariate logistic regression to evaluate for an association between CH (indications as well as 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.

**Calculation of dN/dS ratios.** We used the dNdScv (https://github.com/im3sagner/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 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 multivariate logistic regression to evaluate for an association between CH (indications as well as 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.

**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 NFI, DNMT3A, TET2, IKZF1, RAD21, WT1, KMT2D, SH2B3, TP53, CEBPA, ASXL1, RUNX1, BCR, KDM6A, STAG2, PHH2, KMT2C, PPM1D, ATM, ARID1A, ARID2, ASXL2, CBL, CREBBP, ETV6, EZH2, FBXW7, MGA, MPL, RBL, SETD2, SUZ12 or ZRS2 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, CLEC2, 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 7,880 individuals with hematologic malignancies ≥2 times. We annotated variants as CH-PD if they fulfilled any of the following criteria: (1) any variant noted as oncogenic or likely oncogenic in OncoKB; (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; 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.*

**Mutation calls.** 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 removed deletions and in-frame insertions if the insertion 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.
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 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 were adjusted for differences in strength of the associations 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 was used to test for an 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 sampling 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 T₀ indicate the age of the individual (in days) at the two measurement time points and V and V₀ correspond to the VAF at T and T₀ respectively. We also classified mutations as having increased, decreased or remained with non-DDR mutations within individuals who received cytotoxic therapy. When patients had more than two mutations in the same gene category, we used a test for correlation between the growth rate of mutations in the same person. Among patients with at least one mutation in a DDR 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 in 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 M0F (n = 68); Takahashi et al., abbreviated MDA (n = 67); and Gibson et al., abbreviated DPC (n = 401), studying the effect of CH on tMN risk in patients with cancer. We defined tMN as an AML or MDS diagnosed following exposure to therapeutic radiation or cytotoxic therapy as per the World Health Organization criteria49. 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 M0F. 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 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 with recently diagnosed cancer, 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 cohort 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 package46 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 multivariable 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 SEER51; (4) previously published HR estimates for chemotherapy on the risk of tMN in women with breast cancer from the NCCN50; (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 SEER51. While our IMPACT cohort is not representative of the 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 in blood samples are known to differ by sex, we chose to use the distribution of blood counts from the entire treatment-naïve 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. The 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 (https://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. 45, D941–D947 (2019).
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 B.L., grant no. P50 CA172012 to J.E., grant no. U01 H509315 to V.M.K.), the American Society of Hematology (K.L.B. and E. Papaemmanuel), the EvansMDS Foundation (K.L.B.), the European Hematology Association (E. Papaemmanuel), Gabrielle’s Angels Foundation (E. Papaemmanuel), the V Foundation (E. Papaemmanuel), the Geoffery Beene Foundation (E. Papaemmanuel), 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. Papaemmanuel 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. A.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. A.S., 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 CA08748), 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. Papaemmanuel 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., T.G., S.M.D., A.B., A.S., M.Y., C.C.C., N.M.C., M.W., K.O., Z.S., D.M., J.S., A.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.G., R.L.E., A.L.Y., T.D., N.G., M.B., E. Padron, D.M.H., J.B., L.N., S.G., V.M.K., H.S., D.B., E. Parizio, R.B., M.E.A., M. Ladanyi, D.B.S., M.F.B., M.S.T., M.G.-C., N.C., I.A.D., R.L.L., L.M.M., A.Z. and E. Papaemmanuel 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, Lexo, H3 Biomedicine, Medcape, Octapharma and Pharmacy; and has served as a consultant for AbbVie, Covance, Cowen & Co. and Dendreon Group; and has received institutional research funding from AROG, Gilead, Lexo, 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, Alliera Sciences, Bioramarin, Fortessa 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 Lexo, Bayer, Puma and AstraZeneca. J.P. 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, Galr, 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 Lexo Oncology and Heliosn Therapeutics. D.B.S. has served as a consultant for or received honoraria from Pfizer, Lexo Oncology, Lilly Oncology, Illumina and Vividion 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, Cellarant, Orsenix, ADC Therapeutics and Biosight; serves on the advisory boards of Daiichi-Sankyo, KAHr, rigid, Nobla, 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 (PGDi) and Jounce Therapeutics; is a paid consultant to PGDX and Neoprobe; 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 PGDXs, Jounce Therapeutics, Thrive Earlier Detection and Neoprobe; 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 Lexo, Imagine, C4 Therapeutics and Isolexius, which include equity interest; receives research support from and has consulted for Celgene and Roche and has consulted for Lilly, Jansen, 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. Papaemmanuel receives research funding from Celgene and is a cofounder in Isoplexis, which include equity interest; receives research support from and has consulted for Gilead and has received institutional research funding from AstraZeneca, personal genome diagnostics and mismatch repair deficiency. E. Papaemmanuel 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. A.S., 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 CA08748). 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.

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. Papaemmanuel. 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%)  |
| Ependymoma tumor         | 26 (90%)   | 3 (10%)    |
| Esophageogastric carcinoma| 464 (70%) | 196 (30%)  |
| Ewing sarcoma            | 66 (89%)   | 8 (11%)    |
| Gastrointestinal neuroendocrine tumor | 73 (68%) | 34 (32%) |
Extended Data Table 1 | Clinical characteristics of solid tumor patients assessed for CH (continued)

| Tumor Type                              | 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%)   |
| Gioma                                   | 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    | <1×10⁻⁶ |
| 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  | 1.1-1.2    | 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    | 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  | <1×10⁻⁶ |
| 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 | 1.1–1.2  | 0.000038|

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 | <1x10^-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.
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.

- 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, χ²) 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 | Methods |
|---------------------------------|---------|
| n/a                             | n/a     |
| ☒ Antibodies                    | ☒ ChIP-seq |
| ☒ Eukaryotic cell lines         | ☒ Flow cytometry |
| ☒ Palaeontology                 | ☒ MRI-based neuroimaging |
| ☒ Animals and other organisms   |         |
| ☒ Human research participants   |         |
| ☒ Clinical data                 |         |

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.
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

| Clinical trial registration | NCT01775072 |
|-----------------------------|-------------|

Study protocol

| Study protocol | Details available at ClinicalTrials.gov NCT01775072 or upon request. |
|----------------|------------------------------------------------------------------|

Data collection

| 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.