Evolutionary history of metastatic breast cancer reveals minimal seeding from axillary lymph nodes

Ikram Ullah,[1] Govindasamy-Muralidharan Karthik,[1] Amjad Alkodsi,[2] Una Kjällquist,[1] Gustav Stålhammar,[1] John Lövrot,[1] Nelson-Fuentes Martinez,[3] Jens Lagergren,[4] Sampsa Hautaniemi,[2] Johan Hartman,[1,3] and Jonas Bergh[1,5]

[1]Department of Oncology and Pathology, Karolinska Institute, Stockholm, Sweden. [2]Genome-Scale Biology Research Program Unit, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [3]Department of Clinical Pathology, Karolinska University Hospital, Stockholm, Sweden. [4]Department of Computational Biology, Royal Institute of Technology, Stockholm, Sweden. [5]Radiumhemmet – Karolinska Oncology, Karolinska University Hospital, Stockholm, Sweden.

Metastatic breast cancers are still incurable. Characterizing the evolutionary landscape of these cancers, including the role of metastatic axillary lymph nodes (ALNs) in seeding distant organ metastasis, can provide a rational basis for effective treatments. Here, we have described the genomic analyses of the primary tumors and metastatic lesions from 99 samples obtained from 20 patients with breast cancer. Our evolutionary analyses revealed diverse spreading and seeding patterns that govern tumor progression. Although linear evolution to successive metastatic sites was common, parallel evolution from the primary tumor to multiple distant sites was also evident. Metastatic spreading was frequently coupled with polyclonal seeding, in which multiple metastatic subclones originated from the primary tumor and/or other distant metastases. Synchronous ALN metastasis, a well-established prognosticator of breast cancer, was not involved in seeding the distant metastasis, suggesting a hematogenous route for cancer dissemination. Clonal evolution coincided frequently with emerging driver alterations and evolving mutational processes, notably an increase in apolipoprotein B mRNA–editing enzyme, catalytic polypeptide-like–associated (APOBEC-associated) mutagenesis. Our data provide genomic evidence for a role of ALN metastasis in seeding distant organ metastasis and elucidate the evolving mutational landscape during cancer progression.

Introduction

Breast cancer mortality is almost exclusively a consequence of metastatic spreading of the primary cancer to distant organs (1). Treatment outcomes for patients with primary breast cancers, including those with micrometastatic disease, have substantially improved as a result of the efficient use of adjuvant therapies (2, 3). However, once metastatic disease is established, the response to the same treatment strategy becomes dismal. The difference in treatment response between primary tumors and distant metastases has been attributed, in part, to an ongoing evolution of heterogeneous cancer cell populations giving rise to treatment-resistant clones (4). As we and others previously reported, this is reflected by altered expression of prognostic and therapy-predictive biomarkers during metastasis (5, 6). The discordance of biomarker expression influences survival and changes disease management in 1 of 6 to 7 patients (7, 8). Since cancer evolution can affect therapeutic approaches, there has been strong interest in understanding the dynamics of the genomic evolution and dissemination patterns of metastatic cancer cells during disease progression. Sequencing of spatially and temporally distinct tumor samples in metastatic prostate cancer has identified metastasis-to-metastasis spreading (9), whereby rare subclones develop metastatic capabilities within the primary tumor (10). This opposes the theory that metastatic potential is a property of the entire primary tumor bulk (11, 12). Previous breast cancer studies have reported varying degrees of genomic concordance between metastatic samples and their corresponding primary tumors (13–16). A recent study involving 10 patients with metastatic breast cancer reported both monoclonal and polyclonal origins of metastasis (17). However, none of these studies discussed the role of axillary lymph node (ALN) metastasis in seeding distant organ metastases.

Regional and distant lymph nodes are the most common sites for metastatic invasion in various cancers (18). Metastatic engagement of ALNs is a robust prognostic factor in breast cancer (19, 20). One hypothesis is that lymph nodes are way stations for metastatic seeding via the lymphatic system (21–23). However, genomic evidence either supporting or refuting such a hypothesis is lacking, raising the question of whether distant metastases are primarily seeded lymphatically or hematogenously.

To investigate the evolutionary history of metastatic breast cancer, we performed whole-exome sequencing on 99 samples from 20 breast cancer patients with multiple longitudinal and/or spatially distributed biopsies of primary tumors, local recurrent tumors, ALNs, and distant metastases collected during different therapies. We reconstructed the phylogenetic relationships of primary cancers to their metastatic descendants, which revealed patterns of metastatic spreading and the role of ALN metastasis in subsequent cancer progression. To complement the phylogenetic results, we performed subclonal analy-

Authorship note: IU, GMK, and AA are co–first authors. JH and JB are co–senior authors.
Conflict of interest: The authors have declared that no conflict of interest exists.
Submitted: July 7, 2017; **Accepted:** December 21, 2017.
Reference information: *J Clin Invest*. 2018;128(4):1355–1370.
https://doi.org/10.1172/JCI96149.

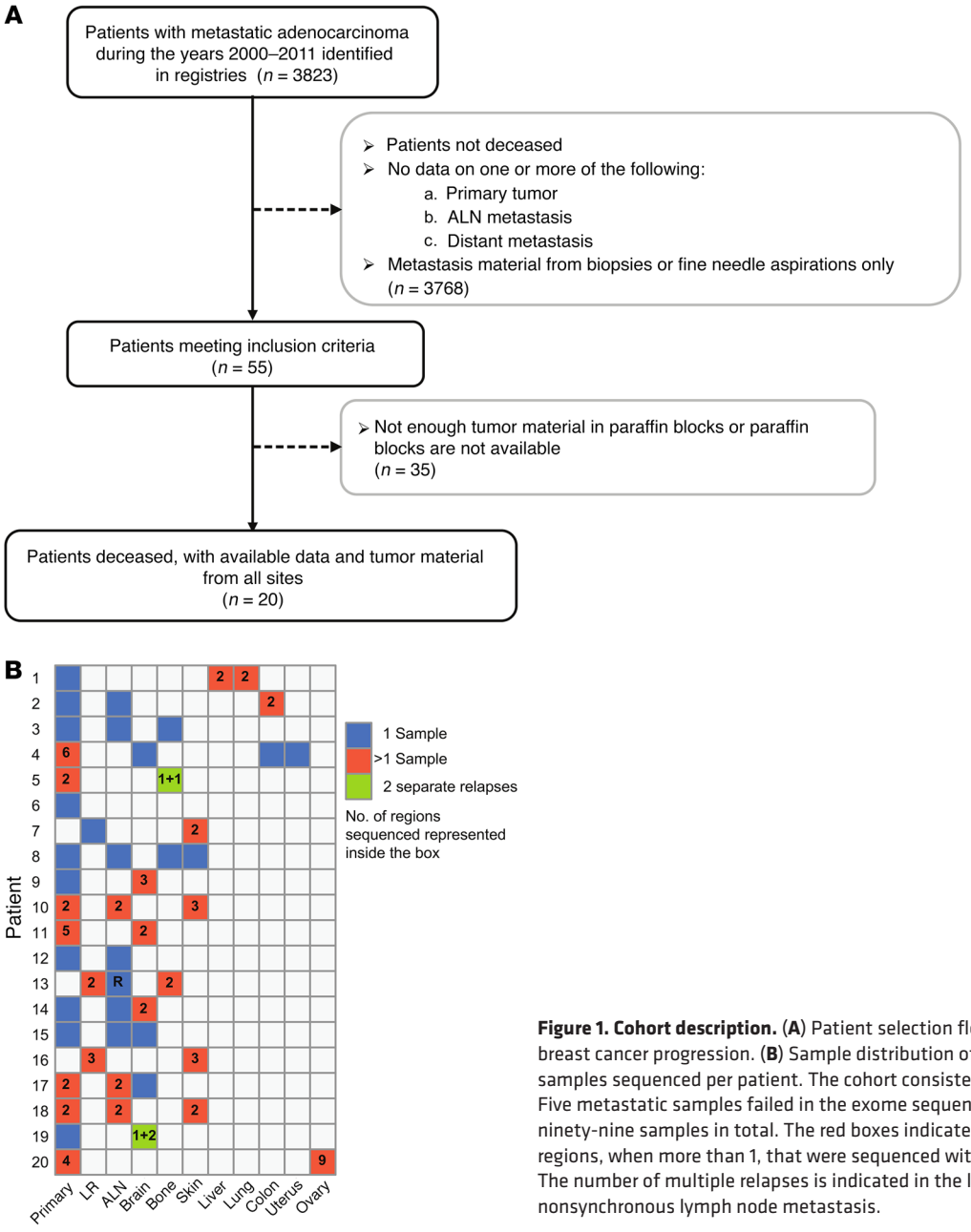


Figure 1. Cohort description. (**A**) Patient selection flow chart to study breast cancer progression. (**B**) Sample distribution of the number of cancer samples sequenced per patient. The cohort consisted of 104 samples. Five metastatic samples failed in the exome sequencing that yielded ninety-nine samples in total. The red boxes indicate the number of regions, when more than 1, that were sequenced within the same tumor. The number of multiple relapses is indicated in the light green boxes. R, nonsynchronous lymph node metastasis.

sis of these samples, which reinforced the phylogenetic results and identified the subclones responsible for seeding successive metastases. Finally, we determined that mutational processes were active during breast cancer progression.

Results

Multiregion sequencing of paired primary and metastatic breast cancers. We performed whole-exome sequencing of 99 samples (formalin-fixed, paraffin embedded [FFPE] tissue blocks) from 20 patients with matched normal controls sampled from normal ALNs (Figure 1A), achieving an average read coverage of 80× (70% targeted regions with >30× coverage) (Supplemental Figure 1 and Supplemental Table 1; supplemental material available online with this article; https://doi.org/10.1172/JCI96149DS1). This included 33 primary tumor samples (from 17 patients, including multiple regions of the same tumors: 2 regions for 4 patients and 4–6 regions for 3 patients); 6 local recurrence (LR) samples (from 3 patients); 12 synchronous ipsilateral ALN metastases (from 9 patients); 1 nonsynchronous ALN relapse; and 47 distant metastasis samples (from 18 patients) including multiple regions of the same metastasis and multiple distant organ metastases from the same patient (Figure 1B). Figure 2 provides an illustration of all the tumor samples and anatomical site information for the sequenced samples. A board-certified surgical pathologist (NFM) verified the origin of the metastatic lesion, the tumor cell percentage, and the Nottingham histological grade for all of the samples sequenced (Supplemental Table 2). We identified a total of 8,859 point mutations (range: 67–1,286; median: 383), of which 3,715 were exonic (range: 40–325; median: 182) and 2,671 were nonsilent, i.e., missense, nonsense, or splicing (range: 27–237; median: 127).

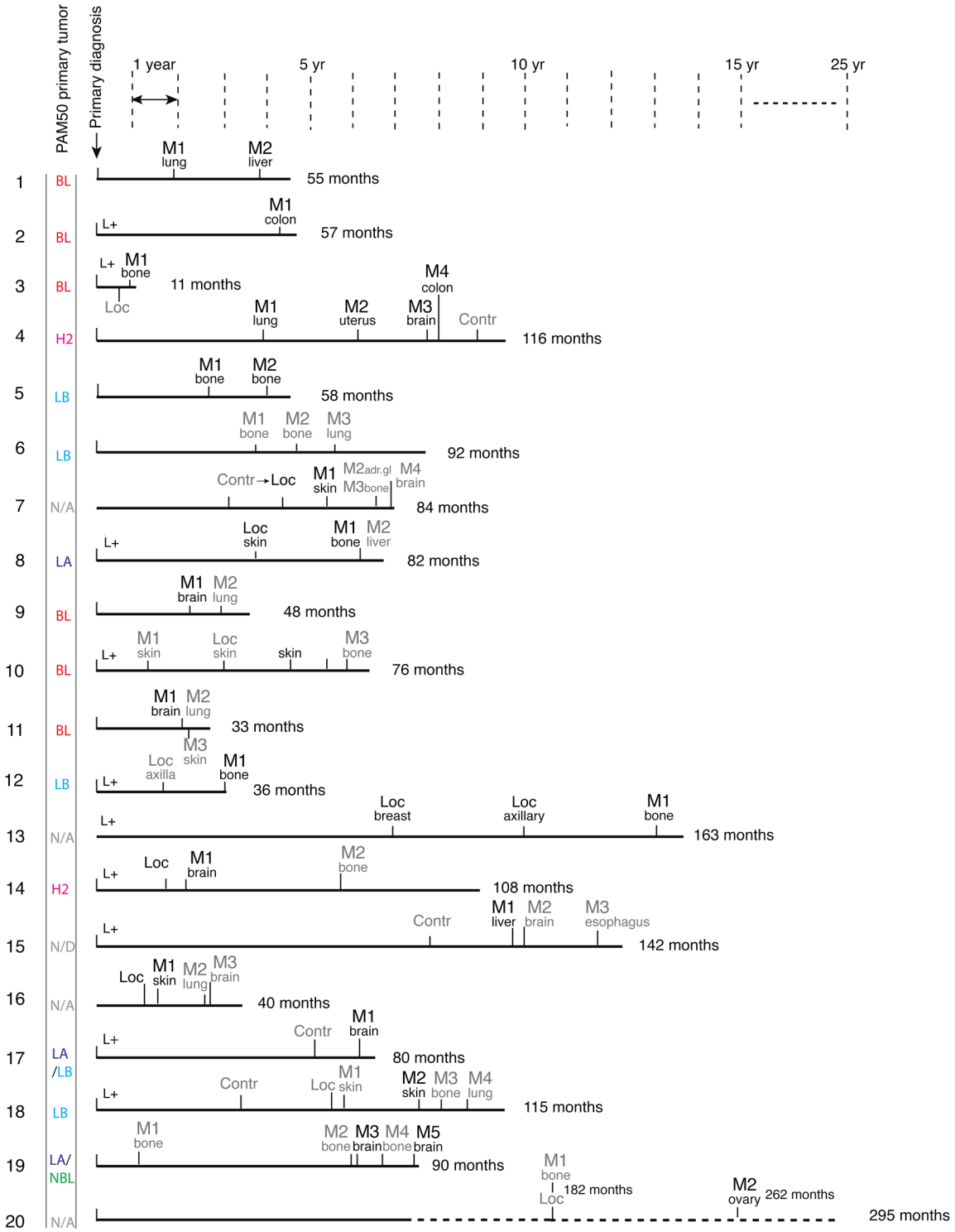


Figure 2. Clinical characteristics and metastatic timeline of the cohort. Timeline illustrating the time of relapse, sequenced cancers (black font), and unsequenced/failed or missing cancer samples (light gray font) for each patient. PAM50-intrinsic molecular subtype information is provided for all primary tumors. Patients with primary tumors that lacked a strong PAM50 classification for 1 subtype were classified into the 2 closest subtypes, such as luminal A/luminal B (patient 17) and luminal A/normal breast–like (patient 19). Each molecular subtype is represented by its own specific color. adr.gl, adrenal gland; L+, positive ALN; Loc, local relapse; Contr, contralateral event; BL, basal-like; LA, luminal A; LB, luminal B; H2, HER2-enriched; NBL, normal breast–like; M1, metastasis 1; M2, metastasis 2; M3, metastasis 3; M4, metastasis 4; N/A, not available; N/D, not determined.

Table 1. Clinical characteristics of the cohort

	Median	Range	
		Minimum	Maximum
Age at primary tumor diagnosis	52.5 yr	34 yr	80 yr
Time to first distant relapse after primary diagnosis	33.5 mo	9 mo	15 yr
Time from first distant relapse to death	24.7 mo	19 d	8.5 yr
Number of relapses per patient	3	1	6
Number of relapses sequenced per patient	2	1	4

Cohort statistics regarding the patient's age at diagnosis, the time to relapse, and the average number of relapses per patient are provided in Table 1. We determined the intrinsic molecular subtypes (PAM50) for all the sequenced samples using the available gene expression data (see Methods). All clinicopathological values including estrogen receptor/progesterone receptor/human epidermal growth factor receptor 2 (ER/PR/HER2) status (reanalyzed by NFM), as well as the intrinsic molecular subtype of primary tumors are provided in Supplemental Table 3. A schematic of the timeline of relapses and relapse tumor characteristics are provided in Supplemental Figure 2 and Supplemental Table 4, respectively. The neoadjuvant, adjuvant, as well as palliative metastatic therapies provided in our cohort are summarized in Supplemental Table 5.

Breast cancer spreads in either a linear or parallel manner. We used the Dollo parsimony method (24) to investigate the evolutionary history of cancer cells across different sites from the same individual. To assess the statistical support of inferred evolutionary relationships in the reconstructed trees, we used nonparametric bootstrapping (25). To infer the progression patterns in metastatic breast cancer, we used the separating property in the phylogenetic trees. To discriminate the patients in whom distant metastatic seeding was driven mainly by the primary tumor and those in whom it was seeded by an earlier distant metastasis, we used the definition of parallel and linear progression as stated previously (26). We validated the phylogenetic results using Lineage Inference for Cancer Heterogeneity and Evolution (LICHeE) (Supplemental Figure 9). In order to investigate the subclonal composition and polyclonal seeding in our cohort, we used a Bayesian clustering method called PyClone (27) (see Methods for details on all evolutionary analyses). To facilitate the comparison of the subclonal relationship vis-à-vis the phylogenetic relationship among samples within each patient, the subclonal information was embedded in the phylogenetic tree (see Supplemental Figure 3 for an illustration of the overall analysis pipeline). Since some patients had multiple blocks from the primary tumor sequenced, we performed an extensive subset analysis to validate the robustness of the phylogenetic inference for progression model analysis, in which different subsets of primary blocks were taken into account (see Supplemental Methods for details and Supplemental Tables 6 and 7 for results).

Five patients (patients 1, 4, 5, 8, and 19) fulfilled the inclusion criteria for the progression model analysis, that is, availability of sequencing data from the primary cancer and more than 1 distant metastasis from the same patient. Four of the five patients (patients 1, 5, 8, and 19) followed a linear progression model of successive metastasis-to-metastasis spreading of tumor cells. Phylogenetic trees, along with the subclonal composition and related information on tumors such as the site of metastasis, time of relapse, and primary cancer characteristics for all 4 patients are illustrated in Figure 3, A–D. The high bootstrap values for the most recent common ancestor (MRCA) of all metastasis pairs in the corresponding phylogenetic trees show strong statistical support for our results (Table 2). To validate the robustness of the results in the presence of only single primary samples, we performed a subset analysis for patient 5 by taking only single primary samples and found that the results did not change (Supplemental Table 6). To investigate sample level similarities, we compared shared and private mutations between pairs of samples in these patients and discovered that the metastasis pairs had greater similarity (a relatively higher percentage of shared mutations and a lower percentage of specific mutations) to each other than they did to the corresponding primary cancer, which validates the observed trees (Supplemental Figure 5). Additionally, subclonal analysis revealed that, in all 4 patients, 1 or more subclones were shared specifically between different metastases, but not with their corresponding primary cancers. Interestingly, the shared subclones harbored predicted deleterious mutations in previously known driver genes, such as a *MED12* missense mutation (p.Q572K) in patient 1, a *STAT3* missense mutation (p.I568M) in patient 8, and a *PALB2* missense mutation (p.G651V) in patient 19.

Unlike the 4 patients discussed above, patient 4, for whom we sequenced 6 different regions of the primary cancer and 3 distant metastases (uterus, brain, and colon), followed a parallel progression model (26). The inference is supported by high bootstrap values for the MRCA of the primary tumor and each metastasis (100%, 100%, and 79% bootstrap support for the MRCA of the primary samples with colon, brain, and uterus, respectively) (Figure 4). Moreover, by evaluating the evolutionary proximity (i.e., shortest distance in terms of the number of edges) of each metastatic site to each primary block across 1,000 bootstrap trees, we found that brain and colon metastases had the closest evolutionary proximity to Primary1 block (probability of 0.604 and 0.71, respectively), while the uterine metastasis had the closest evolutionary proximity to Primary5 block (probability of 0.838) (Table 3). Since 6 paraffin blocks from patient 4 were analyzed, we per-

Table 2. Probability of linear progression from an earlier metastasis to a subsequent metastasis

Patient no.	From	To	Probability
1	Lung (19 mo)	Liver (49 mo)	1
5	Bone.R1 (31 mo)	Bone.R2 (49 mo)	1
8	Skin (49 mo)	Bone (75 mo)	1
19	Brain.R3 (73 mo)	Brain.R5 (88 mo)	1

The probability was computed across 1,000 bootstrap trees according to the blocking property in the phylogenetic trees.

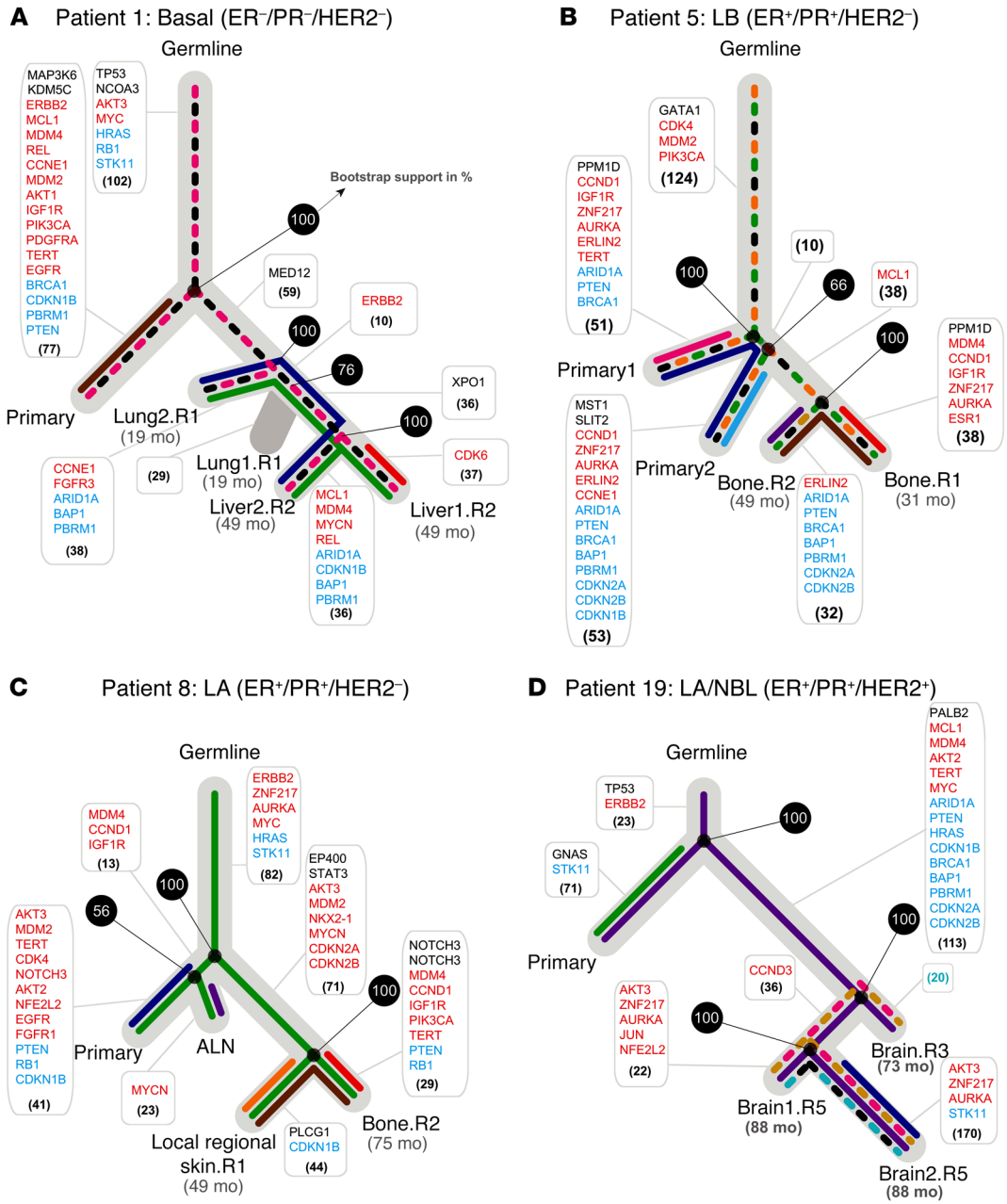


Figure 3. Phylogenetic and subclonal analysis supports a linear progression model in 4 patients. The ER, PR, and HER2 status, when available for primary tumors, is indicated above the tree. Subclonal information is represented by lines in the phylogenetic tree. The solid line represents a single subclone, while the multicolored, dotted line represents multiple subclones. Branch lengths are proportional to the number of substitutions, with the total number for each branch indicated in parentheses. Branches are annotated with known breast cancer gene alterations including somatic mutations (black), amplifications (red), and deletions (blue). Bootstrap support values, computed across 1,000 bootstrapped trees, are shown in the black circles. When subclonal analysis could not be performed for a sample due to low copy number resolution, the respective sample is colored dark gray. The time frames for metastatic relapses after the primary cancer diagnosis are indicated in months. Information on the inferred subclones and their cellular prevalence is provided in the cluster table and density plot, respectively, in Supplemental Figure 6, A, E, G, and Q for patients 1, 5, 8, and 19, respectively. (**A**) Patient 1 had 1 primary tumor, 2 regions of lung relapse (Lung1.R1 and Lung2.R1), and 2 regions of liver relapse (Liver1.R2 and Liver1.R2). (**B**) Patient 5 had 2 regions from the primary tumor (Primary1 and Primary2) and 2 bone relapses (Bone.R1 and Bone.R2). (**C**) Patient 8 had a primary tumor, an ALN metastasis, a skin local regional relapse (Local regional Skin.R1), and a bone metastasis (Bone.R2). (**D**) Patient 19 had 1 primary tumor (Primary), 1 brain relapse 3 (Brain.R3), and 2 blocks from brain relapse 5 (Brain1.R5 and Brain2.R5).

formed an exhaustive subset analysis by taking all possible combinations of primary tumor blocks and then inferring the progression model. The overall results supported parallel progression in patient 4, irrespective of the number and combination of primary samples selected, indicating that the primary tumor had seeded at least 2 of the 3 metastases (Supplemental Table 7). The subclonal analysis results also reflected the parallel progression results, in which, for instance, the red subclone (consisting of 79 mutations including mutations in putative driver genes like *BRCA2*, *DDR2*, and *ROS1*) was shared exclusively between Primary5 and uterus

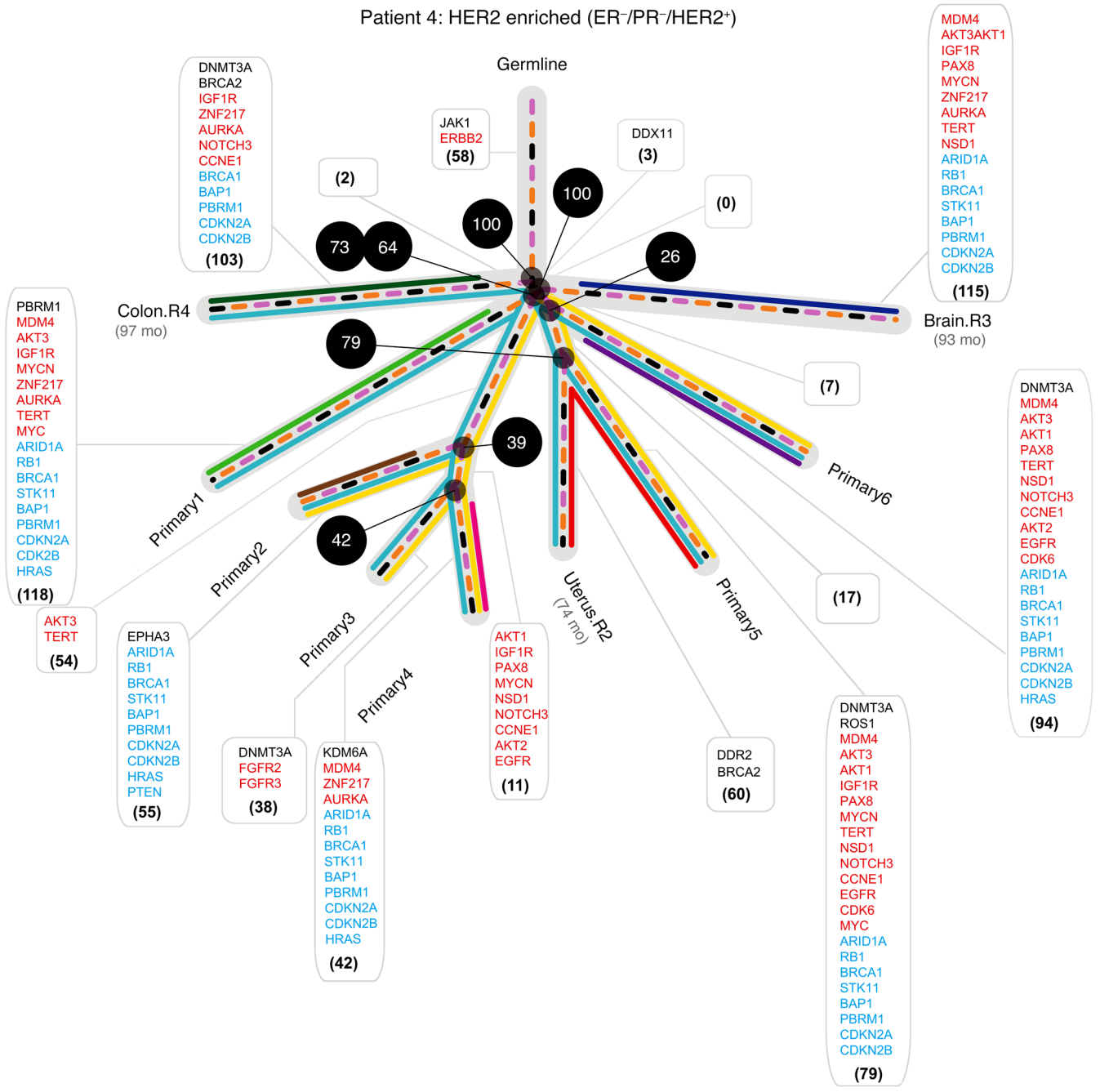


Figure 4. Phylogenetic and subclonal analysis supports a parallel progression model in 1 patient. For patient 4, six different regions of primary tumors (Primary1 to Primary6) and 3 metastatic cancers from uterus relapse 2 (Uterus.R2), brain relapse 3 (Brain.R3), and colon relapse 4 (Colon.R4) were sequenced. Information about the inferred subclones and their cellular prevalence is provided in the cluster table and density plot, respectively, in Supplemental Figure 6D.

(Figure 4 and the density plot in Supplemental Figure 6D). These results suggest that different distant organ metastases were seeded from different regions of the primary tumor rather than from each other. Furthermore, no subclones were shared exclusively, either among all or between any pair of distant metastases, indicating that cancer cells from the primary cancer disseminated and colonized multiple metastatic sites in parallel and then independently accumulated new genetic alterations. Interestingly, the colon metastasis acquired a likely deleterious *BRCA2* mutation (p.K3263Q; variant allele frequency [VAF]: 16 of 58; combined annotation-dependent depletion [CADD] Phred score: 20.2), while the uterus metastasis acquired a different, probably benign, *BRCA2* mutation (p.P721T; VAF: 4 of 52; CADD Phred score: 5.5). The difference in clonality and deleteriousness between the 2 mutations, in addition to the fact that the brain metastasis did not acquire any *BRCA2* mutations, suggests that the 3 metastases had a divergent, rather than convergent, evolution at the driver events level. It is important to mention here that the bulk-sequencing data did not have the required level of resolution to infer complex self-seeding events. That could be obtained more accurately by

Table 3. Probability of different regions of the primary cancer seeding different distant organ metastases in patient 4, computed across 1,000 bootstrap trees

From primary block	To brain	To colon	To uterus
Primary6	0.113	0.071	0.069
Primary4	0.044	0.033	0.031
Primary2	0.056	0.053	0.027
Primary3	0.044	0.04	0.002
Primary1	**0.604**	**0.71**	0.032
Primary5	0.14	0.092	**0.838**

The probability of different regions of the primary cancer seeding different distant organ metastases, computed across 1,000 bootstrap trees, was determined by evaluating the evolutionary proximity of each primary cancer to each metastasis in terms of the number of edges between them in the phylogenetic tree. For each metastasis, the highest probability of being seeded from a particular primary block is indicated in bold font.

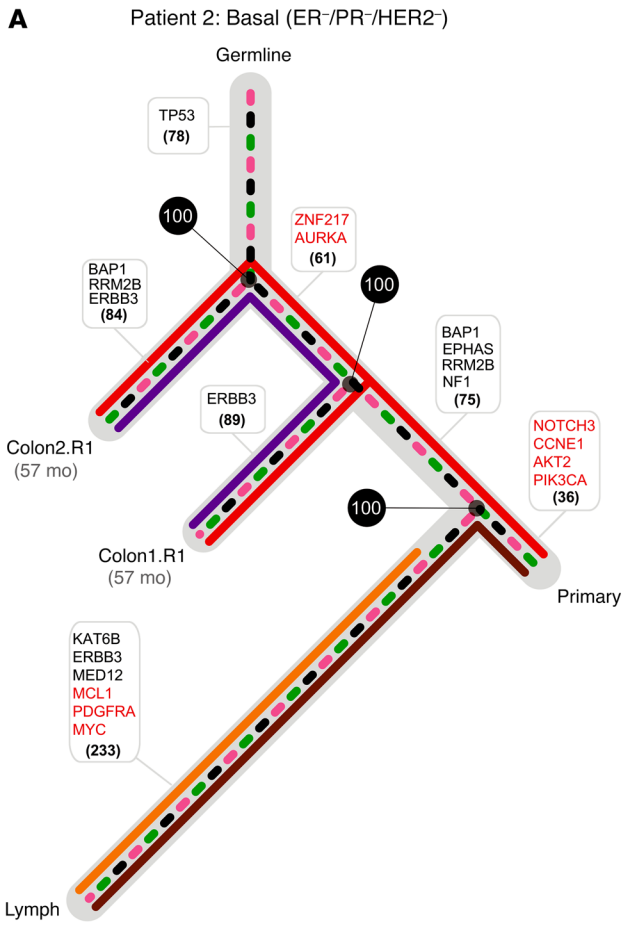
using specialized methods, coupled with single-cell sequencing data. Phylogenetic trees, along with the subclonal density plots for all the patients, are shown in Supplemental Figure 6.

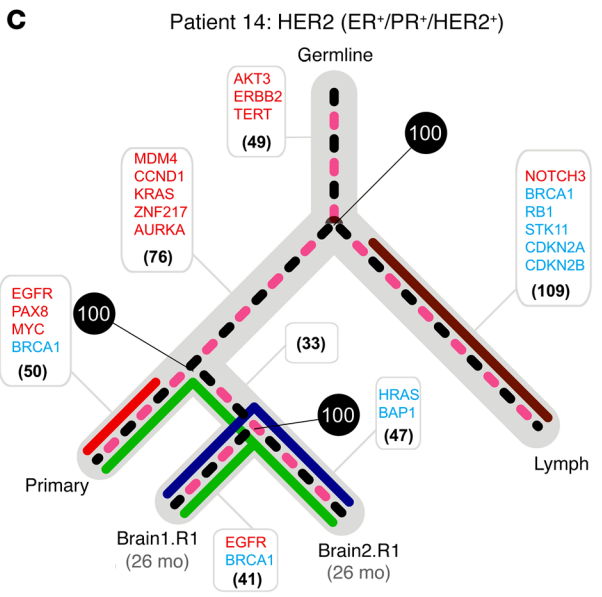
Distant metastases are seeded without involvement of the synchronous ALN metastasis. In order to investigate whether metastatic lymph nodes can secondarily seed distant metastases, we used the separating property in the phylogenetic trees to analyze 8 patients (patients 2, 3, 8, 10, 14, 15, 17, and 18) with primary cancer, ipsilateral ALN metastases, and distant metastasis. Patient 12 was excluded from this analysis because of unavailable distant metastasis sequencing data. Our analysis revealed very low support for ipsilateral ALN–based seeding to distant organ metastases (Figure 5). The highest bootstrap support value for an MRCA of the ALN and a distant metastasis across 8 phylogenetic trees was 23% (patient 8), while the other values were zero or almost zero (Table 4). In 3 patients (patients 2, 10, and 15), we sequenced the only positive ALN (2 blocks sequenced in the case of patient 10), excluding the possibility of a distant metastasis seeded by an unsequenced metastatic lymph node. Subclonal analysis revealed that, except for patient 3, no subclones were shared exclusively among ALN metastases and any distant metastases, thus supporting the phylogenetic results. Even in patient 3, a distant bone metastasis shared a subclone with the primary tumor, making it equally likely that either the primary cancer or the ALN was responsible for the distant organ metastasis. It is important to note that we cannot rule out the possibility of metastatic seeding from a dormant subclone in the ALN metastasis to a distant metastasis, when, after seeding, such a subclone becomes active only in the distant metastasis. Moreover, there is also a possibility that we missed a mutation or a subclone because it was either in an unsequenced part of the ALN or was too rare to be detected by the sequencing coverage of the study. Ideally, one would sequence the entire axillary region, but clinically, it is unethical to conduct such a study. The phylogenetic trees, along with the subclonal composition for patients 2 and 10 (patients with all positive lymph nodes sequenced) and patient 14 (1 of 14 positive lymph nodes sequenced), are shown in Figure 5, A–C, respectively, while similar plots for the other 5 patients are shown in Supplemental Figure 6, C, G, M, O, and P.

Distant metastases are seeded in either a monoclonal or polyclonal manner from primary breast cancer. We studied subclonal propagation from the primary cancer to distant metastases in 15 patients, after excluding 5 patients (patients 6, 7, 12, 13, and 16) for whom the sequencing data from either the primary cancer or the distant metastasis was missing. By identifying subclones shared among different samples from a single patient, we observed both monoclonal (1 subclone) and polyclonal (more than 1 subclone) seeding from primary breast cancers to distant metastases. Four of fifteen patients (patients 8, 15, 17, and 19) had monoclonal seeding (27%), and eleven of fifteen patients (patients 1, 2, 3, 4, 5, 9, 10, 11, 14, 18, and 20) had polyclonal seeding (73%) (Supplemental Table 8). The number of sequenced primary samples did not correlate with the seeding patterns ($P = 0.56$, Fisher's exact test). Interestingly, all 4 of the patients (patients 8, 15, 17, and 19) with monoclonal seeding had primary cancers of a luminal subtype (based on PAM50 and IHC analysis) (Supplemental Table 3). However, we also observed polyclonal seeding in 3 patients (patients 5, 18, and 20) with a luminal subtype, while 8 of 11 patients with polyclonal seeding had nonluminal subtypes (6 basal and 2 HER2 enriched). All of the patients with metastasis-to-metastasis spreading (patients 1, 5, 8, and 19) had polyclonal seeding between successive metastases.

Substantial interindividual genomic diversity among primary tumors and metastases. In order to demonstrate the extent of genomic alterations during breast cancer progression, we categorized the genetic alterations into site-specific categories, i.e., truncal (shared among all samples analyzed, i.e., primary tumors and metastases in a patient), branch (shared by at least 1 primary tumor and 1 metastasis), primary, LR, ALN, and metastasis; the last 4 categories include alterations specific to those samples. We observed large interindividual differences in the number of mutations shared among primary tumors and metastases, indicating varying points of divergence from the primary tumor to distant metastases (Figure 6A). On average, 55% of the primary mutations were retained in the distant metastatic lesions, with considerable disparity among individual patients, ranging from 9% to 88%, and an interquartile range (IQR) of 36% (Figure 6B). To test whether different types of treatment had affected the mutational load, we compared the fraction of mutations privately detected in metastases between treated and untreated patients and found no significant difference for any type of treatment. Copy number variation (CNV) analysis revealed the most altered genomic regions during tumor progression, which included chromosome arms 1q, 8q, and 20q amplifications and 8p and 17p deletions (Figure 7A).

We used a set of putative driver genes in breast cancer compiled by Yates et al. (16) to determine the timing and frequency of driver alterations during breast cancer progression (Figure 7B). Driver alterations such as *TP53*, *PIK3CA*, *PTEN*, and *GATA3* mutations and *MYC* and *ERBB2* amplifications were predominantly early events. However, all these genes, except *GATA3*, gained alterations privately in metastasis in at least 1 patient, indicating secondary late driver events (Figure 7B). For example, the brain metastasis in patient 15 gained a *TP53* mutation (p.R116Q), and another brain metastasis in patient 17 gained a *PIK3CA* mutation (p.H1047R) and 2 frameshift insertions in *PTEN* in different alleles, suggesting biallelic inactivation. While most other putative driver alterations varied in their timing of occurrence, a few were





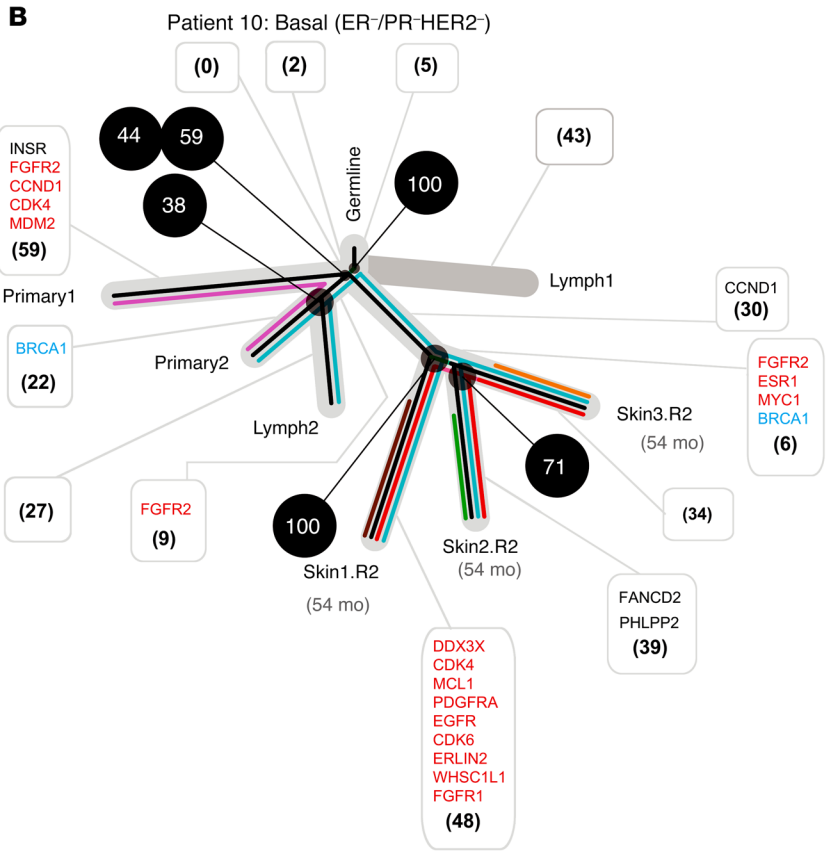


Figure 5. Passive role of ALN metastasis in seeding distant metastases. (**A**) The only positive synchronous ALN metastasis, as well as 1 primary cancer and 2 regions of colon relapse 1 (Colon1.R1 and Colon2.R1) were sequenced for patient 2. (**B**) For patient 10, two regions of primary cancer (Primary1 and Primary2), 2 regions from the only positive lymph node (Lymph1 and Lymph2), and 3 regions of skin relapse 2 (Skin1.R2, Skin2.R2, and Skin3.R2) were sequenced. (**C**) For patient 14, one region of primary cancer, 1 region from 1 of 12 positive ALNs (Lymph), and 2 regions of brain relapse 1 (Brain1.R1 and Brain2.R1) were sequenced. Information about inferred subclones and their cellular prevalence is provided in the cluster table and density plot, respectively, in Supplemental Figure 6, B, I, and L, for patients 2, 10, and 14, respectively. Colon1.R1, colon block 1 relapse 1; Colon2.R1, colon block 2 relapse 1; Lymph1, ALN block 1; Lymph2, ALN block 2; Skin1.R2, skin block 1 relapse 2; Skin2.R2, skin block 2 relapse 2; Skin3.R2, skin block 3 relapse 2; Brain1.R1, brain block 1 relapse 1; and Brain2.R1, brain block 2 relapse 1.

Table 4. Probability of an ALN metastasis seeding distant organ metastases in all 8 patients with 1 or more metastasis-positive ALNs

Patient no.	From	To	Probability	No. of nodes sequenced/ total no. of metastasis-positive nodes
2	Lymph	Colon	0	1/1
3	Lymph	Bone	0.005	1/2
8	Lymph	Bone	0.233	1/3
8	Lymph	Skin	0.233	1/3
10	Lymph1	Skin	0.019	1[A]/1
10	Lymph2	Skin	0.057	1[A]/1
14	Lymph	Brain	0	1/12
15	Lymph	Liver	0.029	1/1
17	Lymph	Brain	0.002	1/3
18	Lymph1	Skin	0	2/17
18	Lymph2	Skin	0	2/17

The probability was computed across 1,000 bootstrap trees according to the blocking property in the phylogenetic trees. [A]Two blocks sequenced per lymph node.

late events that occurred privately in a distant metastasis. These included *BRCA2, ESR1*, and *STAT3* mutations, as well as *AKT2* and *EGFR* amplifications. In total, 963 genes were found to be mutated privately in the metastatic category. Pathway enrichment analysis of these genes identified laminin interactions ($P = 0.0001$, q value = 0.098), nonintegrin membrane–extracellular matrix (ECM) interactions ($P = 0.0011$, q value = 0.372), and degradation of the ECM ($P = 0.0036$, q value = 0.378) as the top significantly mutated pathways (see Supplemental Table 9 for the lists of mutated genes and pathways for all categories analyzed). Taken together, our results suggest that distant metastatic lesions show interindividual disparity in genomic divergence from the primary tumors, with the common occurrence of putative driver alterations during later stages of breast cancer progression.

Activity of mutational processes evolve during breast cancer progression. The repertoire of somatic mutations in the cancer genome carries the imprint of underlying operative mutational processes (28). At least 4 signatures (labeled S1, S2, S3, and S4) were found to be operative in our cohort (Figure 8A), and 3 of them (S1, S2, and S4) were mapped to known mutational processes (Supplemental Figure 7, A–C). The S1 signature is characterized by an elevated number of C>T substitutions in the NpCpG context resulting from spontaneous deamination of 5-methyl-cytosines and was associated with the patient's age at diagnosis (28). The S2 signature has an excess of C>T and C>G mutations in the TpCpN context attributed to the activity of the apolipoprotein B mRNA–editing enzyme, catalytic polypeptide–like (APOBEC) family of cytidine deaminases (29). The S3 signature is a generic signature with unknown etiology that is characterized by slightly elevated C>A, C>T and T>G mutations. Finally, S4 is also a generic signature attributed to deficient homologous recombination (HR) in double-strand break repair that is partly explained by *BRCA1* and *BRCA2* germline mutations (28).

We evaluated signature contributions across site-specific categories in patients to assess their signature contribution separately (Figure 8B). All signatures showed different contributions in different categories ($P < 0.01$, Kruskal-Wallis). We observed increased activity of mutational processes of the APOBEC signature S2 ($P < 0.01$, Mann-Whitney U test with FDR correction), the unknown etiology signature S3 ($P < 0.05$), and the HR signature S4 ($P < 0.05$) in the metastasis-specific category relative to the primary-specific category (Figure 8C). The aging signature S1 did not indicate a significant change between primary-specific and metastasis-specific categories. However, its contribution was depleted in truncal mutations, suggesting a passive role for the S1 signature during clonal expansion. Our evaluation on an individual basis revealed that 14 of 15 patients had a significantly increased contribution of at least 1 of the 3 signatures (S2, S3, and S4) in metastatic lesions relative to their corresponding primary lesions ($P < 0.05$, Fisher's exact test with FDR correction). Interestingly, patient 4, who acquired 2 independent BRCA2 mutations in 2 metastatic lesions, showed a significant increase ($P < 0.05$ or $P = 0.0006$, Fisher's exact test with FDR correction) in the S4 contribution. We found no statistically significant association between the increased activity of a certain mutational process and the molecular subtype or treatment ($P > 0.1$, Fisher's exact test).

Discussion

From a clinical point of view, one of our most important observations was that ipsilateral synchronous ALN metastases were genetically diverse compared with distant organ metastases across all patients. This indicates that ipsilateral ALN metastasis is not crucial for seeding distant metastases; rather, its prognostic value merely reflects the stage migration effect and the acquired capability of cancer cells to survive and proliferate in other organs (30, 31). The role of ALN metastasis has been explained using a speedometer analogy (32). The authors assert that, just as removing the speedometer of a car does not reduce its speed, the removal of positive ALNs does not affect the rate of metastasis. In other words, ALN status is very useful for predicting the tumorigenic capability of the primary tumor but does not drive metastasis per se. Hence, dissection of positive ALNs will not reverse this capability, since spreading to distant sites appears to occur via a hematogenous, rather than a lymphatic, channel. Moreover, gene expression profiles of the primary tumor can predict metastasis location and survival, independent of ALN status, which further supports the idea that ALN metastasis is not the major factor in determining distant metastatic progression (33–35). Finally, the genetic characterization of primary cancers provides important prognostic and predictive information that increasingly supersedes the informative capacity of lymph node status (33, 36). Indeed, several large, randomized clinical trials have demonstrated that ALN resection for limited metastasis conferred minimal or no survival benefit for breast cancer patients (37–40). A few other reports also stated that positive lymph nodes do not metastasize

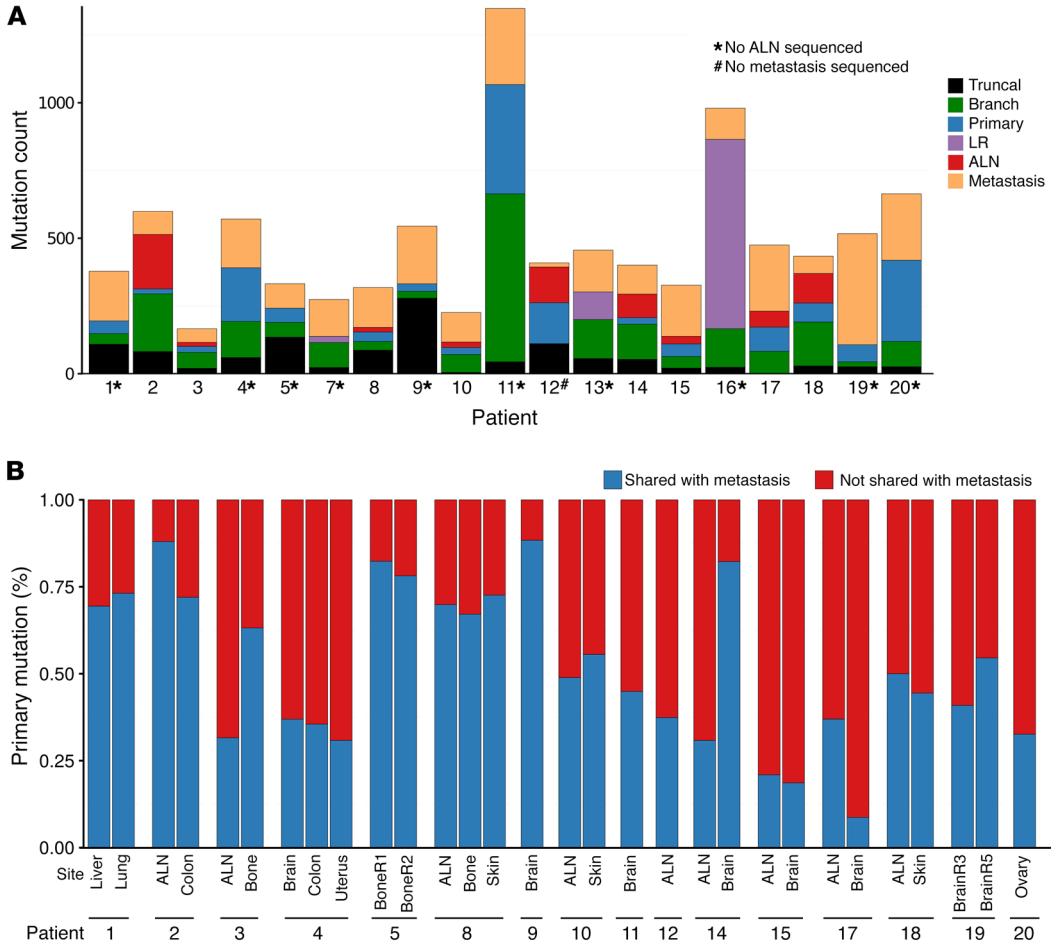


Figure 6. Genomic diversity among primary and metastatic tumors. (**A**) Summary of the number of mutations and indels across the patients. Genomic alterations are categorized as truncal (shared among all samples), branch (shared among 2 or more samples), primary, LR, ALN, and metastasis, in which the last 4 categories contain alterations specific to those samples. Data on patient 6 were excluded, since they involved only 1 primary tumor sample. (**B**) Distribution of the primary tumor's mutation retention percentage in metastatic lesions across all patients. The blue color indicates the percentage of primary tumor mutations shared with distant metastases, while the red color represents the percentage of mutations present only in the primary tumor.

(41, 42), which is in concordance with our observations. A recent study by Brastianos et al. (43) that involved 86 brain metastases, 21 of which were breast cancers, also showed that regional lymph nodes and brain metastasis were genetically distinct and, hence, that the former was not a reliable surrogate for the oncogenic alterations found in the latter.

We demonstrated that primary breast cancer either directly seeds metastases at various distant sites in a parallel manner, or, alternatively, it initially seeds 1 distant metastasis that in turn seeds successive metastases in a linear manner. However, the specific characteristics related to primary breast cancer's favoring one progression model over the other are still unclear. Moreover, multiple other factors including therapy resistance can have an impact on the time frame and seeding pattern of a metastasis. Studies involving larger patient cohorts are needed to accurately quantify the abundance of each progression model and its correlation with molecular subtypes and the treatment provided and would serve to improve the clinical management of metastatic disease. For instance, in a parallel progression model, in which distant metastases evolve independently, biopsies from a single metastatic site for clinical decision making might not be sufficient.

The specific design of our cohort, composed of spatially and temporally separated tumor specimens, allowed us to investigate the subclonal relationship of primary tumors with multiple systemic metastases in unprecedented detail. For instance, in patient 4, we observed that distinct, spatially separated subclones in the primary tumor were independently seeding different distant organ metastases. This suggests that multiple subclones can acquire unique seeding capabilities that enable them to independently colonize different anatomical sites in a "seed and soil" manner. The extent of subclonal heterogeneity in primary tumors corresponding to different molecular subtypes might influence whether the distant metastases are seeded in a monoclonal or polyclonal manner. For instance, all distant metastases of basal and HER2-enriched breast cancer patients were polyclonally seeded from the primary tumors, where multiple subclones acquired metastatic capabilities. On the other hand, distant metastases seeded by a single subclone from the primary tumors were always of the luminal subtype, emphasizing the importance of having different management strategies for different breast cancer subtypes. We also observed that, irrespective of any molecular subtype, subsequent seeding of cancer cells from 1 metastatic organ to another

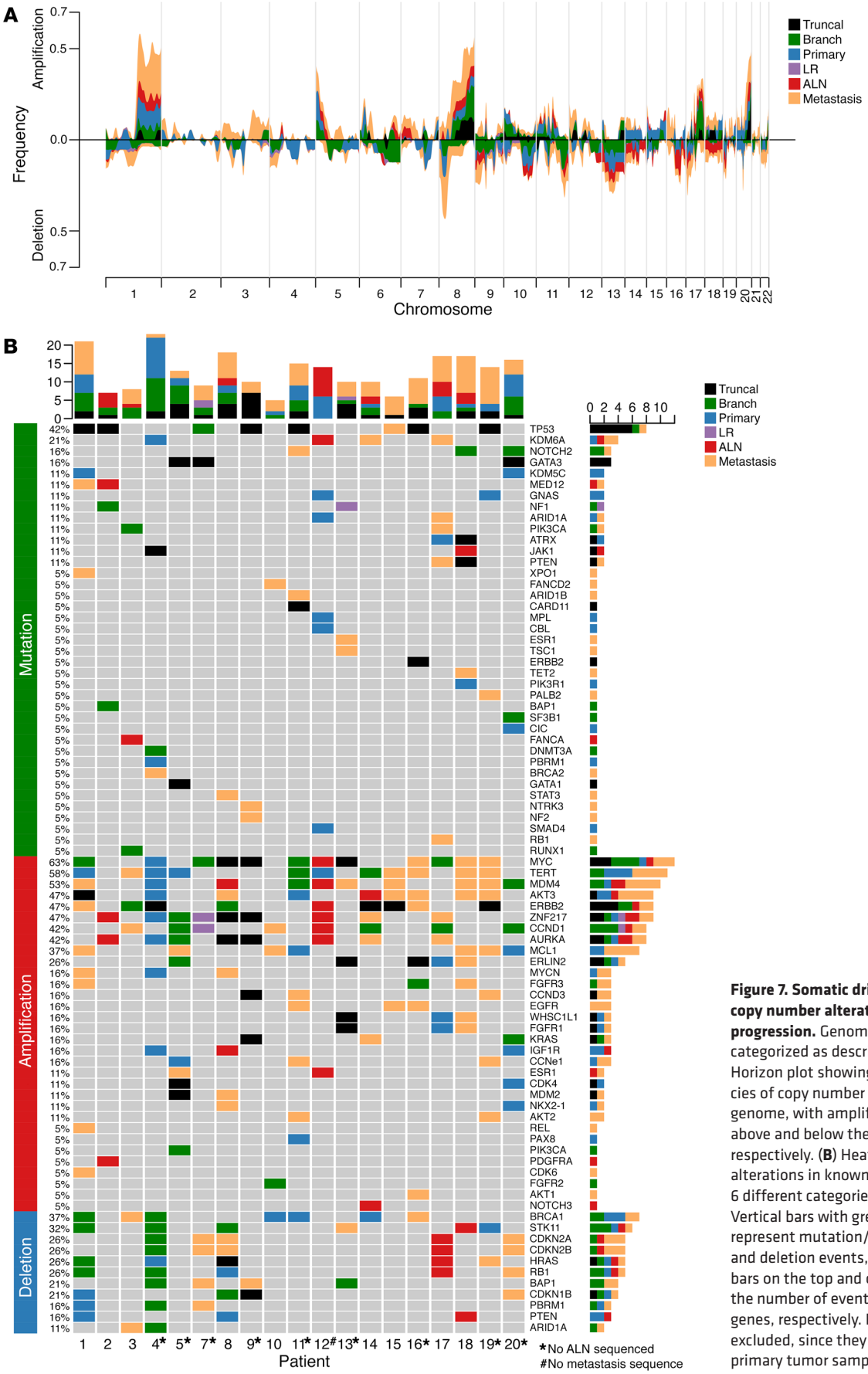


Figure 7. Somatic driver mutations and copy number alterations during cancer progression. Genomic alterations are categorized as described in Figure 6. (**A**) Horizon plot showing smoothed frequencies of copy number alterations across the genome, with amplifications and deletions above and below the horizontal axis, respectively. (**B**) Heatmap showing genomic alterations in known breast cancer genes in 6 different categories across the patients. Vertical bars with green, red, and blue parts represent mutation/indel, amplification, and deletion events, respectively. Stacked bars on the top and on the right represent the number of events across patients and genes, respectively. Data on patient 6 were excluded, since they consisted of only 1 primary tumor sample.

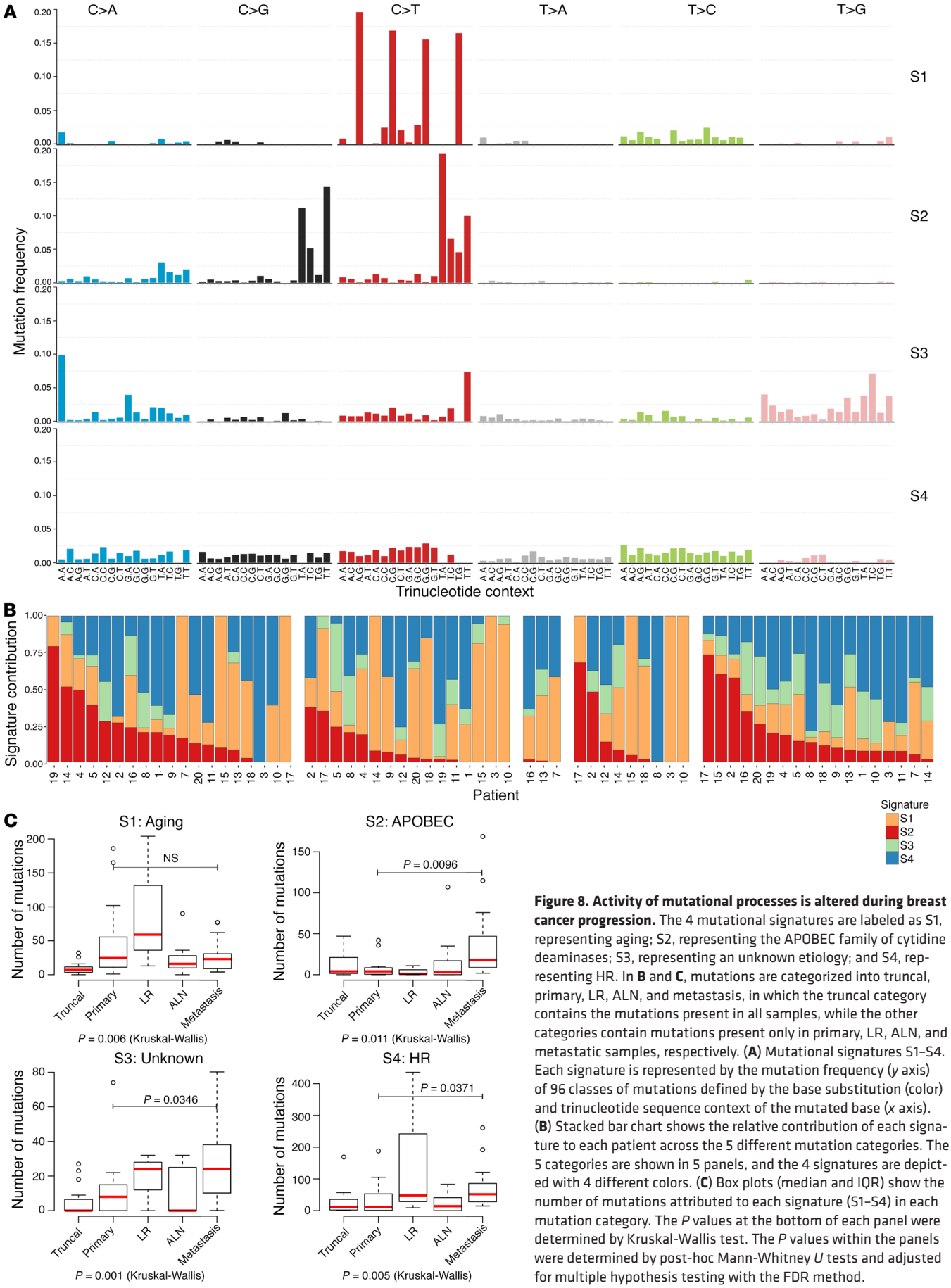


Figure 8. Activity of mutational processes is altered during breast cancer progression. The 4 mutational signatures are labeled as S1, representing aging; S2, representing the APOBEC family of cytidine deaminases; S3, representing an unknown etiology; and S4, representing HR. In **B** and **C**, mutations are categorized into truncal, primary, LR, ALN, and metastasis, in which the truncal category contains the mutations present in all samples, while the other categories contain mutations present only in primary, LR, ALN, and metastatic samples, respectively. (**A**) Mutational signatures S1–S4. Each signature is represented by the mutation frequency (*y* axis) of 96 classes of mutations defined by the base substitution (color) and trinucleotide sequence context of the mutated base (*x* axis). (**B**) Stacked bar chart shows the relative contribution of each signature to each patient across the 5 different mutation categories. The 5 categories are shown in 5 panels, and the 4 signatures are depicted with 4 different colors. (**C**) Box plots (median and IQR) show the number of mutations attributed to each signature (S1–S4) in each mutation category. The *P* values at the bottom of each panel were determined by Kruskal-Wallis test. The *P* values within the panels were determined by post-hoc Mann-Whitney *U* tests and adjusted for multiple hypothesis testing with the FDR method.

always occurred in a polyclonal manner, suggesting a cooperation between different subclones in adapting to the new microenvironment during an advanced disease stage. Furthermore, it demonstrated that polyclonal seeding is a mechanism for preserving the level of heterogeneity when colonizing new distant sites, thereby maximizing the likelihood of the emergence of treatment-resistant clones. Polyclonal seeding of distant metastasis has previously been demonstrated using an experimental breast cancer mouse model (44). A matched primary metastases cohort provided us the opportunity to study the somatic, nonsynonymous mutations present exclusively in distant metastases. Pathway analysis of the genes harboring these mutations revealed that ECM degradation and laminin interaction pathways were highly enriched, suggesting their crucial role in successful colonization at distant sites. Using mouse models, Naba et al. (45) recently showed that ECM composition differs among cancer cells of differing metastatic potential. Thus, genes involved in ECM and laminin interactions should be investigated in detail for their therapeutic and prognostic value in metastatic breast cancer.

Optimal management of metastatic disease requires detailed knowledge of the biological characteristics including therapy-predictive factors and alterations in druggable targets, such as *HER2* and *ESR1*. For instance, patient 1 acquired *HER2* amplification, and patient 5 acquired *ESR1* amplification, specifically in their metastatic lesions but not in their primary cancer. These amplifications were also supported by the immunohistochemical staining of HER2 and ESR1 on the respective metastatic tumor sections. Furthermore, in an ER-positive patient (patient 13), we detected an *ESR1* mutation (p.D538G) in bone and nonsynchronous ALN metastases but not in the earlier LR. These data highlight the importance of characterizing the metastatic lesions for potential clinical interventions. However, metastatic biopsies are not always feasible for this purpose, therefore, the development of liquid biopsy techniques is warranted (46). Additional studies are needed to evaluate whether liquid biopsies are capable of representing the genetic heterogeneity of metastatic lesions in its entirety.

The 4 mutational signatures identified in our cohort were previously reported in primary breast cancers (47), however, their relative activity during breast cancer progression remained uncertain. The relative contribution of the aging signature decreased during cancer progression, giving rise to other mutational signatures. The signature contribution related to homologous recombination (HR) deficiency and a process of unknown etiology changed in both directions in a patient-dependent manner, with a majority of the patients showing increased activity. However, the APOBEC-associated signature consistently showed increased or similar activity across all metastases, indicating its significant role in the progression and selective advantage of its induced mutations. Quantifying the activity of different mutational processes in metastatic cancers and understanding their underlying mechanisms may provide markers for alternative therapeutic strategies such as the use of PARP inhibitors in patients with increased activity of the HR deficiency signature. Increased APOBEC activity during progression has been described in lung cancer (48) and was assumed to influence tumor heterogeneity and a "mutator" phenotype (49). Different strategies have been suggested to limit the development of resistance-causing mutations. For instance, preventing APOBEC activation by accelerating the mutation rate beyond the limit of lethal mutagenesis may be a valuable approach (49, 50). Whether the APOBEC activity is indeed of prognostic value in breast cancer needs to be addressed in a larger patient cohort.

In conclusion, our study demonstrates that ipsilateral ALN metastases, irrespective of their strong prognostic value, which has been frequently demonstrated in multivariate testing, had a consistently passive role in seeding and spreading to distant sites. Second, our study shows a high level of inter-patient heterogeneity in terms of metastasis seeding and spreading modes. Finally, our observations reveal changing activity of the mutagenesis mechanism during cancer progression, accompanied by late induction of driver mutations specific to the metastatic lesions, suggesting an altered tumor biology compared with that of the primary breast cancer.

Altogether, these observations provide a comprehensive overview of breast cancer evolution in existing clinical scenarios that emphasizes the importance of genomic characterization of different metastatic lesions prior to clinical decision making, as these lesions may originate in parallel without notable phenotypic convergence, or successively with an ongoing accumulation of driver events. However, the characterization of metastatic ALNs is not of the same importance, given their less significant role in metastatic seeding. The high prognostic value of ALN status in the clinical management of metastatic breast cancer is generally known and has also been frequently demonstrated in multivariate analyses. Furthermore, both the sequencing of bulk tumors and the mathematical models used to assess such data are still far from perfect. Therefore, we anticipate that follow-up studies will validate our results. A recent large-scale, whole-genome sequencing study of 560 primary breast cancers confirmed the rarity of recurrent fusion genes and noncoding driver mutations (47). This recent study further asserted that the majority of genes containing driver mutations are now known. Thus, sequencing studies within larger patient cohorts as well as ultra-deep sequencing will help to confirm our results regarding progression patterns and the passive role of ALN metastasis in seeding distant organ metastases.

Methods

Patients' material. We assembled a cohort of 20 female patients for this study. For each of these patients, we collected paraffin-embedded material from the primary breast cancer, LR, ipsilateral ALN metastases, and distant metastases. The patients were identified through searches in the IT-support system using the patients' electronic medical records. A board-certified surgical pathologist at the Karolinska University Laboratory diagnosed the lesions as metastatic breast cancer. Further details on the patient selection criteria, sample acquisition procedures, and sequenced samples are provided in the supplemental material.

Tissue microarray and IHC staining. Two tissue microarrays with 43 tumor cores in each from 20 breast cancer patients (both primary and metastatic cancers) were prepared at the accredited clinical laboratory of the Department of Clinical Pathology of Karolinska University Hospital. The paraffin blocks were cut into 3-μm sections and stained for ER, PR, HER2, and Ki-67. A description of the antibodies used and assessments of the IHC-based subtypes are provided in Supplemental Methods.

RNA extraction and PAM50 molecular subtyping after subgroup-specific gene centering. RNA was extracted from two 10-μm sections per FFPE tumor block using an RNeasy FFPE Kit (QIAGEN) according

to the manufacturer's instructions. A SensationPlus FFPE Amplification Kit (Affymetrix) was used to amplify the RNA and profiled in the GeneChip Human Transcriptome Array 2.0 (Affymetrix). Probe intensities were extracted from CEL files, background corrected, normalized, and summarized for probe set expression using the robust multichip average (*rma*) function from the Oligo Package (version 1.30.0) in R (version 3.1.2) from Bioconductor. Further details on the PAM50 subtype classification, the samples analyzed, and the patients whose subtypes changed are provided in the Supplemental Methods.

DNA extraction and preprocessing of whole-exome sequencing reads. We isolated cancer DNA from thick serial sections of FFPE tissues using a QIAamp DNA FFPE Tissue Kit (QIAGEN). We used DNA from normal ALN FFPE tissues as germline controls. In all cases, we followed the manufacturer's recommended protocol. Genomic target capture was performed using the SureSelectXT2 Human All Exon V5 Kit (Agilent Technologies), and captured libraries were whole-exome sequenced on an Illumina HiSeq 2500 Instrument using 2 × 100-bp sequencing reads. Raw sequencing reads were quality and adapter trimmed with Trim Galore (version 0.3.7) with the following parameters: --quality 20 –stringency 2 –length 70 –clip_R1 10 –clip_R2 10. The trimmed reads were aligned to the reference human genome (hg19) using BWA-MEM (version 0.7.12) with default parameters. Aligned reads were sorted and marked for duplicates with Picard (version 1.113). Next, base quality recalibration and realignment around indels were performed using the Genome Analysis Toolkit (GATK version 2.7), which resulted in ready-to-use BAM files. The achieved coverage in target regions was on average 80× (70% targeted regions with >30× coverage) (Supplemental Figure 1 and Supplemental Table 1). All preprocessing and downstream analyses were performed within the Anduril framework for scientific data analysis (51). Sequencing data were deposited in the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG) (accession number EGAS00001002737).

Variant calling, filtering, and copy number alteration detection. We performed point mutation calling using a 2-step approach with MuTect v1.1.4 (52). In the first step, we used MuTect with the high-confidence mode to call somatic variants from individual cancer samples in a matched, cancer-normal setting. In this run, we used single nucleotide polymorphism database (dbSNP) variants, version 138, and Catalogue of Somatic Mutations in Cancer variants, build 68, as inputs to MuTect. In the second step, each detected variant was screened for allele counts in the other samples from the same patient using MuTect, with the option of forcing the output in selected intervals (coordinates of detected variants in the first step). Then, to account for potential artifacts induced by FFPE samples, we filtered C>T/G>A mutations that were private to 1 sample and had a VAF of less than 0.15. To rescue potential real mutations, we excluded from these criteria variants that were reported in the COSMIC database (version 68) and variants with at least 2 reads supporting the variant allele in each strand. Second, we filtered shared variants that had a VAF of less than 0.15 if the respective control sample had any number of reads supporting the variant allele. To test how effective this filtering was at removing FFPE C>T/G>A artifacts, we divided the samples into 2 groups on the basis of sample age and compared the number of C>T/G>A mutations between the 2 groups, both before and after filtering (Supplemental Figure 12).

Somatic indels were identified using 2 steps with VarScan2, version 2.3.6. In the first step, we detected indels in each sample using a minimum cancer sample read depth of 20, a minimum control sample read depth of 15, a minimum number of reads supporting a variant allele in a tumor sample of 4, a maximum number of reads supporting a variant allele in a normal sample of 0, a minimum variant allele frequency of 0.05, and the strand artifact filter turned on. In the second step, we scanned the detected indels in all samples from the same patient using VarScan2 *mpileup2cns*, quantified the VAFs, and discarded multiallelic indels. Functional annotation and the effect prediction of variants were performed using ANNOVAR and RefSeq genes (53). CADD was used for scoring the deleteriousness of variants (54). An absolute estimation of copy number alterations was performed with AscatNGS (Ascat, version2.3), which allowed for the estimation of ploidy and purity values for each sample (55). Genes were assigned the copy number of the most overlapping segment. Genes were called amplified if the assigned absolute copy number was larger than the average ploidy multiplied by 1.5 and were call deleted if the assigned absolute copy number was less than the average sample ploidy multiplied by 0.5.

Phylogenetic tree reconstruction. For phylogenetic tree reconstruction, we used a variant of the parsimony method, named Dollo parsimony (24). The method is based on the assumption that it is harder to gain a mutation than it is to lose it, which is appropriate when considering tumor evolution. Detailed assumptions in the Dollo parsimony method are provided in Supplemental Methods. Dollo parsimony takes as input an S × M binary matrix D, where $D_{ij} = 1$ if the i^{th} sample has a VAF of at least 5% for the j^{th} mutation. The VAF values are normalized for tumor purity and are estimated using AscatNGS. We performed bootstrapping (56) to estimate the statistical confidence of evolutionary relationships among samples. For each patient, the bootstrap support for internal nodes was computed using 1,000 bootstrap samples.

For phylogenetic tree reconstruction, we used the R interface for the Dollo and Polymorphism Parsimony Program (*Rdollop*) function from the R package Rphylip (57), which is a wrapper around the *dollop* program in the PHYLIP library (58). For bootstrapping, we used the Tree Bipartition and Bootstrapping Phylogenies (*boot.phyo*) function from analysis of phylogenetics and evolution (APE) in R package (59). The R code used for phylogenetic reconstruction is provided in the supplemental material.

Separating property in the tumor tree. Once a tumor tree was reconstructed for each patient, we used the separating property to infer the linear or parallel progression. The separating property is described below.

Somatic mutations accumulate as healthy tissue develops and further progress toward cancer. We assumed that any mutation found in a primary cancer had occurred in one of its present or ancestral cells and that all cells of the primary cancer descended from cells of either the primary cancer or the healthy tissue. This implies that, in a tumor tree, any path from a germline sample, or any other sample representing healthy tissue, to a primary cancer sample represents ancestral versions of the primary cancer. Moreover, the mutations along the path have occurred exclusively in the primary cancer. We defined a path P_{st} between 2 samples, *s* and *t*, as being separated if it intersected any germline-to-primary path *P*, i.e., if P_{st} and *P* shared at least 1 vertex.

We used the separating property for inferring archetypical patterns of cancer progression. If the path between 2 nonprimary samples *s* and *t* was separated, then the most recent common ancestor of *s* and *t* in the tumor tree was a vertex representing an ancestral version of the

primary cancer. From this, it followed that *t* could not have seeded *s* and *s* could not have seeded *t* (see Supplemental Figure 4A), i.e., linear progression from *s* to *t*, and vice versa, could be ruled out. In contrast, if the path between *s* and *t* was not separated, then the MRCA of *s* and *t* may represent an ancestral version of the primary cancer, or it may represent an ancestral version of another sample, e.g., *s*, *t*, or another distant metastasis site (see Supplemental Figure 4B). Therefore, this was consistent with linear as well as parallel progression from *s* to *t*, or vice versa. We do, however, generally consider that the nonseparating property suggests linear progression.

If we also assume that a single cell has seeded each of the nonprimary sites, then it follows that all samples from the same site have to form a subtree of the tumor tree. This conclusion turned out to be violated in several of the tumor trees, which implies that multiple seeding had occurred from the primary cancer to distant metastases as well as from the primary cancer to ALNs. The seeding results from our subclonal analysis further reinforce this observation. In fact, polyclonal seeding from the primary cancer to distant metastases was inferred in 73% (11 of 15) patients. Moreover, we observed different copy numbers across samples from the same site (e.g., see ALN1 and ALN2 in the heatmap for patient 18 in Supplemental Figure 8). Nevertheless, every tree was consistent with cancer being initiated in a single cell.

Separating property and ALN-based distant metastasis. In each of the tumor trees containing ALN samples, the path from the ALN to distant metastases was separated, which implies that the former had not seeded the latter (see Supplemental Figure 6, B, C, G, I, L, M, O, and P). Although, it may be hypothesized that ALNs, together with the primary tumor, have seeded distant metastases, our data led us to reject this hypothesis. First, in the evolutionary analysis, we observed that the path from ALNs to distant metastases was separated in every tumor tree (see Supplemental Figure 6, B, C, G, I, L, M, O, and P). Second, with the exception of patient 3, we observed that, based on the subclonal analysis, a subclone was shared between ALNs and distant metastases only if it was also shared with the primary cancer. Conversely, in most of the cases, a distant metastasis shared an exclusive clone with the primary cancer and/or with another metastasis.

Subclonal analysis. We used PyClone, version 0.13.0 (27), for analysis of the subclonal population structure. PyClone is based on a Bayesian clustering method, which uses a Markov chain Monte Carlo–based (MCMC-based) framework to estimate cellular prevalence values using somatic substitution (estimated using Mutect), copy number aberration, and tumor purity data (estimated using AscatNGS). Details on Mutect and AscatNGS analyses are described in the section *Variant calling, filtering, and copy number alteration detection*. PyClone is implemented in the Python programming language. Details on the parameter values used in the PyClone analysis are provided in the supplemental material.

The following criteria were used to filter out low-occurrence clusters: (a) a cluster was considered only if it had 10 or more mutations; and (b) a cluster *sc* in a sample *s* was considered only if the mean cellular prevalence of *sc* was greater than or equal to 0.05, i.e., *sc* was present in at least 5% of the cells in *s*.

Mutational signatures. Extraction of mutational signatures was performed with nonnegative matrix factorization in R, version 3.2.3, using the SomaticSignatures package, version 2.6.1 (60). Details on the mutational signature analysis including determination of the number of signatures and mapping of identified signatures to previously described ones can be found in the supplemental material.

Statistics. All statistical tests were computed in R and were 2 sided. Adjustments for multiple hypothesis testing were applied when needed using the FDR method. A *P* value threshold of 0.05 was considered significant.

Study approval. This study was approved by the ethics committee of Karolinska Institutet. All patients provided written informed consent to participate in the study.

Author contributions

JB and JH conceived the idea and designed and directed the study. IU, GMK, and AA contributed to the conception and design of the study, designed illustrations and figures, and were involved in writing and drafting the manuscript. IU performed phylogenetic and subclonal analyses. GMK summarized clinical information on the cohort and was involved in DNA and RNA extractions from FFPE material. AA performed variant calling, copy number alteration detection, and mutational signature analysis. UK contributed to bioinformatics pipeline optimization and designed the illustrations for patient treatment and relapse timelines. GS helped with the selection of patients' materials and pathological reports. J. Lövrot performed PAM50-intrinsic molecular subtype analysis based on gene expression data. NFM verified the origin of the metastatic lesions and other pathological features for all the included cancer samples. J. Lagergren supervised the phylogenetic and subclonal analyses. SH supervised the mutational signature analysis and provided bioinformatics support. JB, JH, SH, and J. Lagergren reviewed the manuscript.

Acknowledgments

This study was supported by grants from Märit and Hans Rausingg's Fund for Cancer Research; the Swedish Cancer Society; the Swedish Research Council-Starget-Linné; the Academy of Finland (Center of Excellence in Cancer Genetics Research); the Finnish Cancer Foundations; and the Sigrid Jusélius Foundation. JB's research group also receives support from Radiumhemmet; a KI-AZ Support Grant; the KI-Stockholm County Council; the Swedish Society for Medical Research (SSMF); the Strategic Research Programme in Cancer (StratCan); and the Breast Cancer Theme Center (BRECT). We are grateful for the computational infrastructure provided by CSC – Scientific Computing Ltd. and the Swedish National Infrastructure for Computing (SNIC).

Address correspondence to: Johan Hartman, Associate Professor of Pathology, Department of Oncology-Pathology Karolinska Institutet and University Laboratory, S-171 76 Stockholm, Sweden. Phone: 46.0.739.760242; Email: johan.hartman@ki.se.

1. Weigelt B, Peterse JL, van 't Veer LJ. Breast cancer metastasis: markers and models. *Nat Rev Cancer*. 2005;5(8):591–602.
2. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet*. 2005;365(9472):1687–1717.
3. Early Breast Cancer Trialists' Collaborative Group (EBCTCG), et al. Comparisons between different polychemotherapy regimens for early

breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet*. 2012;379(9814):432–444.
4. Greaves M, Maley CC. Clonal evolution in cancer. *Nature*. 2012;481(7381):306–313.
5. Karlsson E, et al. Clonal alteration of breast cancer receptors between primary ductal carcinoma in situ (DCIS) and corresponding local events. *Eur J Cancer*. 2014;50(3):517–524.
6. Lindström LS, et al. Clinically used breast cancer markers such as estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2 are unstable throughout tumor progression. *J Clin Oncol*. 2012;30(21):2601–2608.
7. Thompson AM, et al. Prospective comparison of switches in biomarker status between primary and recurrent breast cancer: the Breast Recurrence In Tissues Study (BRITS). *Breast Cancer Res*. 2010;12(6):R92.
8. Amir E, et al. Prospective study evaluating the impact of tissue confirmation of metastatic disease in patients with breast cancer. *J Clin Oncol*. 2012;30(6):587–592.
9. Gundem G, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*. 2015;520(7547):353–357.
10. Fidler IJ. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nat Rev Cancer*. 2003;3(6):453–458.
11. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. *Nat Genet*. 2003;33(1):49–54.
12. Lee YF, et al. A gene expression signature associated with metastatic outcome in human leiomyosarcomas. *Cancer Res*. 2004;64(20):7201–7204.
13. Ding L, et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*. 2010;464(7291):999–1005.
14. Shah SP, et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*. 2009;461(7265):809–813.
15. Navin N, et al. Tumour evolution inferred by single-cell sequencing. *Nature*. 2011;472(7341):90–94.
16. Yates LR, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med*. 2015;21(7):751–759.
17. Brown D, et al. Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. *Nat Commun*. 2017;8:14944.
18. Disibio G, French SW. Metastatic patterns of cancers: results from a large autopsy study. *Arch Pathol Lab Med*. 2008;132(6):931–939.
19. Colleoni M, et al. Site of primary tumor has a prognostic role in operable breast cancer: the international breast cancer study group experience. *J Clin Oncol*. 2005;23(7):1390–1400.
20. Veronesi U, et al. Risk of internal mammary lymph node metastases and its relevance on prognosis of breast cancer patients. *Ann Surg*. 1983;198(6):681–684.
21. Fisher B. Laboratory and clinical research in breast cancer--a personal adventure: the David A. Karnofsky memorial lecture. *Cancer Res*. 1980;40(11):3863–3874.
22. Hellman S. Karnofsky Memorial Lecture. Natural history of small breast cancers. *J Clin Oncol*. 1994;12(10):2229–2234.
23. Wong SY, Hynes RO. Lymphatic or hematogenous dissemination: how does a metastatic tumor cell decide? *Cell Cycle*. 2006;5(8):812–817.
24. Farris JS. Phylogenetic Analysis Under Dollo's Law. *Systematic Biology*. 1977;26(1):77–88.
25. Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A*. 1996;93(23):13429–13434.
26. Naxerova K, Jain RK. Using tumour phylogenetics to identify the roots of metastasis in humans. *Nat Rev Clin Oncol*. 2015;12(5):258–272.
27. Roth A, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods*. 2014;11(4):396–398.
28. Alexandrov LB, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415–421.
29. Roberts SA, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet*. 2013;45(9):970–976.
30. Feinstein AR, Sosin DM, Wells CK. The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. *N Engl J Med*. 1985;312(25):1604–1608.
31. George S, et al. Will Rogers revisited: prospective observational study of survival of 3592 patients with colorectal cancer according to number of nodes examined by pathologists. *Br J Cancer*. 2006;95(7):841–847.
32. Cady B. Lymph node metastases. Indicators, but not governors of survival. *Arch Surg*. 1984;119(9):1067–1072.
33. van de Vijver MJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*. 2002;347(25):1999–2009.
34. Wang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005;365(9460):671–679.
35. Minn AJ, et al. Genes that mediate breast cancer metastasis to lung. *Nature*. 2005;436(7050):518–524.
36. Liu R, et al. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *N Engl J Med*. 2007;356(3):217–226.
37. Fisher B, Jeong JH, Anderson S, Bryant J, Fisher ER, Wolmark N. Twenty-five-year follow-up of a randomized trial comparing radical mastectomy, total mastectomy, and total mastectomy followed by irradiation. *N Engl J Med*. 2002;347(8):567–575.
38. Galimberti V, et al. Axillary dissection versus no axillary dissection in patients with sentinel-node micrometastases (IBCSG 23-01): a phase 3 randomised controlled trial. *Lancet Oncol*. 2013;14(4):297–305.
39. International Breast Cancer Study Group, et al. Randomized trial comparing axillary clearance versus no axillary clearance in older patients with breast cancer: first results of International Breast Cancer Study Group Trial 10-93. *J Clin Oncol*. 2006;24(3):337–344.
40. Giuliano AE, et al. Axillary dissection vs no axillary dissection in women with invasive breast cancer and sentinel node metastasis: a randomized clinical trial. *JAMA*. 2011;305(6):569–575.
41. Engel J, Emeny RT, Hölzel D. Positive lymph nodes do not metastasize. *Cancer Metastasis Rev*. 2012;31(1–2):235–246.
42. Klein CA. Selection and adaptation during metastatic cancer progression. *Nature*. 2013;501(7467):365–372.
43. Brastianos PK, et al. Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discov*. 2015;5(11):1164–1177.
44. Cheung KJ, et al. Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. *Proc Natl Acad Sci USA*. 2016;113(7):E854–E863.
45. Naba A, Clauser KR, Lamar JM, Carr SA, Hynes RO. Extracellular matrix signatures of human mammary carcinoma identify novel metastasis promoters. *Elife*. 2014;3:e01308.
46. Pantel K, Diaz LA, Polyak K. Tracking tumor resistance using 'liquid biopsies'. *Nat Med*. 2013;19(6):676–677.
47. Nik-Zainal S, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*. 2016;534(7605):47–54.
48. de Bruin EC, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*. 2014;346(6206):251–256.
49. Swanton C, McGranahan N, Starrett GJ, Harris RS. APOBEC enzymes: mutagenic fuel for cancer evolution and heterogeneity. *Cancer Discov*. 2015;5(7):704–712.
50. Fox EJ, Prindle MJ, Loeb LA. Do mutator mutations fuel tumorigenesis? *Cancer Metastasis Rev*. 2013;32(3-4):353–361.
51. Ovaska K, et al. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med*. 2010;2(9):65.
52. Cibulskis K, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213–219.
53. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.
54. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310–315.
55. Van Loo P, et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A*. 2010;107(39):16910–16915.
56. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. VBoca Raton, FL: CRC Press; 1994.
57. Revell LJ, Chamberlain SA. Rphylip: an R interface for PHYLIP. *Methods Ecol Evol*. 2014;5(9):976–981.
58. Baum BR. PHYLIP: Phylogeny Inference Package. Version 3.2. Joel Felsenstein. *Q Rev Biol*. 1989;64(4):539–541.
59. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. 2004;20(2):289–290.
60. Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics*. 2015;31(22):3673–3675.