nature communications

Article

https://doi.org/10.1038/s41467-025-57077-1

A genotype-first approach identifies high incidence of *NF1* pathogenic variants with distinct disease associations

Received: 30 August 2023

Accepted: 10 February 2025

Published online: 01 April 2025

Check for updates

Anton Safonov[1,2,14], Tomoki T. Nomakuchi[3,14], Elizabeth Chao[4], Carrie Horton[4], Jill S. Dolinsky[4], Amal Yussuf[4], Marcy Richardson[4], Virginia Speare[4], Shuwei Li[4], Zoe C. Bogus[5], Maria Bonanni[5], Anna Raper[5], Trust Odia[5,6], Bradley S. Wubbenhorst[5], Elsa Faulders[7], Elisabeth M. Schuth[8], Kate Loranger[9], Jingwen Zhang[9], Carly Bess Scalise[9], Adam ElNaggar[9], Youbao Sha[9], Stephanie A. Felker[10,11], Jeffrey Weitzel[12], Staci Kallish[3,5], Marylyn D. Ritchie[6], Penn Medicine BioBank*, Katherine L. Nathanson[5,6,13,15] ✉ & Theodore G. Drivas[5,6,15] ✉

Loss of function variants in the *NF1* gene cause neurofibromatosis type 1, a genetic disorder characterized by complete penetrance, characteristic physical exam findings, and a substantially increased risk for malignancy. However, our understanding of the disorder is based on patients ascertained through phenotype-first approaches, which estimate prevalence at 1 in 3000. Leveraging a genotype-first approach in multiple large patient cohorts including over one million individuals, we demonstrate an unexpectedly high prevalence (1 in 1,286) of *NF1* pathogenic variants. Half are identified in individuals lacking clinical features of NF1, with many appearing to have post-zygotic mosaicism for the identified variant. Incidentally discovered variants are not associated with classic neurofibromatosis features but are associated with an increased incidence of malignancy compared to control populations. Our findings suggest that *NF1* pathogenic variants are substantially more common than previously thought, often characterized by somatic mosaicism and reduced penetrance, and are important contributors to cancer risk in the general population.

The human disease gene *NF1* encodes the protein neurofibromin, a tumor suppressor and negative regulator of the RAS/MAPK pathway[1–4]. Heterozygous loss of function variants within the *NF1* gene lead to the autosomal dominant disorder neurofibromatosis type 1 (NF1), with an estimated prevalence of 1 in 3000[1–4]. NF1 is often cited as a classic example of a pleiotropic genetic disorder with variable expressivity and complete penetrance; that is, every individual with a germline pathogenic *NF1* variant is expected to meet NF1 diagnostic criteria, although their specific presentation and degree of organ involvement may vary[2–5]. The classic features as defined by the NIH clinical diagnostic criteria include: cutaneous café-au-lait macules (CALMs), neurofibromas, axillary/inguinal freckling, iris hamartoma, and certain skeletal anomalies[6]. Affected individuals also are at increased risk for hypertension, cardiovascular anomalies, and certain malignancies, making early diagnosis and screening a critical aspect of care[4].

The malignancies classically seen in NF1 patients are of neural crest derivation, however patients with NF1 also have been observed to have increased risk for a broad range of different malignancies, often

A full list of affiliations appears at the end of the paper. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: knathans@upenn.edu; theodore.drivas@pennmedicine.upenn.edu

with distinctive disease behavior and prognosis[7,8]. For example, breast cancers in female patients with NF1 are thought to be characterized by an earlier age of onset, increased mortality, and unfavorable prognostic factors, such as estrogen/progesterone receptor negativity and *HER2* amplification[8–10]. With the advent of widespread tumor sequencing in cancer patients without NF1, it has become clear that somatic *NF1* driver mutations are also common in certain tumors, including melanoma, glioblastoma, and breast and ovarian cancer[11].

Somatic mosaic *NF1* variants also have rarely been reported in clonal hematopoiesis (CH)[12,13], an age- and oncologic treatment-related phenomenon characterized by the presence of clonal expanded, genetically distinct subpopulations of the hematopoietic lineage within a single patient[12,14], and in the condition known as segmental neurofibromatosis, a subtype of NF1 with a reported incidence of roughly 1 in 75,000[15]. Segmental NF1 results from a somatic post-zygotic mutation arising early during embryogenesis, with the resultant mutant cell lineage going on to populate limited areas of the body which manifest as foci of affected tissue displaying classic NF1-associated features, and with the causal *NF1* variant detectable only in affected cells. On the other hand, CH results from somatic mutations occurring late in life limited to the blood, and is therefore not expected to be associated with classic NF1 presentations[16,17].

With the advent of broadly applied next-generation sequencing technologies and multi-gene panel testing (MGPT), incidental, mosaic, and otherwise unexpected genetic findings are more commonly identified[18–20]; the finding of an incidental germline pathogenic *NF1* variant in a patient with breast cancer without a clinical NF1 diagnosis would be one such example. Similarly, as the scientific community develops large-scale population-level biobanks[21,22], previously undiagnosed individuals with pathogenic variants (PVs) in disease genes will be identified[20,23]. Medical management of patients with incidentally discovered genetic variants is uncertain as we lack guidelines for care and counseling, and the clinical relevance of an incidentally discovered variant in the absence of a congruent clinical phenotype is unknown.

To investigate the prevalence of *NF1* PVs on a population-scale, we evaluated two cohorts of individuals from independent datasets: the population-level Penn Medicine BioBank[24] (PMBB, $n = 43{,}731$) and a database of patients clinically sequenced for cancer risk evaluation by Ambry Genetics ($n = 118{,}768$). We identified an unexpectedly high prevalence (1 in 450–750) of PVs in the *NF1* gene, more than four times the rate expected given the reported prevalence of NF1. Half of the individuals with *NF1* PVs lacked any evidence of syndromic NF1, and many, but not all, appeared to be mosaic for the identified *NF1* PV. The discovery of an incidental *NF1* PV was not predictive of the presence of classic symptoms of NF1 but was associated with a significantly greater incidence of several malignancies, which was replicated in three additional data sets. Our findings suggest that *NF1* PVs are substantially more common than previously thought, often characterized by somatic mosaicism and reduced penetrance, and are important contributors to cancer risk in the general population.

Results

Patients with incidentally discovered NF1 pathogenic variants have no evidence of NF1

Physicians in the University of Pennsylvania Division of Translational Medicine and Human Genetics evaluated four patients, Cases 1–4 (Supplementary Notes, Supplementary Data 1), for NF1, all of whom had been incidentally diagnosed with an *NF1* pathogenic/likely pathogenic variant (*NF1* PVs). Comprehensive physical exam, medical history, and family history revealed no or very few features consistent with an NF1 diagnosis. None of the four individuals met diagnostic criteria for NF1[25], contrary to the reported complete penetrance of the disorder, and despite genetic data strongly suggestive of heterozygous germline variation for the *NF1* variant in at least one individual.

Frequency of *NF1* PVs in two large patient cohorts

To investigate the frequency and penetrance of *NF1* PVs on a larger scale, we utilized the PMBB, a large academic medical biobank with exome sequencing data on 43,731 individuals, all patients of the University of Pennsylvania Health System (UPHS)[24]. We identified 58 individuals heterozygous for any of 50 unique *NF1* PVs: 43 predicted loss of function (pLOF) variants, five missense variants (unambiguously annotated as pathogenic or likely pathogenic in ClinVar[26]), and two deletions involving the entire *NF1* gene (Fig. 1A, Supplementary Data 2). This prevalence of 1 in 752 (0.13%) is four-fold greater than the reported prevalence of 1 in 2500–3500 for NF1[2–4].

We replicated our analysis in a cohort of 118,769 patients collected by Ambry Genetics, all of whom had undergone MGPT for hereditary cancer predisposition with gene panels including the *NF1* gene from 1/2014-3/2018 (Supplementary Data 3). We identified 281 individuals heterozygous for any of 219 *NF1* variants meeting ACMG criteria[27] for classification as pathogenic or likely pathogenic: 170 pLOF variants, 24 missense variants, 10 exonic deletions/duplications, three single amino acid deletions, and 12 deletions involving the entire *NF1* gene (Fig. 1B, Supplementary Data 4). No patient had multiple *NF1* PVs identified. This prevalence was 1 in 432 individuals (0.24%).

Half of *NF1* PV carriers do not have a clinical NF1 diagnosis

Chart review revealed that only 23 of the 58 *NF1* PV carriers in PMBB (39.7%) had a known diagnosis of NF1. We thus divided the PMBB cohort into two groups: the Clinical-NF1 group (those with an *NF1* PV and a known NF1 diagnosis) and the PV-Only group (those with an *NF1* PV but without a known diagnosis of NF1). Only one of the 35 PV-Only individuals had medical history at all suggestive of NF1 (Supplementary Data 2); individual 39 was noted to have fifteen CALM on skin exam and had upper limb deformities and spina bifida; a diagnosis of NF1 had not been made, although she appears to meet clinical diagnostic criteria. No other patients in the PV-Only group had reported evidence of cafe au lait macules, axillary/inguinal freckling, or neurofibromas. Chart review revealed that individual 32 was the same individual that had been referred to our clinic as Case 1, reported above, with no evidence of NF1 on our own detailed physical exam. In the Clincal-NF1 group, individual 41 had been evaluated in our clinic and was given a clinical diagnosis of NF1 based on widespread CALM and innumerable neurofibromas, with all dermatomes apparently affected. However, clinical genetic testing of the *NF1* gene in this patient returned negative, whereas research-based sequencing in PMBB revealed an *NF1* PV with a variant allele fraction (VAF) of 0.10, suggesting post-zygotic mosaicism for the variant.

Dividing the Ambry cohort into the same Clinical-NF1 and PV-Only groups, we observed very similar results. Although we did not have the same depth of phenotypic information available to us in the Ambry cohort as in PMBB, using a combination of physical exam reports, family history, clinic notes, test requisition forms (TRFs), and outreach to ordering providers (Fig. S1, Supplementary Notes), we were able to classify 152 of the 281 *NF1* PV carriers (54%) as Clinical-NF1, whereas 129 patients (46%), lacking any evidence of a known NF1 diagnosis, were classified as PV-Only (Supplementary Data 4). Thus, 48.7% of *NF1* PV variant carriers across both cohorts, appeared to lack a diagnosis of NF1. The apparent prevalence of Clinical-NF1 in PMBB was found to be 1 in 1807, while the apparent prevalence of Clinical-NF1 in the Ambry dataset was found to be 1 in 781.

Clinical evaluation of PV-Only individuals reveals no features of NF1

We were able to recall four PV-Only patients from PMBB for a detailed history and physical exam (Supplementary Notes). Individual #20 was a White female in her 60 s with an *NF1* pLOF variant with a VAF of 0.27. Individual #26 was a Black female in her 40 s with an *NF1* pLOF variant with VAF of 0.40. Individual #38 was a White female in her 30 s with an *NF1* pLOF variant with VAF of 0.09, who had a history of bilateral

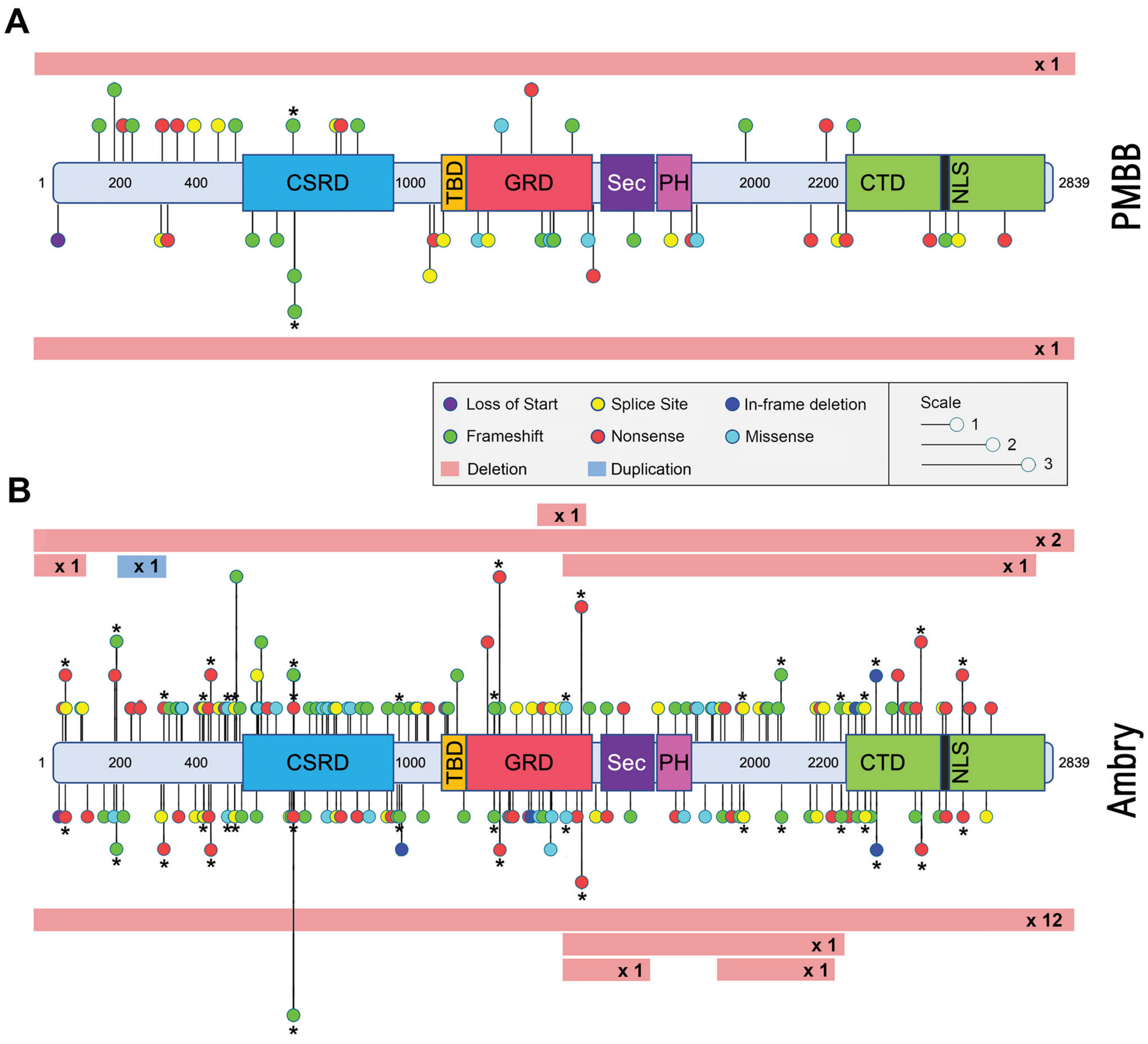

Fig. 1 | *NF1* variants identified in the present study. The *NF1* variants identified in (**A**) the PMBB dataset and (**B**) the Ambry dataset are displayed along a schematic of the NF1 protein. In each case, *NF1* variants identified in the Clinical-NF1 group are indicated along the top of the protein schematic, whereas those identified in the PV-Only group are indicated along the bottom. Variants labeled with an asterisk were identified in both Clinical-NF1 and PV-Only individuals within each dataset. Variants are color-coded by predicted protein effect, with the height of each line segment corresponds to the number of individuals in which the indicated variant was identified. Large deletions and duplications are indicated by red and blue bars, respectively. Amino acid position, based on NP_001035957, are indicated along the protein schematic. NF1 protein domains are indicated, as follows: CSRD Cysteine-and-Serine-Rich Domain, TBD Tubulin-Binding Domain, GRD GAP-Related Domain, Sec Sec14 Homologous Domain, PH Pleckstrin Homologous Domain, CTD C-terminal Domain, NLS Nuclear Localization Signal.

retinoblastoma in childhood due to a de novo *RB1* pathogenic variant for which she underwent enucleation and cryotherapy, but never received any chemotherapy or radiation. Individual #45 was a White female in her 60 s with an *NF1* pLOF variant with VAF of 0.45 with a medical history of papillary thyroid cancer, melanoma, and cervical adenocarcinoma, all of which were treated surgically without chemotherapy, and without radiation. None of these four patients had any features consistent with an NF1 diagnosis on physical exam—more than one café au lait macule, neurofibromas, axillary/inguinal freckling, or Lisch nodules.

Significant demographic differences but no differences in medical comorbidities between the Clinical-NF1 and PV-Only groups

Analysis of demographic and medical history differences between the Clinical-NF1 and PV-Only groups (Supplementary Data 5, 6) revealed that the Clinical-NF1 group was significantly younger than the PV-Only group in both the PMBB and Ambry cohorts (PMBB, mean age of Clinical-NF1 patients 45.0 years, mean age of PV-Only patients 66.1 years, $p = 1.37e\text{-}7$; Ambry, mean age of Clinical-NF1 patients 46.3 years, mean age of PV-Only patients 55.9 years, $p = 1.54e\text{-}08$). In the PMBB cohort, patients in the Clinical-NF1 group were less likely to self-identify as White compared to the PV-Only group ($p = 9.42e\text{-}3$)—this difference did not replicate in the Ambry dataset.

NF1 is classically associated with increased incidence of short stature, hypertension, malignancy, and other medical comorbidities. However, in PMBB we did not identify any significant differences in any medical comorbidities or anthropometrics between the Clinical-NF1 and PV-Only groups on chart review (Supplementary Data 5). However, comparing either group to the overall PMBB population (using

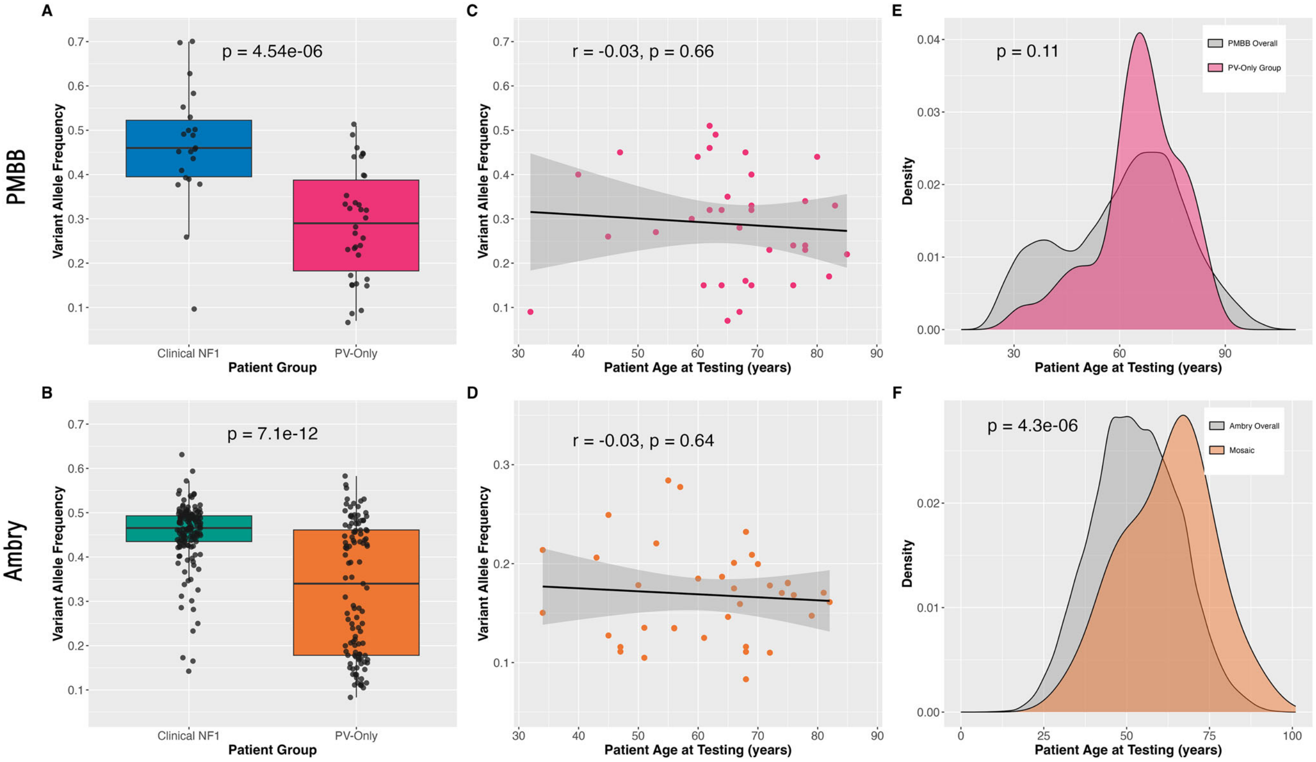

Fig. 2 | Characterization of somatic mosaicism of *NF1* PVs in the PMBB and Ambry datasets. **A** The variant allele fraction (VAF) for each *NF1* PV identified in individuals in the Clinical-NF1 group (left, blue, $n = 22$) and PV-Only group (right, pink, $n = 34$) are displayed for PMBB. Box plots illustrate the median, first and third quartiles, minimum, and maximum for each group. Individuals for whom VAF could nto be determined are excluded. The difference in means between the two groups was statistically significant by 2-sided linear regression ($p = 4.54$e-06). **B** Comparison of *NF1* PV VAF, as in (**A**), but for the Ambry data; individuals in the Clinical-NF1 group ($n = 145$) are shown on the left in green and individuals in the PV-Only group ($n = 112$) are shown on the right in orange. Individuals for whom VAF could not be determined are excluded. Box plots illustrate the median, first and third quartiles, minimum, and maximum for each group. The difference in means between the two groups was statistically significant by 2-sided linear regression ($p = 7.1$e12). **C** 2-sided linear regression of patient age, in years (horizontal axis), against *NF1* PV VAF (vertical axis) for the PMBB PV-Only group. Each point represents a single individual. The regression line is shown, with gray shading illustrating the 95% confidence intervals. There is no significant correlation between the two variables (Pearson correlation coefficient of −0.03, $p = 0.66$). **D** 2-sided linear regression of patient age against *NF1* PV VAF as in (**B**), but for the 47 Ambry PV-Only patients with a mosaic *NF1* PV. Each point represents a single individual. The regression line is shown, with gray shading illustrating the 95% confidence intervals. There is no significant correlation between the two variables (Pearson correlation coefficient of −0.02, $p = 0.64$). **E** The distribution of patient ages at time of genetic testing, in years (horizontal axis) for all 43,559 PMBB participants without an *NF1* PV (gray) and for the 35 individuals in the PMBB PV-Only group (pink). There is no significant difference in the distribution of patient ages between the two groups by the 2-sided Wilcoxon rank sum test ($p = 0.11$). **F** The distribution of patient ages at time of genetic testing for all 118,709 Ambry patients tested with MGPTs containing the *NF1* gene (gray) and the 46 individuals in the Ambry PV-Only group with confirmed mosaic *NF1* PVs (orange). The Ambry PV-Only group with mosaic *NF1* PVs is significantly older than the overall Ambry cohort by the 2-sided Wilcoxon rank sum test ($p = 4.3$e-06).

phenotypes defined by ICD-10 codes extracted from the electronic health record (EHR)), we did identify significant differences (Supplementary Data 5). Both the Clinical-NF1 ($p = 7.8$e-06) and the PV-Only groups ($p = 0.02$) were significantly shorter than the overall PMBB cohort. The rate of malignancy in the Clinical-NF1 (47.8%) and PV-Only groups (48.6%) was higher than in the overall PMBB population (32.4%), but adjusting for age and sex this difference was only significant in the Clinal-NF1 group ($p = 0.01$). Interestingly, the PV-Only group was also found to have a lower rate of hyperlipidemia compared to the overall PMBB population ($p = 0.01$) and a higher rate of cancer-directed radiation therapy ($p = 6.62$e-04). With the caveats that this is a study of small sample sizes, these finding suggests that the Clinical-NF1 and PV-Only groups both have increased risk for certain phenotypes compared to the general population (e.g., short stature), and that the risk of other phenotypes (e.g., hypertension) in patients with Clinical NF1 might have been overestimated in cohorts defined using a phenotype-first approach.

No difference in the types of *NF1* PVs between the Clinical-NF1 and PV-Only groups

Classifying *NF1* PVs by predicted protein effect, we observed a statistically significant enrichment of whole-gene deletions in the PV-Only group compared to the Clinical-NF1 group in the Ambry cohort ($p = 0.008$, Fig. S2A, B, Supplementary Data 6). In total, 10 whole-gene deletions were found among the PV-Only group (representing 8.2% of all variants found in this group), whereas only two were found in the Clinical-NF1 group (representing 1.3% of all variants found in this group). This difference did not replicate in PMBB (Fig. S2A, B, Supplementary Data 5), with only two whole-gene deletions identified, one in the PV-Only group and one in the Clinical-NF1 group. No other significant differences were seen in predicted *NF1* variant effect in either cohort. No differences were seen between the nature of the nucleotide change between the Clinical-NF1 and PV-Only groups in either PMBB or Ambry; in both cases C > T transitions were by far the most common (Fig. S2C, D).

Evidence for somatic mosaicism in the PV-Only groups

In PMBB, the mean VAF of the *NF1* PVs identified in the Clinical-NF1 group was 0.47, consistent with heterozygous germline variation. On the other hand, the PMBB PV-Only group had a mean *NF1* PV VAF of 0.29 (Fig. 2A, Supplementary Data 3). This statistically significant difference ($p = 4.54$e-06) suggests that at least some of the individuals in the PV-Only group carry their *NF1* PVs in the somatic mosaic state

However, not all individuals in the PV-Only group had low VAF *NF1* variants; 17 of the 35 individuals (49%) had *NF1* PVs with VAF ≥ 0.3 (Fig. 2A, Supplementary Data 2), suggestive of heterozygous germline variation. Additionally, two of the 23 Clinical-NF1 individuals (8.7%), including individual 41 discussed above, had an NF1 PV VAF < 0.3 (range: 0.10–0.26), suggestive of mosaicism despite their clinically affected status.

In the Ambry cohort, we found a mean *NF1* PV VAF of 0.45 in the Clinical-NF1 group, again consistent with heterozygous germline variation, whereas the PV-Only group was found to have a mean VAF of 0.35 ($p = 7.1e\text{-}12$), suggesting higher rates of somatic mosaicism in that group (Fig. 2B, Supplementary Data 6). Again, not all individuals in the PV-Only group were found to have a reduced VAF *NF1* PV; the VAF within the PV-Only group had a bimodal distribution (Fig. 2B), suggesting the existence of two populations, with 58 (45%) individuals having VAF ≥ 0.3, 54 (42%) with VAF < 0.3, and 17 (13%) for whom VAF was unavailable (mostly due to technical limitations of copy number variant analysis). Of the 17 individuals for whom VAF was not available, eight were found to be likely mosaic for their *NF1* PV by Sanger sequencing. Of the 53 *NF1* PV-Only samples with VAF < 0.3, 39 (74%) had evidence of mosaicism on confirmatory Sanger sequencing. Integrative Genomics Viewer (IGV)[28] visualizations for three individuals with mosaic *NF1* PVs are shown in Figs. S3–5, demonstrating the likely mosaic nature of their variants. Thus, altogether 47 (36%, 39 with low VAF by NGS, confirmed by Sanger sequencing, eight without VAF from NGS but likely mosaic by Sanger sequencing) of the 129 individuals in the Ambry PV-Only group were likely mosaic for their *NF1* PV and are indicated as "mosaic" in Supplementary Data 5. Additionally, within the Ambry Clinical-NF1 group, seven individuals (4.6%) were found to have an *NF1* PV VAF < 0.30 (range: 0.14–0.29, further clinical and molecular characteristics regarding these cases is included in Supplementary Data 7), but only one of these seven also had evidence of reduced allelic fraction based on Sanger sequencing.

Comparing the predicted protein effect of *NF1* PVs between the germline and mosaic individuals in both PMBB and Ambry (Fig. S2B, C), we observed that *NF1* whole gene deletions were enriched in the mosaic *NF1* PV group, comprising 12.8% of all vs. 2.5% in the heterozygous *NF1* PV group ($p = 0.0015$). No other significant differences were seen. Germline *NF1* deletions are typically characterized by severe NF1 symptomatology[29]; the absence of NF1-associated features in whole gene deletion carriers in the PV-Only group is likely explained by their somatic mosaic state.

Somatic mosaicism of *NF1* PVs and patient age

We investigated the possibility of CH as a driver of *NF1* PV somatic mosaicism in the PMBB cohort, plotting patient age against *NF1* PV VAF for all individuals within the PV-Only group (Fig. 2C). We found no correlation between patient age and *NF1* PV VAF ($r = -0.03$, $p = 0.66$), contrary to what would be expected in CH. Replicating this analysis in the Ambry cohort, we again found no correlation between patient age and *NF1* PV VAF (Fig. 2D, $r = -0.03$, $p = 0.64$). Since the incidence of CH increases sharply with patient age[12,14], we asked whether patients in the PV-Only groups might be older than the overall study populations from which they were drawn. In PMBB, no difference was seen between the ages of the PV-Only individuals and the ages of the 43,559 individuals in the overall PMBB population (Fig. 2E, $p = 0.11$ by Wilcoxon rank sum test), again contrary to what would be expected in CH. On the other hand, within the Ambry cohort, the PV-Only individuals with mosaic *NF1* PVs were significantly older than the overall Ambry population of 118,709 patients (Fig. 2F, Wilcoxon Rank Sum test $p = 4.3e\text{-}06$). Together these data suggest that an age-related process such as CH likely contributes to but is not entirely responsible for the somatic mosaicism that we observed. This finding is corroborated by a recently published study of the UK Biobank and All of Us cohorts, demonstrating no association between age and the frequency of mosaic variants in the *NF1* gene[30].

PheWAS analysis in PMBB for NF1-associated phenotypes

Leveraging the deep phenotypic data available in PMBB, we completed a Phenome-Wide Association Study (PheWAS) across 9030 ICD-10 code-based phenotypes to discover, in an unbiased way, patient phenotypes significantly associated with the presence of an *NF1* PV. We identified 53 significant associations ($p < 5.3e\text{-}6$, Fig. 3A, Supplementary Data 8). The most statistically significant associations were for the ICD-10 codes Q85.00 (Neurofibromatosis, unspecified) and Q85.01 (Neurofibromatosis, type 1). The remaining 51 significant associations, all known features of syndromic NF1, were for ICD-10 codes broadly characterized by benign/malignant neoplasms and skeletal differences (Fig. 3A, Supplementary Data 8). A number of other phenotypes that have previously been suspected of being associated with NF1 came close to reaching significance, including interstitial emphysema ($p = 5.95e\text{-}5$) and functional diarrhea ($p = 6.33e\text{-}5$)[31–33].

PheWAS results considering the 23 Clinical-NF1 individuals only (Fig. 3B, top panel, Supplementary Data 9) identified 43 statistically significant phenotypic associations, of which, 39 (89%) had also been identified in our initial analysis of all 58 *NF1* PV carriers. PheWAS results considering the 35 PV-Only individuals (Fig. 3B, bottom panel, Supplementary Data 10) identified no significant disease associations. With the caveat that this sub-analysis is relatively underpowered, these results suggest that the presence of an incidentally discovered *NF1* PV in blood confers little risk for phenotypes classically associated with syndromic NF1.

Increased incidence of malignancy in both the PV-Only and Clinical-NF1 groups

Within the Ambry cohort, 20 of the 281 individuals we had previously identified as having *NF1* PVs (7.1%) harbored an additional PV in a different cancer predisposition gene (Supplementary Data 11); these individuals were excluded from the following analyses and are reviewed in Supplementary Notes. This rate is higher than the rate of 2.9% that has previously been reported for patients found to have multiple PVs on MGPT for hereditary cancer predisposition[34]. For the following analyses, we defined a control group, the Tested-Negative group, to include all 31,598 patients who had completed MGPT at Ambry with gene panels containing the *NF1* gene with no pathogenic or likely PVs in any cancer predisposition gene.

Altogether 110 individuals (72.4%) in the Clinical-NF1 group, 103 (79.8%) in the PV-Only group, and 21,659 (70.2%) in the Tested-Negative group had a personal history of cancer (Fig. 4A). Adjusting for patient age, no difference in incidence of malignancy between the Clinical-NF1 and PV-only group was found. However, compared to the Tested-Negative group, individuals in both the Clinical-NF1 ($p = 0.004$) and PV-Only groups ($p = 0.03$) were significantly more likely to have a personal history of cancer. Individuals in the Clinical-NF1 group also were found to have a significantly greater number of primary cancers compared to the Tested-Negative group ($p = 6.8e\text{-}05$), whereas no difference was seen in number of primary malignancies between the Tested Negative and PV-Only groups, or between the Clinical-NF1 and PV-Only groups (Fig. 4B). We found no significant difference between the time from cancer diagnosis to time of genetic testing between any of the three groups (Fig. S6A). All of these trends also held true when dividing the Ambry cohort not by NF1 diagnosis status, but by *NF1* PV zygosity (i.e., heterozygous vs mosaic, Figs. S7A, B, S6B).

Older age at cancer diagnosis for patients in the PV-Only group

Individuals in the PV-Only group were significantly older (mean 54.2 years) than both the Clinical-NF1 (mean 44.0 years) and Tested-Negative groups (mean 49.8 years) at the time of first cancer diagnosis

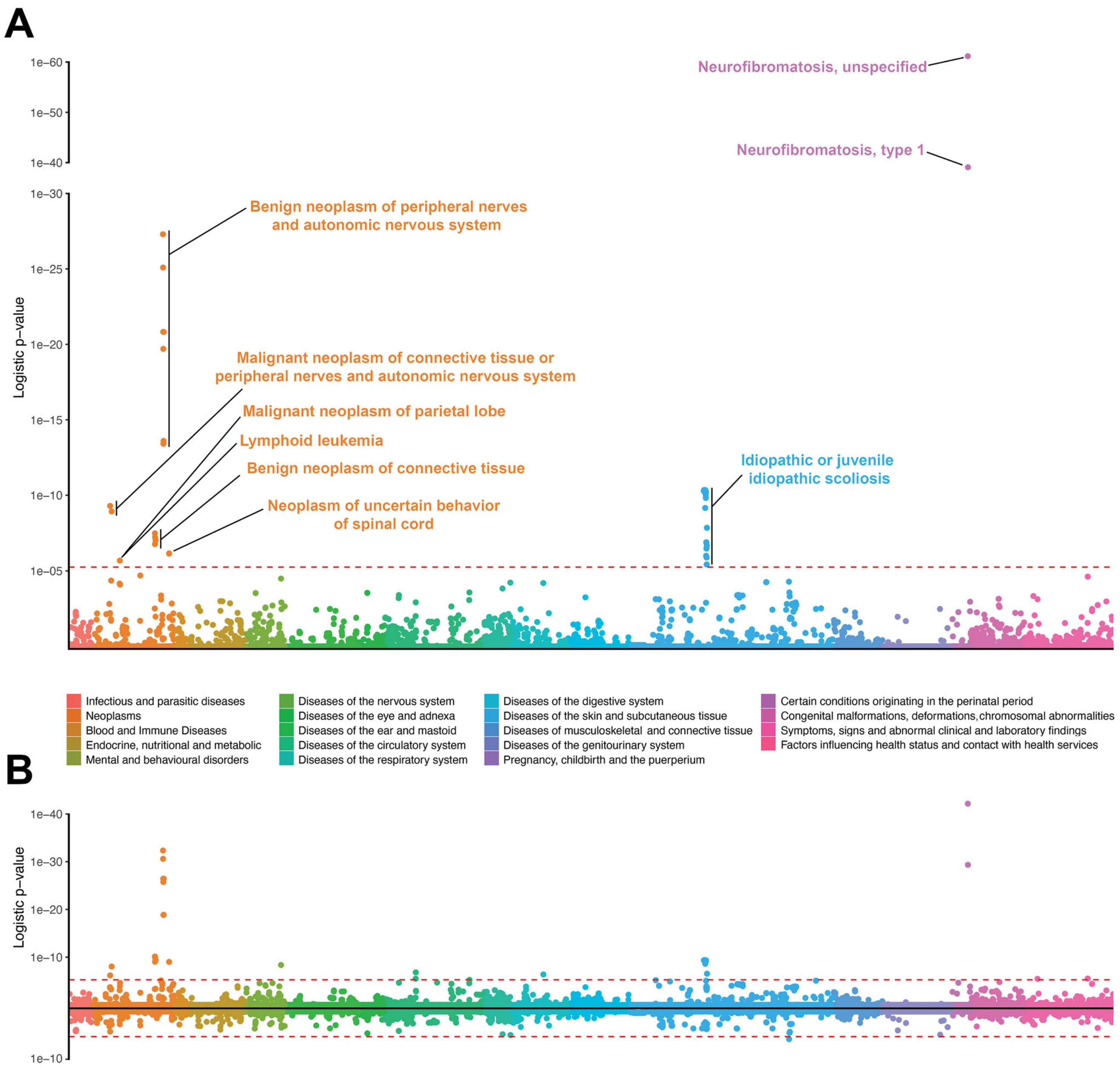

Fig. 3 | PheWAS Results for *NF1* PV carriers in the PMBB cohort. Phenome-wide association studies (PheWAS) were performed to identify ICD-10 code phenotypes significantly enriched within *NF1* PV carriers in PMBB. In (**A**, **B**), individual ICD-10 phenotypes are indicated along the horizontal axis with each point (colored by phenotype group) representing a single ICD-10 code. The height of each point along the vertical axis corresponds to the strength of association for that phenotype with *NF1* PV carrier status, with the *p* value of association (unadjusted) indicated along the vertical axis. The Bonferroni-corrected *p* value significance threshold of 5.5e-6 (correcting for testing across 9436 individual ICD-10 codes) is indicated by red horizontal dashed lines. In both panels, data represent results of two-sided logistic regression. **A** PheWAS results for all 58 *NF1* PV carriers in PMBB are displayed. Phenotype associations surpassing the Bonferroni-corrected significance threshold are labeled (ICD-10 codes with similar descriptions are labeled as groups; more detailed results can be found in Supplementary Data 8). **B** The top portion of the Miami plot illustrates results of PheWAS analysis excluding the 35 *NF1* PV carriers in the PV-Only group; the bottom portion of the plot illustrates results of repeat PheWAS analysis excluding the 23 *NF1* PV carriers in the Clinical-NF1 group. More detailed results can be found in Supplementary Data 9, 10.

($p = 0.022$ and 0.015 respectively, Fig. 4C). Dividing the PV-Only group into two subgroups by *NF1* PV zygosity (i.e., heterozygous vs. mosaic), individuals with a mosaic *NF1* variant trend towards having an older age at first cancer diagnosis compared to individuals with a heterozygous variant (Figs. 4C, S7C).

Increased prevalence of malignancies in both the Clinical-NF1 and PV-Only groups

Dividing cancer diagnoses by type and adjusting for patient age, significant differences were seen between the three groups (Fig. 4D, Supplementary Data 12). Compared to the Tested-Negative group, the Clinical-NF1 group was significantly more likely to be affected by breast cancer ($p = 0.02$), sarcoma ($p = 2.9e\text{-}11$), adrenal cancer ($p < 2e\text{-}16$), central nervous system (CNS) cancers ($p = 4.3e\text{-}07$), and pancreatic cancer ($p = 0.03$), and were significantly less likely to be affected by colorectal cancer ($p = 0.02$) (Fig. 4D). Patients in the PV-Only group were significantly more likely to be affected by ovarian cancer ($p = 0.01$), sarcoma ($p = 0.04$), adrenal cancers ($p = 1.5e\text{-}11$), CNS cancers ($p = 0.04$), and hematologic malignancies ($p = 3.8e\text{-}04$) compared to the Tested-Negative group (Fig. 4D). The increased risk for ovarian

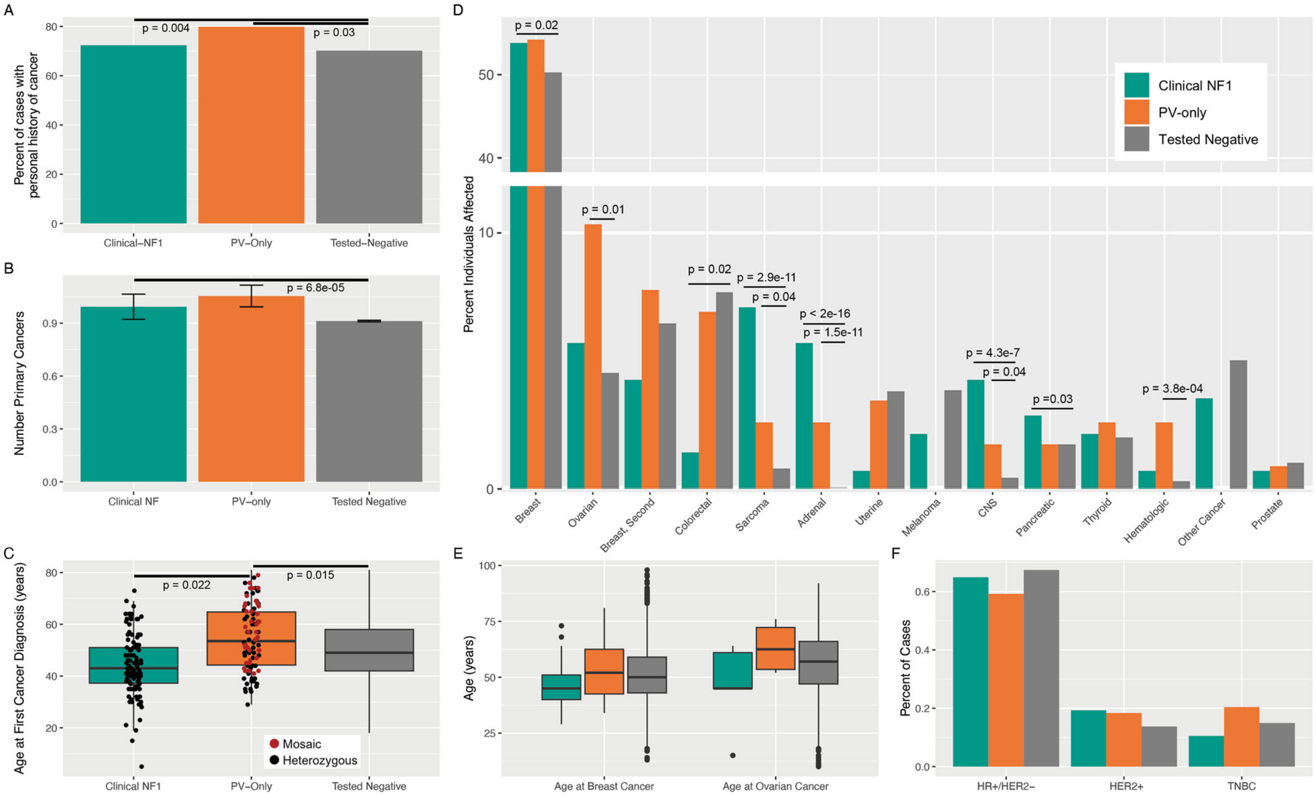

Fig. 4 | Comparison of cancer-related phenotypes between the Clinical-NF1, PV-Only, and Tested-Negative groups in the Ambry cohort. For all panels, statistically significant differences between groups are labeled. In all cases, 2-sided linear (for continuous response variables) or 2-sided logistic regression (for categorical response variables) models were adjusted for patient age. Unadjusted *p* values are shown. **A** Percent of patients within each group (Clinical-NF1 in green, PV-Only in orange, and Tested-Negative in gray) reporting a personal history of malignancy. **B** Mean number of primary malignancies, per patient, across the three groups; Clinical NF1 (*n* = 152), PV-Only (*n* = 129) and Tested Negative (*n* = 31,599). Error bars represent standard deviation. **C** Comparison of age at first cancer diagnosis between each group. Box plots illustrate the median, first and third quartiles, minimum, and maximum for each group, Clinical NF1 (*n* = 152), PV-Only (*n* = 129), and Tested Negative (*n* = 31,599). For the Clincal-NF1 and PV-Only group, individual data points are shown. PV-Only individuals with mosaic *NF1* PVs are shown in red. **D** Percent of individuals, per group, affected by each of 14 different malignancies. Note that for 28 independent comparisons, a strict Bonferroni-corrected significance threshold of 0.0018 should be considered. **E** Mean age at first breast cancer (Clinical NF1 n = 84; PV-Only *n* = 74; Tested Negative *n* = 15,466) and ovarian cancer (Clinical NF1 *n* = 9; PV-Only *n* = 13; Tested Negative *n* = 1430) diagnosis across the three groups. Error bars represent standard deviation. **F** Incidence, across the three groups, of different breast cancer receptor statuses among patients for whom a diagnosis of breast cancer was reported and sufficient receptor status information was provided. HR hormone receptor, HER2 human epidermal growth factor receptor 2, TNBC triple negative breast cancer.

and hematologic malignancies, the rates of which were more than double what was observed in the Tested-Negative group, was unique to the PV-Only group, and was not seen in the Clinical-NF1 group. No significant differences were seen in rates of specific malignancies between the Clinical-NF1 and PV-Only groups. Again, these trends held true when dividing the Ambry cohort by *NF1* variant zygosity rather than NF1 diagnosis status (Fig. S7D).

Receptor status and age at diagnosis for breast and ovarian cancer

Individuals in the Clinical-NF1 group were diagnosed with breast cancer at younger ages (mean 46.0 years) compared to both the PV-Only (mean 53.3 years) and Tested-Negative groups (mean 51.0 years); however, when adjusted for patient age at time of testing, these differences were not statistically significant (Fig. 4E). Similarly, no statistically significant differences were seen in age at ovarian cancer diagnosis between any of the three groups (Fig. 4E). When comparing breast cancer receptor status between the three groups, no significant differences were seen (Fig. 4F, Supplementary Data 13). These trends all held generally true when dividing the Ambry cohort by *NF1* variant zygosity rather than NF1 diagnosis status (Fig. S7E, F), with the additional finding that patients with mosaic *NF1* PVs were significantly more likely to be affected by HER2+ breast cancers, compared to the Tested-Negative group (p = 0.01).

Replication of increased incidence of malignancy in both the PV-Only and Clinical-NF1 groups in multiple additional data sets

To further explore the association between *NF1* PV status and personal history of malignancy that we identified in the Ambry data, we replicated our analysis of personal history of malignancy in four additional cohorts: the UK Biobank[35], PMBB[24], All of Us[36], and the data maintained by the clinical genetic testing company Natera[37]. Altogether, this data included more than one million individuals who had had genetic testing including the *NF1* gene. We identified a total of 784 individuals with an *NF1* PV (1 in 1286 (0.078%) of all individuals, Supplementary Data 14). We again divided individuals into Clinical-NF1, PV-Only, and Tested-Negative groups (based on TRFs in the Natera data, ICD diagnosis codes in UK Biobank and PMBB, and SNOMED CT codes in All of Us). Across all data sets, 314 Clinical-NF1 individuals were identified (1 in 3211 (0.031% of all individuals, Supplementary Data 14)), and 470 PV-Only individuals were identified (1 in 2145 (0.047% of all individuals, Supplementary Data 14)). In all cases, both the Clinical-NF1 and PV-Only cohorts had increased odds of having a personal history of malignancy compared to the Tested-Negative group, adjusting for sex and age at

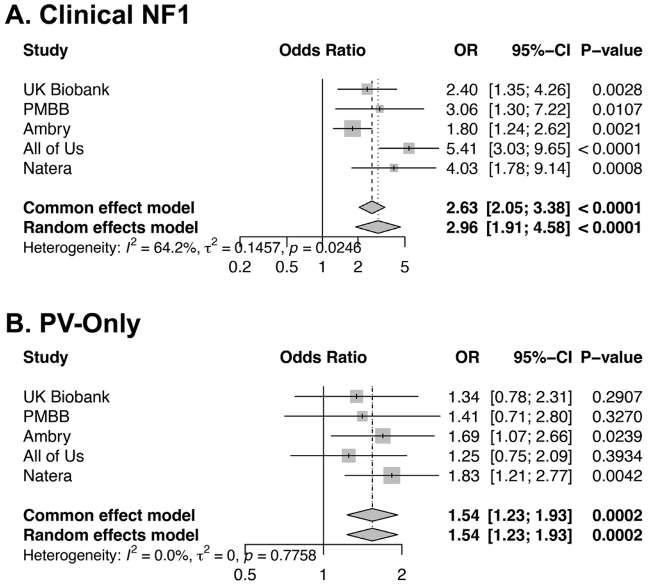

Fig. 5 | Forest plot of odds ratios for the association of *NF1* mutations with personal history of malignancy across multiple cohorts for both the Clinical-NF1 and PV-Only groups. A Forest plot showing the odds ratios (OR, indicated as a vertical tick), 95% confidence intervals (95%–CI, indicated as a horizontal line), and unadjusted *p* values, derived from 2-sided logistic regression, for personal history of malignancy within in the Clinical-NF1 group, compared to the Tested-Negative group, from five different cohorts: UK Biobank, PMBB, Ambry, All of Us, and Natera. The size of the boxes represents the weight of each study in the meta-analysis, with larger boxes corresponding to studies with higher precision (i.e., smaller standard errors). The overall combined OR and CI are shown at the bottom of the plot as diamonds, summarizing the meta-analysis results under both fixed-effect and random-effects models. The tau-squared (τ^2) and I^2 statistics indicate the level of heterogeneity between the studies, with I^2 describing the percentage of total variation across studies due to heterogeneity rather than chance. **B** Forest plot showing the odds ratios (OR, indicated as a vertical tick), 95% confidence intervals (95%–CI, indicated as a horizontal line), and unadjusted *p* values, derived from 2-sided logistic regression, for personal history of malignancy within in the PV-Only group, compared to the Tested-Negative group, from five different cohorts: UK Biobank, PMBB, Ambry, All of Us, and Natera. The size of the boxes represents the weight of each study in the meta-analysis, with larger boxes corresponding to studies with higher precision (i.e., smaller standard errors). The overall combined OR and CI are shown at the bottom of the plot as diamonds, summarizing the meta-analysis results under both fixed-effect and random-effects models. The tau-squared (τ^2) and I^2 statistics indicate the level of heterogeneity between the studies, with I^2 describing the percentage of total variation across studies due to heterogeneity rather than chance. In both cases, the meta-analysis shows a significantly increased odds of personal history of malignancy in association with the presence of an *NF1* PV.

time of testing (Supplementary Data 14). Meta-analyzing our results across all cohorts (Fig. 5), we found that the Clinical-NF1 group had an almost three-fold increased odds of having a personal history of malignancy as compared to the Tested-Negative group (p = 5.50e-16, OR 2.96 [1.91–4.58]). For the PV-Only group, the odds were also significantly increased at 1.5-fold (p = 1.77e-04, OR 1.54 [1.23–1.93]), further confirming that the presence of an *NF1* PV, even in the absence of an NF1 diagnosis, is associated with an increased odds of malignancy.

Discussion

NF1 is a classic Mendelian disorder with a characteristic phenotypic presentation[4]. However, our understanding of the disorder is mainly based on patients ascertained through phenotype-first approaches[3,5,6]. Our study, leveraging a genotype-first approach in two unique large patient cohorts, with replication of key findings in three additional data sets, revealed surprising results: *NF1* PVs are significantly more common than would be expected given the prevalence of clinical NF1; half of all patients with *NF1* PVs are apparently unidentified; 15–30% of all *NF1* PVs appear to be present in the somatic mosaic state; and incidentally discovered *NF1* PVs are significantly associated with an increased incidence of malignancy.

The incidental discovery of PVs is a unique challenge brought on by the broad application of genetic testing in larger patient populations. Although incidentally discovered medically-actionable variants are detected in 3–6% of individuals undergoing broad genetic sequencing[38–40], some studies suggest that as few as 18% of individuals with incidentally discovered medically-actionable variants will have any medical history related to these genetic findings[41]. Our results are in line with these observations, suggesting that *NF1* PVs are substantially more common than previously thought and often discovered in apparently unaffected individuals.

Our data suggest that somatic mosaicism (both post-zygotic and clonal hematopoiesis), incomplete penetrance, and missed diagnoses all contribute to the high prevalence of *NF1* PVs that we observed. Reports on the prevalence and penetrance of Mendelian disorders, such as NF1, suffer from ascertainment bias[42]; patients with more subtle presentations go undiagnosed, artificially decreasing estimates of disease/genetic variant prevalence. However, missed diagnoses cannot completely explain our findings. The possibility of reduced penetrance is suggested by our experience with Cases 1-4 and with the four PMBB PV-Only individuals who we recalled for physical exam and history, many of whom were found to have *NF1* PVs with VAFs consistent with germline variation. However, we cannot rule out the possibility of mosaicism, rather than reduced penetrance, in these patients without testing other tissues or other family members. The true population prevalence of such cases is difficult to estimate from our data, a limitation of our study.

However, the largest contributor to the high prevalence of *NF1* PVs that we observed was somatic mosaicism. In some cases, individuals meeting clinical diagnostic criteria of NF1 were observed with low VAF *NF1* PVs. For example, we identified one individual (individual 41), who met NF1 clinical diagnostic criteria with all areas of skin affected, but with a VAF for the *NF1* PV identified of only 0.10, with previous clinical sequencing being reported as negative. This phenomenon is well documented in Tuberous Sclerosis Complex, with multiple individuals meeting clinical diagnostic criteria having extremely low VAFs in the blood[43]. At the same time, in many other individuals with low VAF *NF1* PVs, no signs of NF1 were found on exam. As we are limited in our ability to confidently call mosaic variants in all sequenced individuals due to read depth limitations and having access to only a single tissue per patient, the prevalence of somatic mosaicism that we report is likely an underestimate.

Somatic mosaic *NF1* variants in the blood have previously been reported in the context of hematologic malignancy[44–46], and rarely in the context of CH[12,14]. The patients that we identified with mosaic *NF1* PVs were not known to have active hematologic malignancies, arguing against this etiology as a major driver of the mosaicism that we observed. Although post-zygotic mosaicism and CH are indistinguishable in the assays used here, clinical context provides multiple lines of evidence arguing against an age-related process like CH as the exclusive underlying etiology for the mosaicism we encountered. This finding is in line with a recent report of widespread mosaic genetic variants found in blood that are not clearly attributable to malignancy or CH, but likely due to post-zygotic mosaicism[47,48]. Thus, our data suggest that much of the apparent mosaicism that we observe for *NF1* PVs may represent post-zygotic mosaicism rather than CH, and that mosaicism for *NF1* PVs is common.

Most importantly, our results suggest that all *NF1* PVs, incidentally discovered or otherwise, are associated with increased incidence of malignancy; in the Ambry cohort, both the Clinical-NF1 and PV-only groups showed increased rates of malignancy, even when adjusting for patient age. Although our PheWAS analysis in PMBB PV-Only individuals did not specifically replicate this finding, the overall rate of malignancy in the PMBB PV-Only group was 45.7%, higher than the reported lifetime cancer risk of 39.7% in the United States[49] and higher than the malignancy rate of 39.1% observed in the Clinical-NF1 group (Supplementary Data 5), a population that is known to be at increased risk of malignancy. The PMBB PV-only individuals who had a cancer diagnosis had a wide range of malignancy types, many of which are known be associated with somatic PVs in *NF1*, further supporting the link to cancer predisposition in this group[11]. The increased incidence of malignancy we observed in the PV-Only group held true in every cohort we examined; both the Clinical-NF1 and PV-Only groups were consistently found to have an increased incidence of malignancy compared to control populations. The odds ratio of association was larger in the Clinical-NF1 group, with a nearly 3-fold increased odds of having malignancy, compared to the one and a half-fold increased odds in the PV-Only group, which is consistent with the hypothesis that the burden of cells with an *NF1* PV may directly correlate with the malignancy risk. These findings, along with our observation that both the Ambry and PMBB PV-Only groups are enriched for mosaic *NF1* PVs and are significantly older at the time of first cancer diagnosis, suggests that mosaic and incidental *NF1* PVs confer a real but perhaps attenuated cancer predisposition compared to inherited/germline *NF1* PVs in patients with syndromic NF1.

Our study using the Ambry data set, representing the largest known cancer-focused cohort of patients with *NF1* PVs, confirms some previous observations about associations between NF1 and malignancy and contradicts others. Consistent with prior studies, we observe a younger age at first cancer diagnosis among those with heterozygous *NF1* PVs and increased incidence of breast cancer among patients with Clinical NF1 but did not find an earlier age of diagnosis for breast cancer or an enrichment of HER2-amplified or triple negative breast cancers in patients with NF1[8–10]. These inconsistencies with previous work may be explained by sampling bias. The patient population that we studied, referred for genetic testing at Ambry, is likely enriched for individuals with earlier and more aggressive breast cancers, and thus may not completely reflect the pattern of breast cancer in the general population. Our data corroborate the increased incidence of sarcoma, adrenal cancers, CNS malignancies, and pancreatic cancers in patients with clinical NF1. Increased incidence of these same malignancies was observed in the PV-Only group, suggesting similar mechanisms of oncogenesis and predispositions to malignancy in patients with incidentally-discovered *NF1* PVs. We also observed an increased incidence of hematologic and ovarian cancers specifically within the PV-only group. Our data do not allow us to dissect the causal relationship between these observations. It is possible that incidentally discovered *NF1* PVs in the blood confer increased risk for ovarian and hematologic malignancy, as both tumor types are known to often harbor somatic mutations in *NF1*, but it also possible that these variants developed secondarily due to treatment-related CH, reflecting the relatively high rates of chemotherapy employed early in the treatment course of these malignancies[50].

Our study has several limitations. Utilizing EHR and TRFs means that the phenotyping data are based on observations of clinicians not specifically assessing for NF1. Even with optimized data abstraction efforts the distinction between the Clinical-NF1 and PV-only groups is challenging. Similarly, our classification of *NF1* PVs as mosaic or germline was based on bulk sequencing of only a single tissue, and thus we are limited in our ability to confidently call mosaic variants or distinguish between post-zygotic mosaicism and CH. Although our study is the largest to investigate the incidence of malignancy in patients with *NF1* PVs, these variants are still relatively rare, and we are likely underpowered to detect small differences in patient phenotypes. Lastly, with the possible exception of the UK Biobank, the populations that we study in this report represent relatively sick patient populations. The Ambry and Natera cohorts are likely enriched for patients with a personal and/or family history of malignancy. The PMBB and All of Us cohorts, while more reflective of the general population, are drawn primarily from populations of patients seeking care at major academic medical centers and is may therefore be enriched for patients with disease. Thus, more study is required to validate the generalizability of our findings to other patient populations.

Despite these limitations, our work in multiple distinct and complementary patient cohorts suggests that *NF1* PVs are substantially more common than previously appreciated, often characterized by somatic mosaicism and reduced penetrance, and associated with increased incidence of malignancy even in patients without syndromic NF1. Although our work begins to suggest a framework for the clinical interpretation of incidentally identified PVs in the *NF1* gene, the establishment of optimal screening and management strategies will require further research and clinical efforts. It is inevitable that the continued expansion of broad genetic sequencing in large patient populations will identify incidental variants in many other genes, some of which will also be found to have important clinical associations. Our data are consistent with a growing literature supporting the importance of post-zygotic mosaicism in disease causality[43,51–53]. A broader understanding of the true population-level frequency and pathogenicity of germline and somatic mosaic variants in Mendelian disease genes across the genome will likely lead to a transformation in our understanding of the genetic architecture of both rare and common human genetic disease.

Methods

Informed consent

All patients signed informed consent forms for inclusion in the respective biobanks that our data were drawn from. For the recall study in PMBB specifically, all individuals signed informed consent for participation in the recall study, including detailed physical exam, history, and publication of fundings.

PMBB patient recruitment and exome sequencing

The PMBB[24] is a University of Pennsylvania academic biobank which recruits patient-participants from the UPHS around the greater Philadelphia area in the United States. Appropriate consent was obtained from each participant regarding storage of biological specimens, genetic sequencing, access to all available EHR data, and permission to recontact for future studies. The study was approved by the Institutional Review Board of the University of Pennsylvania and complied with the principles set out in the Declaration of Helsinki. This study included the subset of 43,731 individuals enrolled in PMBB who had previously undergone exome sequencing and genotyping array. Briefly, for each individual, DNA was extracted from stored buffy coats and exome sequences were generated by the Regeneron Genetics Center (Tarrytown, NY) and mapped to GRCh38 as previously described[54]. Sample-level filtering was as follows: individuals with low exome sequencing coverage (less than 75% of targeted bases achieving 20× coverage) or with high missingness (greater than 5% of targeted bases) were removed from analysis, leaving 43,612 samples after sample-level filtering. Variant-level filtering was as follows: in each sample, all single nucleotide variants (SNVs) with a total read depth <7 were changed to “no-call”, and similarly all insertion/deletion variants with a total read depth <10 were changed to “no-call.” Relatedness across all samples was calculated in PLINK[55] using a minimum PI_HAT cutoff of 0.09375 to capture out to 3rd degree relationships. None of the *NF1* PV carriers that we identified were 3rd degree relatives or closer.

PMBB variant annotation

The genetic variants in PMBB exome sequencing data were subset to include only the *NF1* gene locus and were subsequently annotated using the Ensembl Variant Effect Predictor (VEP, version 102)[56] with the plugin LOFTEE (version 0.3)[57] to specifically annotate predicted loss of function (pLOF variants) and dbNSFP (version 4.2)[58,59] to specifically annotate all single nucleotide variants within the *NF1* gene. Only variants affecting the NCBI RefSeq canonical *NF1* transcript (NM_001042492.2) were considered. Copy number variants (CNVs) were annotated, starting with the same exome sequencing data as described above, using version 1.3 of the CLAMMS pipeline[60]. Standard quality control measures were taken both at the sample level (samples with >40 CNVs or with >40,000 exons called as CNVs were removed) and chromosome level (for samples with >10% of a chromosome covered by >1 CNV call, that chromosome was removed). QC levels ranging from 0-3 were assigned to each CNV call based on Q_non_dip, Q_exact, and allele balance and heterozygosity metrics. Only CNVs meeting the most stringent QC threshold of 3 were included in the analysis. The resultant CNV calls at the *NF1* locus were manually reviewed to identify any potentially suspicious annotations—no CNVs were excluded after manual review. The data available in PMBB did not permit us to determine VAF or potential mosaicism for CNVs. PVs were defined to include: (1) all nonsense variants, frameshift insertions/deletions, disruption of canonical splice site dinucleotides, or gain/loss of the start or stop codon that were not predicted to escape nonsense-mediated decay; (2) any variant unambiguously annotated as "pathogenic" or "likely pathogenic" in the ClinVar database; and (3) any whole-gene deletion of *NF1*.

PMBB chart review

We performed manual chart review of complete EHR data for each carrier of an *NF1* pathogenic variant that we identified in PMBB. All 58 charts were reviewed by a single clinician blinded to the results of exome sequencing, and detailed information was extracted. At a minimum for each individual, we gathered: any mention of the terms "NF1" "neurofibromatosis" or "von Recklinghausen;" EHR-reported race/ethnicity and sex (which typically represents a combination of self-reported and provider-assigned values); height at most recent measurement; number of encounters with UPHS providers (if fewer than five); whether the individual had been seen by a clinical geneticist; all documented skin exams; and all documentation regarding malignancies. The clinician subsequently generated a summary paragraph describing each individual's overall health issues and diagnoses. This information is documented in Supplementary Data 2 and was used to generate the summary statistics listed in Supplementary Data 3.

PMBB recall by genotype study

We obtained IRB approval to complete a recall by genotype study in PMBB. In this study, all living PMBB participants with *NF1* pLOF variants identified on exome sequencing but without an NF1 diagnosis on chart review were invited to come in for a detailed history and physical exam and compensated $100 for their time. Twenty-four participants were contacted, and four agreed to come in as described in the Supplementary Notes section. We obtained informed consent from each of these four individuals for history, physical, and publication of findings. A full personal medical history, family history, and physical exam was obtained by a physician board-certified in Internal Medicine and Medical Genetics. If any concerning medical findings were identified, appropriate referrals were made.

PMBB phenotype generation

ICD-9 and ICD-10 disease diagnosis codes and procedural billing codes were extracted from patient EHRs cross the entire PMBB dataset. ICD-9 encounter diagnoses were mapped to ICD-10 using the Center for Medicare and Medicaid Services 2017 General Equivalency Mappings (https://www.cms.gov/Medicare/Coding/ICD10/2017-ICD-10-CM-and-GEMs.html) with unmappable ICD-9 codes dropped from further analysis. Participants were determined as having a phenotype if they had the corresponding ICD diagnosis code on at least one date, while phenotypic controls consisted of individuals who never had the ICD code used. ICD codes S00-T98 (within the group "Injury, poisoning and certain other consequences of external causes") and V01-Y98 (within the group "External causes of morbidity and mortality") were excluded from PheWAS analysis. Only ICD phenotypes with a total case count of 20 or more across all 43,612 individuals in the analysis were included in our study.

PMBB Phenome-wide association studies (PheWAS)

Using a PheWAS[61] approach, all *NF1* PVs in PMBB were collapsed into a single binned genotype and tested for association with each ICD code-derived phenotype (described above) using the program BioBin[62] with a logistic regression model adjusted for age, sex and the first ten principal components of genetic ancestry. Our association analyses considered only phenotypes with at least 15 cases across all of PMBB, as described above, leading to the interrogation of 9032 total phenotypes. Subsequent PheWAS analyses were completed using the same approach, but excluding from the analysis either the patients in the Clinical-NF1 group or PV-Only group. Q-Q plots were generated to assess for *p* value inflation (Fig. S8). Note—24 individuals in PMBB were identified as having ICD code Q85.01 (Neurofibroatosis, type 1), but only 18 of these were found to have an NF1 PV in our analysis. Of the six who were not identified to have an *NF1* PV by our analysis, five had clinical diagnoses of NF1 without genetic testing, while one had a clinical NF1 diagnosis, with genetic testing revealing only an *NF1* variant with conflicting reports of pathogenicity (VUS, Likely Pathogenic, Pathogenic), and thus would not have been identified as definitively pathogenic in our analysis. More information about these six individuals is included in Supplementary Data 15. The other five individuals with ICD code Q85.01 but without any *NF1* variants identified may have had smaller CNVs not detectable by our methodology, missense variants not classified as pathogenic/likely pathogenic, deep intronic variants missed on exome sequencing, or even ICD codes entered in error.

Ambry cohort definition and molecular testing

We queried a cohort of 118,768 patients from the Ambry Genetics (Aliso Viejo, CA) laboratory database from 1/1/2014 through 3/31/2018 who had undergone clinical Next-Generation Sequencing (NGS) to identify individuals with pathogenic or likely PVs in *NF1*. This cohort contained sequencing data obtained from clinical multigene panel testing, including BreastNext, CancerNext, CancerNext Expanded, OvaNext, CustomNext, PGLNext, and BrainTumorNext, covering between two and 67 genes related to hereditary cancer risk (Supplementary Data 4). All patients were clinician referred; ordering standards were based on clinician judgment or practice-specific thresholds. Sequencing, variant calling, and variant annotation was performed at Ambry Genetics as previously described[34] and all identified pathogenic/likely pathogenic *NF1* variants were classified as described in Supplementary Data 5. Pathogenic and likely PVs were defined to include deletions (including whole gene deletions or smaller deletions encompassing at least one exon of the *NF1* gene), exonic duplications, frameshift variants, nonsense variants, canonical splice site variants, and missense variants meeting ACMG/AMP criteria for pathogenicity[27]. Regions with <20× coverage on NGS were followed up with Sanger analysis. In addition, variants in regions complicated by pseudogene interference, variant calls not satisfying depth of coverage (100×) and variant allele frequency (40%) quality thresholds, reportable small insertions and deletions, and potentially homozygous variants were verified by Sanger sequencing. Standard protocols for clinical testing were used for automated variant calling in Sanger sequencing (SeqPilot, JSI Medical Systems) analysis. Where variants were not called by the software but clearly visible at low levels on visual inspection by qualified personnel, variants were designated as mosaic.

Ambry cohort patient phenotyping

For the Ambry data set, demographic and clinical information including race/ethnicity and sex (as reported on the TRF, which typically represents a combination of self-reported and provider-assigned values), personal cancer history, family cancer history, and family history of NF1 was obtained through review of laboratory TRF and, where available, clinic notes and pedigrees. For cases lacking documentation of a clinical diagnosis of NF1, referring providers were contacted to obtain further clinical details. Patients were subsequently categorized as "Clinical-NF1" or "PV-only" (pathogenic variant-only) using a three-tiered classification process (Fig. S1, Supplementary Notes). First, when physical exam data was documented, cases of Clinical-NF1 were defined using the NIH NF1 diagnostic criteria[63]. Second, if physical examination data were not recorded but the patient had a documented first-degree family member with NF1, the patient was included in the Clinical-NF1 group. Third, if neither physical exam nor family history data were available, patients were categorized according to clinician-provided description on the TRF. PV-Only cases were ideally defined by comprehensive physical exam data documenting lack of concordance with NIH criteria (as illustrated in the top branch of Fig. S1). For most cases, however, PV-Only cases were defined by information in the TRF.

Replication analysis in UK Biobank

Starting from the UK Biobank 500k Whole Genome release from November 2023, standard QC metrics were applied as described above for PMBB and the subset of individuals with NF1 PVs was identified using an approach identical to the approach described above for PMBB. Basic demographic information, including age, sex (as reported in data field 31, representing a combination of self-reported and provider-assigned values), and ancestry (as reported in data field 22006, which records genetically-informed ancestry and self-reported "white British" ancestry) were obtained. Given the substantial preponderance of individuals of white British ancestry in UK Biobank, we limited our analysis to only those individuals of self-reported white British ancestry. The Clinical-NF1 group ($n = 51$) was defined by the presence of the ICD10 code Q85.01 in any patient with an NF1 PV, while the PV-Only group ($n = 71$) was defined by the absence of this code in any individual with an NF1 PV. The Tested-Negative group ($n = 472{,}326$) was defined as all individuals lacking both an NF1 PV and lacking the Q85.01 ICD10 code. Personal history of malignancy was defined based on the data contained in the national cancer registry (excluding benign, non-malignant and unspecified cancers), considering any patient with a reported malignancy a case, and any individual without any reported malignancy as a control. Statistical analysis was performed as described below, adjusting models for patient age and sex.

Replication analysis in All of Us

Variant call data from 182,459 individuals with short read whole genome sequencing and EHR data was obtained from the *All of Us* Research Program Controlled Data Repository v7 using the *All of Us* Research Program Researcher Workbench. Variants in *NF1* transcript (NM_001042492.2) were annotated using the methodology described in "PMBB variant annotation," and *All of Us* Research Program Participants with variants at positions with a read depth over 14 were extracted. The EHR of the *All of Us* Research Program cohort was screened for individuals with SNOMED concept terms and/or daughter concept terms for cancer conditions as defined by Aschebrook-Kilfoy et al.[64], and individuals with NF1 were identified using NF1 SNOMED concept terms or daughter concept terms (92824003, NF1; 403816002, Multiple CALMs due to neurofibromatosis; 403817006, Multiple neurofibromas in neurofibromatosis). Data regarding race/ethnicity and sex were derived from data using self-identification on participant-facing surveys. Statistical analysis was performed as described below, adjusting models for patient age and sex. The workspace to analyze this data is titled "Malignancy in Individuals with Incidental NF1 P/LP Variants" on the *All of Us* Researcher Workbench.

Replication analysis in Natera

Genetic testing was performed at Natera, Inc laboratory, and clinical information was extracted from the accompanying TRFs. This stratification allowed for comparative analysis between individuals with genetically confirmed NF1, those with genetic findings but no reported clinical features, and those without identifiable PVs. The study cohort was stratified into three distinct groups based on genetic testing results and clinical information provided on the TRFs: (1) Clinical NF group: This group comprised individuals who tested positive for a pathogenic or likely pathogenic (P/LP) variant in the NF1 gene and had clinical NF1 findings indicated on their test requisition form; (2) NF1 positive only group: This group included individuals who tested positive for a P/LP variant in the NF1 gene but lacked clinical NF1 findings on their test requisition form; (3) Tested negative group: This group consisted of individuals who underwent genetic testing with any gene panel that included the NF1 gene but had no P/LP variants identified in any of the genes tested. For all groups, demographic data was characterized including age at the time of testing, race/ethnicity and sex (as reported on the TRF which typically represents a combination of self-reported and provider-assigned values), and personal history of malignancy (yes/no). A logistic regression model was performed taking personal history of malignancy as the outcome and using the clinical groups at predictor variables with age and sex as covariates.

Statistical analysis

All statistical analyses were performed with R software, version 4.2.2. Within both the PMBB and Ambry cohorts, differences in *NF1* PV VAF and patient age between the Clinical-NF1 and PV-Only groups were compared by linear regression. Differences in the distributions of individual ages between the *NF1* PV-Only group and the overall PMBB/Ambry datasets were compared using the Wilcoxon rank sum test. For the two whole-gene deletions identified in the PMBB cohort, and for 25 individuals in the Ambry cohort, VAF could not be determined, and these individuals were excluded from analyses requiring VAF. For all categorical traits and phenotypes listed in Supplementary Data 5 and Supplementary Data 6, statistical analyses were completed by logistic regression (logistic regression models were adjusted for patient age and sex as indicated). For the continuous traits of VAF, read depth, height, and age, statistical analyses were accomplished via linear regression. The proportions of each class of genetic alteration and mutational spectrum were compared between groups with logistic regression. Within the Ambry cohort, comparisons between the Clinical-NF1, PV-Only, and Tested-Negative groups, or between the heterozygous *NF1* PV, mosaic *NF1* PV, and Tested-Negative groups, were completed as follows. Age at first cancer diagnosis, number of cancer primaries, age at first breast cancer diagnosis, and first ovarian cancer diagnosis were compared between the groups by linear regression, adjusting models for individual age at the time of testing. The incidence of personal history of cancer, of each specific malignancy, and breast cancer receptor status, were compared by logistic regression adjusting the model for individual age at the time of testing. Time between first cancer diagnosis and genetic was compared using the Wilcoxon rank sum test.

Reporting summary

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

Data availability

The summary data supporting the findings of this study have been made openly available in the supplemental tables. Individual-level data, aside from what has been included in the manuscript and supplemental materials, cannot be shared due to patient privacy concerns and as patient participants have not been broadly consented for dissemination of individual-level data, except where otherwise noted. The PennMedicine BioBank (PMBB) data used in this study were generated previously, and PMBB data are available to researchers through collaboration with PMBB via the PMBB website at https://pmbb.med.upenn.edu/investigators.php. The UK Biobank (UKBB) data used in this study were generated previously, and UKBB data are available through the UKBB website upon request at https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access. As a part of our agreement to use the data contained within UKBB, we are not allowed to share the raw data ourselves, but individuals who are interested can request access. All of Us data used in this study are available to registered researchers of the *All of Us* Researcher Workbench at https://www.researchallofus.org/register/. The data provided from Ambry and Natera that were used in this study are available through collaboration with each corporation.

References

1. Kiuru, M. & Busam, K. J. The NF1 gene in tumor syndromes and melanoma. *Lab. Investig.* **97**, 146–157 (2017).
2. Sabbagh, A. et al. Unravelling the genetic basis of variable clinical expression in neurofibromatosis 1. *Hum. Mol. Genet.* **18**, 2768–2778 (2009).
3. Karaconji, T., Whist, E., Jamieson, R. V., Flaherty, M. P. & Grigg, J. R. B. Neurofibromatosis type 1: review and update on emerging therapies. *Asia Pac. J. Ophthalmol.* **8**, 62–72 (2019).
4. Stewart, D. R., Korf, B. R., Nathanson, K. L., Stevenson, D. A. & Yohay, K. Care of adults with neurofibromatosis type 1: a clinical practice resource of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **20**, 671–682 (2018).
5. Evans, D. G. et al. Birth incidence and prevalence of tumor-prone syndromes: estimates from a UK family genetic register service. *Am. J. Med. Genet. A* **152A**, 327–332 (2010).
6. Gutmann, D. H. et al. The diagnostic evaluation and multi-disciplinary management of neurofibromatosis 1 and neurofibromatosis 2. *JAMA* **278**, 51–57 (1997).
7. Landry, J. P. et al. Comparison of cancer prevalence in patients with neurofibromatosis type 1 at an academic cancer center vs in the general population from 1985 to 2020. *JAMA Netw. Open* **4**, e210945 (2021).
8. Sharif, S. et al. Women with neurofibromatosis 1 are at a moderately increased risk of developing breast cancer and should be considered for early screening. *J. Med. Genet.* **44**, 481–484 (2007).
9. Uusitalo, E. et al. Breast cancer in neurofibromatosis type 1: overrepresentation of unfavourable prognostic factors. *Br. J. Cancer* **116**, 211–217 (2017).
10. Pearson, A. et al. Inactivating NF1 mutations are enriched in advanced breast cancer and contribute to endocrine therapy resistance. *Clin. Cancer Res.* https://doi.org/10.1158/1078-0432.ccr-18-4044 (2019).
11. Philpott, C., Tovell, H., Frayling, I. M., Cooper, D. N. & Upadhyaya, M. The NF1 somatic mutational landscape in sporadic human cancers. *Hum. Genom.* **11**, 13 (2017).
12. Robertson, N. A. et al. Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects. *Nat. Med.* **28**, 1439–1446 (2022).
13. Feusier, J. E. et al. Large-scale identification of clonal hematopoiesis and mutations recurrent in blood cancers. *Blood Cancer Discov.* **2**, 226–237 (2021).
14. 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.* https://doi.org/10.1016/j.stem.2017.07.010 (2017).
15. Victor, F. C. Segmental neurofibromatosis. *Dermatol Online J.* **11**, 20 (2005).
16. Ruggieri, M. & Huson, S. M. The clinical and diagnostic implications mosaicism in the neurofibromatoses. *Neurology* **56**, 1433–1443 (2001).
17. Ejerskov, C., Raundahl, M., Gregersen, P. A. & Handrup, M. M. Clinical features and disease severity in patients with mosaic neurofibromatosis type 1: a single-center study and literature review. *Orphanet J. Rare Dis.* **16**, 180 (2021).
18. Maani, N. et al. Incidental findings from cancer next generation sequencing panels. *NPJ Genom. Med.* **6**, 63 (2021).
19. Heald, B. et al. Unexpected actionable genetic variants revealed by multigene panel testing of patients with uterine cancer. *Gynecol. Oncol.* **166**, 344–350 (2022).
20. Damrauer, S. M. et al. FBN1 coding variants and non-syndromic aortic disease. *Circ. Genom. Precis Med.* **12**, e002454 (2019).
21. Coppola, L. et al. Biobanking in health care: evolution and future directions. *J. Transl. Med.* **17**, 172 (2019).
22. Paskal, W., Paskal, A. M., Dębski, T., Gryziak, M. & Jaworowski, J. Aspects of modern biobank activity—comprehensive review. *Pathol. Oncol. Res.* **24**, 771–785 (2018).
23. Forrest, I. S. et al. Population-based penetrance of deleterious clinical variants. *JAMA* **327**, 350–359 (2022).
24. Verma, A. et al. The Penn Medicine BioBank: towards a genomics-enabled learning healthcare system to accelerate precision medicine in a diverse population. *J. Pers. Med.* **12**, 1974 (2022).
25. Legius, E. et al. Revised diagnostic criteria for neurofibromatosis type 1 and Legius syndrome: an international consensus recommendation. *Genet Med.* **23**, 1506–1513 (2021).
26. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
27. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
28. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
29. Bettegowda, C. et al. Genotype-phenotype correlations in neurofibromatosis and their potential clinical use. *Neurology* **97**, S91–S98 (2021).
30. Vlasschaert, C. et al. A practical approach to curate clonal hematopoiesis of indeterminate potential in human genetic data sets. *Blood* **141**, 2214–2223 (2023).
31. Agaimy, A., Vassos, N. & Croner, R. S. Gastrointestinal manifestations of neurofibromatosis type 1 (Recklinghausen's disease): clinicopathological spectrum with pathogenetic considerations. *Int. J. Clin. Exp. Pathol.* **5**, 852–862 (2012).
32. Ejerskov, C., Krogh, K., Ostergaard, J. R., Joensson, I. & Haagerup, A. Gastrointestinal symptoms in children and adolescents with neurofibromatosis type 1. *J. Pediatr. Gastroenterol. Nutr.* **66**, 872–875 (2018).
33. Nguyen, K. A., Elnaggar, M., Gallant, N. M. & Tanios, M. Neurofibromatosis type 1: a case highlighting pulmonary and other rare clinical manifestations. *BMJ Case Rep.* **2018**, bcr2017222614 (2018).
34. LaDuca, H. et al. A clinical guide to hereditary cancer panel testing: evaluation of gene-specific cancer associations and sensitivity of genetic testing criteria in a cohort of 165,000 high-risk patients. *Genet. Med.* https://doi.org/10.1038/s41436-019-0633-8 (2019).
35. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
36. All of Us Research Program Investigators. The 'All of Us' research program. *N. Engl. J. Med.* **381**, 668–676 (2019).

37. Westbrook, L. et al. Hereditary cancer testing in a diverse sample across three breast imaging centers. *Breast Cancer Res. Treat.* **203**, 365–372 (2024).
38. Lawrence, L. et al. The implications of familial incidental findings from exome sequencing: the NIH Undiagnosed Diseases Program experience. *Genet. Med.* **16**, 741–750 (2014).
39. Jalkh, N., Mehawej, C. & Chouery, E. Actionable exomic secondary findings in 280 Lebanese participants. *Front. Genet.* **11**, 208 (2020).
40. Gordon, A. S. et al. Frequency of genomic secondary findings among 21,915 eMERGE network participants. *Genet. Med.* **22**, 1470–1477 (2020).
41. van Rooij, J. et al. Reduced penetrance of pathogenic ACMG variants in a deeply phenotyped cohort study and evaluation of ClinVar classification over time. *Genet. Med.* **22**, 1812–1820 (2020).
42. Cassa, C. A., Tong, M. Y. & Jordan, D. M. Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum. Mutat.* **34**, 1216–1220 (2013).
43. Treichel, A. M. et al. Diagnosis of mosaic tuberous sclerosis complex using next-generation sequencing of subtle or unusual cutaneous findings. *JID Innov.* **3**, 100180 (2023).
44. Eisfeld, A.-K. et al. NF1 mutations are recurrent in adult acute myeloid leukemia and confer poor outcome. *Leukemia* **32**, 2536–2545 (2018).
45. Parkin, B. et al. NF1 inactivation in adult acute myelogenous leukemia. *Clin. Cancer Res.* **16**, 4135–4147 (2010).
46. Side, L. E. et al. Mutations of the NF1 gene in children with juvenile myelomonocytic leukemia without clinical evidence of neurofibromatosis, type 1. *Blood* **92**, 267–272 (1998).
47. Weinstock, J. S. et al. The genetic determinants of recurrent somatic mutations in 43,693 blood genomes. *Sci. Adv.* **9**, eabm4945 (2023).
48. Kessler, M. D. et al. Common and rare variant associations with clonal haematopoiesis phenotypes. *Nature* **612**, 301–309 (2022).
49. Howlader N. et al. (eds) SEER Cancer Statistics Review, 1975–2012, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/archive/csr/1975_2012/, based on November 2014 SEER data submission, posted to the SEER web site, April 2015. *SEER* https://seer.cancer.gov/archive/csr/1975_2012/index.html.
50. Liu, Y. L. et al. Pre-operative neoadjuvant chemotherapy cycles and survival in newly diagnosed ovarian cancer: what is the optimal number? A Memorial Sloan Kettering Cancer Center Team Ovary study. *Int. J. Gynecol. Cancer* **30**, 1915–1921 (2020).
51. Corrigan, R. R., Mashburn-Warren, L. M., Yoon, H. & Bedrosian, T. A. Somatic mosaicism in brain disorders. *Annu. Rev. Pathol.* https://doi.org/10.1146/annurev-pathmechdis-111523-023528 (2024).
52. Chen, M. H. et al. Contributions of germline and somatic mosaic genetics to thoracic aortic aneurysms in nonsyndromic individuals. *J. Am. Heart Assoc.* **13**, e033232 (2024).
53. Morren, M.-A. et al. Mosaic RASopathies concept: different skin lesions, same systemic manifestations? *J. Med. Genet.* **61**, 411–419 (2024).
54. Park, J. et al. Exome-wide evaluation of rare coding variants using electronic health records identifies new gene-phenotype associations. *Nat. Med.* **27**, 66–72 (2021).
55. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
56. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
57. GitHub—konradjk/loftee. https://github.com/konradjk/loftee.
58. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* **12**, 103 (2020).
59. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
60. Packer, J. S. et al. CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics* **32**, 133–135 (2016).
61. Denny, J. C. et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
62. Moore, C. B., Wallace, J. R., Frase, A. T., Pendergrass, S. A. & Ritchie, M. D. BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. *BMC Med. Genom.* **6**, S6 (2013).
63. Ferner, R. E. et al. Guidelines for the diagnosis and management of individuals with neurofibromatosis 1. *J. Med. Genet.* **44**, 81–88 (2007).
64. Aschebrook-Kilfoy, B. et al. An overview of cancer in the first 315,000 All of Us participants. *PLoS ONE* **17**, e0272522 (2022).

Acknowledgements

We would like to thank members of the Ritchie Lab who provided scripts or analysis support for aspects of these analyses. We acknowledge the Penn Medicine BioBank (PMBB) for providing data and thank the patient-participants of Penn Medicine who consented to participate in this research program. We would also like to thank the Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for analysis; complete contribution lists for both are included in the supplemental information section. The PMBB is approved under IRB protocol# 813913 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878. This research has been conducted using data from UK Biobank, a major biomedical database, under Application Number 69929. The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants. S.A.F. is supported by 5T32HG008961. T.G.D. is supported in part by the NIH K08 1K08DK127247 and by the Burroughs Wellcome Fund. M.D.R. is supported in part by NIH R01 AI077505, GM115318, AI116794. K.L.N. is supported in part by the Breast Cancer Research Foundation, Basser Center for BRCA, the Gray Foundation, and the National Institutes of Health—R01 CA176785 and R01 CA192393. J.W. is supported by the Breast Cancer Research Foundation. T.T.N. is supported by 5T32GM008638.

Author contributions

A.S. and T.T.N. are equal contributors to this work and designated as co-first authors. K.L.N. and T.G.D. conceived and planned the experiments and are equal contributors to this work and are designated as co-last authors. A.S., T.T.N., T.O., B.S.W. Y.S., S.A.F., and T.G.D. carried out the analyses. E.C., C.H., J.S.D, A.Y., M.R., V.S., S.L., K.L., J.Z., C.B.S., A.E., Y.S., J.W., K.L.N., and T.G.D. verified the analytical methods. A.S., T.T.N., E.C., C.H., J.S.D, A.Y., M.R., V.S., S.L., K.L., J.Z., C.B.S., A.E., Y.S., and J.W. contributed to data preparation and analysis. E.C., C.H., J.S.D, A.Y., M.R., V.S., S.L., K.L., J.Z., C.B.S., A.E., Y.S., and J.W. contributed to sample preparation. A.S., T.T.N, E.C., C.H., J.S.D, A.Y., M.R., V.S., S.L., S.K., and

M.D.R. contributed to the interpretation of the results. K.L.N. and T.G.D. supervised the project. T.T.N., Z.C.B., M.B., A.R., E.F., E.M.S., S.K., K.L.N., and T.G.D. carried out clinical evaluations and documentation. T.G.D. took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

Competing interests

E.C., C.H., J.S.D., A.Y., M.R., V.S., and S.L. are employees of Ambry Genetics. K.L., J.Z., C.B.S., A.E., and Y.S. are employees of Natera Inc. J.W. is a consultant for Natera Inc. The authors declare no other competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57077-1.

Correspondence and requests for materials should be addressed to Katherine L. Nathanson or Theodore G. Drivas.

Peer review information *Nature Communications* thanks Pedro Quiros, Heiko Runz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

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

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

[1]Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [2]Breast Medicine Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [3]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [4]Department of Clinical Diagnostics, Ambry Genetics, Aliso Viejo, CA, USA. [5]Division of Translational Medicine and Human Genetics, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [6]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [7]College of Arts and Sciences, Oberlin College, Oberlin, OH, USA. [8]College of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA. [9]Natera Inc., Austin, TX, USA. [10]University of Alabama in Birmingham, Heersink School of Medicine, Department of Genetics, Birmingham, AL, USA. [11]HudsonAlpha Institute of Biotechnology, Huntsville, AL, USA. [12]Division of Precision Prevention, Department of Medicine, University of Kansas School of Medicine, Kansas City, KS, USA. [13]Basser Center for BRCA and Abramson Cancer Center, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [14]These authors contributed equally: Anton Safonov, Tomoki T. Nomakuchi. [15]These authors jointly supervised this work: Katherine L. Nathanson, Theodore G. Drivas. ✉e-mail: knathans@upenn.edu; theodore.drivas@pennmedicine.upenn.edu

Penn Medicine BioBank

Marylyn D. Ritchie [6] & Theodore G. Drivas [5,6,15] ✉