PLOS ONE

RESEARCH ARTICLE

The genomic landscape of metastatic breast cancer: Insights from 11,000 tumors

Jacob Rinaldi[1]*, Ethan S. Sokol[2], Ryan J. Hartmaier[2], Sally E. Trabucco[2], Garrett M. Frampton[2], Michael E. Goldberg[2], Lee A. Albacker[2], Anneleen Daemen[1], Gerard Manning[1]*

1 Department of Bioinformatics & Computational Biology, Genentech Inc., South San Francisco, CA, United States of America, **2** Foundation Medicine, Cambridge, MA, United States of America

* jacob.rinaldi@gmail.com (JR); manning@manninglab.org (GM)

Citation: Rinaldi J, Sokol ES, Hartmaier RJ, Trabucco SE, Frampton GM, Goldberg ME, et al. (2020) The genomic landscape of metastatic breast cancer: Insights from 11,000 tumors. PLoS ONE 15(5): e0231999. https://doi.org/10.1371/journal.pone.0231999

Editor: Krishna Rani Kalari, Mayo Clinic Rochester, UNITED STATES

Received: October 28, 2019

Accepted: April 3, 2020

Published: May 6, 2020

Data Availability Statement: This study involved next generation sequencing (NGS)-based genomic profiling of breast tumors. The study was approved by Western Institutional Review Board (Protocol NO. 20152817), who granted a waiver of informed consent and a HIPAA waiver of authorization as the authors did not have access to potentially identifying information. The samples used in this study were submitted to Foundation Medicine and were limited to patients who consented to anonymized research; patients who provided tumor samples to Foundation Medicine did not

Abstract

Background

Metastatic breast cancer is the leading cause of cancer death in women, but the genomics of metastasis in breast cancer are poorly studied.

Methods

We explored a set of 11,616 breast tumors, including 5,034 metastases, which had undergone targeted sequencing during standard clinical care.

Results

Besides the known hotspot mutations in *ESR1*, we observed a metastatic enrichment of previously unreported, lower-prevalence mutations in the ligand-binding domain, implying that these mutations may also be functional. Furthermore, individual *ESR1* hotspots are significantly enriched in specific metastatic tissues and histologies, suggesting functional differences between these mutations. Other alterations enriched across all metastases include loss of function of the CDK4 regulator *CDKN1B*, and mutations in the transcription factor *CTCF*. Mutations enriched at specific metastatic sites generally reflect biology of the target tissue and may be adaptations to growth in the local environment. These include *PTEN* and *ASXL1* alterations in brain metastases and *NOTCH1 alterations* in skin. We observed an enrichment of *KRAS*, *KEAP1*, *STK11* and *EGFR* mutations in lung metastases. However, the patterns of other mutations in these tumors indicate that these are misdiagnosed lung primaries rather than breast metastases.

Conclusions

An order-of-magnitude increase in samples relative to previous studies allowed us to detect novel genomic characteristics of metastatic cancer and to expand and clarify previous findings.

consent to releasing raw sequence data. Therefore, associated raw sequence data cannot be shared. However, variants from 2575 samples used in this analysis have been deposited in the Genomic Data Commons (accession #phs001179), allowing complementary analyses to be performed by others. Some of the initial findings based on a smaller sub-cohort did not replicate in the full cohort of >11,000 breast cases. Additionally, some of the findings in this manuscript cannot be replicated with smaller cohort sizes because of a lack of statistical power. Inquiries about access should be directed to client.services@foundationmedicine.com.

Funding: All authors were employees and may have held shares of Genentech/Roche or Foundation Medicine at the time this work was completed but the companies did not have any additional role in the study design, data analysis, or preparation of the manuscript. Both companies gave consent to publish. Data was collected by Foundation Medicine during standard clinical sequencing. The specific roles of these authors are articulated in the 'author contributions' section."

Competing interests: All authors were employees and may have held shares of Genentech/Roche or Foundation Medicine at the time this work was completed. Both companies are involved in the development of cancer drugs and clinical assays for metastatic breast cancer, and have multiple patents in related areas. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

Abbreviations: AKT1, AKT serine/threonine kinse 1; ASXL1, ASXL transcriptional regulator 1; BRCA1/2, BRCA 1/2 DNA repair associated; CBFB, core-binding factor subunit beta; CTFC, transcriptional repressor CTFC; CCND1, cyclin D1; CDH1, cadherin 1; CDK4, cyclin-dependent kinase 4; CDKN1B, cyclin dependent kinase inhibitor 1b; CDKN2A/B, cyclin dependent kinase inhibitor 2a/b; CTCF, CCCTC-binding factor; DNA, deoxyribonucleic acid; DNMT3A, DNA methyltransferase 3 alpha; EGFR, epidermal growth factor receptor; ERBB2, erb-b2 receptor tyrosine kinase 2; ESR1, estrogen receptor 1; FFPE, formalin-fixed paraffin-embedded; FGFR, fibroblast growth factor receptor; GATA3, GATA binding protein 3; H&E, hematoxylin and eosin stain; HER2, erb-b2 receptor tyrosine kinase 2; KEAP1, kelch like ECH associated protein 1; JAK/STAT, Janus kinase/signal transduce and activator of transcription; KMT2D, lysine methyltransferase 2D; KRAS, KRAS proto-oncogene; MAF, minor allele frequency; MAPK, mitogen activated kinase-like protein; MAP3K1, mitogen-activated protein kinase

Background

Breast cancer is the most commonly diagnosed malignancy, and the leading cause of cancer death in women [1]. Virtually all breast cancer deaths are due to metastatic disease [2]. The process via which cancer cells disseminate from the primary tumor, colonize distal sites, and adapt to novel tumor microenvironments has not been fully characterized, but recent work implicates a cascade of genetic and epigenetic events that drive active degradation of the extracellular matrix, induce angiogenesis, enhance motility, promote immune evasion, and co-opt the epithelial-mesenchymal transition [3,4]. There is no cure for metastatic breast cancer and median survival is 18 to 24 months, representing an enormous unmet medical need [5]. Success in developing better treatments for breast cancer will primarily be determined by our ability to impede the metastatic process and treat metastatic disease, which is largely unchecked by current therapies [6,7]. A key component of this effort will be understanding how specific oncogenic events in individual patients lead to metastasis.

Breast cancer is the archetype for the use of molecular profiling to guide treatment decisions and develop targeted therapies [8]. The observation that a subset of breast cancers express the estrogen receptor (ER) led to the development of aromatase inhibitors, selective estrogen receptor degraders (SERDs) and selective estrogen receptor modulators (SERMs), which now make up a core component of standard clinical care for patients with ER positive disease [9–11]. Similarly, the discovery of epidermal growth factor receptor 2 (HER2/ERBB2) overexpression on the surface of breast cancer cells led to the development of the monoclonal antibodies trastuzumab and pertuzumab, now part of standard care for HER2 positive patients [12,13]. Because cancer is a genomic disease, genomic profiling holds the promise of extending the results of molecular profiling by further refining personalized treatment decisions and catalyzing the development of additional targeted therapies.

Most large-scale sequencing efforts in breast cancer have focused on primary tumors, where thousands of cancer genomes have been analyzed across multiple studies [14–17]. The most frequently reported alterations include mutations in *TP53*, *PIK3CA*, *GATA3*, *MAP3K1*, *AKT1*, and *CBFB*; amplification of *HER2*, *MYC*, *FGFR1* and *FGF3/4*; deletion of *PTEN*, *RB1* and *CDKN2A/B*; and oncogenic germline polymorphisms in *BRCA1/2* [14,17]. This mountain of genomic information has led to targeted therapies, improved prognostic and predictive models, and innovative clinical trials that stratify patients by genomic phenotype. Clinical trials are ongoing for drugs that target patients with *BRCA1/2* mutations, *AKT1* mutations, *PIK3CA* mutations, and *FGFR* amplification, illustrating the utility of genomics to define patient populations and guide drug discovery [18–23].

Significantly less work has been done to characterize genomic alterations in metastatic disease because of the difficulty in gaining access to samples and the efficacy of adjuvant therapies. Initial efforts uncovered mutations in the ligand-binding domain of the estrogen receptor (*ESR1*) in 10–30% of metastatic breast cancer patients, an alteration that is largely absent in primary disease [24–27]. Recent work in small cohorts of ~100–1000 metastatic tumors suggests that the majority of alterations are shared between primary tumors and metastases, and that JAK/STAT and SWI/SNF pathways are dysregulated at higher rates in metastatic disease; these studies have also implicated genes involved in DNA damage repair, the MAPK pathway, and epigenetic regulators [28–32].

Here we demonstrate the utility of large-scale sequencing of metastatic breast cancer, in an unprecedented cohort of 4,512 local and 5,034 metastatic breast tumors (with an additional 1,357 lymph node biopsies and 713 tumors with ambiguous metastatic status) collected during standard clinical care. We hypothesized that relative to local disease, the genomic fingerprints of metastatic tumors are enriched for (i) mechanisms of acquired resistance (due to treatment

kinase kinase 1; OCR, optical character recognition; PIK3CA, phosphatidylinositol-4,5-biphosphate 3-kinase catalytic subunit alpha; PTEN, phosphatase and tensin homolog; RB1, transcriptional corepressor 1; SERM, selective estrogen receptor modulator; SERD, selective estrogen receptor degrader; STK11, serine/threonine kinase 11; SWI/SNF, Switch/Sucrose non-fermentable; TP53, tumor protein p53.

history) and (ii) alterations that induce or accelerate metastasis. We find evidence for both classes of alterations, complementing recent metastatic profiling efforts in small cohorts with clinical annotation and/or matched primary and metastatic samples.

Methods

Tumor samples and sequencing

Samples were submitted to a CLIA-certified, New York State-accredited, and CAP-accredited laboratory (Foundation Medicine, Cambridge, MA) for next-generation sequencing (NGS)-based genomic profiling. The pathologic diagnosis of each case was confirmed by review of hematoxylin and eosin (H&E) stained slides, and all samples that advanced to nucleic acid extraction contained a minimum of 20% tumor cells. The samples used in this study were not selected and represent "all comers" to Foundation Medicine genomic profiling. Samples were processed in the protocol defined by solid tumors and hematological cancers as previously described [33,34]. A brief description is provided below.

For solid tumors, DNA was extracted from formalin-fixed, paraffin-embedded (FFPE) 10 micron sections. Adaptor-ligated DNA underwent hybrid capture for all coding exons of 287 or 395 cancer-related genes plus select introns from 19 or 31 genes frequently rearranged in cancer. Genes are listed in **S8 Table.**

Captured libraries were sequenced to a median exon coverage depth of >500x using Illumina sequencing, and resultant sequences were analyzed for base substitutions, small insertions and deletions (indels), copy number alterations (focal amplifications and homozygous deletions) and gene fusions/rearrangements, as previously described [33,34]. Frequent germline variants from the 1000 Genomes Project (dbSNP142) were removed. To maximize mutation-detection accuracy (sensitivity and specificity) in impure clinical specimens, the test was previously optimized and validated to detect base substitutions at a ≥5% mutant allele frequency (MAF), indels with a ≥10% MAF with ≥99% accuracy, and fusions occurring within baited introns/exons with >99% sensitivity [33]. Known confirmed somatic alterations deposited in the Catalog of Somatic Mutations in Cancer (COSMIC v62) are called at allele frequencies ≥1% [35].

Statistical analyses

Enrichment analyses were conducted by performing a logistic regression to predict the variable of interest (for instance, local or met status) and then a Wald test on the coefficient of interest. In all cases, predicted probability of ER-positivity, HER2 status, and mutation load per megabase were used as additional covariates. Statistical analysis of count tables (ESR1 mutations by site, subtype by site) was performed using Fisher's exact tests. Sets of probabilities (output of machine learning algorithms) were compared using Kolmogorov-Smirnov tests. To control for additional covariates when comparing outputs of the skin classifier a beta regression was performed with a Wald test on the coefficient of interest. Corrected p-values for enrichment analyses were calculated by permuting the variable of interest 1000 times and selecting the most significant genomic alteration. P-values from the true dataset were then compared with these 1000 iterations to estimate the probability that any p-value across all genomic alterations would be deemed as significant by chance. Alteration rates in local disease and metastases were compared using Mann-Whitney tests.

Machine learning algorithms

All classifiers used the random forest algorithm [36], with an input feature set consisting of the mutation (short variant), copy number, and structural rearrangement status of each gene on

the panel–excluding enriched genes for the tissue of origin classifiers, as well as the per-sample mutation load per megabase. In each case, not all features impact prediction–the training process selects a subset of features that are useful for the classification task. The random forest classifier was implemented using R 3.1.0 with the randomForest 4.6–7 and caret 6.0–30 packages [37–39]. Random forests are ensembles of decision trees, where each tree is fit on a random sample with replacement of the training set and each candidate split is made on a random subset of the features. This produces models that generalize well to unseen data. Ensemble methods work by combining many weak models to produce a prediction that is more accurate than the prediction made by a single strong model. Each tree has been trained on a different subset of the training samples and a different subset of features, making them relatively independent and representative of different hypotheses about how features map to class labels. By combining these different hypotheses, which will be accurate on different unseen samples, we can generate a combined prediction that is much less susceptible to overfitting noise in the training set. To make a prediction on an unseen sample, the mode of the predictions of all the trees in the ensemble is used–in other words, every tree votes and the majority wins. The percentage of votes given to one class is the probability that the model assigns to a given prediction. Mutation calls were summarized such that any gene harboring at least one mutation (regardless of functional impact) was considered "mutated" for that gene. The *mtry* parameter, which determines the number of variables available for sampling at each tree node, was set using 10-fold cross-validation—and a new model with the optimal *mtry* value was trained after this cross-validation phase to avoid over-fitting. The random forest algorithms consisted of 5,000 trees, each fit with a stratified resampling of the data to rebalance classes.

For the molecular and histological subtype predictors, all available samples were used for training (ER status was generated with an algorithm to infer status from pathology reports with 95% accuracy, see Pathology Report Parser below)– 1405 samples for molecular subtype and 3959 for histological subtype. Five hundred samples in each class were used for indication prediction. All reported accuracies are on held-out test data. We assessed the importance of each feature to model performance by independently permuting each feature and assessing the resultant decrease in out-of-bag accuracy.

SGZ (Somatic-Germline-Zygosity) determination

Somatic vs. germline origin and homozygous vs. heterozygous or sub-clonal state of variants identified was determined without a matched normal control as described previously [40]. Briefly, for each patient we generate a segmented genome-wide copy number model and calculate the minor allele frequency (MAF) based on the patient SNP profiles. We then model the copy number of the MAF taking the observed noise into account. Goodness of fit was assessed with a Gibbs sampling-based Markov chain Monte Carlo algorithm and a grid-sampling approach.

Somatic/germline/ambiguous prediction was calculated using a 2-tailed binomial test and α cutoff of 0.01. Mutations were called homozygous if all copies in the tumor carried the mutant allele, heterozygous if both the reference and the mutant alleles were present, or sub-clonal somatic if the somatic allele frequency was significantly lower than the expected allele frequency.

Pathology report parser

We parsed the hormone status from patients' pathology reports using optical character recognition (OCR) software and a set of scripts employing natural language processing techniques.

We received electronic scan images of pathology reports stored as PDFs, and extracted and stored the text from the images using ABBYY Fine Reader Engine 11. To parse the hormone status from the text, we built a set of python scripts (Python 2.7) that used regular expressions and filtering to find the most likely hormone receptor status. Within several lines of the string "ER" or "Estrogen", we searched for the closest string representing a possible status. Strings associated with ER-positive staining included "positive", "detect", and "expression". Strings associated with negative staining included "negative" and "rare". We also detected negation and switched the status accordingly (e.g., "ER staining was detected" vs. "ER staining was not detected"). We filtered for text commonly found in pathology reports associated with incorrect matches, such as text explaining the interpretation of stain statuses or bibliographical entries. Finally, we set a threshold for the maximum distance in number of characters between the strings "ER" or "Estrogen" and the status after finding that distance negatively correlated with accuracy in our training dataset (130 pathology reports from the breast carcinoma cohort from TCGA and 25 from FMI, both redacted of protected health information).

We validated and calculated the accuracy of our parsing strategy by testing on a set of 100 new, internal pathology reports. A trained pathologist hand-analyzed each report and provided us with a "gold-standard" dataset. By comparing the ER status parsed from the report to the gold-standard, hand-analyzed status, we found that our parsing strategy was 94% accurate.

Visualization

All visualizations were created using ggplot2 2.2.1 in R 3.1.0, except for the lollipop visualization, which was created using Lollipops, and the cell diagrams, which were made using ComplexHeatmap 1.14 in R 3.1.0 [37,41–43].

Ethics approval and consent to participate

Approval for this study, including a waiver of informed consent and a HIPAA waiver of authorization, was obtained from the Western Institutional Review Board (Protocol NO. 20152817)

Results

Overview of clinical data

Samples from 11,616 breast cancer patients were submitted for targeted sequencing as part of standard clinical care. Thirty-nine percent were biopsied from local breast tumors (primary tumors or local recurrences), and 12% from lymph node metastases (**Fig 1A**). Notably, 5,034 (43%) were from distal metastases, comprising the largest collection of genomic profiles from metastatic breast cancer assembled to date. The median age was 55, with 1,343 patients under 40. Pathology reports containing ER status were available for 1,405 (12%) samples and were annotated with ER status using an automated algorithm. In the remaining cases ER status was imputed with an accuracy of 76% from mutation and copy number data using a machine learning algorithm trained on the samples with known ER status (**S1 Fig**). HER2 positivity was defined as *HER2* amplification as measured by the sequencing assay. Fifty-five percent of samples were scored as ER+/HER2-, with a significantly higher prevalence in metastatic samples (64% in metastases vs. 48% in local disease, $p < 2.2e\text{-}16$; **Fig 1B**), similar to the pattern seen in samples with clinical annotation (60% in metastases vs. 48% in local disease; **Fig 1C**). We believe the lower prevalence of non-metastatic ER+ samples in our cohort relative to traditional prevalence estimates is caused by treatment landscape and prognosis, which both impact the utilization of genomic profiling in standard clinical practice. The prevalence of HER2 amplification was similar in metastatic and local tumors (9.4% in metastases vs. 8.7% in local

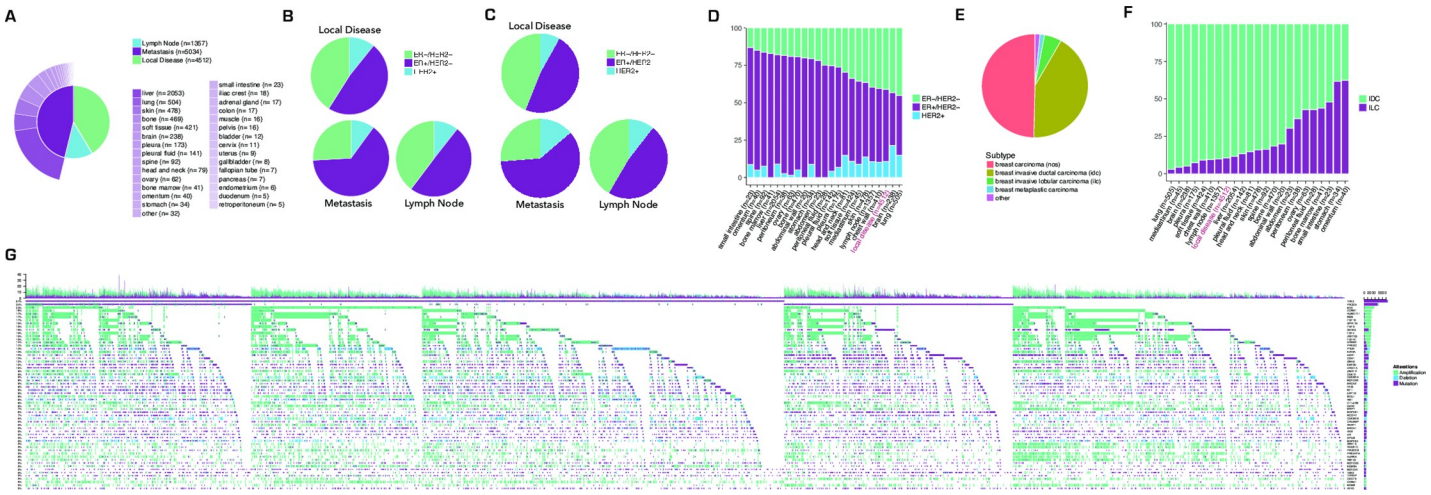

Fig 1. Overview of clinical and genomic data. (a) Frequency of biopsy sites for 10,903 sequenced breast cancer samples. An additional 713 tumors had an ambiguous classification. (b) Prevalence of *HER2* amplification and inferred ER status by biopsy site in 10,903 samples. ER status was inferred using a machine learning algorithm trained on 1,405 samples for which subtype was known–out-of-bag predictions are plotted for samples in training set, true values for these samples are shown in C (see S1 Fig). (c) Prevalence of *HER2* amplification and ER IHC status by biopsy site in 1,405 samples with complete clinical annotation. (d) Prevalence of inferred molecular subtype by metastatic site (n = 11293, note that some samples have a tissue biopsy site that confers ambiguous primary/met/ln status and 98 samples were from unknown sites). (e) Prevalence of histological subtype for all sequenced samples (n = 11616). (f) Prevalence of histological subtype by biopsy site Histological subtype was inferred using a machine learning algorithm trained on annotated data (see S1 Fig, n = 11293). (g) Landscape of genomic alterations in the cohort. Each cell represents the status of one gene in one patient, colored by alteration type (mutation, amplification, or deletion). Genes (rows) are sorted by alteration rate. Barplot shows alterations per sample, colored by type.

https://doi.org/10.1371/journal.pone.0231999.g001

disease; Fig 1B). ER and HER2 status differed significantly across metastatic sites, with higher rates of ER positivity in liver and bone, lower rates in brain and lung, and with high prevalence of HER2 amplification in brain ($p<1e-6$ for association between subtype and site; Fig 1D) [44].

Histological subtype information was available for 50% of samples. Of these, 4,896 (84%) were categorized as invasive ductal carcinoma (IDC) and 611 (10%) as invasive lobular carcinoma (ILC) with the remainder from several rare subtypes including metaplastic carcinoma (n = 158), neuroendocrine carcinoma (n = 43), inflammatory carcinoma (n = 29), adenoid cystic carcinoma (n = 22), phyllodes tumors (n = 21), and mucinous carcinoma (n = 21). These constitute a substantial expansion of genomic profiling for rare breast cancers (Fig 1E). The genomic landscapes of these rare subtypes differ substantially from IDC and ILC, and can be found in S2 Table. Histological subtype was significantly associated with metastatic biopsy site, with ILC metastasizing at higher rates to the ovary and gastrointestinal tract ($p<1e-6$; Fig 1F), although the full extent of all metastatic sites within an individual cannot be determined from this data. For the 5,770 tumors without histological subtype, we inferred the probability that a tumor was IDC or ILC with 95% accuracy using a machine learning algorithm (S1 Fig).

Overview of genomic alterations

All samples underwent targeted sequencing using the FoundationOne® assay, which interrogates the coding sequences of cancer-associated genes as well as introns from genes frequently rearranged in solid tumors [33]. This assay provides a sensitive and specific readout of mutations, including oncogenic germline polymorphisms, copy number alterations, and structural rearrangements. Analyses were restricted to a set of 287 genes included on all versions of the assay (S8 Table). Because this targeted assay is used during standard clinical care we had access to a large patient population, though targeted sequencing does not provide the complete

portrait created by whole exome or whole genome sequencing. The high *TP53* mutation rate relative to prior studies and the presence of *ESR1* mutations suggest that both the local and metastatic tumors in this dataset are enriched for patients with poor prognosis [14,45,46]. Samples had an average of 6.3 mutations (6.7 in metastases vs 5.6 in local disease, p<2e-16), 5.1 copy number alterations (5.3 in metastases vs 4.9 in local disease, p = 0.003), and 0.55 structural rearrangements (0.53 in metastases vs 0.56 in local disease, p = 0.07). One-hundred sixteen samples harbored mutations in more than 25 genes, with a significant enrichment of this hypermutated phenotype in metastatic samples relative to local samples (p<1e-5) and in ER + samples (p<1e-08) [47]. A diverse set of genomic alterations were observed at high frequency across the 11,616 samples (**Fig 1G**), including mutations in *TP53* (55.9% of samples), *PIK3CA* (32.4%), *CDH1* (11%), *GATA3* (10.9%), *ESR1* (10.2%), and *KMT2D* (9.5%); amplifications of *MYC* (22.8%), *CCND1* (17.4%), and *HER2* (9.9%); deletions of *PTEN* (5.7%), *CDKN2A/B* (5%), and *RB1* (2.5%); and structural rearrangements of *HER2* (1.5%) and *FGFR1* (1.3%). *BRCA1/2* sequence variants (including deleterious mutations, variants of unknown significance, and deleterious germline variations) were also highly prevalent, at frequencies of 5.6% for *BRCA1* and 7.2% for *BRCA2*. These alterations have been consistently associated with breast cancer in prior reports [14,15,17,24,45,48]. Here we show that most of these frequently altered genes are enriched in a particular subtype based on ER/HER2 status (**S2 Fig**, **S1 Table**) or histology (**S2 Table**).

Genomic alterations enriched in metastatic breast cancer

To search for alterations enriched in metastatic breast cancer we compared local and metastatic tumors while accounting for subtype (ER/HER2) and mutation load. Mutation load can explain the majority of the observed increase in mutation frequency for some genes, particularly large genes without clear hotspots. We confirmed the significant enrichment for *ESR1* mutations in metastatic tumors (18.3% in metastases vs. 2.2% in local disease, p<3e-80; **Fig 2A**, **S3 Table**, **S5**–**S7 Tables**). Beyond this principal feature of metastatic breast cancer, we found a previously unreported enrichment for *CTCF* mutations in metastatic samples (2% in metastases vs. 0.9% in local disease, p<2e-5; **Fig 2A**). Mutations in *CTCF* and at CTCF binding sites have previously been associated with multiple forms of cancer, putatively disrupting the epigenetic regulation of proliferation [49,50]. Furthermore, CTCF has been associated with epithelial-to-mesenchymal transition—a developmental process via which cells gain migratory and invasive properties that can be hijacked during cancer metastasis [51]. As such, we speculate that *CTCF* mutations are a metastatic driver in up to 2% of metastatic breast cancers.

In addition, we observed a significantly higher rate of *CDKN1B* ($p27^{kip1}$) *amplification* in local tumors (1.3% in metastases vs. 3.6% in local disease, p<2e-5) (**Fig 2A**). Strengthening this result, the opposite alterations, *CDKN1B deletions* (0.2% in metastases vs. 0.1% in local disease, p = 0.09) and *mutations* (1.9% in metastases vs. 1.1% in local disease, p = 0.05), trend toward significant enrichment in metastatic tumors. *CDKN1B* controls cell cycle progression at G1 via inhibition of CDK4/6, and high expression of CDKN1B is a positive prognostic biomarker in early-stage disease [52]. A series of CDK4/6 inhibitors have recently emerged for the treatment of late-stage ER+ breast cancer patients that are particularly effective when used in conjunction with hormone therapy [53,54]. One possible interpretation of our data is that *CDKN1B* amplification in primary tumors acts to slow the rate of tumor proliferation and metastasis, which would point toward the possible utility of CDK4/6 inhibitors as a means of delaying progression to metastatic disease.

We also observed significantly higher rates of amplification of the FGFR ligands *FGF3*, *FGF4* and *FGF19* in metastases relative to local tumors, both for ER+ disease (27% in

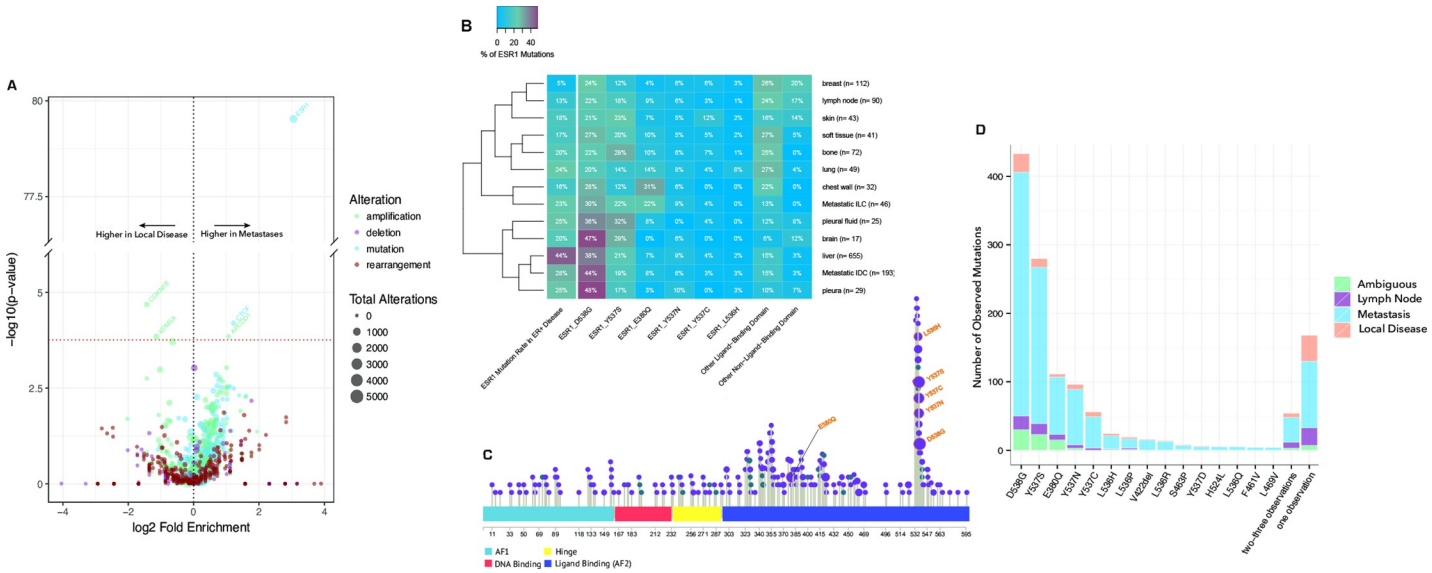

Fig 2. Comparison of metastatic tumors and local disease. (a) Enrichment analysis for alterations occurring at different rates in metastatic tumors vs. local disease, controlling for mutation load and molecular subtype (ER status and *HER2* amplification). (b) Prevalence of *ESR1* hotspot mutations by metastatic site and histological subtype. The far-left column represents the percent of patients with at least one *ESR1* mutation. All other columns represent the percentage of *ESR1* mutations of a certain type that we observe within a specific patient stratification (i.e., the top-left corner shows that 27% of the *ESR1* mutations we see in soft tissue samples are D538G). *ESR1* hotspots occur at significantly different frequencies at different metastatic sites, a result that we do not observe for other genes. (c) Distribution of *ESR1* mutations in the cohort, sized by prevalence. The majority of mutations occur within the ligand-binding domain. (d) Number of *ESR1* mutations by hotspot. All mutations observed 4 or more times are shown. Mutations observed 2–3 times, or 1 time, were pooled for analysis.

https://doi.org/10.1371/journal.pone.0231999.g002

metastases vs. 17% in local disease, p = 0.004, 0.0007, 0.0015 for *FGF3*, *FGF4*, *FGF19*, resp.; S6 Table) and ER- disease (13% in metastases vs. 4% in local disease, p = 0.0008, 0.0001, 0.0002, resp.; S7 Table), when analyzing 1,405 samples with known ER status (*FGF* amplification did not reach significance in the full cohort because of a strong association with ER status). FGF signaling has been previously implicated in resistance to endocrine therapy [55]. Lastly, when only considering variants with known or likely tumorigenic potential, we found a significant enrichment of *KRAS* and *NF1* mutations in metastatic tumors (*KRAS*—1.9% in metastases vs. 1.1% in local disease, p = 0.0045; *NF1*–4.7% vs. 3.3%, p = 0.013, S5 Table), which suggests a potential role for the ras/MAPK pathway in metastasis [30].

Site and subtype association of ESR1 hotspot mutations

Next we provided a comprehensive portrait of the prevalence and diversity of *ESR1* mutations in metastatic breast cancer. *ESR1* mutations are found in 1,183 tumors, 8.9% of those carry more than one *ESR1* mutation, and 922 (78%) are metastases. The *ESR1* mutation rate is highest in ER+ liver metastases (44%), followed by pleura (25%), lung (24%) and bone (20%) (Fig 2B). The mutation rate in ER+ brain metastases was 20%, contrary to the absence of *ESR1* mutations at this site in prior studies [56]. The most prevalent *ESR1* mutations are gain-of-function mutations in the ligand-binding domain, which have been shown to confer constitutive activity in the absence of estrogen (Fig 2C): D538G – 33.2% of all *ESR1* mutations; Y537S – 21.4%; E380Q – 8.5%; Y537N – 8.0%; Y537C – 4.3%; L536H –1.9%; and V422del – 1.4%. We also see enrichment for rare *ESR1* mutations in metastases relative to local disease (p<1e-04 for the set of mutations seen twice; p<1e-05 for the set of mutations seen once) (Fig 2D). This effect is confined to the ligand-binding domain of *ESR1* (p = 0.004 for the set of mutations seen once in the ligand-binding domain vs. p = 0.9 for the set of mutations seen

once outside the ligand-binding domain; p = 0.04 comparing enrichment between the two sets, demonstrating that the difference is not attributable to the larger number of mutations within the ligand-binding domain), suggesting that a long tail of mutations in the *ESR1* ligand-binding domain represents additional resistance mechanisms to aromatase inhibition.

The prevalence of these hotspot mutations further varies by site of metastasis (p<7e-7) and the histological subtype of the tumor (p = 0.003) (Fig 2B). In terms of metastatic site, visceral tissue (liver, pleura, pleural fluid, brain, and lung [57]) has significantly more D538G mutations (20–48% of ESR1 mutations) than all other *ESR1* hotspot mutations, including Y537S (p = 0.007). Bone metastases (non-visceral), on the other hand, have significantly more Y537S mutations (29%) relative to D538G (22%). Peripheral tissue (chest wall) has increased prevalence of E380Q mutations (31%), and both local breast tumors and lymph node metastases have higher rates of likely passenger mutations outside the ligand-binding domain. In terms of histological subtype, the *ESR1* hotspot mutations in invasive ductal carcinoma reflect visceral disease (44% are D538G). Those in invasive lobular carcinoma are enriched for E380Q mutations (22% are E380Q).

Genomic alterations enriched at specific metastatic sites

We next searched for site-specific metastatic alterations by comparing the genomic profiles of tumors from *specific* metastatic sites with local tumors while controlling for subtype and mutation load. We find multiple significant associations between genomic alterations and the site of metastasis (Fig 3A, S4 Table), the most intriguing of which include an enrichment of *ASXL1* amplifications (4.2% vs. 0.8% in local disease, p<2e-05) and *PTEN* deletions (11.8% vs. 5%,

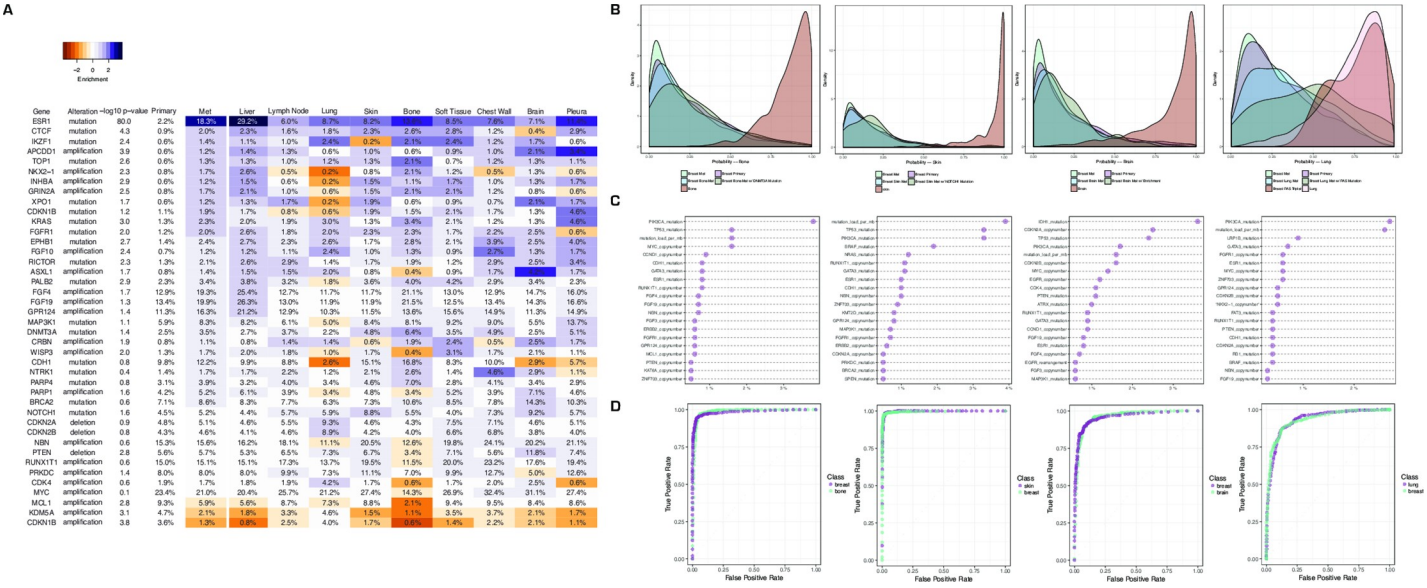

Fig 3. Mutations associated with specific metastatic sites. (a) Alterations enriched at specific metastatic sites. P-values represent comparison between all metastases and local disease. Each cell represents the rate of a specific alteration at a specific metastatic site, colored by enrichment or depletion relative to all local tumors. (b) Probability distributions for a panel of validation samples, using machine learning algorithms trained to differentiate breast tumors from bone, skin, brain, and lung tumors (left to right) using genomic features. Alterations enriched at each metastatic site were not included in the classifiers, but were used to stratify patients—to determine if any of the observed enrichments could be explained by misdiagnosis of new primary tumors. (c) Variable importance for the machine learning algorithms used in (b). The x-axis represents the mean decrease in accuracy of the classifier when a variable is permuted and indicates how useful a specific alteration is in determining the tissue of origin of a tumor from the set of genomic alterations it harbors. (d) ROC curves for the machine learning algorithms used in (b). Each point represents the true and false positive rate for one indication using one threshold on the model output to make a classification decision. A larger area under the curve represents a more accurate model, and we achieve high accuracy in all cases (93.8%, 96.6%, 91.7%, and 85.9% from left to right).

https://doi.org/10.1371/journal.pone.0231999.g003

p<1e-04) in brain metastases; an enrichment of *DNMT3A* mutations in bone metastases (6.4% vs. 2.5%, p<2e-5); an enrichment of *NOTCH1* mutations in skin metastases (8.8% vs. 4.5%, p = 5e-4); and enrichments of *KRAS*, *KEAP1*, *STK11* and *EGFR* mutations in lung metastases (2.6–3.2% vs. 1.0–2.1%, p = 0.004–0.14). The latter enrichment strengthened when only considering known and likely driver mutations (2.8, 1.2, 2.4, 1.8% vs. 1, 0.3, 1, 0.6%, p = 0.0003, 0.003, 0.007, 0.004, respectively).

Most of these enriched alterations have been associated with primary tumors at the site of metastasis. *KRAS* is the most common driver mutation in lung cancer, and is highly co-occurrent with *KEAP1* and *STK11* mutations [58,59], *NOTCH1* has a putative role in skin cancer [60], loss of *PTEN* is the most common genomic alteration in glioblastoma [61], and *DNMT3A* has been associated with leukemia and myelodysplasia [62,63]. This suggests that these alterations serve to mimic primary tumors at the site of metastasis and confer adaptation to the local tumor microenvironment. To further support this hypothesis, we ensured that these tumors are not misdiagnosed *de novo* primary tumors occurring in patients with a prior history of breast cancer. For this purpose, we developed machine learning algorithms to differentiate breast tumors from primary tumors of the brain, bone, lung, and skin (**Fig 3B, 3C and 3D**). The algorithms were trained using genomic data from 500 tumors of each indication, sequenced during standard care using the FoundationOne assay, with the enriched alterations associated with each indication masked (e.g., *PTEN* deletions excluded from the brain classification). In each case, we achieved high accuracy on held-out test data for the prediction of skin vs. breast (96.6%), lung vs. breast (85.9%), bone vs. breast (93.8%), and brain vs. breast (91.7%) (**Fig 3D**). We then applied the algorithms to the metastatic breast tumors containing the enriched alterations, to confirm that those tumors are equally likely to originate from breast tissue relative to all other metastatic breast tumors in the dataset. Fourteen percent of breast tumors with *DNMT3A* mutations that metastasized to bone were classified as bone, statistically indistinguishable from the 12% of all breast metastases classified as bone (p = 0.39). Similarly, 7% of brain metastases that harbor *ASXL1* amplifications or *PTEN* deletions were classified as brain, less than the 9% of all breast metastases classified as brain. Skin metastases that harbor *NOTCH1* mutations were classified as skin cancer at a higher rate (18% vs 5%, p = 0.052), but the effect was primarily driven by differences in subtype prevalence between local tumors and skin metastases and was not significant after controlling for this (p = 0.78). These results are consistent with the hypothesis that the enrichments for *DNMT3A* mutations in bone metastases, *NOTCH1* mutations in skin metastases, and *ASXL1* amplifications/*PTEN* deletions in brain metastases are biologically relevant, either as adaptations that arise in response to the local tumor microenvironment or as drivers of site-specific patterns of metastasis. An alternative hypothesis for the enrichment of *DNMT3A* mutations is clonal hematopoiesis of unknown potential, a process in which somatic mutations in hematopoietic stem cells lead to the outgrowth of distinct subclones that have been associated with cancer [64–66]; we provide evidence for this hypothesis in S3 Fig.

We observed a different pattern in breast tumors that metastasize to lung and harbor a mutation in *KEAP1*, *KRAS*, *STK11*, or *EGFR* (**Fig 3B**). The lung classifier, which does not consider mutations in these four genes, scored a significantly larger fraction of these tumors as lung compared to all breast metastases (50% vs. 18%, p = 2e-5, n = 52). Furthermore, we observed that breast cancer metastases at *any* site that harbor mutations in *KEAP1*, *KRAS*, and *STK11*, a triplet commonly associated with lung cancer, genomically resemble lung tumors (100% vs. 18%, p = 2e-8, n = 13). Consistent with these findings, we found enrichment for additional lung-associated alterations in this set of tumors including mutations in *LRP1B* (21.2% vs. 8.2%, p = 0.03) and deletions of *CDKN2A/B* (11.5% vs. 4.5%, p = 0.02). These findings are highly suggestive of misdiagnosis (primary lung tumors diagnosed as breast

metastases and lung cancer metastases diagnosed as breast cancer metastases). Misdiagnosis has substantial implications for treatment choice and efficacy and illustrates the potential of genomic profiling to complement other modalities in selecting effective treatments for individual patients.

Discussion

Large-scale sequencing of cancer genomes holds the promise of delivering novel therapeutic targets and personalized treatment for patients. Initial efforts to characterize and understand the cancer genome focused on primary disease, and led to the development of multiple targeted therapies with substantive impact on patients' lives. We have conducted a comprehensive analysis of real-world breast cancer metastases sequenced during standard clinical care and show enrichment for (i) mechanisms of acquired resistance and (ii) alterations that may induce or accelerate metastasis. It is our hope that a deeper understanding of these classes of alterations will ultimately lead to new treatments for metastatic disease, which is the cause of most deaths in breast cancer and represents a substantial unmet medical need.

Mutations in the *ESR1* ligand-binding domain are the principal feature of hormone receptor-positive metastatic breast cancer, arising in response to aromatase inhibition and allowing the tumor to progress in the absence of estrogen [67,68]. Therapies that modulate the mutant receptor are in clinical trials, and *ESR1* mutations in circulating tumor DNA are a promising biomarker for disease progression [26,69]. We have shown that specific *ESR1* hotspot mutations are associated with specific metastatic niches and disease histologies, suggesting the possibility of cofactor interactions or biological contexts that could be druggable; prior work has demonstrated functional differences between ESR1 hotspot mutations *in vitro* and shown that hotspots differentially respond to drug [56,70]. The development of targeted therapy against ESR1 mutations is an active area of research with great promise for treating metastatic disease, and a deeper understanding of how specific mutations function *in vivo* will be crucial for optimizing patient outcomes. In addition, we provide evidence for a long tail of ligand-binding domain mutations that appear to be functional given their enrichment in metastatic disease. Patients with these rare mutations may be resistant to traditional hormone therapy and should be monitored closely for disease progression.

The enrichment for *CTCF* mutations that we see in metastatic disease is not an anticipated resistance mechanism. *CTCF* binding site mutations are enriched in multiple cancers, and *CTCF* mutations in cancer have been shown to specifically alter interactions with promoters or insulators of genes associated with proliferation [50,51]. The observed enrichment is consistent with the hypothesis that multiple steps in the metastatic cascade involve epigenetic transitions that allow tumor cells to co-opt biological processes to promote growth and metastasis.

In order to colonize a distal site, tumor cells need to disseminate, evade the immune system, and adapt to a novel microenvironment. It is an open question whether this metastatic potential is present in primary disease or is enabled by additional genetic and epigenetic events. We have shown multiple instances in which breast cancer metastases are enriched for alterations that occur in primary tumors at the site of metastasis. This is consistent with the hypothesis that, in some contexts, metastases mimic primary tumor biology to adapt to novel microenvironments, but experimental follow-up will be necessary to fully understand this process.

When a patient with a history of cancer presents with a malignancy at a new site, the decision whether the lesion is a metastasis or a novel primary tumor has a dramatic impact on prognosis and treatment. Diagnosis is complicated by the possibility that the metastasis has arisen from a subclone of the primary tumor. Microsatellite analysis has been used for differential diagnosis, and traditional methods include histology and interval to cancer formation

[71]. We have shown the potential of an orthogonal approach, leveraging machine learning and large-scale sequencing data to classify the most probable tissue of origin of a metastatic lesion. Using this method, we found multiple instances of likely primary lung tumors misdiagnosed as breast metastases, though we cannot exclude the possibility that these represent breast tumors that strongly mimic lung cancer biology.

Conclusions

Besides the known hotspot mutations in *ESR1*, we observed a metastatic enrichment of previously unreported, lower-prevalence mutations in the ligand-binding domain, implying that these mutations may also be functional. Furthermore, individual *ESR1* hotspots are significantly enriched in specific metastatic tissues and histologies, suggesting functional differences between these mutations. Other alterations enriched across all metastases include loss of function of the CDK4 regulator *CDKN1B*, and mutations in the transcription factor *CTCF*. Mutations enriched at specific metastatic sites generally reflect biology of the target tissue and may be adaptations to growth in the local environment. These include *PTEN* and *ASXL1* alterations in brain metastases and *NOTCH1 alterations* in skin. We observed an enrichment of *KRAS*, *KEAP1*, *STK11* and *EGFR* mutations in lung metastases. However, the patterns of other mutations in these tumors indicate that these are misdiagnosed lung primaries rather than breast metastases.

A core implication of this paper, as well as other recent work on cancer metastasis in different indications and utilizing different assays, is that there are perhaps fewer specific genomic alterations that drive metastasis than was anticipated [28,29]. This suggests several non-exclusive possibilities. First, the majority of primary tumors may harbor metastatic potential without the need to incur additional genomic alterations. Second, changes in cell state during the process of metastasis may be primarily epigenetic rather than genetic in nature. Third, metastasis may be driven by a large number of alterations with small individual effects.

Cancer is a heterogeneous disease—each patient presents with a unique constellation of genetic and epigenetic alterations that have transformed healthy tissue into a malignancy. We are in the early stages of embracing and understanding this complexity, but doing so will allow us to develop new therapies and target the right patients with the right drugs. This work, which utilizes real-world data that lacks extensive clinical annotation but provides enormous scope and scale, is complementary to efforts that generate curated data in small cohorts. Each has a comparative advantage in tackling specific questions, but both may be necessary to realize the promise of genomic medicine, deliver effective personalized oncology, and ultimately improve outcomes for patients.

Supporting information

S1 Fig. Machine learning algorithms to classify molecular and histological subtype. (a) Probability distributions for the output of machine learning algorithms trained to infer molecular (left) or histological (right) subtype from the set of genomic alterations harbored by a tumor. See Fig 3B legend for more details. (b) Variable importance for the machine learning algorithms used in (a). The x-axis represents the mean decrease in accuracy of the classifier when a variable is permuted and indicates how useful a specific alteration is in determining the subtype of a tumor from the set of genomic alterations it harbors. (c) ROC curves for the machine learning algorithms used in (a). See Fig 3B–3D legend for more details.
(TIF)

S2 Fig. Landscape of genomic alterations by molecular subtype. Landscape of genomic alterations in (a) ER+, (b) ER-, and (c) HER2+ disease. Each cell represents the status of one gene

in one patient, colored by alteration type. ER status was determined by pathology report. HER2 status was determined by *HER2* copy number.
(TIF)

S3 Fig. Evidence for clonal hematopoiesis. Clonal hematopoiesis is a process via which somatic mutations in hematopoietic stem cells lead to the outgrowth of distinct subclones [64]. Clonal hematopoiesis is observed in 10% of adults over 65 years of age, but in only 1% of those under 50, and has been associated with cancer [65,72]. *DNMT3A* mutations are the most frequently observed mutation in clonal hematopoiesis of indeterminate potential (CHIP) [64], and have not previously been associated with breast cancer. As such, we speculated that the observed enrichment of *DNMT3A* mutations in bone metastases might be a consequence of clonal hematopoiesis and not of alterations harbored by the tumor. Consistent with this hypothesis, we observe an increasing mutation rate with patient age (a) that cannot be explained by changes in histological and molecular subtype (c) and a decreasing fraction of reads associated with the mutant allele that we do not observe in other genes (b). The enrichment is not specific to bone metastases, but the rate at which clonal hematopoiesis may be present varies by biopsy site (d). (a) Frequency of mutation by patient age, normalized to the observed frequency in patients aged 20–39, for genes that show the strongest association with patient age. Most effects can be explained by changing proportions of histological and molecular subtype, seen in Fig 1D and 1F. *DNMT3A* mutations increase with age and show a unique pattern. (b) Fraction of reads associated with the mutant allele in patients that harbor a mutation for *PIK3CA*, *TP53*, and *DNMT3A*. The read fraction for *DNMT3A* decreases with patient age, consistent with CHIP. (c) Prevalence of histological and molecular subtype by patient age. (d) *DNMT3A* mutation rate by patient age and biopsy site.
(TIF)

S1 Table. Top alterations by molecular subtype, as defined by *HER2* copy number and ER status from pathology report, in 1,405 samples with complete clinical annotation. Pathology reports were scored by an algorithm with 95% accuracy.
(XLSX)

S2 Table. Top alterations by histological subtype in male patients and patients under 40.
(XLSX)

S3 Table. Alterations enriched in metastatic tumors relative to local disease (primary tumors and local recurrences). Corrected p-values were calculated by permuting the met/local status of samples 1000 times, reflecting the probability of observing a more significant enrichment by chance.
(XLSX)

S4 Table. Alterations enriched by site of metastasis relative to local disease (primary tumors and local recurrences). Corrected p-values were calculated by permuting the tissue of samples 1000 times. Results for the 9 most common biopsy sites are shown, for alterations that occurred at least ten times at the metastatic site.
(XLSX)

S5 Table. Mutations enriched in metastatic tumors relative to local disease (primary tumors and local recurrences) after filtering out variants of unknown significance. Corrected p-values were calculated by permuting the met/local status of samples 1000 times, reflecting the probability of observing a more significant enrichment by chance.
(XLSX)

S6 Table. Mutations enriched in ER+ metastatic tumors relative to ER+ local disease (primary tumors and local recurrences) as defined by IHC for samples with available IHC (n = 719). Corrected p-values were calculated by permuting the met/local status of samples 1000 times, reflecting the probability of observing a more significant enrichment by chance. (XLSX)

S7 Table. Mutations enriched in ER- metastatic tumors relative to ER- local disease (primary tumors and local recurrences) as defined by IHC for samples with available IHC (n = 532). Corrected p-values were calculated by permuting the met/local status of samples 1000 times, reflecting the probability of observing a more significant enrichment by chance. (XLSX)

S8 Table Genes included on FoundationOne Panels.
(XLSX)

Acknowledgments

We thank Sophia Maund for excellent data and collaboration management; Juliann Chmielecki, Laurie Gay, Julia Elvin, Shakti Ramkissoon, and Bob Trucci for development of the pathology report parser (PReP); Oleg Mayba for statistical consulting; and Ciara Metcalfe, Craig Cummings and Tim Wilson for feedback on the manuscript

Author Contributions

Conceptualization: Jacob Rinaldi, Anneleen Daemen, Gerard Manning.

Data curation: Jacob Rinaldi, Ethan S. Sokol, Ryan J. Hartmaier, Sally E. Trabucco, Garrett M. Frampton, Michael E. Goldberg, Lee A. Albacker.

Methodology: Jacob Rinaldi.

Software: Lee A. Albacker.

Supervision: Garrett M. Frampton, Anneleen Daemen, Gerard Manning.

Visualization: Jacob Rinaldi.

Writing – original draft: Jacob Rinaldi, Anneleen Daemen, Gerard Manning.

Writing – review & editing: Jacob Rinaldi, Ethan S. Sokol, Ryan J. Hartmaier, Sally E. Trabucco, Anneleen Daemen, Gerard Manning.

References

1. Torre L. A., Islami F., Siegel R. L., Ward E. M. & Jemal A. *Global cancer in women: burden and trends*. (AACR, 2017).
2. Gupta G. P. & Massagué J. Cancer metastasis: building a framework. *Cell* 127, 679–695 (2006). https://doi.org/10.1016/j.cell.2006.11.001 PMID: 17110329
3. Valastyan S. & Weinberg R. A. Tumor metastasis: molecular insights and evolving paradigms. *Cell* 147, 275–292 (2011). https://doi.org/10.1016/j.cell.2011.09.024 PMID: 22000009
4. Scully O. J., Bay B.-H., Yip G. & Yu Y. Breast cancer metastasis. *Cancer Genomics-Proteomics* 9, 311–320 (2012). PMID: 22990110
5. Jatoi I. & Rody A. *Management of Breast Diseases*. (Springer, 2016).
6. Tevaarwerk A. J. et al. Survival in patients with metastatic recurrent breast cancer after adjuvant chemotherapy. *Cancer* 119, 1140–1148 (2013). https://doi.org/10.1002/cncr.27819 PMID: 23065954
7. Steeg P. S. Targeting metastasis. *Nat. Rev. Cancer* 16, 201–218 (2016). https://doi.org/10.1038/nrc.2016.25 PMID: 27009393

8. Sledge G. W. et al. Past, present, and future challenges in breast cancer treatment. *J. Clin. Oncol.* 32, 1979–1986 (2014). https://doi.org/10.1200/JCO.2014.55.4139 PMID: 24888802

9. Group E. B. C. T. C. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *The Lancet* 365, 1687–1717 (2005).

10. Smith I. E. & Dowsett M. Aromatase inhibitors in breast cancer. *N. Engl. J. Med.* 348, 2431–2442 (2003). https://doi.org/10.1056/NEJMra023246 PMID: 12802030

11. Howell A. et al. Results of the ATAC (Arimidex, Tamoxifen, Alone or in Combination) trial after completion of 5 years' adjuvant treatment for breast cancer. *Lancet* 365, 60–62 (2005). https://doi.org/10.1016/S0140-6736(04)17666-6 PMID: 15639680

12. Romond E. H. et al. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N. Engl. J. Med.* 353, 1673–1684 (2005). https://doi.org/10.1056/NEJMoa052122 PMID: 16236738

13. Smith I. et al. 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial. *The lancet* 369, 29–36 (2007).

14. Network C. G. A. Comprehensive molecular portraits of human breast tumors. *Nature* 490, 61 (2012). https://doi.org/10.1038/nature11412 PMID: 23000897

15. Ciriello G. et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* 163, 506–519 (2015). https://doi.org/10.1016/j.cell.2015.09.033 PMID: 26451490

16. Burstein M. D. et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin. Cancer Res.* (2014).

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

18. A Study of Ipatasertib in Combination With Paclitaxel as a Treatment for Participants With PIK3CA/AKT1/PTEN-Altered, Locally Advanced or Metastatic, Triple-Negative Breast Cancer or Hormone Receptor-Positive, HER2-Negative Breast Cancer—Full Text View—ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT03337724. (Accessed: 13th November 2017)

19. AZD5363 in Patients With Advanced Solid Tumors Harboring AKT Mutations—Full Text View—ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT03310541. (Accessed: 13th November 2017)

20. Efficacy and Safety of Treatment With Alpelisib Plus Endocrine Therapy in Patients With HR+, HER2-negative aBC, With PIK3CA Mutations, Whose Disease Has Progressed on or After CDK 4/6 Treatment With an Aromatase Inhibitor (AI) or Fulvestrant—Full Text View—ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT03056755. (Accessed: 13th November 2017)

21. Fulvestrant, Palbociclib and Erdafitinib in ER+/HER2-/FGFR-amplified Metastatic Breast Cancer—Full Text View—ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT03238196. (Accessed: 13th November 2017)

22. Pembrolizumab in Advanced BRCA-mutated Breast Cancer—Full Text View—ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT03025035. (Accessed: 13th November 2017)

23. PARP Inhibition for Triple Negative Breast Cancer (ER-/PR-/HER2-)With BRCA1/2 Mutations—Full Text View—ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT01074970. (Accessed: 13th November 2017)

24. Robinson D. R. et al. Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nat. Genet.* 45, 1446–1451 (2013). https://doi.org/10.1038/ng.2823 PMID: 24185510

25. Chandarlapaty S. et al. Prevalence of ESR1 mutations in cell-free DNA and outcomes in metastatic breast cancer: a secondary analysis of the BOLERO-2 clinical trial. *JAMA Oncol.* 2, 1310–1315 (2016). https://doi.org/10.1001/jamaoncol.2016.1279 PMID: 27532364

26. Schiavon G. et al. Analysis of ESR1 mutation in circulating tumor DNA demonstrates evolution during therapy for metastatic breast cancer. *Sci. Transl. Med.* 7, 313ra182–313ra182 (2015). https://doi.org/10.1126/scitranslmed.aac7551 PMID: 26560360

27. Spoerke J. M. et al. Heterogeneity and clinical significance of ESR1 mutations in ER-positive metastatic breast cancer patients receiving fulvestrant. *Nat. Commun.* 7, (2016).

28. Zehir A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* (2017).

29. Yates L. R. et al. Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* 32, 169–184. e7 (2017). https://doi.org/10.1016/j.ccell.2017.07.005 PMID: 28810143

30. Razavi P. et al. The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell* 34, 427–438. e6 (2018). https://doi.org/10.1016/j.ccell.2018.08.008 PMID: 30205045

31. Angus Lindsay et al. The genomic landscape of 501 metastatic breast cancer patients. (2018).

32. Andre Fabrice et al. Genomic characterization of metastatic breast cancer. (2018).
33. Frampton G. M. et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* 31, 1023–1031 (2013). https://doi.org/10.1038/nbt.2696 PMID: 24142049
34. He J. et al. Integrated genomic DNA/RNA profiling of hematologic malignancies in the clinical setting. *Blood* 127, 3004–3014 (2016). https://doi.org/10.1182/blood-2015-08-664649 PMID: 26966091
35. Forbes S. A. et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* 39, D945–D950 (2010). https://doi.org/10.1093/nar/gkq929 PMID: 20952405
36. Ho, T. K. Random decision forests. in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on* 1, 278–282 (IEEE, 1995).
37. Team R. C. *R: a language and environment for statistical computing. Version 3.1. 3.* Vienna, Austria: R Foundation for Statistical Computing; 2015. (2013).
38. Kuhn M. Caret package. *J. Stat. Softw.* 28, 1–26 (2008). https://doi.org/10.18637/jss.v028.i07 PMID: 27774042
39. Liaw A. & Wiener M. Classification and regression by randomForest. *R News* 2, 18–22 (2002).
40. Sun J. X. et al. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLoS Comput. Biol.* 14, e1005965 (2018). https://doi.org/10.1371/journal.pcbi.1005965 PMID: 29415044
41. Jay J. J. & Brouwer C. Lollipops in the clinic: information dense mutation plots for precision medicine. *PloS One* 11, e0160519 (2016). https://doi.org/10.1371/journal.pone.0160519 PMID: 27490490
42. Wickham H. *ggplot2: elegant graphics for data analysis*. (Springer, 2016).
43. Gu Z., Eils R. & Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32, 2847–2849 (2016). https://doi.org/10.1093/bioinformatics/btw313 PMID: 27207943
44. Kennecke H. et al. Metastatic behavior of breast cancer subtypes. *J. Clin. Oncol.* 28, 3271–3277 (2010). https://doi.org/10.1200/JCO.2009.25.9820 PMID: 20498394
45. Pharoah P. D. P., Day N. E. & Caldas C. Somatic mutations in the p53 gene and prognosis in breast cancer: a meta-analysis. *Br. J. Cancer* 80, 1968–1973 (1999). https://doi.org/10.1038/sj.bjc.6690628 PMID: 10471047
46. Hartmaier R. J. et al. High-throughput genomic profiling of adult solid tumors reveals novel insights into cancer pathogenesis. *Cancer Res.* canres. 2479.2016 (2017).
47. Lefebvre C. et al. Mutational profile of metastatic breast cancers: a retrospective analysis. *PLoS Med.* 13, e1002201 (2016). https://doi.org/10.1371/journal.pmed.1002201 PMID: 28027327
48. Daemen A. An update on the genomic landscape of breast cancer: new opportunity for personalized therapy? *Transl. Cancer Res.* 1, 279–282 (2012).
49. Filippova G. N. et al. Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter its DNA-binding specificity. *Cancer Res.* 62, 48–52 (2002). PMID: 11782357
50. Katainen R. et al. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* 47, 818 (2015). https://doi.org/10.1038/ng.3335 PMID: 26053496
51. Bhattacharya A., Hur J. & Dhasarathy A. The role of CCCTC binding factor (CTCF) in epithelial to mesenchymal transition (EMT). *FASEB J.* 31, 593.8–593.8 (2017).
52. Chu I. M., Hengst L. & Slingerland J. M. The Cdk inhibitor p27 in human cancer: prognostic potential and relevance to anticancer therapy. *Nat. Rev. Cancer* 8, 253 (2008). https://doi.org/10.1038/nrc2347 PMID: 18354415
53. Finn R. S., Aleshin A. & Slamon D. J. Targeting the cyclin-dependent kinases (CDK) 4/6 in estrogen receptor-positive breast cancers. *Breast Cancer Res.* 18, 17 (2016). https://doi.org/10.1186/s13058-015-0661-5 PMID: 26857361
54. Otto T. & Sicinski P. Cell cycle proteins as promising targets in cancer therapy. *Nat. Rev. Cancer* 17, 93 (2017). https://doi.org/10.1038/nrc.2016.138 PMID: 28127048
55. Tomlinson D. C., Knowles M. A. & Speirs V. Mechanisms of FGFR3 actions in endocrine resistant breast cancer. *Int. J. Cancer* 130, 2857–2866 (2012). https://doi.org/10.1002/ijc.26304 PMID: 21792889
56. Toy W. et al. Activating ESR1 Mutations Differentially Affect the Efficacy of ER Antagonists. *Cancer Discov.* 7, 277–287 (2017). https://doi.org/10.1158/2159-8290.CD-15-1523 PMID: 27986707

57. Robertson J. F. et al. Fulvestrant 500 mg versus anastrozole 1 mg for hormone receptor-positive advanced breast cancer (FALCON): an international, randomised, double-blind, phase 3 trial. *The Lancet* 388, 2997–3005 (2016).

58. Bhattacharya S., Socinski M. A. & Burns T. F. KRAS mutant lung cancer: progress thus far on an elusive therapeutic target. *Clin. Transl. Med.* 4, 35 (2015). https://doi.org/10.1186/s40169-015-0075-0 PMID: 26668062

59. Skoulidis F. et al. Co-occurring genomic alterations define major subsets of KRAS-mutant lung adenocarcinoma with distinct biology, immune profiles, and therapeutic vulnerabilities. *Cancer Discov.* 5, 860–877 (2015). https://doi.org/10.1158/2159-8290.CD-14-1236 PMID: 26069186

60. Reichrath J. & Reichrath S. Notch-signaling and nonmelanoma skin cancer: an ancient friend, revisited. *Notch Signal. Embryol. Cancer* 265–271 (2012).

61. Ohqaki H., Dessen P. & Jourde B. Genetic pathways to glioblastoma: a population2based study. *Cancer Res* 64, 689226899 (2004).

62. Ley T. J. et al. DNMT3A mutations in acute myeloid leukemia. *N. Engl. J. Med.* 363, 2424–2433 (2010). https://doi.org/10.1056/NEJMoa1005143 PMID: 21067377

63. Walter M. J. et al. Recurrent DNMT3A mutations in patients with myelodysplastic syndromes. *Leukemia* 25, 1153 (2011). https://doi.org/10.1038/leu.2011.44 PMID: 21415852

64. Steensma D. P. et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* 126, 9–16 (2015). https://doi.org/10.1182/blood-2015-03-631747 PMID: 25931582

65. Coombs C. C. et al. Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes. *Cell Stem Cell* 21, 374–382. e4 (2017). https://doi.org/10.1016/j.stem.2017.07.010 PMID: 28803919

66. Severson E. A. et al. Detection of clonal hematopoiesis of indeterminate potential in clinical sequencing of solid tumor specimens. *Blood* blood-2018-03-840629 (2018).

67. Desmedt C. et al. ESR1 mutations in metastatic lobular breast cancer patients. *NPJ Breast Cancer* 5, 9 (2019). https://doi.org/10.1038/s41523-019-0104-z PMID: 30820448

68. Clatot F., Augusto L. & Di Fiore F. ESR1 mutations in breast cancer. *Aging* 9, 3 (2017). https://doi.org/10.18632/aging.101165 PMID: 28130553

69. A Study of GDC-9545 Alone or in Combination With Palbociclib and/or Luteinizing Hormone-Releasing Hormone (LHRH) Agonist in Locally Advanced or Metastatic Estrogen Receptor-Positive Breast Cancer—Full Text View—ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT03332797. (Accessed: 14th November 2017)

70. Bahreini A. et al. Mutation site and context dependent effects of ESR1 mutation in genome-edited breast cancer cell models. *Breast Cancer Res.* 19, 60 (2017). https://doi.org/10.1186/s13058-017-0851-4 PMID: 28535794

71. Leong P. P. et al. Distinguishing Second Primary Tumors From Lung Metastases in Patients With Head and Neck Sauamous Cell Carcinoma. *J. Natl. Cancer Inst.* 90, 972–977 (1998). https://doi.org/10.1093/jnci/90.13.972 PMID: 9665144

72. Genovese G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* 371, 2477–2487 (2014). https://doi.org/10.1056/NEJMoa1409405 PMID: 25426838