The prognostic effects of somatic mutations in ER-positive breast cancer

Authors:

Obi L Griffith, PhD[1,2,3,4,*], Nicholas C Spies, BSc[1,*], Meenakshi Anurag, PhD[5,6*], Malachi Griffith, PhD[1,2,3,4], Jingqin Luo, PhD[3,6], Dongsheng Tu, PhD[8], Belinda Yeo, PhD[9], Jason Kunisaki, BSc[1], Christopher A Miller, PhD[1,2], Kilannin Krysiak, PhD[1,2], Jasreet Hundal, MSc[1], Benjamin J Ainscough, BSc[1], Zachary L Skidmore, MEng[1], Katie Campbell, BSc[1], Runjun Kumar, BSc[2], Catrina Fronick, BSc[1], Lisa Cook, BSc[1], Jacqueline E Snider, BSc[2], Sherri Davies, PhD[2], Shyam M Kavuri, PhD[5,6], Eric C Chang, PhD[5,6], Vincent Magrini, PhD[1,4,10], David E Larson, PhD[1], Robert S Fulton, MSc[1,4], Shuzhen Liu, MSc[8], Samuel Leung, MSc[8], David Voduc, MD[8], Ron Bose, MD, PhD[2], Mitch Dowsett PhD, FMedSci[9], Richard K Wilson, PhD[1,3,4], Torsten O Nielsen, MD, PhD[8], Elaine R Mardis, PhD[1,3,4,10,†], Matthew J Ellis MB, BChir, PhD[5,6†]

Affiliations:

1. McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO
2. Department of Medicine, Division of Oncology, Washington University School of Medicine, St. Louis, MO
3. Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO
4. Department of Genetics, Washington University School of Medicine, St. Louis, MO
5. Lester and Sue Smith Breast Center and Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX
6. Department of Medicine, Baylor College of Medicine, Houston, TX
7. Division of Biostatistics, Washington University School of Medicine, St. Louis MO
8. Genetic Pathology Evaluation Centre, University of British Columbia, Vancouver, Canada
9. Institute of Cancer Research, London, UK
10. Current address: Nationwide Children's Hospital and Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH

* These authors contributed equally.

[†] Corresponding authors. matthew.ellis@bcm.edu; elaine.mardis@nationwidechildrens.org

Author emails:

obigriffith@wustl.edu, nspies@wustl.edu, anurag@bcm.edu, mgriffit@wustl.edu, jingqinluo@wustl.edu, dtu@ctg.queensu.ca, belinda.yeo@onjcri.org.au, kunisakijh@wustl.edu, c.a.miller@wustl.edu, kkrysiak@wustl.edu, jhundal@wustl.edu, b.ainscough@wustl.edu, zskidmor@wustl.edu, katiecampbell@wustl.edu, runjunkumar@wustl.edu, cfronick@wustl.edu, cooklisa@wustl.edu, jsnider@dom.wustl.edu, sdavies@dom.wustl.edu, meghashyam.kavuri@bcm.edu, echang1@bcm.edu, vincent.magrini@nationwidechildrens.org, delarson@wustl.edu, rfulton22@wustl.edu, shuzhensuzanne.liu@vch.ca, samuel.leung@vch.ca, dvoduc@bccancer.bc.ca, rbose@dom.wustl.edu, mitchell.dowsett@icr.ac.uk, richard.wilson@nationwidechildrens.org, torsten@mail.ubc.ca, elaine.mardis@nationwidechildrens.org, matthew.ellis@bcm.edu

Abstract

DNA from primary breast cancer samples from 625 postmenopausal (UBC-TAM series) and 328 premenopausal (MA12 trial) hormone receptor-positive (HR+) patients were subjected to targeted sequencing of 83 genes to determine interactions between somatic mutation and prognosis. Independent validation of prognostic interactions was achieved using data from the METABRIC study. Previously established associations between MAP3K1 and PIK3CA with luminal A status/favorable prognosis and TP53 mutations with Luminal B/non-luminal tumors/poor prognosis were observed, validating the methodological approach. As observed in UBC-TAM, *NF1* frame-shift nonsense *(FS/NS)* mutations were also a METABRIC-validated poor outcome driver. For MA12, poor outcome associated with PIK3R1 mutation was also reproducible in METABRIC. DDR1 mutations were strongly associated with poor prognosis in UBC-TAM despite stringent false-discovery correction (q=0.0003). In conclusion, uncommon recurrent somatic mutations should be further explored to create a more complete explanation of the highly variable outcomes that typifies ER+ breast cancer.

Introduction

While recent genomic studies have provided a comprehensive catalog of genes that accumulate somatic point mutations and small insertions/deletions (indels) in estrogen receptor-positive (ER+) breast cancer, there remains considerable uncertainty as to how these newly discovered mutations relate to disease outcomes [1, 2, 3]. Most genomic discovery cohorts were neither uniformly treated nor followed long enough. For ER+ disease in particular, prognostic studies require prolonged observation since relapses often occur after 5 years [4]. Uniform treatment was a feature of a whole genome sequencing study of samples accrued from a neoadjuvant aromatase inhibitor (AI) clinical trial for ER+ clinical stage 2 or 3 disease, although only short-term anti-proliferative response to AI were reported. This investigation identified that mutations in *MAP3K1*, a tumor suppressor gene involved in stress kinase activation, were associated with indolent biological features and low proliferation rates [5]. The resulting hypothesis was that *MAP3K1* mutation would be associated with favorable outcomes. In contrast, *TP53* mutations associated with poor prognosis features and high proliferation rates.

To more comprehensively address the relationships between somatic mutations and outcomes in ER+ breast cancer, we developed an approach to detect somatic mutations in DNA isolated from formalin fixed tumor blocks that were over 20 years old. After curating existing mutational data from breast cancer genomics discovery studies (Supplementary Data 1), 83 genes were chosen for analysis (Supplementary Table 1). We applied DNA hybrid capture, sequencing and somatic analysis to three ER+ breast cancer discovery cohorts with contrasting clinical characteristics: An older cohort treated with adjuvant tamoxifen and no chemotherapy (UBC-TAM series[6]), a premenopausal cohort uniformly treated with chemotherapy and randomized to tamoxifen versus observation (NCIC MA12 clinical trial[7]); and a third mixed cohort that was used only to expand the mutational landscape analysis (POLAR) (Supplementary Table 2). An analytical pipeline was developed to identify somatic variants while compensating for the lack of matched normal DNA, which is generally unavailable in the setting of older formalin-fixed tumor material. Somatic mutations were analyzed for association with standard clinical variables, wherein mutated *TP53* and *MAP3K1* served as *a priori* hypotheses for poor and good outcome, respectively. Additional objectives were to identify new mutational hotspots, assess interactions with PAM50-based intrinsic subtypes and to determine mutation frequencies for therapeutic targets. Validation was possible by comparing our results to those in cBioPortal where the genes sequenced in the METABRIC cohort overlapped with the 83 genes investigated in the study described herein.

Results

Sequencing and final study cohorts

University of British Columbia Tamoxifen Series (UBC-TAM): These cases were drawn from a well-annotated cohort of patients treated with adjuvant tamoxifen without chemotherapy [6]. A total of 625 of 632 (98.8%) patient samples that fully met study criteria passed a minimum sequencing quality cutoff of at least 80% of targeted exonic bases covered at greater than 20X (mean coverage: 133X) with other quality metrics described in the supplementary data (Supplementary Figure 1-5 and Supplementary Data 2). Mean depth was correlated with input DNA and negatively correlated with time since diagnosis (approximate age of sample) and duplication rates were negatively correlated with input DNA and positively correlated with sample age. However, despite these trends, overall metrics were excellent with an average of 135.8X coverage and 3.0% duplicate rate despite the generally low input amounts and old sample age. The final patient population had an average age of 67 at diagnosis (range: 40-89+). All were treated with five years of adjuvant tamoxifen, and were primarily postmenopausal, grade 2 or 3 cancers, of ductal histologic subtype (Supplementary Table 2). All were ER+ (>1% cells positive by IHC) and at least 88.6% were clinically HER2- (13/625 unknown). A subset of 463 of these patients had PAM50 subtyping data available from a previous study [6]. The median follow up in the cohort examined was 25 years and one month.

NCIC-MA12 Trial cohort: These cases were drawn from a clinical trial in premenopausal women treated with a standard adjuvant chemotherapy regimen and randomized to tamoxifen versus observation. A total of 459 patient samples passed the minimum sequencing quality threshold (mean coverage: 200X), of which 328 were hormone receptor positive (HR+; >1% cells positive for ER or PR by IHC), and only the HR+ cohort were included here for most analyses. The majority were premenopausal (mean age of 45). All patients received chemotherapy, and 48% were treated with 5 years of adjuvant tamoxifen. A subset of 255 of these patients had PAM50 subtyping data available. The median follow up in the cohort examined was 9.7 years

POLAR cohort: This patient series was a case-control study of ER+ (>1% cells positive by IHC) breast tumors, 175 of 194 (90.2%) patient samples passed minimum sequencing quality thresholds (mean coverage: 75X). A case was defined as any patient who relapsed during follow-up, and controls were defined as lacking relapse through a similar follow-up duration. Based on these definitions, there were 91 cases and 84 controls. Of the cases, 43 were early relapses (<5 years since diagnosis) and 48 were late relapses (>5 years). Patients were only included if they received adjuvant endocrine therapy, but chemotherapy was not an exclusion criterion, nor was menopausal status. Because the POLAR study was a case-control design, outcome data could not be easily integrated into prognostic analysis. Therefore, these cases were used in the mutation landscape and hotspot analyses only.

Across the three cohorts, there were 1,259 patient samples that passed minimum sequencing quality thresholds and 1,128 of these were ER+ (UBC-TAM and POLAR) or ER and/or PgR+ (HR+) (MA12).

Variant calling and filtering

A total of over 62 million variants were identified in UBC-TAM. After extensive filtering against a set of nearly 70,000 unmatched normal samples and manual review to eliminate common polymorphisms and false positives (see methods), 1,991 putative somatic variants were identified (0 to 26 variants per patient). A set of 1,693 mutations was defined as the “non-silent” set for further analysis that excluded sequencing variants in splice regions (except proximal splice site), RNA genes (except *MALAT1*), UTRs, introns, and all silent mutations. Finally, a set of 408 frameshift or nonsense mutations was defined. The same filtering method was applied to both the POLAR and MA12 datasets. A total of 540 putative somatic mutations (436 non-silent, 145 FS/NS) were identified in POLAR, and 2,104 (1,753 non-silent, 610 FS/NS) in MA12. Full details on these variants are included in Supplementary Data 3 and summarized for key genes in Supplementary Figure 6.

Mutation landscape analysis

In 1128 samples passing quality control standards, considering only non-silent mutations, 17 genes were mutated at a rate greater than 5%, and 6 at a rate greater than 10%; *PIK3CA* was the only gene mutated in greater than 20% of samples (**Figure 1A**). The order from most recurrent to least for the 10 most frequently mutated genes was: *PIK3CA* (41.1%), *TP53* (15.5%), *MLL3* (13.4%), *MAP3K1* (12.0%), CDH1 (10.5%), MALAT1 (10.0%), GATA3 (9.1%), MLL2 (8.7%), ARID1A (7.2%), and BRCA2 (6.6%). This list correlates well with previously reported recurrently mutated genes. For example, the top 4 most significantly mutated (non-silent) genes in the ER+ subset of TCGA breast project[3] were *PIK3CA* (24.0%), *TP53* (14.6%), *GATA3* (8.6%) and *MAP3K1* (6.1%). Considering METABRIC ER+ patients, the most recurrently mutated genes were PIK3CA (~46%), TP53 (~21%), GATA3, MLL3, CDH1, and MAP3K1 (all ~12-14%) demonstrating slightly higher but very similar frequencies. The overall average non-silent mutation frequency was estimated as 1.6 per MB of coding sequence (range: 0.5 to 5.8 mutations per MB, excluding samples with no mutations called). In order to determine whether mutations in any gene pair were mutually exclusive or co-occurring in this dataset, a pairwise Chi-squared or Fisher's exact test was performed. Mutations in PIK3CA and MAP3K1 were significantly more likely to co-occur (after BH FDR correction) in TAM dataset, and were near significance in MA12 although not after correction (p = 0.08). These results are summarized in Supplementary Data 4.

Hotspot analysis

As anticipated [8], mutations in *PIK3CA* at *E542K*, *E545K*, and *H1047R* were highly recurrent in this study with 69/1259 (5.5%) E542K, 104 (8.3%) E545K, and 181 (14.4%) H1047R mutations (Supplementary Figure 6C). Mutations in the ligand binding domain of *ESR1* (1.1%) were extremely rare [3, 9, 10] (Supplementary Figure 6A). To uncover novel hotspots in these data, both Chi-squared and Fisher's exact tests were performed using mutation frequencies from previous sequencing studies as the expected values (see Methods for definition of multi-study MAF file) (Supplementary Table 3). The most notable novel finding was in *CBFB* (**Figure 1B**). At least 6 different genomic alterations were observed in 15 patients (Supplementary Data 3) that affected the donor splice site of exon 2. Manual review of this splice site identified at least two additional patients with evidence for mutations at this location. The predicted effect of these mutations is skipping of exon 2 or alternate donor site usage, each likely resulting in loss-of-function of the *CBFB* protein. Additional splice site mutations were observed at the exon 2, exon 4 and exon 5 acceptor sites of *CBFB*. ErbB2 exhibited the anticipated profile of activating mutations from earlier publications[11] with 22/1259 (1.7%) samples harboring known activating mutations and another 6 variants of unknown significance in the kinase domain or at the S310 residue (**Figure 8C**).

Somatic mutation association with PAM50-based intrinsic subtype

PAM50 intrinsic subtype calls were obtained from previously published analyses to compare to their mutational profiles for UBC-TAM and MA12 (HR+ only) studies. In both studies about half the patients had luminal A tumors. However, the MA12 cohort had a higher proportion of non-luminal subtypes, with 19.8% HER2-E and 6.6% basal and fewer luminal B tumors (25.1% versus 42.4%) (**Figure 2A-B**). As expected, patients with the HER2-E intrinsic subtype were enriched for HER2+ve status compared to other subtypes (Fisher's exact test p<0.0001). Of interest, in the HER2-enriched group there were 51 tumors that were not HER2 amplified and of these 4 were HER2 mutant (~8%), indicating that HER2 mutation could be an occasional explanation for a HER2-E subtype assignment in the absence of HER2 amplification. For NF1 FS/NS mutations, there was also a statistically significant association with the HER2-E subtype (P=0.002) (supplementary Figure 7B, also supplementary data 5). Notably NF1 non-silent mutations were enriched in the HER2-E non-HER2 amplified subgroup, where they were present in 8/51 cases (16%). Compared to the frequency in all other subtypes 12/582 (2%), this enrichment was significant (Fishers exact test p<0.0001) (Supplementary 7A right panel). This association could be reproduced in the METABRIC data with an NF1 non-silent mutation incidence in the HER2-E non HER2 amplified group of 8/80 (10%) versus 35/1283 (3%) in the rest of the subtypes (p=0.003) (Supplementary 7A right panel). Age density plots by subtype serve to emphasize the large difference in the median age between the two sample cohorts (43 versus 65), and also the influence of age with respect to the intrinsic subtype incidence. Namely, in the younger MA12 cohort, there is a younger peak incidence with basal-like breast cancer than Luminal A disease (**Figure 2D**). In contrast in the older UBC-TAM cohort, an influence of age on intrinsic subtype was not observed (**Figure 2C**). Relationships between intrinsic subtype and mutation patterns were also explored, classifying mutation positive status as "non-silent", "missense", nonsense/frame-shift (FS/NS) or FS/NS+splice site (Supplementary Data 5). The FDR corrected p-value (q-value) took into account that 83 genes were examined. However, this level of false discovery detection could be viewed as overly conservative in an exploratory analysis. Therefore, any gene mutation with q-value association of <0.2 was therefore considered reportable for the purposes of subsequent validation efforts [12, 13, 14]. For MA12, non-silent TP53 mutation was highly subtype-associated because of the very high incidence in non-luminal versus luminal subtypes. PIK3CA and MAP3K1 mutations were associated with Luminal A disease in both cohorts (Supplementary Figure 7B). Finally, there was a strong association between Luminal B status and non-silent (Supplementary Figure 7A) as well as FS/NS mutations in GATA3 (Supplementary Data

5, q value = 0.006) for MA12 (but not UBC-TAM). GATA3 mutations were present in 28-30% of Luminal B cases and less so in luminal A cases (5%). Considering q values of <0.2 the associations between FS/NS and non-silent mutations in ATM and Luminal B tumors in MA12 (8-13%) suggests that ATM disruption is also a possible luminal B driver (Supplementary Figure 7C), at least in younger women (MA12). Relationships between age and mutation incidence were therefore also explored (Supplementary Figure 7D**)**, with the finding that both ATM mutation and GATA3 mutations were associated with an earlier age of onset within the luminal B category (**Figure 2E and 2F**). Some ATM mutations are likely to be germline (see discussion below), which could partially explain the association with younger age.

Survival analysis according to somatic mutation.

For the UBC-TAM Series (**Figure 3A**) univariate analysis, favorable prognostic associations for breast-cancer-specific survival (BCSS) were detected for non-silent mutations in *MAP3K1*, *ERBB3*, XBP1 and PIK3CA (**Figure 3B**, Supplementary Data 6). Adverse prognostic effects were observed for non-silent mutations in *DDR1* and *TP53*, as well as for frame-shift and nonsense (FS/NS) mutations in NF1. An analysis for recurrence free survival (RFS) produced similar results, except for ARID1B, which was marginally associated with more favorable outcome. A multivariate Cox model was applied to put each gene in the context of clinical parameters (grade, tumor size and node status). These analyses indicated that the prognostic effects of non-silent DDR1, PIK3CA, GATA3 FS/NS, TP53 and MAP3K1 mutations were independent of grade and pathological stage (**Figure 3C**). Multiple correction testing, yielded DDR1 as the only gene that remained significant with a q-value of 0.0003 (Supplementary Data 5). For the MA12 clinical trial cohort (**Figure 4A**) we focused on overall survival associations, as this was the primary endpoint of the study and the most robust endpoint. A number of rarely mutated genes were associated with poor outcome in univariate analysis as displayed in **Figure 4B**. Multiple testing corrections indicated none of these findings could be considered significant [12, 13, 14]. However, in multivariate analysis, based on the uncorrected p value, the prognostic effects of mutations in ErbB2, ErbB4, LTK FS/NS, MAP3K4, PIK3R1, RB1, RELN and TGFB2 were independent of pathological stage and grade (**Figure 4B**).

Verification of Prognostic effects of Mutations in METABRIC data.

While few genes were significant in univariate analysis after multiple testing correction, their identification provides valuable hypotheses for further testing and validation. We therefore sought additional data in the public domain to further assess the uncorrected p value-based findings in our data set. The METABRIC consortium have reported somatic mutations in cBioPortal [15] with co-reported detailed hormone receptor status, age at diagnosis (median age=64 years for ER+ patients), mean follow up of >8 years and disease-specific (breast-cancer-specific) outcome [16, 17]. This data set provided the opportunity to conduct a validation exercise for overlapping genes in the two data sets. For the UBC-TAM series (**Figure 3**), 9 genes with a univariate p value of <0.05 were brought forward for validation (**Figure 5**). Of the 6 overlapping genes also examined in METABRIC, consistent prognostic effects independent of clinical variables were observed for non-silent mutations in three genes, *MAP3K1* (favorable), *TP53* (unfavorable) and *NF1* FS/NS mutations (unfavorable). In order to maintain coherence in discovery and validation patient cohorts, a similar analysis was carried out restricting the patient pool to postmenopausal patients only. No significant variation in hazard ratio for candidate genes where observed (Supplementary Table 4). For the MA12 series (**Figure 4**), 5 shared genes were identified with univariate p values of <0.05, yet only *PIK3R1* mutations (non-silent or FS/NS) showed consistent adverse prognostic effects (**Figure 6**). The Kaplan Meier survival plots for the consistent adverse prognostic effects of *NF1* FS/NS (TAM vs METABRIC) and non-silent *PIK3R1* (MA12 vs METABRIC) mutations are illustrated in **Figure 7A-D**. Copy number aberrations and chromosomal instability have been associated with prognosis across multiple cancer types, including ER-positive (ER+) breast cancer[16, 18, 19]. To gauge the confounding nature of commonly amplified genes in breast cancer, we further performed multivariate analysis on the candidate genes with cases of amplification of MYC, FGFR1, CCND1 and ERBB2 (Supplementary Table 5). We did not observe a significant change in the hazard ratio reported in Figure 5B and 6B).

Prognostic interactions between PIK3CA and MAP3K1.

Since PIK3CA and MAP3K1 mutations co-associate, the combined effect of non-silent mutations in these genes was examined. Patients with tumors exhibiting both genes mutated have a more favorable clinical course than either singly mutant cases or cases without either gene mutated. While the prognostic effects were strongest in the UBC-TAM series, this result was also reproduced in the METABRIC data (**Figure 7E-F**).

Mutation Analyses for Uncommon Targetable Kinases.

Of the 83 genes analyzed, at least 8 are directly targetable with small molecules or antibodies that are either FDA approved or in late-stage development (**Figure 8**). Pre-existing data on these mutations is summarized (Supplementary Data 7). PIK3CA is not further discussed here, since the mutation spectrum is well described and large therapeutic studies are already underway. An examination of the 23 mutations in ERbB2 revealed locations that were, as expected, clustered in 2 major domains, with 2 of 23 having extracellular domain mutations at residue 310 and 21 of 23 having kinase domain mutations between residues 755-842 [11, 20]. To further investigate the preliminary finding of an adverse prognostic effect for ErbB2 mutation in the MA12 series, an examination of the METABRIC data indicated that known activating mutations in ErbB2 were associated with a near significant adverse effect (HR=1.71, P=0.075) (Supplementary Figure 8).

For ERBB3, 2 known-activating mutations were identified (V104L and E928A)[21]. The DDR1 kinase domain mutation, R776W, is possibly homologous to EGFR hot spot mutation L858R, but the remaining DDR1 variants are of unknown significance. For the mutations in JAK1, 3 of 12 are loss of function mutations (frame shift or non-sense) and the S816* mutation has been reported in a lung adenocarcinoma sequencing data set [22]. The loss of function mutations in JAK1 have been shown to associate with immunotherapy resistance [23, 24]. A few mutations identified in ERBB4, MET, and PDGFRA have been previously reported but those reported here have not been functionally tested.

Discussion

The strength of this investigation includes the prolonged follow up, controlled adjuvant treatment and the relatively large number of genes and patients studied. Weaknesses include the lack of treatment prediction because endocrine treatment in UBC Tam was uniform but not randomized. In MA12 the use of tamoxifen was randomized, but the numbers were too small to examine treatment interactions. The landscape of recurrently mutated genes in ER+ breast cancer observed in this study is consistent with reports where matched germline samples were available, indicating that our variant filters were effective for somatic mutation detection in a research setting. Overall, mutation frequencies were higher in our cohort (e.g., for *PIK3CA*, *MLL3*, *MAP3K1*) than the TCGA cohort, but were also lower for a few specific genes (e.g., *TP53* and *GATA3*). Due to higher sequencing data coverage of recurrently mutated target genes than TCGA and the use of a different hybrid capture reagent, we were likely able to detect mutations that were missed with lower-depth exome or whole genome sequencing data. Differences in patient populations may also be a factor. Frequencies were much closer to reported values for METABRIC which also used a targeted sequencing approach. It is also possible that in some instances we overestimated somatic mutation frequency, due to the lack of matched normal samples and imperfections in our germline polymorphism filtering. In particular, a significant number of *BRCA1*, *BRCA2*, and *ATM* mutations are likely *de novo* germline mutations that we would not be able to easily distinguish from somatic mutations. Of the 117 non-silent *BRCA1/2* mutations observed (from 110/1128 patients across all 3 cohorts; 7 patients had two hits) 74 were observed at a VAF greater than 40% and 31 were greater 60%. Additionally, of the 61 non-silent *ATM* mutations (from 58/1128 samples; 3 samples had 2 hits) 39 had VAF greater than 40% and 18 had VAF greater than 60 (Supplementary Data 9). Variants with VAFs this high are less likely to be somatic given the general expectation of impure tumor samples and heterozygous mutations. Indeed, the VAFs for *BRCA1/2* and *ATM* non-silent mutations (mean=46.0%) were significantly higher than for other genes (mean=36.7%, p=5.92e-09). Even when considered separately, the VAFs for *BRCA1* (mean=46.6%), *BRCA2* (mean=43.8%) and *ATM* (mean=48.2%) were significantly higher than the other genes (p=0.002, p=0.0015, and 5.27e-5 respectively). Among the *BRCA1/2* variants, there were 8 known pathogenic (ENIGMA expert reviewed) mutations according to a search of the BRCA Exchange database (http://brcaexchange.org, Nov 12, 2017) and another 37 assumed pathogenic (FS/NS) mutations. Of the remaining, 4 were benign according to expert review (ENIGMA), and 8 benign, 15 likely benign and 45 variants of unknown significance according to all public sources. Out of the 61 *ATM* variants queried in ClinVar, 4 were designated as pathogenic, 3 were pathogenic/likely pathogenic, and 2 were likely pathogenic. Another 7 were frameshift mutations and assumed pathogenic. Additionally, 23 variants had uncertain significance, 8 variants had conflicting interpretations of pathogenicity (any combination of benign, likely benign, or uncertain significance), and the remaining 14 variants had no data. *ATM* variants were also queried in the Leiden Open Variation Database (LOVD) [25], which identified 1 variant that affects function (designated as likely pathogenic by ClinVar), 10 variants with unknown effect, and 1 variant that probably does not affect function (uncertain significance in ClinVar). The remaining variants had no data in LOVD. Given these complexities the prognostic effects of somatic versus germline BRCA1/2 and ATM mutations remain unresolved, however attention should clearly be paid to therapeutic strategies for these patients. The ATM findings deserve a particular highlight because of the younger age/luminal B association and the current lack of studies devoted to this population.

The discovery of a novel recurrent *CBFB* (core binding factor subunit beta) splice site mutation in this cohort illustrates a limitation of exome capture reagents. The affected bases in exon 2 of CBFB display reduced sequence coverage, possibly due to high GC content, in the breast TCGA exome dataset (Supplementary Figures 9-10). This site was mutated in at least 1.5% of ER+ breast cancers sequenced, bringing the overall rate of CBFB mutations to nearly 6%, which should drive further investigation of this gene in ER+ breast cancer pathogenesis. *CBFB* functions as a subunit in a heterodimeric core binding transcription factor that interacts with *RUNX1*[26]. Consistent with this model, *CBFB* mutants were mutually exclusive from *RUNX1* mutants in this cohort with only a single sample harboring non-silent mutations in both *CBFB* and *RUNX1*.

The UBC-TAM and MA12 studies revealed different lists of potentially prognostic mutations. Prognostic effects are likely to be strongly affected by the use of systemic therapy as well as by patient age at diagnosis. The UBC-TAM series is the simplest study to interpret from a drug resistance perspective since the only systemic therapy was tamoxifen. Thus, the consistent adverse effect of NF1 FS/NS mutation on prognosis is intriguing as this result is consistent with results from an *in vitro* screen for tamoxifen resistance[27]. Understanding why only FS/NS mutations predict poor outcome, rather than missense or other non-silent mutations, will require further investigation. The association with the HER2-E, non-HER2 amplified subset with non-synonymous NF1 mutations was observed in both the discovery and validation (METABRIC) data sets. It is a logical proposition that mutations that activate RAS, like NF1 mutation, could create a tumor with a similar transcriptional phenotype as some HER2 amplified breast cancers. PIK3R1 mutation also emerged as a consistent poor prognosis mutation from the MA12 analysis, with validation in METABRIC. The proposed favorable prognostic effects of PIK3CA mutation were observed in the UBC-TAM series, but were not found to be independent of stage and grade, and PTEN mutations were neutral.

According to our validation results, NF1, PIK3R1, PIK3CA and TP53 are therefore likely to be prognostic drivers that are independent of clinical variables. In postmenopausal women treated with adjuvant endocrine therapy, DDR1, PRKDC and XBP1 should be further studied and of these DDR1 is the strongest candidate because it was significant despite strict false discovery correction. DDR1 is a collagen-binding receptor expressed in epithelial cells that stabilizes E-cadherin–mediated intracellular adhesion[28]. *DDR1* mutations also occur in endometrial cancer[29], acute leukemia[30] and lung cancer[31]. Loss of DDR1 (DDR1-null mice) produces hyper-proliferation and abnormal branching of mammary ducts, suggesting DDR1 is a breast tumor suppressor[32]. Mutations in PRKDC will potentially produce a defective ATM response/low ATM levels [33] which is interesting in the context of the finding herein that ATM mutations are a potential luminal B driver gene. The significance of a defective ATM pathway as a cause of endocrine resistance is highlighted by the recent finding that dysregulation of the MutL complex (MLH1, PMS1 and PMS2) causes failure of ATM/CHK2-based negative regulation of CDK4/6 [34]. Prognostic candidate mutations revealed by the MA12 analysis were different from the UBC TAM series, likely reflecting the different patient profiles and adjuvant treatments illustrated in **Figure 2**. The prognostic effects of mutations ERBB2, ERBB4, JAK1, LTK, MAP3K4, MET, PDGFRA, RB1, RELN, TGFB2, all await further study with even larger sample sizes.

A limitation of this study is that the mutation datasets we generated for UBC-TAM and MA12 cohorts lack comprehensive assessment of copy number signatures that have been associated with prognosis in ER+

breast cancer[16, 18, 19]. While multivariate analysis considering key CNVs did not appear to affect our prognostic associations, future studies may be needed to completely understand the interplay between simple and large-scale variation for prognostic prediction. Another limitation to this study was the heterogeneity in the datasets in terms of age, treatment, and other factors that limited direct comparison and made validation with METABRIC somewhat challenging. The collection of sufficiently large, uniformly treated populations with long-term follow-up for discovery and validation remains a challenge that must be addressed to fully characterize the prognostic significance of somatic mutations, especially low-frequency mutations.

In conclusion, we have successfully utilized clinically well-annotated, uniformly treated patient samples using DNA from archival material greater than 20 years old without a matched normal to explore the prognostic effects encoded by the mutational landscape of ER+ breast cancer. We were able to confirm our prospective hypothesis from our earlier studies [5] that MAP3K1 is associated with indolent disease and TP53 with adverse outcomes. We also associated NF1 FS/NS mutations with strong adverse effects on prognosis. Similarly, PIK3R1 mutations were associated with an adverse prognosis, in contrast to PIK3CA mutation which were weakly favorable. This suggests somatic mutations in these two physically interacting gene products are not biologically equivalent with respect to PI3 kinase pathway activation and resistance effects. The possibility that the long tail of low frequency mutation events in luminal type breast cancer may harbor multiple molecular explanations for poor outcomes should spur new collaborative efforts to thoroughly screen thousands of properly annotated cases. Only after these iterative efforts of proposing and confirming candidates will a clinically useful and comprehensive somatic mutation-based classification of ER+ breast cancer emerge. In the meantime, functional studies should be pursued to understand the biological effects of low frequency somatic mutations, prioritizing these studies according to whether the mutations are driving an adverse prognostic effect and whether their disruption creates a therapeutic vulnerability.

Methods

For the UBC-TAM series, an institutional review board approved study was based on formalin-fixed paraffin embedded (FFPE) primary tumor blocks from 947 female patients diagnosed with estrogen receptor positive invasive breast cancer in the province of British Columbia in Canada between 1986 and 1992[6, 35, 36, 37]. The sample flow and analysis are provided in a REMARK summary (**Figure 3A**). DNA was isolated from tumor-rich regions using the Qiagen blood and tissue kit, which yielded sufficient DNA in 645 samples, of which 625 met all study criteria and had sufficient sequence coverage. Similarly, approved studies provided 194 and 454 HR+ patient samples for the POLAR and MA12 (**Figure 4A**) cohorts. A total of 175 POLAR and 459 (328 HR+) MA12 samples yielded sufficient DNA and had sufficient sequence coverage for analysis. Detailed descriptions of the patient data sets are provided in Supplementary Table 3. A meta-analysis of six existing published large-scale breast cancer sequencing studies [1, 2, 3, 5, 38, 39] was performed to identify genes with recurrent coding region somatic mutations in breast cancer (Supplementary Data 1). Additional drug targets[40] and genes with relevance to breast cancer from targeted sequencing[41], copy-number studies[16] or knowledge relating to somatic or germline mutations (e.g., *BRCA1*, *BRCA2*, *ERBB2*, *ESR1* and *PRLR*) were also included. This resulted in a final list of 83 breast-cancer-related genes (Supplementary Table 1). These genes were targeted comprehensively with 3,029 complementary probes for hybridization-based enrichment (Supplementary Data 8). Sequencing libraries were constructed, hybridized with capture probes, multiplexed and run on a single flow cell with up to 96 samples per pool per lane yielding approximately 375 Mb of DNA sequence per sample from an Illumina HiSeq paired end 2 X 100bp (TAM) or 2 X 125bp (POLAR, MA12) sequencing run following manufacturer's protocols.

Variant calling was performed with the Genome Modeling System as previously described[42]. Specifically, sequence data were aligned to reference sequence build GRCh37 using BWA[43] and de-duplicated with Picard. SNVs and indels were detected using the union of samtools[44] and VarScan2[5] and annotated using Ensembl version 70. Variants were restricted to the coding regions of targeted genes and filtered for false positives and germline polymorphisms against a database of nearly 70,000 unmatched normals from the ExAC consortium[45], 1000 Genomes[46], NHLBI exomes[47] and TCGA data sets[3, 48]. A binomial probability model was then applied to the variants using VAF and total coverage to determine a log-likelihood ratio of being a somatic variant as previously described[49] (See Supplementary Methods). After filtering, all remaining variants were manually reviewed. To ensure that variants of known clinical relevance were not missed by automated variant calling approaches, a knowledge-based variant calling strategy was performed focused on the mutations in the Database of Curated Mutations[50].

Patient groups were defined by mutation status or truncating mutation status for each gene. Fisher's exact and Chi-squared tests were used for hotspot analysis, mutual exclusivity or co-occurrence, and other categorical clinical statistics (e.g., mutation status vs. intrinsic subtype) as appropriate. Univariate Kaplan-Meier and Cox survival analyses were performed for breast-cancer-specific survival (BCSS), relapse free survival (RFS), or overall survival (OS) with non-silent or truncating mutation status as a factor. Significant survival differences between the groups were determined by log rank (Mantel-Cox) test. The Benjamini-Hochberg method was performed for multiple testing corrections to report the false discovery rate adjusted p-value (q-value). A multivariate Cox proportional hazard model was fitted to BCSS and RFS separately on gene mutation status, node status, grade and tumor size and adjusted hazard ratios were calculated with Wald test p-values. All statistical analyses were performed in the R statistical programming language with core, 'survival' and 'multtest' libraries. Genomic visualizations were created with ProteinPaint[51] and GenVisR[52].

Data Availability

All mutation calls are made available as a MAF file with this publication. The raw sequence data from UBC-TAM patients are available in the database of Genotypes and Phenotypes (dbGaP) under accession number [dbGAP:phsxxxx.x]. Raw sequence data from MA12 and POLAR could not be deposited in public repository due to patient consent issues and complexities of institutional certification. However, these data are available from the authors (contact Obi Griffith and Matthew Ellis). Primary clinical outcome data for UBC-TAM and MA12 can be made available to qualified researchers through application to the Canadian Cancer Trials Group. Primary clinical outcome data for POLAR can be made available to qualified researchers through application to Mitch Dowsett at the Ralph Lauren Centre for Breast Cancer Research.

Acknowledgements

Research reported in this publication was primarily supported by Susan G. Komen Promise grant (PG12220321 to MJE), a Cancer Prevention and Research Institute of Texas (CPRIT) Recruitment of Established Investigators award (RR140033 to MJE). Dr. Ellis is a McNair Medical Institute Investigator and a Susan G. Komen Scholar. The study was also supported by DOD BCRP award No. W81XWH-16-0538 to MJE and EC. SMK was supported by a Komen CCR award (CCR16380599). The MA12 analysis was supported by research grants from Canadian Cancer Society Research Institute to the NCIC Clinical Trials Group (021039 and 015469). OLG was supported by the National Cancer Institute (NIH NCI K22CA188163 and NIH NCI U01CA209936).

Contributions

O.L.G., T.O.N., M.J.E., and E.R.M. designed the experiments; N.C.S, M.A., M.G., J. K., C.A.M., K.K., J.H., B.J.A., Z.L.S., K.C., R.K., C.F., L.C., J.E.S., S.D., V.M., D.E.L., R.S.F., S.L., and R.K.W. generated the sequencing data. T.O.N., B.Y., M.D., S.L., and D.V. orchestrated the sample pipeline, M.A., O.L.G. and N.C.S. prepared the figures and tables. O.L.G., N.C.S., M.A., J.L., and D.T. provided statistical analysis. S.M.K., R.B., and E.C.C. provided functional annotations. T.O.N. provided pathology analysis. M.J.E., N.C.S., M.A., and O.L.G. wrote the manuscript. E.R.M., T.O.N., and M.D., critically read and commented on the manuscript.

Conflict of Interest

Dr. Ellis and Dr. Mardis report income on patents on the PAM50 intrinsic subtype algorithm. Dr. Ellis reports ownership in Bioclassifier LLC that licenses PAM50 patents to Nanostring for the Prosigna breast cancer prognostic test. Commercial platforms and algorithms were not used in the analyses reported in this paper.

Figure Legends

Figure 1. Mutation recurrence and novel splice site mutation
A) The overall mutation recurrence rate ranged from 41.1% of samples for *PIK3CA* to 0.0% for *PIN1*. The figure depicts non-silent mutations for all 1128 patients for the top 17 most recurrently mutated genes (>5% recurrence). If a patient had multiple mutations it is colored according to the “most damaging” mutation following the order presented in the Mutation Type legend (vertical color bar). Mutations per MB were calculated using the total number of mutations observed over the total exome space corresponding to the tiled space from “SeqCap EZ Human Exome Library v2.0”. A correction factor was applied to account for genes not assayed using the expected number of additional mutations based on ER+ TCGA data. The coverage histogram (top sidebar) shows the percent of targeted exonic bases with at least 20X, 30X and 40X coverage. B) Mutation recurrence frequencies (amino acid level) in this study were compared to previously reported mutation frequency from a multi-study MAF file of six reported breast cancer sequencing studies (Supplementary Data 1). An entirely novel mutation “hot spot” was discovered affecting the exon 2 splice (donor) site of *CBFB* in at least 15 patients. Six different single nucleotide substitutions, insertions and deletions were observed, all affecting either the first or second base of the donor splice site. These mutations were most likely missed in previous studies because of a lack of sequencing coverage due to the GC-rich nature of exons 1 and 2 of *CBFB* (Supplementary Figures 9-10). Such mutations are predicted to significantly alter the canonical donor site and result in either alternate donor usage or skipping of one or more exons of *CBFB*.

Figure 2. Cross-cohort age and subtype analysis
A-B) Percentage composition of samples by intrinsic subtype of the tumor in the two discovery cohorts for UBC-TAM (A) and MA12 (B) cohorts. C-D) Age-density plots for patients categorized by intrinsic subtype in UBC-TAM (C) and MA12 (D) cohorts. The overall median age shows that UBC-TAM is constituted mostly of post-menopausal patients (median age=65), in contrast to MA12, which has younger patients (median age=43). E-F) Younger luminal B subtype patients harbor GATA3 (E) and ATM (F) mutations in the combined set of UBC-TAM and MA12 Luminal B cases (median age=52, p=0.01; median age=58, p=0.03 for GATA3 and ATM respectively).

Figure 3. Candidate discovery from UBC-TAM cohort and prognosis evaluation
(A) DNA was extracted from tumor specimens from 947 patients with ER+ breast cancer treated with tamoxifen monotherapy for 5 years. 632 samples with adequate yield were sequenced for 83 genes known to be recurrently mutated or breast cancer relevant. A total of 625 samples passed minimum quality checks and were sequenced to an average of 135.8X coverage. A total of ~62 million variants from the reference genome were identified. Extensive filtering and manual review reduced this list to 1,991 putatively somatic variants. Survival analysis was applied to non-silent and truncating gene mutation status versus disease outcome (relapse or breast-cancer-specific death). In addition, mutations were analyzed for novel hotspots, patterns of mutual exclusivity or co-occurrence and association with clinical variables. (B) Forest plot of impact of mutations in candidate genes, identified using the UBC-TAM population, on breast-cancer-specific survival (red) and recurrence-free survival (blue). The variant types are characterized based on non-silent or nonsense/frameshift (FS/NS) mutations. The box size is relative to frequency of mutations in the analysis, with larger boxes representing higher incidence mutations. (C) Multivariate forest plot of effect of mutations in UBC-TAM candidate genes on breast-cancer-specific survival when assessed together with clinical factors including Tumor Grade, Node positivity and Tumor Size (>5cm).

Figure 4. Candidate discovery from MA12 cohort and prognosis evaluation
(A) DNA was extracted from tumor specimens and 470 samples with adequate yield were sequenced for 83 genes known to be recurrently mutated or breast cancer relevant. A total of 459 (328 HR+) samples passed minimum quality checks and were sequenced to an average of 272.6X coverage. A total of 406 million variants from the reference genome were identified. Extensive filtering and manual review reduced this list to 2104 putatively somatic variants. Survival analysis was applied to non-silent and truncating gene mutation status versus overall survival. (B) Forest plot showing effect of mutation in candidate genes on overall survival (univariate - blue, multivariate - orange), along with the clinical factors used in the multivariate analysis (black),

tumor grade, node positivity and tumor size (>5cm). The box size is relative to frequency of mutations in the analysis, with larger boxes representing higher incidence mutations. Note: a few boxes are not shown if their hazard ratios were greater than 4.0.

Figure 5. Validation of UBC-TAM candidates in ER+ METABRIC

A) Six out of nine candidate genes from UBC-TAM analysis had mutations reported in the METABRIC cohort. 1,060 ER+ samples with breast-cancer-specific survival information were used to test the effect of mutations in the candidate genes on prognosis. B) Forest plot shows effect of mutated candidate genes on breast-cancer-specific survival in METABRIC ER+ cohort with univariate cox proportional-hazard ratio in blue and multivariate in orange. The clinical factors used in the multivariate analysis, namely tumor grade, node positivity and tumor size (>5cm), are shown in black. The box size is relative to frequency of mutations in the analysis, with larger boxes representing higher incidence mutations. The # cases/CNV column shows the total number of cases with the SNV/Indel variant surrounded by a ring chart indicating the proportion of total cases with CNV alterations.

Figure 6. Validation of MA12 candidates in ER+ METABRIC

A) Five out of eleven candidates from MA12 analysis had mutations reported in the METABRIC cohort. 1,415 ER+ samples with overall survival information were used to test the effect of mutations in the candidate genes on prognosis. B) Forest plot shows effect of mutated candidate genes, shortlisted based on MA12 mutation analysis, on overall survival in METABRIC ER+ breast cancer patients. Univariate (blue) and multivariate (orange) cox proportional-hazard ratio depict the independent prediction of survival outcomes for the six candidate genes. The box size is relative to frequency of mutations in the analysis, with larger boxes representing higher incidence mutations. The # cases/CNV column shows the total number of cases with the SNV/Indel variant surrounded by a ring chart indicating the proportion of total cases with CNV alterations.

Figure 7. Kaplan-Meier plots

A-B) Kaplan-Meier graphs showing the prognostic role of NF1 mutations, separated by variant type – Missense (MUT MS, green), Frameshift/Nonsense (MUT FS/NS, blue) in ER+ breast cancer patients from A) UBC-TAM and B) METABRIC cohort establishing the association between FS/NS mutations in NF1 with poor prognosis. C-D) Kaplan-Meier graph showing the prognostic role of PIK3R1 in C) MA12 and D) METABRIC ER+ breast cancer patients, categorized based on tumors with wildtype (WT, black) or mutated PIK3R1 non-silent mutations (MUT, red). E-F) Kaplan-Meier graph demonstrating co-occurrence of non-silent mutations in MAP3K1 and PIK3CA (red) in E) UBC-TAM and F) METABRIC associates with better survival when compared against tumors with mutations exclusively in MAP3K1 (blue) or PIK3CA (green) or wildtype for both MAP3K1 and PIK3CA (black). p, log rank (Mantel-Cox) test p-value.

Figure 8. Mutation profiles for selected genes

Mutation frequency plots illustrate all non-silent mutations (TAM, POLAR, and MA12; n=1259) for representative transcripts for several kinase genes of interest. The domains belonging to A) DDR1 (RefSeq ID: NM_013994) and B) JAK1 (NM_002227) are indicated below the schematic diagram of each gene. The ECD (extracellular domain), TM (transmembrane domain), and kinase domain are depicted as green, red, and orange bars respectively for C) ERBB2 (NM_004448), D) ERBB3 (NM_001982), E) ERBB4 (NM_005235), F) MET (NM_000245), and G) PDGFRA (NM_006206). The variant counts across the three datasets for each gene are provided below the gene's name. Note, in the mapping from Ensembl (**Supplementary Data 3**) to RefSeq annotations (required for use of ProteinPaint tool) a small number of variants annotations may have changed or been lost, despite selecting the most similar representative transcript possible.

References

1. Banerji S*, et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405-409 (2012).

2. Stephens PJ*, et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400-404 (2012).

3. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).

4. Kennecke HF*, et al.* Late risk of relapse and mortality among postmenopausal women with estrogen responsive early breast cancer after 5 years of tamoxifen. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO* **18**, 45-51 (2007).

5. Ellis MJ*, et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353-360 (2012).

6. Nielsen TO*, et al.* A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **16**, 5222-5232 (2010).

7. Chia SK*, et al.* A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. *Clinical cancer research : an official journal of the American Association for Cancer Research* **18**, 4465-4472 (2012).

8. Samuels Y*, et al.* High frequency of mutations of the PIK3CA gene in human cancers. *Science* **304**, 554 (2004).

9. Toy W*, et al.* ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nat Genet* **45**, 1439-1445 (2013).

10. Li S*, et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep* **4**, 1116-1130 (2013).

11. Bose R*, et al.* Activating HER2 mutations in HER2 gene amplification negative breast cancer. *Cancer discovery* **3**, 224-237 (2013).

12. Amar D, Shamir R, Yekutieli D. Extracting replicable associations across multiple studies: Empirical Bayes algorithms for controlling the false discovery rate. *PLoS Computational Biology* **13**, e1005700 (2017).

13. Capanu M, Seshan VE. False discovery rates for rare variants from sequenced data. *Genetic epidemiology* **39**, 65-76 (2015).

14. Efron B. Size, Power and False Discovery Rates. *The Annals of Statistics* **35**, 1351-1377 (2007).

15. Gao J*, et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* **6**, pl1 (2013).

16. Curtis C*, et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-352 (2012).

17. Pereira B*, et al.* The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature communications* **7**, 11479 (2016).

18. Endesfelder D*, et al.* Chromosomal instability selects gene copy-number variants encoding core regulators of proliferation in ER+ breast cancer. *Cancer research* **74**, 4853-4863 (2014).

19. McGranahan N, Burrell RA, Endesfelder D, Novelli MR, Swanton C. Cancer chromosomal instability: therapeutic and diagnostic challenges. *EMBO reports* **13**, 528-538 (2012).

20. Ma CX*, et al.* Neratinib Efficacy and Circulating Tumor DNA Detection of HER2 Mutations in HER2 Nonamplified Metastatic Breast Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **23**, 5687-5695 (2017).

21. Jaiswal BS*, et al.* Oncogenic ERBB3 mutations in human cancers. *Cancer cell* **23**, 603-617 (2013).

22. Imielinski M*, et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107-1120 (2012).

23. Shin DS*, et al.* Primary Resistance to PD-1 Blockade Mediated by JAK1/2 Mutations. *Cancer discovery* **7**, 188-201 (2017).

24. Zaretsky JM*, et al.* Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *The New England journal of medicine* **375**, 819-829 (2016).

25. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. *Human mutation* **32**, 557-563 (2011).

26. Lukasik SM*, et al.* Altered affinity of CBF beta-SMMHC for Runx1 explains its role in leukemogenesis. *Nature structural biology* **9**, 674-679 (2002).

27. Mendes-Pereira AM*, et al.* Genome-wide functional screen identifies a compendium of genes affecting sensitivity to tamoxifen. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 2730-2735 (2012).

28. Yeh YC, Wu CC, Wang YK, Tang MJ. DDR1 triggers epithelial cell differentiation by promoting cell adhesion through stabilization of E-cadherin. *Molecular biology of the cell* **22**, 940-953 (2011).

29. Rudd ML*, et al.* Mutational analysis of the tyrosine kinome in serous and clear cell endometrial cancer uncovers rare somatic mutations in TNK2 and DDR1. *BMC cancer* **14**, 884 (2014).

30. Loriaux MM*, et al.* High-throughput sequence analysis of the tyrosine kinome in acute myeloid leukemia. *Blood* **111**, 4788-4796 (2008).

31. Ding L*, et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-1075 (2008).

32. Vogel WF, Aszodi A, Alves F, Pawson T. Discoidin domain receptor 1 tyrosine kinase has an essential role in mammary gland development. *Molecular and cellular biology* **21**, 2906-2917 (2001).

33. Peng Y*, et al.* Deficiency in the catalytic subunit of DNA-dependent protein kinase causes down-regulation of ATM. *Cancer research* **65**, 1670-1677 (2005).

34. Haricharan S*, et al.* Loss of MutL Disrupts CHK2-Dependent Cell-Cycle Control through CDK4/6 to Promote Intrinsic Endocrine Therapy Resistance in Primary Breast Cancer. *Cancer discovery* **7**, 1168-1183 (2017).

35. Cheang MC*, et al.* Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *Journal of the National Cancer Institute* **101**, 736-750 (2009).

36. Liu S*, et al.* Prognostic significance of FOXP3+ tumor-infiltrating lymphocytes in breast cancer depends on estrogen receptor and human epidermal growth factor receptor-2 expression status and concurrent cytotoxic T-cell infiltration. *Breast cancer research : BCR* **16**, 432 (2014).

37. Parker JS*, et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 1160-1167 (2009).

38. Kan Z*, et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869-873 (2010).

39. Shah SP*, et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395-399 (2012).

40. Griffith M*, et al.* DGIdb: mining the druggable genome. *Nature methods* **10**, 1209-1210 (2013).

41. Chanock SJ*, et al.* Somatic sequence alterations in twenty-one genes selected by expression profile analysis of breast carcinomas. *Breast cancer research : BCR* **9**, R5 (2007).

42. Griffith M*, et al.* Genome Modeling System: A Knowledge Management Platform for Genomics. *PLoS Comput Biol* **11**, e1004274 (2015).

43. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

44. Li H*, et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

45. Lek M*, et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).

46. Genomes Project C*, et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).

47. Fu W*, et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-220 (2013).

48. Cancer Genome Atlas Research N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England journal of medicine* **368**, 2059-2074 (2013).

49. Krysiak K*, et al.* A genomic analysis of Philadelphia chromosome-negative AML arising in patients with CML. *Blood cancer journal* **6**, e413 (2016).

50. Ainscough BJ*, et al.* DoCM: a database of curated mutations in cancer. *Nature methods* **13**, 806-807 (2016).

51. Zhou X*, et al.* Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat Genet* **48**, 4-6 (2016).

52. Skidmore ZL*, et al.* GenVisR: Genomic Visualizations in R. *Bioinformatics* **32**, 3012-3014 (2016).

Figure 1.

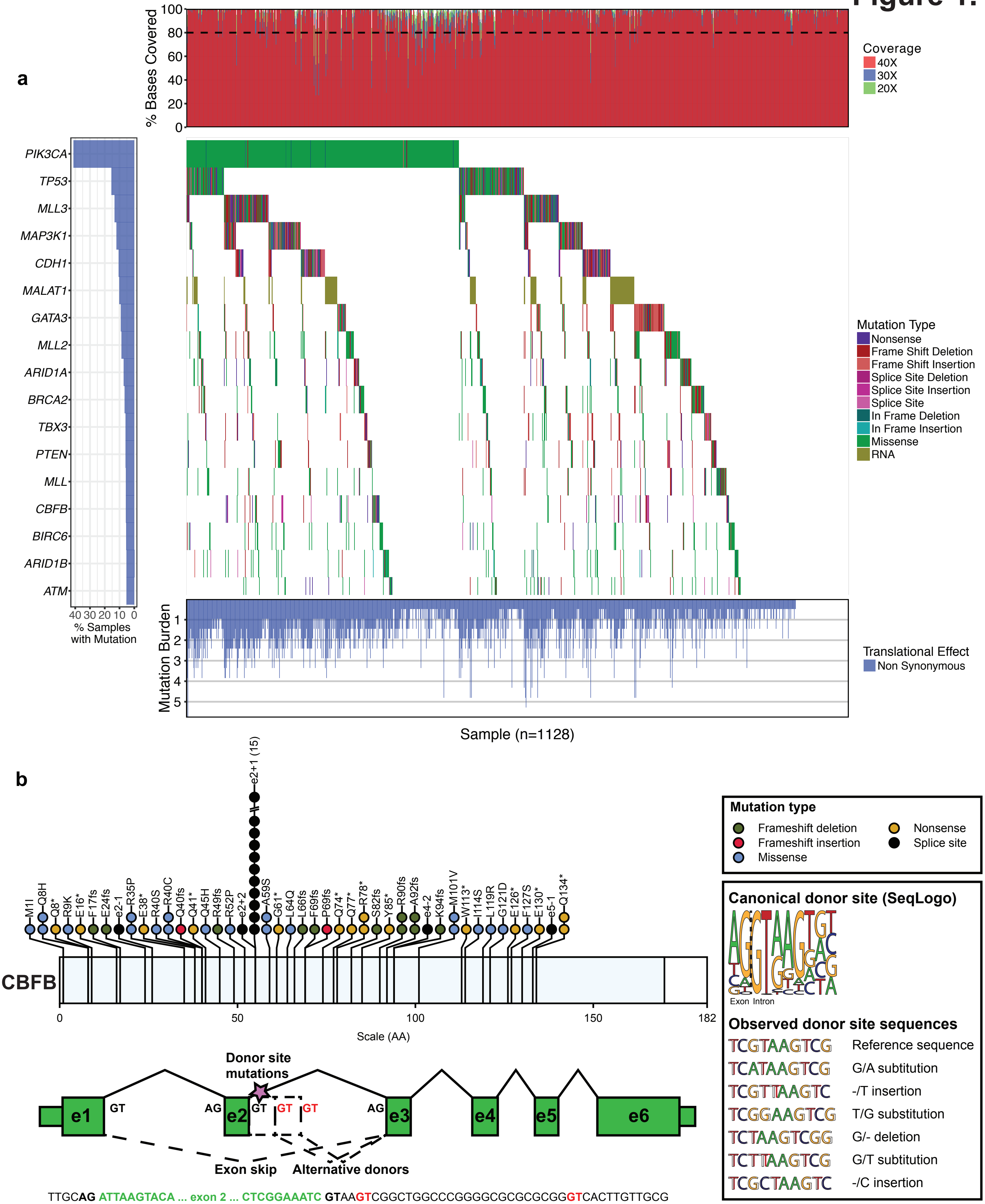
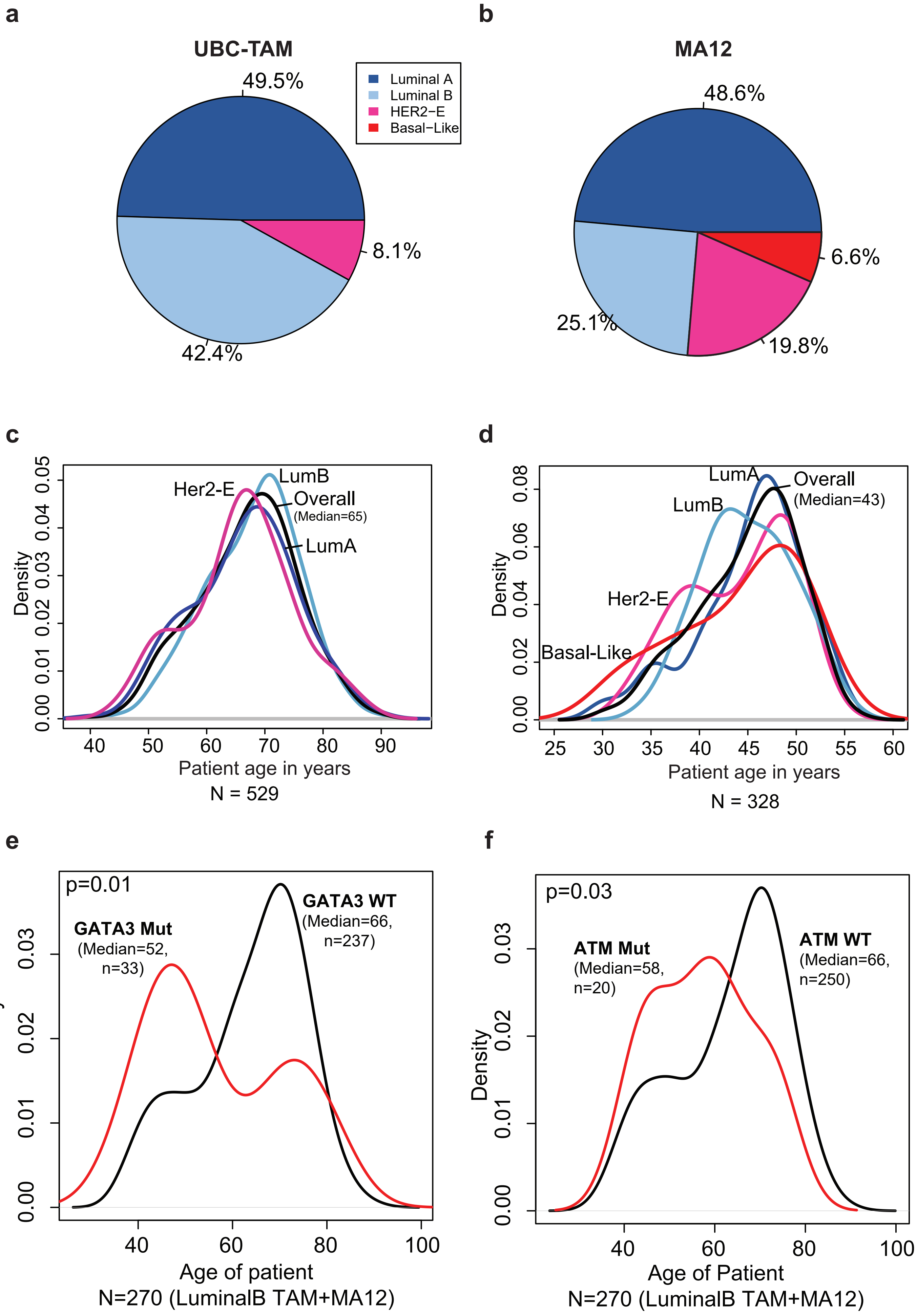
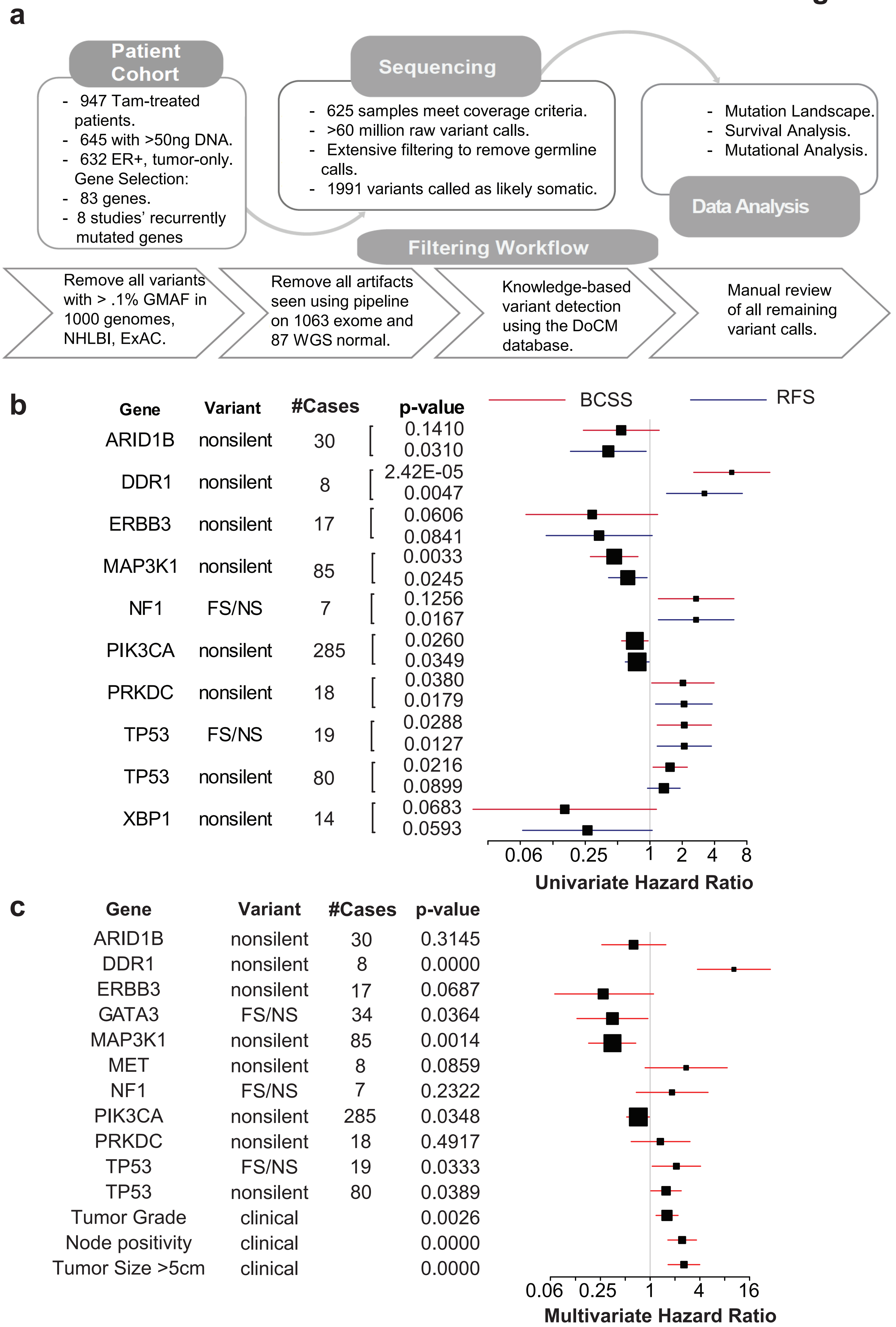

a
% Bases Covered
100
80
60
40
20
0
Coverage
40X
30X
20X
PIK3CA
TP53
MLL3
MAP3K1
CDH1
MALAT1
GATA3
MLL2
ARID1A
BRCA2
TBX3
PTEN
MLL
CBFB
BIRC6
ARID1B
ATM
40 30 20 10 0
% Samples with Mutation
Mutation Type
Nonsense
Frame Shift Deletion
Frame Shift Insertion
Splice Site Deletion
Splice Site Insertion
Splice Site
In Frame Deletion
In Frame Insertion
Missense
RNA
Mutation Burden
1
2
3
4
5
Translational Effect
Non Synonymous
Sample (n=1128)
b
e2+1 (15)
M1I
Q8H
Q8*
R9K
E16*
F17fs
E24fs
e2-1
R35P
E38*
R40S
R40C
Q40fs
Q41*
Q45H
R49fs
R52P
e2+2
A59S
G61*
L64Q
L66fs
F69fs
P69fs
Q74*
Q77*
R78*
S82fs
Y85*
R90fs
A92fs
e4-2
K94fs
M101V
W113*
I114S
L119R
G121D
E126*
F127S
E130*
e5-1
Q134*
CBFB
0
50
100
150
182
Scale (AA)
Mutation type
Frameshift deletion
Frameshift insertion
Missense
Nonsense
Splice site
Donor site mutations
e1
GT
AG
e2
GT
GT
GT
AG
e3
e4
e5
e6
Exon skip
Alternative donors
TTGCAG ATTAAGTACA ... exon 2 ... CTCGGAAATC GTAAGTCGGCTGGCCCGGGGCGCGCGCGGGTCACTTGTTGCG
Canonical donor site (SeqLogo)
Exon Intron
Observed donor site sequences
TCGTAAGTCG Reference sequence
TCATAAGTCG G/A substitution
TCGTTAAGTC -/T insertion
TCGGAAGTCG T/G substitution
TCTAAGTCGG G/- deletion
TCTTAAGTCG G/T subtitution
TCGCTAAGTC -/C insertion

Figure 2.



a
UBC-TAM
49.5%
8.1%
42.4%
Luminal A
Luminal B
HER2−E
Basal−Like
b
MA12
48.6%
6.6%
25.1%
19.8%
c
Density
0.00
0.01
0.02
0.03
0.04
0.05
LumB
Her2-E
Overall
(Median=65)
LumA
40
50
60
70
80
90
Patient age in years
N = 529
d
Density
0.00
0.02
0.04
0.06
0.08
LumA
Overall
(Median=43)
LumB
Her2-E
Basal-Like
25
30
35
40
45
50
55
60
Patient age in years
N = 328
e
p=0.01
GATA3 Mut
(Median=52,
n=33)
GATA3 WT
(Median=66,
n=237)
Density
0.00
0.01
0.02
0.03
40
60
80
100
Age of patient
N=270 (LuminalB TAM+MA12)
f
p=0.03
ATM Mut
(Median=58,
n=20)
ATM WT
(Median=66,
n=250)
Density
0.00
0.01
0.02
0.03
40
60
80
100
Age of Patient
N=270 (LuminalB TAM+MA12)

Figure 3.



a

Patient Cohort

- 947 Tam-treated patients.
- 645 with >50ng DNA.
- 632 ER+, tumor-only.

Gene Selection:

- 83 genes.
- 8 studies' recurrently mutated genes

Sequencing

- 625 samples meet coverage criteria.
- >60 million raw variant calls.
- Extensive filtering to remove germline calls.
- 1991 variants called as likely somatic.

Data Analysis

- Mutation Landscape.
- Survival Analysis.
- Mutational Analysis.

Filtering Workflow

Remove all variants with > .1% GMAF in 1000 genomes, NHLBI, ExAC.

Remove all artifacts seen using pipeline on 1063 exome and 87 WGS normal.

Knowledge-based variant detection using the DoCM database.

Manual review of all remaining variant calls.

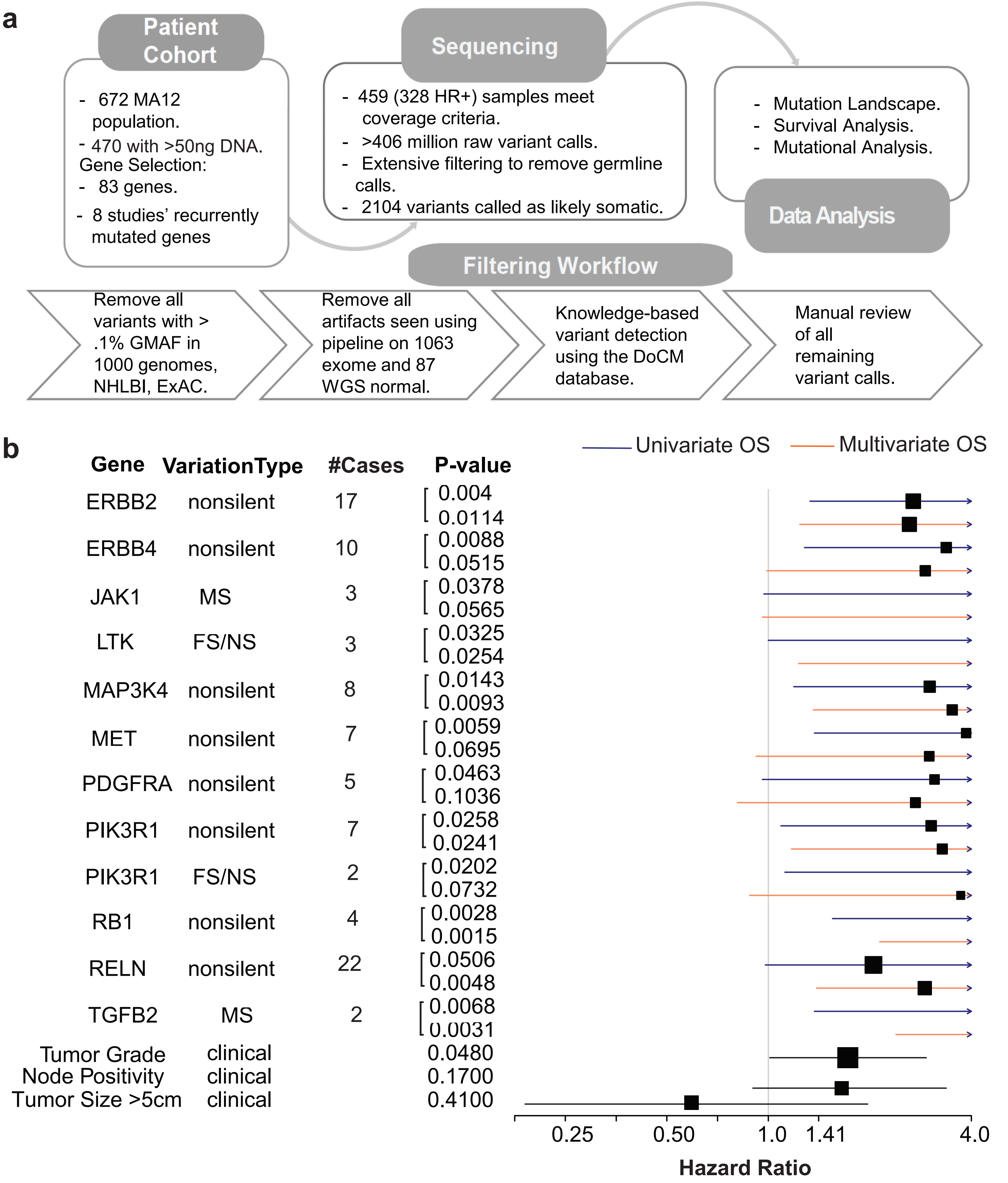
b

Gene	Variant	#Cases	p-value
ARID1B	nonsilent	30	0.1410 0.0310
DDR1	nonsilent	8	2.42E-05 0.0047
ERBB3	nonsilent	17	0.0606 0.0841
MAP3K1	nonsilent	85	0.0033 0.0245
NF1	FS/NS	7	0.1256 0.0167
PIK3CA	nonsilent	285	0.0260 0.0349
PRKDC	nonsilent	18	0.0380 0.0179
TP53	FS/NS	19	0.0288 0.0127
TP53	nonsilent	80	0.0216 0.0899
XBP1	nonsilent	14	0.0683 0.0593

c

Gene	Variant	#Cases	p-value
ARID1B	nonsilent	30	0.3145
DDR1	nonsilent	8	0.0000
ERBB3	nonsilent	17	0.0687
GATA3	FS/NS	34	0.0364
MAP3K1	nonsilent	85	0.0014
MET	nonsilent	8	0.0859
NF1	FS/NS	7	0.2322
PIK3CA	nonsilent	285	0.0348
PRKDC	nonsilent	18	0.4917
TP53	FS/NS	19	0.0333
TP53	nonsilent	80	0.0389
Tumor Grade	clinical		0.0026
Node positivity	clinical		0.0000
Tumor Size >5cm	clinical		0.0000

Figure 4.



a

Patient Cohort

- 672 MA12 population.
- 470 with >50ng DNA.

Gene Selection:

- 83 genes.
- 8 studies' recurrently mutated genes

Sequencing

- 459 (328 HR+) samples meet coverage criteria.
- >406 million raw variant calls.
- Extensive filtering to remove germline calls.
- 2104 variants called as likely somatic.

Data Analysis

- Mutation Landscape.
- Survival Analysis.
- Mutational Analysis.

Filtering Workflow

Remove all variants with > .1% GMAF in 1000 genomes, NHLBI, ExAC.

Remove all artifacts seen using pipeline on 1063 exome and 87 WGS normal.

Knowledge-based variant detection using the DoCM database.

Manual review of all remaining variant calls.

b

Gene	VariationType	#Cases	P-value
ERBB2	nonsilent	17	0.004 0.0114
ERBB4	nonsilent	10	0.0088 0.0515
JAK1	MS	3	0.0378 0.0565
LTK	FS/NS	3	0.0325 0.0254
MAP3K4	nonsilent	8	0.0143 0.0093
MET	nonsilent	7	0.0059 0.0695
PDGFRA	nonsilent	5	0.0463 0.1036
PIK3R1	nonsilent	7	0.0258 0.0241
PIK3R1	FS/NS	2	0.0202 0.0732
RB1	nonsilent	4	0.0028 0.0015
RELN	nonsilent	22	0.0506 0.0048
TGFB2	MS	2	0.0068 0.0031
Tumor Grade	clinical		0.0480
Node Positivity	clinical		0.1700
Tumor Size >5cm	clinical		0.4100

Figure 5.

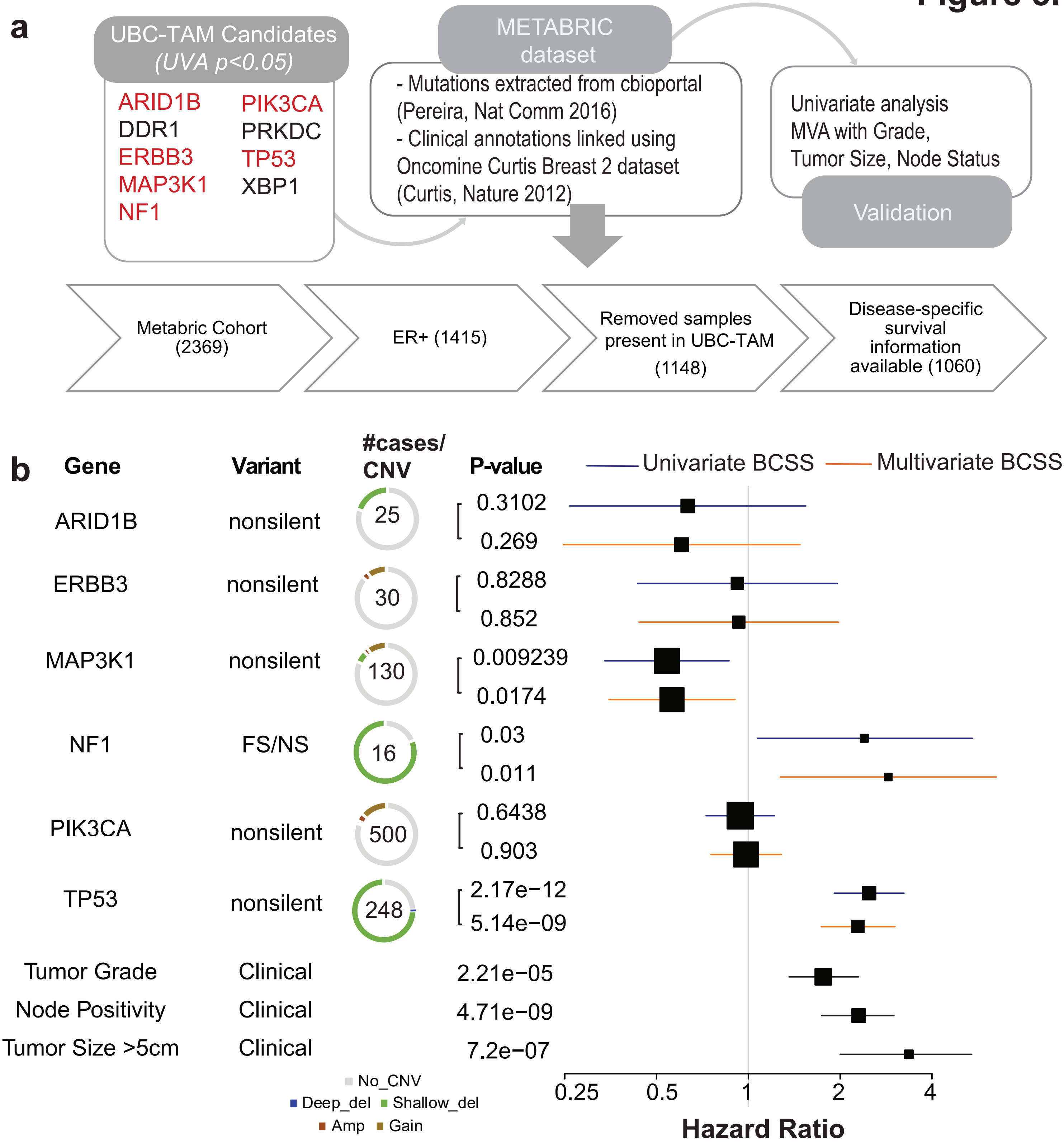

a
UBC-TAM Candidates
(UVA p<0.05)
ARID1B
DDR1
ERBB3
MAP3K1
NF1
PIK3CA
PRKDC
TP53
XBP1
METABRIC dataset
- Mutations extracted from cbioportal (Pereira, Nat Comm 2016)
- Clinical annotations linked using Oncomine Curtis Breast 2 dataset (Curtis, Nature 2012)
Univariate analysis
MVA with Grade, Tumor Size, Node Status
Validation
Metabric Cohort (2369)
ER+ (1415)
Removed samples present in UBC-TAM (1148)
Disease-specific survival information available (1060)
b
Gene
Variant
#cases/ CNV
P-value
Univariate BCSS
Multivariate BCSS
ARID1B
nonsilent
25
0.3102
0.269
ERBB3
nonsilent
30
0.8288
0.852
MAP3K1
nonsilent
130
0.009239
0.0174
NF1
FS/NS
16
0.03
0.011
PIK3CA
nonsilent
500
0.6438
0.903
TP53
nonsilent
248
2.17e−12
5.14e−09
Tumor Grade
Clinical
2.21e−05
Node Positivity
Clinical
4.71e−09
Tumor Size >5cm
Clinical
7.2e−07
No_CNV
Deep_del
Shallow_del
Amp
Gain
0.25
0.5
1
2
4
Hazard Ratio

Figure 6.

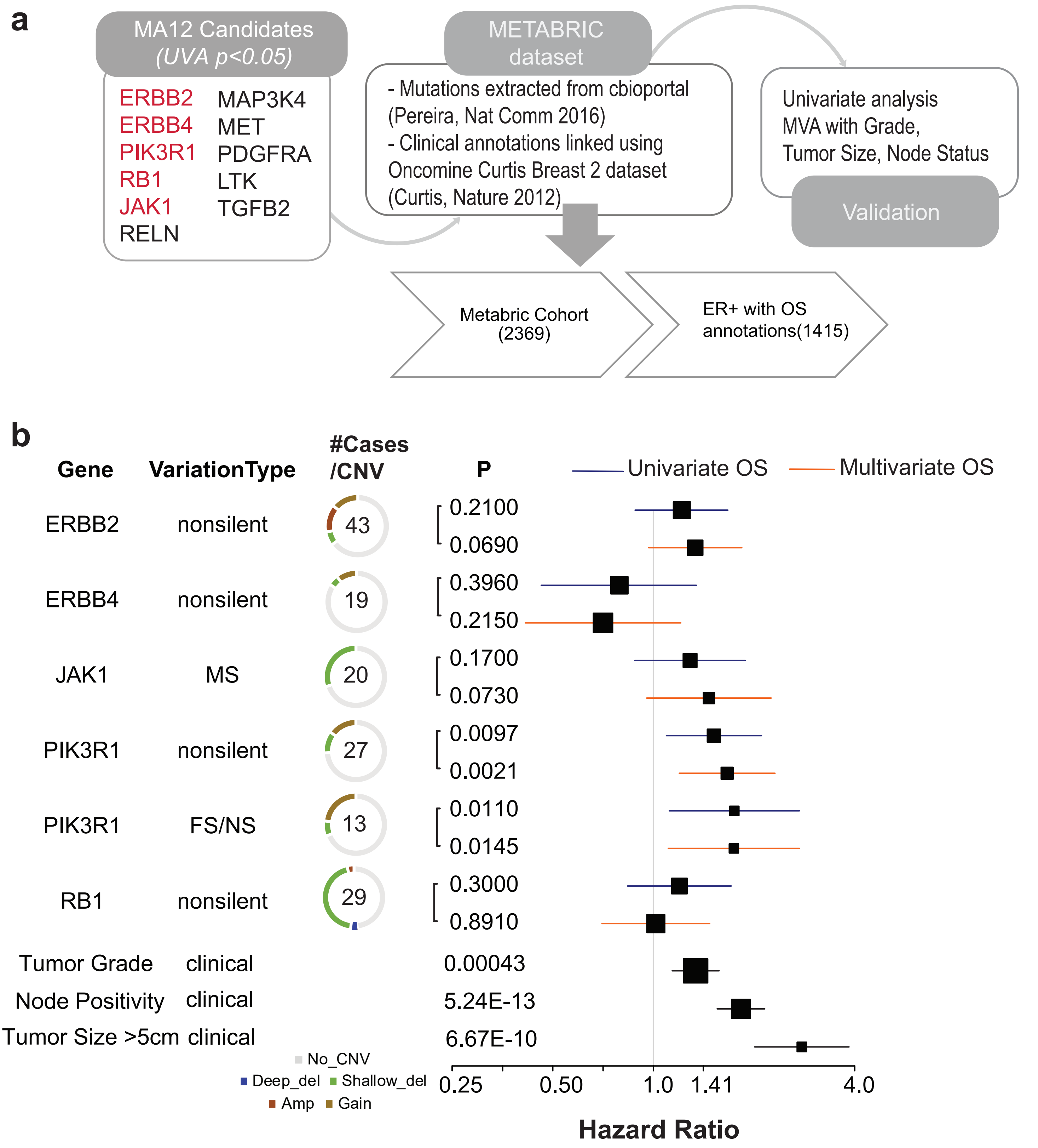

a
MA12 Candidates
(UVA p<0.05)
ERBB2
ERBB4
PIK3R1
RB1
JAK1
RELN
MAP3K4
MET
PDGFRA
LTK
TGFB2
METABRIC dataset
- Mutations extracted from cbioportal (Pereira, Nat Comm 2016)
- Clinical annotations linked using Oncomine Curtis Breast 2 dataset (Curtis, Nature 2012)
Univariate analysis
MVA with Grade,
Tumor Size, Node Status
Validation
Metabric Cohort (2369)
ER+ with OS annotations(1415)
b
Gene
VariationType
#Cases /CNV
P
Univariate OS
Multivariate OS
ERBB2
nonsilent
43
0.2100
0.0690
ERBB4
nonsilent
19
0.3960
0.2150
JAK1
MS
20
0.1700
0.0730
PIK3R1
nonsilent
27
0.0097
0.0021
PIK3R1
FS/NS
13
0.0110
0.0145
RB1
nonsilent
29
0.3000
0.8910
Tumor Grade
clinical
0.00043
Node Positivity
clinical
5.24E-13
Tumor Size >5cm
clinical
6.67E-10
No_CNV
Deep_del
Shallow_del
Amp
Gain
0.25
0.50
1.0
1.41
4.0
Hazard Ratio

Figure 7.

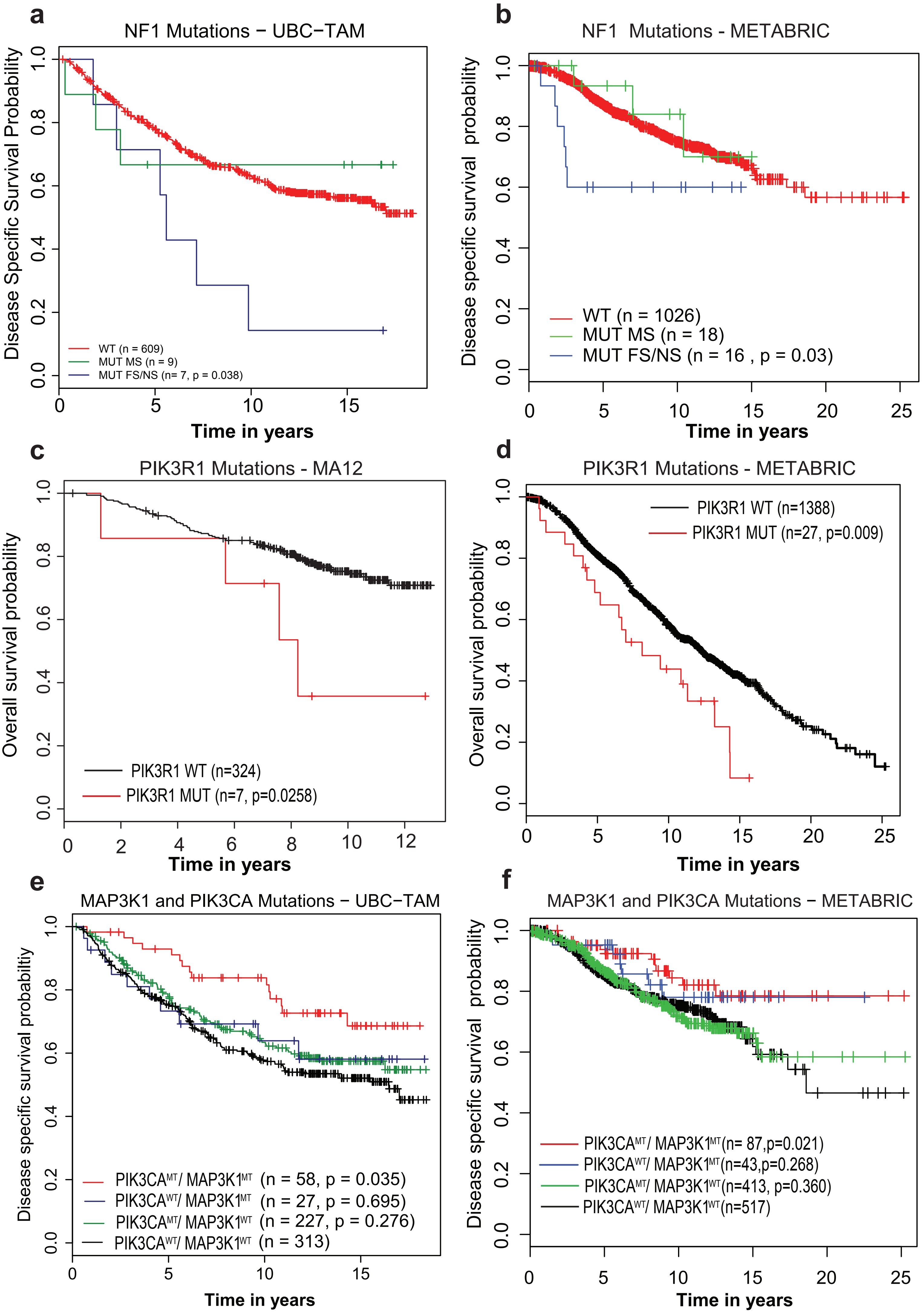
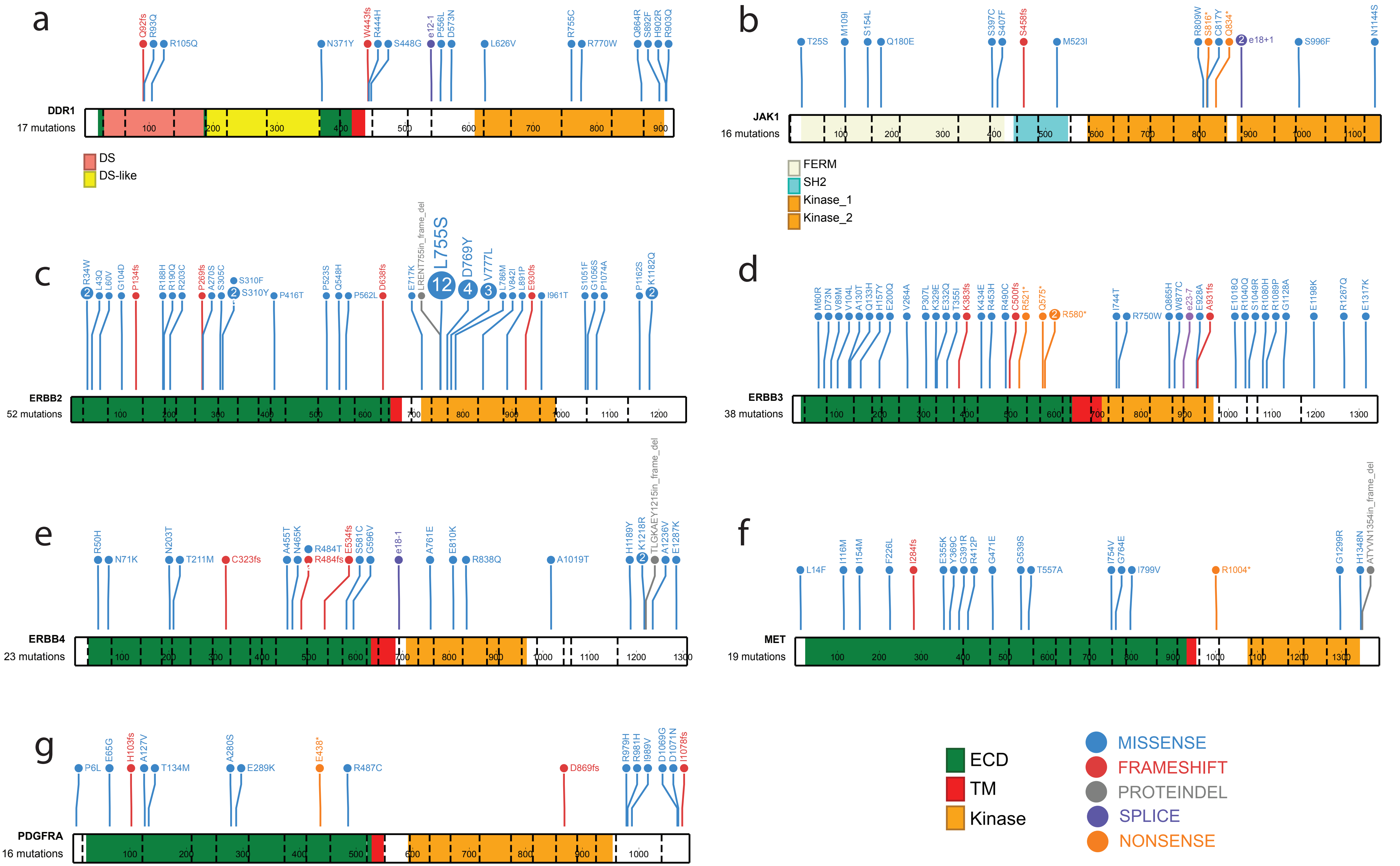

a
NF1 Mutations − UBC−TAM
Disease Specific Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
Time in years
WT (n = 609)
MUT MS (n = 9)
MUT FS/NS (n= 7, p = 0.038)
b
NF1 Mutations - METABRIC
Disease specific survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
Time in years
WT (n = 1026)
MUT MS (n = 18)
MUT FS/NS (n = 16 , p = 0.03)
c
PIK3R1 Mutations - MA12
Overall survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
Time in years
PIK3R1 WT (n=324)
PIK3R1 MUT (n=7, p=0.0258)
d
PIK3R1 Mutations - METABRIC
Overall survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
Time in years
PIK3R1 WT (n=1388)
PIK3R1 MUT (n=27, p=0.009)
e
MAP3K1 and PIK3CA Mutations − UBC−TAM
Disease specific survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
Time in years
PIK3CA^{MT}/ MAP3K1^{MT} (n = 58, p = 0.035)
PIK3CA^{WT}/ MAP3K1^{MT} (n = 27, p = 0.695)
PIK3CA^{MT}/ MAP3K1^{WT} (n = 227, p = 0.276)
PIK3CA^{WT}/ MAP3K1^{WT} (n = 313)
f
MAP3K1 and PIK3CA Mutations − METABRIC
Disease specific survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
Time in years
PIK3CA^{MT}/ MAP3K1^{MT}(n= 87,p=0.021)
PIK3CA^{WT}/ MAP3K1^{MT}(n=43,p=0.268)
PIK3CA^{MT}/ MAP3K1^{WT}(n=413, p=0.360)
PIK3CA^{WT}/ MAP3K1^{WT}(n=517)

Figure 8.


a
Q92fs
R93Q
R105Q
N371Y
W443fs
R444H
S448G
e12-1
P556L
D573N
L626V
R755C
R770W
Q864R
S892F
H902R
R903Q
DDR1
17 mutations
100
200
300
400
500
600
700
800
900
DS
DS-like
b
T25S
M109I
S154L
Q180E
S397C
S407F
S458fs
M523I
R809W
S816*
C817Y
Q834*
2
e18+1
S996F
N1144S
JAK1
16 mutations
100
200
300
400
500
600
700
800
900
1000
1100
FERM
SH2
Kinase_1
Kinase_2
c
2
R34W
L43Q
L60V
G104D
P134fs
R188H
R190Q
R203C
P269fs
A270S
S305C
S310F
2
S310Y
P416T
P523S
Q548H
P562L
D638fs
E717K
LRENT755in_frame_del
12
L755S
4
D769Y
3
V777L
L786M
V842I
L891P
E930fs
I961T
S1051F
G1056S
P1074A
P1162S
2
K1182Q
ERBB2
52 mutations
100
200
300
400
500
600
700
800
900
1000
1100
1200
d
M60R
D73N
V89M
V104L
A130T
Q133H
H157Y
E200Q
V264A
P307L
K329E
E332Q
T355I
K383fs
K434E
R453H
R490C
C500fs
R521*
Q575*
2
R580*
I744T
R750W
Q865H
W877C
e23-7
E928A
A931fs
E1018Q
R1040Q
S1049R
R1080H
R1089P
G1128A
E1198K
R1267Q
E1317K
ERBB3
38 mutations
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
e
R50H
N71K
N203T
T211M
C323fs
A455T
N465K
R484T
R484fs
E534fs
S581C
G596V
e18-1
A761E
E810K
R838Q
A1019T
H1189Y
2
K1218R
TLGKAEY1215in_frame_del
A1236V
E1287K
ERBB4
23 mutations
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
f
L14F
I116M
I154M
F226L
I284fs
E355K
Y369C
G391R
R412P
G471E
G539S
T557A
I754V
G764E
I799V
R1004*
G1299R
H1348N
ATYVN1354in_frame_del
MET
19 mutations
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
g
P6L
E65G
H103fs
A127V
T134M
A280S
E289K
E438*
R487C
D869fs
R979H
R981H
I989V
D1069G
D1071N
I1078fs
PDGFRA
16 mutations
100
200
300
400
500
600
700
800
900
1000
ECD
TM
Kinase
MISSENSE
FRAMESHIFT
PROTEINDEL
SPLICE
NONSENSE