Evolution of chromosome arm aberrations in breast cancer through genetic network rewiring

Elena Kuzmin[1,2,3*], Toby M. Baker[4], Tom Lesluyes[4], Jean Monlong[5,6], Kento T. Abe[7,8], Paula P. Coelho[1,2], Michael Schwartz[1,2], Dongmei Zou[1], Genevieve Morin[2], Alain Pacis[6,9], Yang Yang[5], Constanza Martinez[1,10,11], Jarrett Barber[7,12], Hellen Kuasne[1], Rui Li[5,6], Mathieu Bourgey[6,9], Anne-Marie Fortier[1], Peter G. Davison[13,14], Atilla Omeroglu[10], Marie-Christine Guiot[10], Quaid Morris[5,12,15], Claudia L. Kleinman[5,16], Sidong Huang[1,2], Anne-Claude Gingras[7,8], Jiannis Ragoussis[5,6], Guillaume Bourque[5,6,9], Peter Van Loo[4,17,18], Morag Park[1,2,11,19*]

Affiliations:

[1] Rosalind and Morris Goodman Cancer Institute, Montreal, QC, H3A 1A3, Canada
[2] Department of Biochemistry, McGill University, Montreal, QC, H3G 1Y6, Canada
[3]Present: Department of Biology, Centre for Applied Synthetic Biology, Centre for Structural and Functional Genomics, Concordia University, Montreal, QC, H4B 1R6, Canada; Department of Human Genetics, McGill University, Montreal, QC, H3A 0C7, Canada
[4]The Francis Crick Institute, London, NW1 1AT, UK
[5]Department of Human Genetics, McGill University, Montreal, QC, H3A 0C7, Canada
[6]McGill Genome Centre, Montreal, QC, H3A 0G1, Canada
[7]Department of Molecular Genetics, University of Toronto, Toronto, ON, M5S 1A8
[8]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Sinai Health, Toronto, ON, M5G 1X5, Canada
[9]Canadian Centre for Computational Genomics (C3G), McGill University, Montreal, QC, H3A 0G1, Canada
[10]Department of Pathology, McGill University, Montreal, QC, H3A 2B4, Canada
[11]Gerald Bronfman Department of Oncology, McGill University, Montreal, QC, H4A 3T2, Canada
[12]Vector Institute, Toronto, M5G 1M1, Canada; Ontario Institute for Cancer Research, Toronto, Ontario, M5G 0A3, Canada; Computational and Systems Biology, Sloan Kettering Institute, New York City, NY 10065, USA
[13]Department of Surgery, McGill University, Montreal, QC, H3G 1A4, Canada
[14]McGill University Health Centre, Montreal, QC, H4A 3J1, Canada
[15]Department of Computer Science, University of Toronto, Toronto, M5S 2E4, Canada
[16]Lady Davis Institute for Medical Research, Montreal, QC, H3T 1E2, Canada
[17]Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[18]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[19]Lead Contact
[*]Correspondence: elena.kuzmin@mcgill.ca, morag.park@mcgill.ca

Summary

The basal breast cancer subtype is enriched for triple-negative breast cancer (TNBC) and displays consistent large chromosomal deletions. Here, we characterize the evolution and maintenance of chromosome 4p (chr4p) loss in basal breast cancer. TCGA data analysis showed recurrent deletion of chr4p in basal breast cancer. Phylogenetic analysis of a unique panel of 23 primary tumor/patient-derived xenograft basal breast cancers revealed early evolution of chr4p deletion. Mechanistically we show that Chr4p loss is associated with enhanced proliferation. Gene function studies identified an unknown gene, *C4orf19,* within chr4p, which suppressed proliferation when overexpressed and is a novel member of a PDCD10-GCKIII kinase module, we name as *PGCA1*. Genome-wide pooled overexpression screens using a barcoded library of human open reading frames, identified chromosomal regions, including chr4p, that suppress proliferation when overexpressed in a context-dependent manner implicating network interactions. Together this sheds light on the early emergence of complex aneuploid karyotypes involving chr4p and adaptive landscapes shaping breast cancer genomes.

Keywords

Cancer genomics, Cancer evolution, Basal breast cancer, Triple-negative breast cancer, Aneuploidy, Chromosomal arm copy number aberrations, Chromosome 4p, PDCD10, GCK-III

Introduction

Breast cancer is a heterogeneous disease comprising several clinical and molecular subtypes. Therapeutic strategies have been devised for patients based on biomarkers, such as hormone (estrogen and progesterone) receptor expression or human epidermal growth factor 2 (HER2) receptor amplification [1]. However, triple-negative breast cancer (TNBC), constituting 10-20% of all breast cancers, lacks these receptors, thus, lacks precision therapies targeting them and is predominantly treated by chemotherapy. Currently, due to limited therapeutic options TNBC has the most aggressive behavior and worst prognosis (5-year relative survival percent), leading to a large percentage of breast cancer deaths [2-4]. The basal breast cancer molecular subtype constitutes ~80% of TNBC and shows a complex mutational spectrum without common oncogenic drivers [5-7]. Notably basal breast cancers frequently display consistent large chromosomal deletions [8,9] that are thought to play an important role in pathogenesis but the consequence of which are poorly understood.

An important hallmark of cancer cells is genomic instability, which generates mutations and chromosome alterations that confer selective advantage on subclones of cells and lead to their growth and dominance in a local tissue [10,11]. Considerable effort has been invested into identifying which oncogenes, the increased activity, or tumor suppressor genes, the loss of function of which, drive cancer development [10]. With the advent of genomic technologies, it has become possible to generate a detailed map of genetic changes in cancer [6,12-14]. Genomic analyses revealed that chromosome arm somatic copy number aberrations are more common than whole chromosome somatic copy number aberrations and certain chromosomal arms are preferentially lost or gained, suggesting that these events are selected because they are advantageous during cancer progression [15,16]. Evolutionary analyses of Pan-Cancer Analysis of Whole Genomes (PCAWG) data on 38 types of cancer showed that chromosomal arm copy number losses occur early and typically precede gains, indicating their selective advantage in tumor onset and progression [17]. Recent findings suggest that chromosomal arm aberrations occur in bursts enabling genome diversification and preferential clonal expansion in TNBC [18]. Although, they have been implicated in some cancers in increasing cell growth [19] and evading immune system detection [20], the functional consequences of chromosomal arm deletions remain poorly understood.

We previously established that a large deletion on chromosome 5q in basal breast cancer leads to a loss of function of *KIBRA*, encoding a multi-domain scaffold protein, activating oncogenic transcription factors, *YAP/TAZ* [21]. Chromosome 8p loss in breast cancer alters fatty acid and ceramide metabolism, leading to invasiveness and tumor growth under stress conditions due to increased autophagy, thus contributing to resistance to chemotherapeutic agents [22]. Another study found that cooperative effects, resulting from genes co-deleted within a region harboring *TP53* on chromosome 17p, lead to more aggressive lymphoma than individual mutations [23]. Interestingly, a recent study reported differences in specific chromosome arm losses, such as chromosome 3p loss, which positively correlated with immune signatures suggesting that specific chromosomal regions can exert selective pressures rather than overall aneuploidy level [24].

In this study, we identified chromosome 4p loss as a frequently recurrent chromosome arm loss in the basal subtype of breast cancer and established that this occurred as an early clonal event functionally associated with an enhanced proliferative state. Scanning genes on chr4p by functional assays, we identified genes whose elevated expression suppressed proliferation in human breast epithelial cells. This included an unknown gene, *C4orf19,* within chr4p, we show suppresses proliferation and is a member of the programmed cell death 10 (PDCD10)-germinal centre kinase III (GCKIII) module (which we call *PGCA1*). Genome-wide pooled overexpression screens using a barcoded library of human open reading frames identified chr4p and other chromosomal regions that suppress proliferation when overexpressed in a context-dependent manner. Together this provides new insight into TNBC, a hard to treat cancer, for which the current standard therapeutic options have not significantly changed the overall survival rate.

Results

Loss of chromosome 4p is recurrent and functionally significant in basal breast cancer.

We investigated the frequency of chromosome arm losses in the breast cancer basal subtype[8]. The three most frequent chromosomal arm losses in basal breast cancer were 8p, 5q and 4p, occurring collectively in ~65-76% of cases (Fig. 1B& S1A, Table S1). We detected chromosome 4p (chr4p) loss in ~65% of basal breast cancer cases, which has not been studied previously. Regions of chr4p loss span the entire short arm of chromosome 4 without an apparent minimal deleted region, suggesting that its selective advantage is conferred by the loss of multiple genes residing within the chromosome arm rather than a single tumor suppressor gene (Fig. 1C).

Approximately 80% of genes (133 of 156) within chr4p showed reduced gene expression in patients with chr4p loss, indicating that this loss is functionally significant (Fig. 1D, Table S2). Among the ten genes with the most substantial reduction in gene expression upon chr4p deletion are *SLC34A2* and *RHOH,* that are known tumor suppressor genes in other cancers but have not been previously implicated in breast cancer[25]. Notably the worse survival of cases exhibiting chr4p loss is not due to the enrichment of *TP53* loss of function mutations in chr4p loss since their distribution among groups was not significantly different (Fig. 1E, Methods). Hence, chr4p loss may be clinically prognostic. We also detected statistically significant chr4p deletion in multiple cancer types, including Lung Squamous Cell Carcinoma, Testicular Germ Cell Tumors and Ovarian Serous Carcinoma, indicating that chr4p loss is broadly observed in other cancers (Fig. 1F, Table S3).

Next, we interrogated global transcriptomic changes associated with chr4p loss (Fig. 1G&H, Table S4). Basal breast cancers with chr4p loss showed an elevation of genes with roles in DNA replication ($p = 6.9x10^{-7}$) such as *GINS4* encoding a member of the GINS complex, which plays an essential role in initiation of DNA replication and progression of DNA replication forks [26] and cell cycle ($p = 4.1x10^{-4}$), such as *STK33* encoding a serine/threonine protein kinase, which activates ERK signaling pathway [27]. Together these terms suggest of a proliferative advantage conferred by chr4p loss likely through a combined effect of expression changes of multiple genes belonging to these pathways. Elevated expression of genes with a role in microtubule cytoskeleton organization ($p = 1.1x10^{-6}$) and protein translation ($p = 1.4x10^{-3}$), suggest the involvement of cellular plasticity, which enables cancer cell adaptation to stress [28]. Chr4p loss was associated with decreased expression of genes annotated to positive regulation of immune response, cytokine-mediated signaling, T cell and B cell activation, interferon gamma mediated signaling and natural killer cell-mediated immunity consistent with TNBC displaying immune evasion and poorer outcome. This differential gene expression was not due to general differences in arm-level and chromosome-level copy number changes ($p = 0.074$) (Fig. S2, Table S5) [20]. These findings support that these global transcriptomic changes are specific to chr4p loss rather than general differences in aneuploidy between tumors and highlight the importance of specific chromosome arm losses in basal breast cancer.

Chromosome 4p loss is an early event in basal breast cancer evolution. Evolution of basal breast cancer genomes can be reconstructed from somatic mutations detected by whole genome sequencing (WGS)[17]. To understand the evolutionary timing of chr4p loss, we performed phylogenetic reconstruction on bulk WGS of our primary tumor / patient-derived xenograft (PT/PDX) panel of 23 paired samples annotated to PAM50 basal breast cancer subtype, which we previously collected[29]. To our knowledge this is the largest available phylogeny for TNBC to date. Aggregated single-sample ordering revealed a typical timing of chromosome arm aberrations and other genetic events (Fig. 2A, Table S6). Coding mutations in *TP53* had a high likelihood of being clonal and thus occurring early in tumor progression, consistent with it being a known driver event in basal breast cancer; as well as preceding whole genome duplication consistent with *TP53* function and recent findings from single-cell, single-molecule DNA-sequencing of 8 human triple-negative breast cancers and four cell lines [18]. We also observed that chromosomal arm losses occurred before chromosomal gains, and chr17p loss harboring *TP53* occurred with similar timing to *TP53* coding mutations, likely driving biallelic inactivation of *TP53*. Most frequent and clonal chromosomal arm losses included 4p, 17q, 3p, 17p, 15q, 14q, and 5q, the majority of which occurred after *TP53* coding mutations. Even though the median age of diagnosis for this PT/PDX cohort was 54, whole genome duplication occurred ~6 years prior, around 48 years of age in 16 samples (~70%) and the most recent common ancestor was observed on average ~2 years before diagnosis, around 52 years of age (Fig. 2B, Table S6). Unlike early events, there was no apparent subclonal structure since it tended to be distinct among patients, suggesting that even though early events are shared among basal breast cancer patients, late events diverge. These findings are consistent with a recent study on gastric cancer evolution, which showed that *TP53* leads to chromosome arm-level aneuploidy in a temporally preferred order [30]. Thus, it is important to focus on clonal events for therapeutic application.

Since we detected *TP53* coding mutations occurring earlier than chr4p loss and no single known cancer gene was associated with chr4p loss, it is possible that a combination of cancer gene mutations in concert with *TP53* leads to a selective advantage of chr4p loss. We also observed that chr8p loss, which was detected in 76% of the TCGA basal breast cancer samples, was detected in ~20% PT/PDX basal breast cancer cohort, suggesting that it is related to ancestry differences. The PT/PDX cohort largely consists of a Québec patient population, which has been reported to have distinct genomic characteristics compared to European ancestry or other backgrounds present in the TCGA dataset [31]. To understand the selection pressures that maintain chr4p loss, we decided to focus on an individual patient (PT/PDX1735) whose trajectory revealed an early clonal chr4p loss, a clonal *TP53* coding mutation and WGD (Fig. 2C) and for which we established a PDX and PDX-derived cell line. Our scDNAseq analysis of PDX1735 confirmed that chr4p loss was an early event in this basal breast cancer progression (Fig. S3, Table S7).

Chromosome 4p loss is associated with a proliferative state. To understand the functional effect of chr4p loss, we leveraged single cell RNAseq (scRNAseq) data from PDX1735, established from a basal breast cancer primary tumor, as reported in our previous study (Fig. 3A top left) [32]. To infer copy number status at a single cell resolution to identify transcriptional programs associated with cells harboring chr4p deletion (Fig. 3), we employed a method, which detects consistent variation in gene expression of consecutive genes across chromosomal regions [33]. To obtain a normal gene expression baseline, we performed scRNAseq on breast tissue samples from two patients undergoing bilateral mammoplasty reduction (Figure S4, Table S8). We computed a z-score relative to the baseline and called copy number aberrations using a Hidden Markov model (HMM) with three states: neutral copy number, loss, and gain. In this manner, we identified four stable 'communities', groups of cells with a shared pattern of inferred copy number profiles. Three communities (1-3) harbor chr4p deletion and community 4 harbors a chr4p copy neutral state (Fig. 3A top right, bottom, Table S9). The relatively small size of community 4 (~ 9%), which is copy neutral for chr4p, is consistent with chr4p loss being an early clonal event, as revealed by our timing analysis (Fig. 2C&S3).

To understand the selection pressures that maintain chr4p loss, we compared the distributions of chr4p copy neutral and deletion communities across different cellular clusters associated with distinct transcriptional programs. We previously described six cellular clusters embedded within PDX1735, which included proliferation, nuclear/mitochondria, antigen presentation, basal, mesenchymal/stem and boundary based on differential gene expression [32]. As expected, the inferred copy number communities did not overlap with any specific gene expression cluster, since the normalized expression was smoothed using a rolling median approach to reduce the effect of single-gene outliers (Fig. S4B). Thus, cells belonging to chr4p deletion communities (community 1+2+3), which comprised most cells of the PDX1735, did not show a preferential distribution across any cellular cluster (Fig. 3B). However, cells belonging to the chr4p copy neutral community (community 4) were significantly strongly depleted in the proliferation and to a lesser extent in the mesenchymal cluster (Fig. 3B). The proliferation cluster was characterized by proliferative cells since a large proportion of cells in this cluster (~95%) were cycling and exhibited inferred G2/M and S cell cycle states based on the relative gene expression of G1/S and G2/M gene sets [34] as well a high expression of cell cycle genes, such as *MKI67* (Fig. 3C, Table S9).

To functionally validate the findings from scRNAseq data, we performed immunofluorescence staining combined with RNA-*in situ* hybridization (RNA ISH). The staining of a paraffin-embedded fixed tissue section of PDX1735 used a combination of immunofluorescence for Ki67 and RNA ISH for *RBPJ*. *RBPJ* was selected as a marker of chr4p copy number state because of the availability of a probe for RNA ISH and a consistent gene expression difference in our basal breast cancer PT/PDX panel between chr4p copy neutral and deletion samples. This analysis revealed that there was an inverse relationship between Ki67 abundance and *RBPJ* gene expression. Breast cancer cells with chr4p deletion and thus low *RPBJ* expression showed a high abundance of Ki67 and thus were more proliferative than chr4p copy neutral cells (Fisher exact test, $p = 8.7\text{x}10^{-9}$) (Fig. 3D). Together these findings suggest that chr4p loss confers on basal breast cancer cells a proliferative advantage.

Suppression of proliferation by overexpression of chromosome 4p genes is context-dependent. To determine if chr4p deletion in basal breast cancer is selected due to a proliferative advantage, we tested whether the overexpression of genes within this region elicits a proliferation defect. Chr4p copy neutral normal breast epithelial cell line, MCF10A, chr4p copy neutral basal breast cancer PDX-derived cell line, GCRC1915, or chr4p deletion basal breast cancer PDX-derived cell line, GCRC1735 were used to generate stable cell populations overexpressing candidate chr4p genes using lentivirus-mediated integration of constructs from the human ORF collection (Fig. 4A) [35]. The candidate genes resided within a high confidence chr4p deletion region in GCRC1735 according to whole exome sequencing (WES) from our previous study [32], encompassing about half of the chromosome arm and which contained 30 genes, for about half of which our collection contained lentiORF overexpression vectors.

Surprisingly, overexpressing a large fraction of chr4p genes suppressed proliferation in a context-dependent manner, whereby proliferation suppression was only observed in a basal breast cancer PDX-derived cell line which is deleted for chr4p, PDX1735, and not in cell lines that were chr4p copy neutral, MCF10A or GCRC1915. Of note GCRC1915 displays LOH within chr4p it is copy neutral with no change in chr4 copy number (Fig. 4B left). This extent of suppression was further exacerbated when two random genes within chr4p were overexpressed (Fig. 4B right). GCRC1915 and GCRC1735 are characterized by many other distinct SNVs and CNAs [29]. The context-dependency is likely not due to *TP53* mutation status since both GCRC1915 and GCRC1735 harbor a coding mutation in *TP53*, whereas MCF10A carries a wild-type copy of *TP53* [29]. On the other hand, GCRC1735, unlike GCRC1915 and MCF10A, habours a germline *BRCA1* mutation [29] as well as early clonal losses of chr5q, chr8p, chr9p and chr19q

(Table S6) which may underlie the context dependent suppression of proliferation of chr4p gene overexpression observed in GCCR1735 and not in GCRC1915 or MCF10A. Hence, the observed context-dependent suppression of proliferation may be due to a genetic interaction with another genetic aberration, which rewires the genetic network sensitizing chromosome 4p region to overexpression and thus maintaining it in a deletion state. Further studies in isogenic model systems should be conducted to test which combinations of genetic events interact with chr4p loss to confer a proliferative advantage, since it was previously found that individual chromosome arm losses lead to growth defects [24].

To determine whether the context-dependent suppression of proliferation was specific to chromosome 4p and to identify other such regions, we conducted a genome-wide overexpression screen using pooled TRC3 LentiORF collection as previously described (Fig. 4C, Table S11) [36]. Two cell lines with different copy number states of chr4p, MCF10A (chr4p copy neutral) and GCRC1735 (chr4p deletion), were transduced with pooled lentiORF library at multiplicity of infection (MOI) of 0.3 to ensure one integration event per cell. After puromycin selection cells were maintained for 6-8 doublings and 1000x coverage was maintained at each step of the experiment. Next generation sequencing (NSG) was used to capture barcode abundance, which served as a proxy for cell growth rate. Genome-wide pooled lentiORF overexpression screen uncovered genes that suppressed proliferation in both cell lines and were previously identified as STOP genes in another study, such as epithelial tumor suppressor *ELF3*, transcription factor *EBF1* and a DNA repair protein *RAD51* [19]. The screen also revealed regions that suppressed proliferation in a context-dependent manner. The context-dependent regions that suppressed proliferation when overexpressed in GCRC1735 but not in MCF10A included chr4p and 13q. These regions were also deleted in GCRC1735 and not in MCF10A (Fig. 4D), suggesting that this mode of selection is not specific to chr4p loss and likely exerts the selection pressure early in tumor progression since our evolutionary timing analysis revealed that both are clonal events (Fig. 2C). These observations suggest that the dosage of these genes exerts a selection pressure to maintain this chromosomal region deleted in a specific genomic context of basal breast cancer.

Overexpression of C4orf19 suppresses proliferation and reveals an interaction with PDCD10-GCKIII kinase module. We observed that *C4orf19* suppressed proliferation when overexpressed in multiple contexts, such as MCF10A and GCRC1735 cells (Fig. 4B). To understand the biological role of *C4orf19*, we analyzed the sequence of the protein it encodes. C4orf19 is an uncharacterized protein 314 amino acids in length, which has orthologs in mouse and rat according to the Alliance of Genome Resources ortholog inference [37]. Functional analysis of its protein sequence using InterPro revealed that it belongs to the protein family domain unknown function DUF4699 and is predicted to contain two consensus disorder regions (36-142, 267-291). We mined BioGRID [38] for previously identified protein-protein interactions: high throughput methods, such as affinity capture-MS and yeast two-hybrid both revealed programmed cell death 10, PDCD10 (Fig. 5A) [39,40]. Additionally, all three members of the germinal center kinases (GCK-III) subfamily, STK24, STK25, STK26 that directly interact with PDCD10 in a mutually-exclusive heterodimer [41] were reported in the affinity capture-MS method [40]. We heterologously expressed C4orf19-v5 and 3xFLAG- PDCD10, 3xFLAG-STK25 and 3xFLAG-STK26 in MCF10A cells (Fig. 5B). Co-immunoprecipitation assay using Anti-V5 for pull-down showed the presence of 3xFLAG- PDCD10, 3xFLAG-STK25 and 3xFLAG-STK26, confirming previously identified high-throughput interactions, indicating that C4orf19 interacts with PDCD10 and its associated GCK-III kinases (Fig. 5C).

To gain more insight into the biological role of C4orf19, we performed proximity-dependent biotinylation of proteins coupled to mass spectrometry (miniTurboID) [42] to reveal the comprehensive physical neighbourhood in which C4orf19 resides. MiniTurbo biotin ligase was fused to C4orf19 and expressed in MCF10A (alongside negative controls; bait expression was verified by western blotting), and biotinylation of proximal proteins was induced by the addition of biotin (Fig. S5A&B). Biotinylated proteins were recovered by streptavidin-affinity chromatography and identified by mass spectrometry. Reduction in proliferation was observed 48 hr after induction of *C4orf19* overexpression with doxycycline (Fig. 5D). Using the SAINTexpress computational tool, we identified 370 high-confidence (Bayesian FDR < 5%) proximal interactors, which included PDCD10, STK24, STK25 and STK26 (Fig. 5A, Table S10). The analysis of gene ontology molecular function (GO MF) of C4orf19 proximal interactors showed enrichment of proteins at the plasma membrane, suggesting that C4orf19 is localized to the cell periphery (Fig. 5E). The subcellular localization of C4orf19 at the cell periphery was further validated by immunofluorescence of C4orf19. The signal intensity of C4orf19 relative to area of the inner or outer cell section indicates that C4orf19 abundance is higher at the cell periphery although it was not uniformly distributed (Fig. 5F). This finding is consistent with Human Protein Atlas, which showed that across tissues C4orf19 is at the cell periphery in some cancers, such as breast lobular carcinoma and RT4 cells derived from a urinary bladder transitional cell papilloma [43,44]. Similarly, the Alliance of Genome Resources computationally predicted its localization to cell junctions based on GO annotation and orthology [37]. We did not observe C4orf19 in the nucleus, which is provided as a secondary predicted localization of this protein, which may be due to cell type specificity.

These findings support that C4orf19 is associated with a subset of PDCD10 and GCK-III kinases localized to the cell periphery, and we propose to rename *C4orf19* to *PGCA1* (*P*DCD10-*GC*K-III Kinase *A*ssociated). *STK24* resides on chr13q and *STK25* on chr2q, both of which show a deletion state in GCRC1735 and suppress proliferation when overexpressed (Fig. 4D), suggesting of copy number change as a selection for stoichiometric balance of the members of the PDCD10-GCKIII kinase module. Overall, the deletions of chr13q and chr4p are both early clonal events in the basal PT/PDX panel highlighting a potential importance of the stoichiometric balance of the PDCD10-GCKIII kinase module as an additional common evolutionary mechanism in this breast cancer subtype.

Discussion

This study identified chromosome 4p loss as a frequently recurrent chromosome arm loss in the basal breast cancer molecular subtype, affecting ~65% of TNBC patients. Our data indicate that chr4p loss is functionally significant. It is associated with reduced expression of most genes within the region and poor prognosis. An evolutionary timing analysis revealed that chr4p loss is an early clonal event. Through multiple approaches we showed that the deletion of chr4p is associated with enhanced proliferation. Targeted single and dual gene overexpression assays of genes within chr4p uncovered *C4orf19,* which suppressed proliferation and was identified to be associated with the PDCD10-GCKIII kinase module, and we propose to rename this gene to

PGCA1. However, most genes within chr4p suppressed proliferation when overexpressed in a context-dependent manner associated with chr4p deletion. Genome-wide pooled overexpression screens identified other chromosome arms whose suppression of proliferation was context-dependent and associated with copy number loss. Our findings support chr4p loss confers a proliferative advantage in basal breast cancer, and multiple genes within chr4p suppress proliferation when overexpressed in chr4p loss but not copy-neutral cells. Together this provides a unique understanding of the early emergence of complex aneuploid karyotypes involving chr4p and the adaptive landscapes shaping breast cancer genomes.

We found that chromosome arm losses are hemizygous, likely due to the presence of core essential genes residing within them (Fig. S1A) [45,46]. The apparent lack of a clear minimal deletion region of chr4p in basal breast cancer suggests that this event is not driven by the loss of a single tumor suppressor gene but rather by multiple genes and/or genetic elements at multiple, spatially separated loci whose deletion together yields a proliferative advantage. While there were four tumor suppressors genes within chr4p, none of them has been implicated in breast cancer[25]. These include *SLC34A2,* which encodes a pH-sensitive sodium-dependent phosphate transporter [47] and *N4BP2*, which encodes 5'-polynucleotide kinase, playing a role in DNA repair, which have been implicated in lung cancer [48]. *RHOH* encodes a member of the Ras superfamily of guanosine triphosphate (GTP)-metabolizing enzymes and has been implicated in non-Hodgkin's lymphoma [49]. *PHOX2B* is a transcription factor involved in neuroblastoma [50]. In contrast to *PHOX2B*, which is recessive, the tumor suppressive effects of *SLC34A2* and *RHOH* are dominant [25], indicating that the perturbation of one copy of these genes is sufficient to contribute to carcinogenesis and their combined effect due to chr4p loss may be providing a selective advantage in TNBC. Surprisingly, the expression of several genes is elevated, likely due to the loss of *cis*-transcriptional repressors. These include *UCLH1*, which is a ubiquitin hydrolase previously shown to be highly expressed in metastatic estrogen receptor negative (ER–) and triple negative breast cancer subtypes [51].

Simultaneous gene expression silencing using RNAi of multiple combinations of up to three genes along chromosome 8p inhibited tumorigenesis in a mouse model of hepatocellular carcinoma (HCC) indicating that multiple TSGs show a greater capacity to promote tumorigenesis than individual genes [52]. The effect of chromosome 3p loss in lung cancer is also attributed to an alteration in a combination of genes [24,53]. The apparent lack of a minimal region in a large chromosomal aberration has been also previously observed in human embryonic stem cells and induced pluripotent stems cells when screening for genetic changes occurring in cell culture to evaluate their tumorigenicity, which reported a recurrent gain of chromosome 1, 12 and 17 without any frequently repetitive minimal amplicon [54]. The gain of chromosome 12 or its short arm 12p has been also shown to arise rapidly during the reprogramming process and is associated with gene expression changes indicating its functional significance in conferring selective growth advantage [55].

The functional significance of chr4p loss as assessed by the reduction in gene expression of a large portion of genes within the region, as well as the association of the chromosome arm loss with a proliferative state and mesenchymal state, are consistent with previous observations related to chr3p loss, whereby hallmark sets of cell cycle and epithelial-mesenchymal transition genes were upregulated when chr3p was lost [24]. Chr4p loss in basal breast cancer was also associated with a decrease in immune signature suggesting that specific partial aneuploidy of chr4p rather than general differences in aneuploidy between tumors may be important for immune system evasion, which were previously reported in a pan-cancer analysis [20]. This is especially important since it was previously shown that triple-negative breast cancer with an "immune-cold" microenvironment characterized by the absence of CD8+ T cells in the tumor resulted in poor outcomes [56]. The finding of chromosome arm specific aneuploidy exerting distinct effects on the immune system is likely due to different immure-related genes residing there which has recently emerged, such as a finding that chr3p loss is associated with increased immune activity [24].

It is thought that changes in the copy number of specific genes due to large chromosomal variants lead to increased cell fitness. We previously established in a murine model that 5q loss of heterozygosity leads to a loss of function of *KIBRA*, which encodes a multi-domain scaffold protein that inhibits oncogenic transcriptional co-activators YAP/TAZ that mediate mechanotransduction signals [21]. Chr5q also harbors *RAD17*, *RAD50* and *RAP80* genes that are important for *BRCA1*-dependent DNA repair, and their loss impairs *BRCA1*-pathway function critical for DNA damage control, contributing to increased genomic instability and cancerous phenotype [57]. Since noncoding genes including lncRNA and miRNA reside within chr4p, it is possible that their loss leads to overexpression of certain target genes contributing to a proliferation state. The observation that chr4p loss is prevalent in *HER2* amplified group but is not associated with decreased survival (Fig. S1A&B), likely indicates that *HER2* amplification, a key event in these tumors, masks the effect, whereby the double mutant carrying both chr4p loss and *HER2* amplification resembles the more extreme phenotype of the single mutant *HER2* rather than the combined effect of both aberrations. The timing analysis showed a preferred order for chromosome arm-level CNA in basal breast cancer as reported in a study on gastric cancer organoids [30]. In addition, in both studies, these evolutionary early alterations were observed in similar chromosome arms, such as the loss of chr3p, chr4p and chr4q and suggest convergent mechanisms for regulating tumor emergence in gastric and breast cancer, which may be applicable to other cancers.

Our study revealed that *PGCA1* (*C4orf19*) is physically associated with PDCD10 and a subfamily of GCK-III: STK24, STK25, and STK26. It is possible that PGCA1 is a direct binder of PDCD10 since the likelihood of an interaction between two human proteins in a yeast two-hybrid assay when at least one is not nuclear, PDCD10, is very high to be direct [39]. PGCA1 is likely excluded from the STRIPAK complex, a large multiprotein assembly, the striatin-interacting phosphatase and kinase (STRIPAK) complex, which was initially characterized in HEK293 cells [58], since there is no evidence of striatins in the PGCA1 miniTurboID data or other publicly available protein-protein interaction data. This finding further supports the direct binding of PGCA1 to PDCD10 and suggests that PGCA1 competes with striatins for binding to PDCD10. The interface that is mediating this interaction between PGCA1 and PDCD10 may thus be the same as the interface required for binding of striatins to PDCD10 or other partners which was previously shown to be occurring in a mutually exclusive manner [41,59]. Thus, we propose that PGCA1 is a protein that can tether the GCK-III kinases through the PDCD10 adapter to the plasma membrane bridging the GCK-III kinases to a substrate in this locale.

PDCD10 is also known as *CCM3* is a causative gene of cerebral cavernous malformation, a neurovascular disease that is characterized by vascular malformations [60]. In addition to interacting with and controlling signals emanating from the CCM2/CCM1 pathway in cytoskeletal organization, PDCD10 also regulates STRIPAK, and potentially other pathways also implicated in vascular integrity [60]. PDCD10 is also pro-apoptotic and controls the cell cycle, entry into senescence, apoptotic response to oxidative damage, inflammation and DNA damage repair as well as cell migration [60]. Its many functions are thought to be enabled by its multiple subcellular localizations, including cell-to-cell junctions and the Golgi apparatus. Our findings suggest that since PGCA1 is at the cell periphery, it likely interacts with PDCD10 at the cell-to-cell junctions. PDCD10 heterodimerizes with GCK-III kinase subfamily and modulates cell migration by regulating Golgi assembly which is mediated by its interaction with STK25 [61], regulates exocytosis through its interaction with STK24, which when lost results in oxidative damage and dismantling of the adherens junctions [62], as well as maintains ion homeostasis through its interaction with STK26 [63]. Thus, PGCA1, through its interaction with PDCD10, may tether GCK-III: STK24, STK25 and STK26 to the membrane. This could affect diverse processes that decrease cell proliferation when *PGCA1* is overexpressed and contribute to the proliferation and mesenchymal transcriptomic signatures observed in cell populations with chr4p loss. The similarity of the context-dependent suppression of proliferation of chr4p genes to chr13q, a region deleted in ~45% basal breast cancers, which also harbors a PDCD10 heterodimerization partner, STK24, highlights the important role of PDCD10-kinase module stoichiometric balance in exerting selection pressures on copy number evolution of breast cancer genome.

It has been previously proposed that the cancer genome is shaped by sensitivity to a change in gene dosage caused by chromosome arm loss or gain [64]. The gene dosage balance hypothesis postulates that the balance in the ratio of oncogenes to tumor suppressor genes exerts a selective pressure on the cancer cell. Thus, chromosome arm loss will be favoured if the number of tumor suppressor genes is higher than oncogenes and *vice versa* in the case of chromosome arm gain. Our observation of context dependent suppression of proliferation of a frequently lost chromosome arm suggests that additional mechanisms exist that maintain cancer cells with specific chromosome arm losses. A recent pan-cancer evolutionary study noted recurrent early genetic events and the broadening of this set in later stages, suggesting a preference for these genomic changes in early tumor evolutionary stages and potential genetic interactions that constrain the evolution [17]. This is consistent with previous studies that suggested that despite aneuploidy resulting in a growth disadvantage due to proteotoxic and metabolic stress, it may lead to increased selective pressure on cells to acquire growth-promoting genetic alterations and genetic interactions between aneuploidies have been recently suggested to be involved in cancer genome evolution and have been reported in model organisms[65-68].

The suppression of proliferation of chr4p genes when overexpressed suggests a mechanism of negative selection of chr4p gain in cancers. We showed that chr4p loss was broadly observed across multiple cancer types. The recurrence of chr4p loss in ovarian serous carcinoma is not surprising since this cancer type shares many molecular features and was suggested to have similar etiology to basal breast cancer [8]. A recent pan-cancer analysis of cancer aneuploidy similarly detected chr4p loss in squamous, gynecological and gastrointestinal tumors [24]. Since chr4p loss contains multiple genes that promote tumorigenesis when co-deleted, their simultaneous loss may result in vulnerabilities that cannot be identified by studying single genes and thus could provide potential novel therapeutic avenues for patients with triple-negative breast cancer and other cancers.

Figure 1. Loss of chromosome 4p in basal breast cancer is recurrent and functionally significant. (A) Experimental and analytic pipeline. **(B)** The Cancer Genome Atlas (TCGA) Invasive Breast Carcinoma single nucleotide polymorphism (SNP) array dataset was used to investigate the frequency of chromosome arm losses among the basal subtypes. The three most frequent chromosomal arm losses in basal breast cancer are shown, whereby chr4p loss occurs in ~65% patients. **(C)** Regions of chr4p loss span a large fraction of the chromosome 4p. Dark blue denotes stringent threshold deletion segmented mean < -0.3, light blue denotes lenient threshold -0.3 < deletion segmented mean < -0.1, light red denotes lenient threshold 0.1 < deletion segmented mean < 0.3, red denotes stringent threshold deletion segmented mean > 0.3, white denotes copy neutral state. **(D)** TCGA basal breast cancer gene expression dataset was used to show that ~80% of genes along chr4p decrease in expression upon its copy number loss. **(E)** Overall survival of basal breast cancer patients with copy neutral or deletion status of chr4p, $p <$ 0.0997. **(F)** Chr4p copy number status across pan-cancer TCGA datasets. Gene Set Enrichment Analysis (GSEA) showing representative terms that are enriched for genes displaying **(G)** elevated or **(H)** decreased expression due to chr4p loss in TCGA basal breast cancer.

Figure 2. Chromosome 4p loss is an early event in basal breast cancer evolution. Basal breast cancer primary tumor / patient-derived xenograft (PT/PDX) panel was used for the phylogenetic reconstruction. **(A)** Aggregated single-sample ordering reveals typical timing of chromosome arm aberrations. Preferential ordering diagrams show probability distributions revealing uncertainty of timing for specific events in the cohort. The prevalence of the event type in the cohort is displayed as a bar plot on the right. **(B)** Timeline representing the length of time, in years, between the fertilized egg and the median age of diagnosis for breast cancer. Real-time estimates for major events, such as whole genome doubling (WGD) and the emergence of the most recent common ancestor (MRCA), are used to define early, variable, late and subclonal stages of tumor evolution approximately in chronological time. Driver mutations and copy number alterations (CNA) are shown in each stage according to their preferential timing, as defined by relative ordering. Events occurring in > 40% of all cases are depicted. **(C)** An example of individual patient (PT/PDX1735) trajectory (partial ordering relationships), the constituent data for the ordering model process.

Figure 3. Chromosome 4p loss is associated with a proliferative state. (A) Single cell RNA sequencing data of PDX1735 (left panel) from a previous study[32] were used to infer copy number status (right panel) and displayed using tSNE plots. Groups of cells with shared pattern of gene expression profiles or inferred copy number profiles are colored in the same color. Heatmap (bottom panel) shows the inferred copy number profile of four communities which identified three communities (1-3) harboring chr4p deletion and community 4 harboring chr4p copy neutral state. chr denotes chromosome, dashed line denotes centromere, solid line denotes start/end of chromosome. Loss is in blue, copy neutral is in white, gain is in red. Likelihood of inferred copy number change is represented with Wilcoxon test $-\log_{10} p$ value, with darker shade reflecting higher confidence. **(B)** Frequency of cells inferred to harbor chromosome 4p deletion or copy neutral state across different cellular clusters with distinct transcriptional programs. The

"transcriptional program" category received a count for any combination in which a cell belonged both to a specific inferred copy number community and a specific gene expression cluster. The size of the circle assigned to each "transcriptional program" element reflects the fold increase over the background fraction of all cells in a specific gene expression cluster. Significance was assessed with a hypergeometric test; $p < 0.05$. Solid black circles represent significant depletion; open black circles represent significant enrichment; grey denote no significant change. **(C)** Distribution of cells across (left panel) cell cycle phases (light grey denotes G1 phase, medium grey denotes S phase, black denotes G2/M phase) and (right panel) Ki67 gene expression based on cell clusters with distinct transcriptional programs as shown in A. **(D)** Staining of paraffin embedded fixed tissue section of PDX1735 using a combination of immunofluorescence (IF) for Ki67 (marker of proliferation transcriptional program cell cluster), RNA-*in situ* hybridization (ISH) for *RBPJ* (marker of chr4p), DAPI staining for nuclei, pan Cytokeratin (PanCK) IF for epithelial cancer cells and hematoxylin and eosin (H&E) staining for cancer histology. Significance was assessed by Fisher exact test.

Figure 4. Overexpression of chromosome 4p genes leads to context-dependent suppression of proliferation. (A) A schematic of gene overexpression strategy. Chr4p copy neutral normal breast epithelial cell line (MCF10A), chr4p copy neutral basal breast cancer patient derived xenograft (PDX) -derived cell line (GCRC1915) or chr4p deletion basal breast cancer PDX-derived cell line (GCRC1735) were used to generate stable cell populations using lentivirus-mediated integration. The resulting mutant cell line populations overexpressed candidate genes residing within chr4p region, which is found in a high confidence chr4p deletion region in GCRC1735 according to whole exome sequencing (WES) from a previous study [32]. Lentiviral constructs for single and dual gene overexpression were obtained from the ORF collection[35]. **(B)** Chr4p gene overexpression confers a proliferation defect in GCRC1735 with chr4p loss but not in chr4p copy neutral cell lines (MCF10A, GCRC1915). Dual overexpression exacerbates the proliferation defect observed in GCRC1735. Blue denotes proliferation defect, black no change relative to control. **(C)** Schematic of the genome-wide pooled lentiORF overexpression screen. MCF10A and GCRC1735 cell lines were transduced with pooled lentiORF library at multiplicity of infection (MOI) of 0.3. After puromycin selection cells were maintained for 6-8 doublings and 1000x coverage was maintained at each step of the experiment. Next generation sequencing (NSG) was used to capture barcode abundance which served as a proxy for cell growth rate. **(D)** Genome-wide pooled lentiORF overexpression screen revealed genomic regions with context dependent suppression of proliferation. For example, there was a growth defect in GCRC1735 and not in MCF10A, including chr4p and 13q that are deleted in GCRC1735 and not in MCF10A. Blue denotes proliferation defect, black no change relative to T0 control.

Figure 5. C4ORF19 (PGCA1) **is associated with the PDCD10-GCK-III module. (A)** Summary of co-IP assay results from (B), miniTurbo ID conducted in MCF10A cells expressing C4rf19-miniTurbo and literature curation using BioGRID. **(B)** Western blot using whole cell lysate shows heterologous expression of C4orf19-v5 and PDCD10-GCK-III module 3xFLAG-PDCD10, 3xFLAG-STK25 and 3xFLAG-STK26 in MCF10A cells. **(C)** Co-immunoprecipitation assay using Anti-V5 for pull-down shows the presence of 3xFLAG-PDCD10, 3xFLAG-STK25 and 3xFLAG-STK26 in MCF10A cells indicating that C4orf19 interacts with PDCD10 and associated GCK-III kinases. **(D)** Brightfield microscopy images of MCF10A cells overexpressing GFP control or C4orf19 48 hrs post induction with doxycycline. **(E)** Analysis of

gene ontology molecular function (GO MF) of proteins in proximity to C4orf19 from miniTurbo ID from (C) shows enrichment of proteins at the plasma membrane suggesting that C4rof19 is localized to the cell periphery. **(F)** Immunofluorescence of C4orf19 indicates that its subcellular localization is at the cell periphery. Signal intensity relative to area of the inner or outer cell section indicates that C4orf19 abundance is higher at the cell periphery, n = 27. Significance was assessed using Wilcoxon rank sum test.

Acknowledgements: We thank Anie Monast and Virginie Pilon for assistance with animal experiments, Valentina Munoz Ramos for lab organization and management, Margarita Souleimanova and Nicholas Bertos with handling mammoplasty reduction surgery tissue, Chris Law and the Center for Microscopy and Cellular Imaging at Concordia University for guidance and help with the image analysis. Single cell RNA and DNA sequencing library preparation and sequencing was performed by Yu Chang Wang at the McGill Genome Centre, confocal microscopy was performed at the Centre for Microscopy and Cell Imaging, Concordia University and Advanced BioImaging Facility, McGill University. Histology sections were prepared and imaged in the Histology Innovation Platform, lentiviral ORF collections were obtained from McGill Platform for Cell Perturbation and flow cytometry was conducted at the Flow Cytometry Core Facility, Goodman Cancer Institute. Proteomics was performed at the Network Biology Collaborative Centre at the Lunenfeld-Tanenbaum Research Institute (supported by the Canada Foundation for Innovation, the Ontario Government, Genome Canada, and Ontario Genomics [OGI-139]). The breast tissue bank at McGill University (from which PDXs and cell lines were generated) is supported by the Réseau de Recherche en Cancer of the Fonds de Recherche du Québec-Santé and the Québec Breast Cancer Foundation (to M.P.). This work was supported by the Cancer Research Society (M.P.), the Canadian Institutes for Health Research (FDN 143301 to A.-C.G. and FDN 143281 to M.P.), the Terry Fox Research Institute (to A.-C.G.), Genome Canada Genomic Platform grant (CFI 33406 and 40104 to J.R.; CFI MSI 35444 to J.R.) and Worldwide Cancer Research (16-0402 to M.P.). T.M.B., T.L. and P.V.L. were supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (CC2008), the UK Medical Research Council (CC2008), and the Wellcome Trust (CC2008). For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute. P.V.L. is a CPRIT Scholar in Cancer Research and acknowledges CPRIT grant support (RR210006). Additional support was provided by Fonds de Recherche du Quebec-Sante Postdoctoral Fellowship to E.K., L'Oréal Canada Women in Science Research Excellence Fellowship to E.K., Canadian Institutes of Health Research Banting Postdoctoral Fellowship to E.K. and Cancer Research Society Scholarship for the Next Generation of Scientists Postdoctoral Fellowship to E.K., and CIHR Doctoral Fellowship to K.T.A. T.M.B. is supported by a PhD fellowship from Boehringer Ingelheim Fonds.

Author contributions: Conceptualization: E.K., and M.P.; Methodology and investigation: E.K., K.T.A., P.P.C., M.S., D.Z., G.M., H.K., A.M.F. and J.R.; Formal analysis: E.K., T.M. B., T.L., J.M., K.A., G.M., A.P., Y.Y., J.B., R.L., M.B., M.C.G., A.O., Q.M., C.K., S.H., A.C.G., J.R., G.B., P.V.L., and M.P.; Resources: C.R.M., P.G.D.; Writing – original draft: E.K., and M.P.; Writing – review and editing: E.K., T.B., T.L., J.M., K.A., P.P.C., C.M., J.B., H.K., R.L., M.B., A.M.F., A.O., Q.M., C.L.K., S.H., A.C.G, J.R., G.B., P.V.L., and M.P.; Supervision: M.P.; Funding acquisition: A.C.G., J.R., P.V.L. and M.P.

Declaration of interest: The authors declare no competing interest.

STAR Methods:

Isolation of Normal Breast Epithelial Single Cell Suspension

All tissue was collected with informed consent under REB-approved protocols at the McGill University Health Centre. Two patients age of 46 and 18 years old undergoing bilateral mammoplasty reduction due to hypertrophy of the breast with diagnosis and/or management at McGill University Health Centre, Montreal, QC, Canada were recruited for this study. 2,000–3,000 mm^3 surgically removed breast epithelial tissue was harvested and kept on ice in transport medium: RPMI 1640, 50 µg/mL gentamycin, 100 U/mL Pen/Strep, 2.5 µg/mL Fungizone until sample processing. The tissue dissociation was achieved as previously described [33]. Briefly, the tissue fragment was minced in ~1 mL cold DMEM and MidiMACS Starting Kit (LS) was used as per manufacturing instructions. Minced tissue was collected in a sterile gentleMACS C tube and enzyme A, enzyme R and enzyme H were added from the Tumor Dissociation Kit. GentleMACS Octo Dissociator with Heaters, program was run (1 h, mild speed, 37 °C) to begin the mechanical and enzymatic digestion process. The mix was incubated on ice for 3 min to allow for gravity sedimentation and the oily layer was aspirated. 3 ml of cell suspension was run through 70 µm strainer and collected. The remaining undigested tissue fragments were loaded into the gentleMACS Octo Dissociator, program (1 min, high speed, room temperature) to continue the mechanical and enzymatic digestion process. The cell suspension was passed through 70 µm strainer and collected. The strainer was washed with 10 mL PBS and combined with the filtrate, centrifuge for 10 min, 1,500 rpm. The cell pellet was resuspended in 500 µL Complete DMEM and Trypan Blue staining was used to quantify cell number and viability. Red blood cells were removed by aspirating the supernatant and adding 3 mL ACK lysing buffer and incubating at room temperature for 5 min. Then, 7 mL PBS were added and the suspension was centrifuged at 1,200 rpm for 4 min. The cell pellet was then resuspended in 500 µL complete DMEM. Single cell RNA sequencing was conducted if the cell viability exceeded 60%.

Cell culture

MCF10A cell line was obtained from the ATCC and cultured in DMEM/Ham's F12 medium, 20 ng/ml hEGF, 100 ng/ml cholera toxin, 10 µg/ml bovine insulin, 500 ng/ml hydrocortisone, 5% horse serum (HS), 50 µg/ml gentamicin. HEK293T cell line was cultured in DMEM with 10% fetal bovine serum (FBS).

Patient-derived xenograft derived cell lines (GCRC1735, GCRC1915) were isolated from the respective PDXs. PDX tumor fragments were minced and digested as previously described [33]. Cancer epithelial cells were established using a Conditional reprogramming protocol as previously described [69]. Briefly, after tumor fragments digestion, single-human epithelial cancer cells were transferred to a dish containing lethally irradiated 3T3-J2 cells (1 × 106 cells) and cultured with F- media (DMEM (Gibco) and F-12 Nutrient Mixture (Ham) (Gibco-) (1:4), 5% FBS (Life Technologies), 0.4 ug/mL Hydrocortisone (Sigma-Aldrich), 5 ug/mL Insulin (Gibco-), 8.4 ng/mL Cholera toxin (Sigma-Aldrich), 10 ng/mL Epidermal growth factor (BPS bioscience), 10 umol/L Y-27632 (Abmole), 50 ug/mL Gentamicin (Gibco), 1% P/S (Thermo Fisher Scientific,), Amphotericin B (1 ug/ml) (Thermo Fisher Scientific). After five passages of coculture, murine irradiated 3T3-J2 cells were removed using a Feeder Removal MicroBeads kit (Miltenyi), and epithelial cancer cells were expended in F-media. PDX-derived cell lines were cultured in F media: 5% fetal bovine serum, 400 ng/ml hydrocortisone, 5 μg/ml insulin, 8.4 ng/ml cholera toxin, 10 ng/ml hEGF, 10 μM Y-27632 (ROCK inhibitor), 50 μg/ml gentamicin.

All cell lines used were routinely tested for Mycoplasma (Lonza Mycoalert and EZ-PCR Mycoplasma Detection Kit) and were authenticated using short tandem repeat analysis. The human origin of PDX-derived cell lines was validated by flow cytometry using FITC anti-human EpCAM antibody clone VU-1D9 and (Thermo Fisher Scientific, #A15755) and PE/Cy7 anti-mouse H2Kd antibody clone SF1-1.1 (Biolegend, # 116622). All cells were maintained at 37°C, 5% CO_2.

Single and dual-gene overexpression assay

For generation of stable *c4orf19, RBPJ, SEPSECS, SEL1L3, KCNIP4, TBC1D19, RELL1, SLC34A2, STIM2, LGI2, PI4K2B* and *CCKAR* overexpression cells, lentiviral ORF vectors were retrieved from the arrayed MGC premier human lentiviral ORF (Sigma) (ccsb ID, blasticidin resistant) and MISSION® TRC3 Human ORF collection (Sigma) (TRCN ID, puromycin resistant) obtained from the McGill Platform for Cell Perturbation (MPCP). The following lentiviral ORF vectors were used: *c4orf19* (ccsbBroad304_03572, TRCN0000469204), *RBPJ* (ccsbBroad304_06435, TRCN0000470066), *SEPSECS* (ccsbBroad304_11945), *SEL1L3* (ccsbBroad304_11701, TRCN0000479888), *KCNIP4* (ccsbBroad304_09030, TRCN0000474912), *TBC1D19* (TRCN0000468467), *RELL1* (TRCN0000476648), *SLC34A2* (TRCN0000476745), *STIM2* (TRCN0000477969), *LGI2* (TRCN0000481617), *PI4K2B* (TRCN0000489163), *CCKAR* (TRCN0000489014, TRCN0000491970), *CDKN1A* (ccsbBroad304_00282, TRCN0000471863), *CDKN1B* (ccsbBroad304_05980, TRCN0000475049) and GFP control vector pLX317-GFP and pLX304-GFP. Viral particles were produced by co-expressing ORF or control constructs with packaging plasmids psPAX2 and pMD2.G in HEK-293T cells using lipofectamine 2000 transfection protocol. Media containing viral particles was collected and passed through a 0.45 μm filter. Cells were treated with virus in media containing 8 μg/ml polybrene. Twenty-four hours after transduction cells were recovered for another 24 hrs and then MCF10A were selected in 3 μg/ml puromycin dihydrochloride (Sigma) for 48 hrs or 10 μg/ml blasticidin (Gibco) for 72 hrs for; GCRC1735 in 5 μg/ml puromycin dihydrochloride for 48 hrs or 7.5 μg/ml blasticidin for 72 hrs and GCRC1915 in 5 μg/ml puromycin dihydrochloride for 48 hrs or 10 μg/ml blasticidin for 72 hrs. MOI was determined for each construct for each cell line and ~ 0.3 MOI was used for all constructs ensuring one integrant per cell. Viral transductions with the respective vectors were carried out sequentially. Overexpression was confirmed by Western Blot using Anti-V5 antibody (Abcam #27671, 1:1000). All single gene overexpression mutant cells were constructed such that they overexpressed each ORF under one of the selection markers and GFP was then under the second selection marker. All double gene overexpression mutant cells were constructed such that they overexpressed each ORF under one of the selection markers and the second ORF was then under the second selection marker. Confluence was measured by IncuCyte. Briefly, 4000 MCF10A and GCRC1915 cells and 2000 GCRC1735 cells were plated per well representing 20% confluence at the start of the experiment. Imaging was done at 4 hr intervals for a duration of 5-6 days. Each mutant cell line was plated in three wells per plate for a total of three technical replicates. The experiment was repeated for a total of three independent biological replicates. *CDKN1A* and *CDKN1B* were included as positive controls, which are known to lead to a severe proliferation defect when overexpressed[19].

Genome-wide pooled overexpression screen

The pooled MISSION TRC3 LentiORF collection (Sigma) provided by the McGill Platform for Cell Perturbation (MPCP) was used to infect MCF10A (10^8) and GCRC1735 ($1.5x10^8$) cell lines. Cells were treated with virus in media containing 8 µg/ml polybrene for 24 hrs. Viral supernatant was removed and the media was refreshed recovering the cells for 48 hrs and then MCF10A and GCRC1735 were selected in 3 or 5 µg/ml puromycin dihydrochloride for 48 hrs, respectively. MOI ~ 0.3 MOI was used for the screens and 1000x coverage was maintained at each step of the screen for both cell lines. Following 6-8 doublings, genomic DNA was isolated using the Roche High Pure PCR Template Preparation Kit followed by an RNase A treatment. One microgram of DNA was then used in 48 2-step PCR reactions with barcoded Illumina sequencing primers and then with P5/P7 primers. The reactions were then purified using the Roche PCR Purification Kit. Samples were then sequenced at The Center for Applied Genomics at Toronto Sick-Kids hospital on the Illumina HiSeq 2500 platform. The 50-base kit with 62 cycles and single-end reads was used to obtain the exact read-length needed for the library vector. Sequences were then deconvoluted. For all downstream analyses, we only included genes with a read count higher than 100 in T0 samples (MCF10A_T0 and GCRC1735_T0). Raw counts were normalized using edgeR's TMM algorithm (Robinson et al., 2010) and were then transformed to log2-counts per million (logCPM) using the voom function implemented in the limma R package (Ritchie et al., 2015). To assess differences in gene expression levels, we fitted a linear model using limma's lmfit function. Nominal p-values were corrected for multiple testing using the Benjamini-Hochberg method. Genomic heatmaps of log2 fold-changes were created using CNVkit (Talevich et al., 2016).

TCGA data computational analysis

Patient copy number data were obtained from a TCGA Breast Invasive Carcinoma (BRCA) (n = 2199) using Firehose Broad GDAC (https://gdac.broadinstitute.org/; accessed on 31 July 2016). Frequencies of gene deletions were derived from the single nucleotide polymorphism array dataset (genome_wide_snp_6-segmented_scna_minus_germline_cnv_hg19) and analyzed by GISTIC2.0 (Mermel et al., 2011). Parameters used for analysis were: reference genome build hg19; amplification threshold 0.1; deletion threshold -0.1; join segment size 4; qv threshold 0.25; remove X chromosome yes; cap value 1.5; confidence level 95; broad analysis yes; broad length cut-off 0.5; maximum samples per segments per sample 2000; arm peel-off yes. PAM50 annotation was obtained from a previous study [8]. TP53 mutations were obtained from cBioPortal and missense mutations were annotated by IARC filename: functionalAssessmentIARC TP53 Database, R18.xlsx. LOF mutations were considered: Frame_Shift_Del, Frame_Shift_Ins, Nonsense_Mutation, Splice_Site and Missense_Mutation if there were more cases of LOF than GOF.

Patient copy number data were obtained from a TCGA Lung Squamous Cell Carcinoma (LUSC, n = 1032)**,** Testicular Germ Cell Tumors (TGCT, n = 304), Ovarian Serous Carcinoma (OV, n = 1168), Esophageal carcinoma (ESCA, n = 373), Cervical Squamous Cell Carcinoma and

Endocervical Adenocarcinoma (CESC, n = 586), Mesothelioma (MESO, n = 172), Rectum adenocarcinoma (READ, n = 316), Stomach Adenocarcinoma (STAD, n = 904), Breast Invasive Carcinoma (BRCA, n = 2199), Colon Adenocarcinoma and Rectum Adenocarcinoma (COADREAD, n = 1234), Colon Adenocarcinoma (COAD, n = 918), Head and Neck Squamous Cell Carcinoma (HNSC, n = 1089) and Uterine Corpus Endometrial Carcinoma (UCEC, n = 1089) (https://gdac.broadinstitute.org/; accessed on 31 July 2016). Frequencies of gene deletions were derived from the single nucleotide polymorphism array dataset (genome_wide_snp_6-segmented_scna_minus_germline_cnv_hg19) and analyzed by GISTIC2.0 (Mermel et al., 2011). Parameters used for analysis were: reference genome build hg19; amplification threshold 0.1; deletion threshold -0.1; join segment size 4; qv threshold 0.25; remove X chromosome yes; cap value 1.5; confidence level 95; broad analysis yes; broad length cut-off 0.5; maximum samples per segments per sample 2000; arm peel-off yes. Higher amplification and deletion thresholds than above were used to increase stringency for the pan-cancer analysis.

TCGA gene expression data set from breast cancer invasive ductal carcinoma was used for differential gene expression. For all downstream analyses, excluded lowly expressed genes with an average read count lower than 10 were excluded from all samples. Raw counts were normalized using edgeR's TMM algorithm and were then transformed to log2-counts per million (logCPM) using the voom function implemented in the limma R package. To assess differences in gene expression levels, we fitted a linear model using limma's lmfit function. Nominal p-values were corrected for multiple testing using the Benjamini-Hochberg method. Gene-Set enrichment analysis based on pre-ranked gene list was performed using the R package fgsea (http://bioconductor.org/packages/fgsea/). Default parameters were used.

Immunoprecipitation and Western Blot

MCF10A cells were plate at a density of 1.5x10^6/10cm-dish and transfected with 4µg of candidate interactors using Lipofectamine (ThermoFisher, 18324012) and Plus (ThermoFisher, 11514015) transfection reagents. For this DNA constructs were incubated with 8µl Plus reagent in 500µl Opti-MEM (ThermoFisher,11058-021), Lipofectamine reagent was incubated separately in another 500µl Opti-MEM, for 15min; following initial incubation, the solutions were mixed and incubated together for another 15min. The transfection mixture was added dropwise to pre-washed cells containing 2ml Opti-MEM, and incubated for another 3h at 37˚C, at which point the transfection solution was removed, and cells were returned to normal growth media. Following 24hrs, cells were harvested in lysis buffer (50 mm HEPES, 150 mm NaCl, 1.5 mm $MgCl_2$, 1 mm EGTA, 1% Triton X-100, 10% glycerol 1 mM PMSF, 1 mM Na_3VO_4, 1 mM NaF, 10 µg/ml aprotinin and 10 µg/ml leupeptin, pH 7.4). Lysates were pre-cleared with 30µl of either protein-A-sepharose (GE Healthcare, 17-5280-01) or protein-G-sepharose beads (GE Healthcare, 17-0618-01) for 1h at 4˚C. 1500µg of protein was then incubated with either 1.8µl (~*5µg*) V5-tag antibody (Abcam, ab27671), 1.25µl (~*5µg*) FLAG-tag antibody (Sigma, F3165) or 1.5µl (~*5µg*) mouse-IgG negative control and either 40µl of protein-A or protein-G-sepharose beads overnight at 4°C. Beads with bound proteins were washed three times in lysis buffer plus inhibitors, and eluted by boiling in SDS sample buffer. Eluted proteins and 50 µg of protein from whole cell lysate were resolved in 4-15% NuPAGE gradient gel (ThermoFisher, NP0335) using MOPS running buffer (ThermoFisher, NP000102). Proteins were transferred on PVDF Odyssey membranes (MilliporeSigma) using a Mini Trans-Blot System from Bio-Rad. Detection and quantification of protein levels was performed on the Odyssey IR imaging System (Li-COR

Biosciences) using fluorescently labeled secondary antibodies, anti-mouse-680 (Mandel Scientific, LIC-926-68070) or anti-rabbit-800 (Mandel Scientific, LIC-926-32211).

miniTurboID

Gateway cloning was used to clone c4orf19 (ccsbBroadEn_03572) from pDONR223 to pSTV6-miniTurbo. MCF10A cells were transduced with lentivirus backbone containing pSTV6-C4orf19-3xFLAG-miniTurbo or pSTV6-GFP-3xFLAG-miniTurbo and selected in media containing puromycin as previously described in the "Cell Culture" section. Western blot was conducted to ensure C4orf19-3xFLAG-miniTurbo expression and biotinylation as described above. Anti-GAPDH primary antibody (Santa Cruz Biotechnology #sc-25778; 1:1000) was used as described above. Streptavidin-HRP conjugate (Millipore Sigma #RPN1231VS, 1:5000) was used and visualized using Immobilon Forte Western HRP substrate (Millipore Sigma #WBLUF0500).

Cells were grown to ~70% confluency and bait expression and biotin labeling was induced simultaneously (0.5 µg/ml doxycycline, 40 µM biotin). After 4 h, cells were rinsed and scraped into 1 mL of PBS. Cells were collected by centrifugation (500 × g for 3 min) and stored at −80 °C until further processing.

Cell pellets were thawed on ice and resuspended in lysis buffer containing 50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1% Nonidet P-40 substitute (NP40; IGEPAL-630), 0.4% SDS, 1 mM $MgCl_2$, 1 mM EGTA, 0.5 mM EDTA, 0.4 % sodium deoxycholate, benzonase & protease inhibitors at a ratio of 10:1 (w/v). Cells were lysed with 15 seconds of sonication (5 sec on, 3 sec off) at 30% amplitude on a Q500 Sonicator with an 1/8" Microtip and were rotated end-over-end at 4 °C for 20 min. Cell debris was pelleted via centrifugation at 15,000 × g for 15 min at 4 °C. Supernatants were incubated with 25 µL (packed bead volume) of streptavidin-Sepharose beads (GE) with rotation for 3 hr at 4 °C. Beads were pelleted at 500 × g for 2 min, transferred to new tubes and resuspended in 500 µL of fresh lysis buffer.

Beads were washed once with SDS wash buffer (50 mM Tris-HCl, pH 7.5, 2% SDS), twice with lysis buffer, once with TNNE wash buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.1 % NP-40), and thrice with 50 mM ammonium bicarbonate, pH 8.0 (ABC). Each wash consisted of bead resuspension in 500 µl of each buffer, pelleting of beads at 500 × g for 30 sec and aspiration of supernatant. On-bead digestion was performed by resuspending beads in 100 µL of ABC containing 1 µg of sequencing grade trypsin (T6567, Sigma-Aldrich). Samples were gently mixed at 37 °C overnight. Samples were spiked with 1 µg of fresh trypsin and digested further for 3 hr. The supernatant, containing digested peptides, was transferred to new tubes. Beads were washed twice with HPLC-grade water to wash off peptides, and these were pooled with the collected supernatant. Peptides were vacuum centrifuged until dry.

Mass spectrometry acquisition

Each sample was resuspended in 5 % formic acid and loaded onto an equilibrated high-performance liquid chromatography column (800 nL/min). Peptides were eluted with a 90 min gradient generated by a Eksigent ekspert™ nanoLC 425 (Eksigent, Dublin CA) nano-pump and analyzed on a TripleTOF™ 6600 instrument (AB SCIEX, Ontario, Canada).

The MS acquisition method has been described previously on identical instrumentation [70]. The gradient was delivered at 400 nL/min and consisted of three steps: sample delivery, column cleanup and column equilibration. The gradient used to pass sample over the column took place over 90 min starting with 2% acetonitrile (ACN) + 0.1 % formic acid (FA) and ending with 35 % ACN + 0.1 % FA. Cleanup was performed by passing 80 % ACN + 0.1 % FA over the column for 15 min, and the column was equilibrated back to 2 % ACN + 0.1 % FA over 15 min.

Instrument calibration was performed on bovine serum albumin reference ions to adjust for mass drift and verify peak intensity before samples were analyzed in data-dependent acquisition (DDA) mode. One 250 ms MS1 TOF (time of flight) survey scan (over mass range 400 - 1800 Da) was performed and was followed by 10 × 100 ms MS2 candidate ion scans (100 - 1800 Da). Ions that exceeded a threshold of 300 counts per second and had a charge of 2+ to 5+ were selected for MS2. Precursors were excluded for 7 sec after one occurrence.

Data-dependent acquisition data search

The ProHits laboratory information management system was used to analyze proteomics data [71]. WIFF files were converted with the WIFF2MGF converter and to a mzML format using ProteoWizard (V3.0.10702) and the AB SCIEX MS data converter (V1.3 beta). Converted files were searched with Mascot (2.3.02) [72] & Comet (2016.01 rev.2) [73]. Spectra were searched against a collection of 72,482 entries comprised the following: human and adenovirus sequences (version 57, January 30th, 2013), common contaminants [Max Planck Institute (http://maxquant.org/contaminants.zip) & Global Proteome Machine (GPM; ftp://ftp.thegpm.org/fasta/cRAP/crap.fasta)], reversed sequences, bait tags (eg. BirA or GFP) and streptavidin. Search parameters were set to search for trypsinized peptides allowing for two missed cleavages. For precursors, a mass tolerance of 35 parts per million was set, and peptides of +2 to +4 charges were allowed with a tolerance of ± 0.15 amu for fragment ions. Variable modifications included deamidated asparagine and glutamine as well as oxidized methionine. Search results were analyzed with the Trans-Proteomic Pipeline (v4.7 POLAR VORTEX rev 1) and iProphet pipeline [74].

SAINT analysis

An iProphet probability score > 0.95 and more than two unique peptides were required for protein identification. SAINTexpress [version 3.6.1 [75]] was used to score proximity interactions from DDA data using default parameters. Bait runs, run in biological duplicate, were compared against four negative control runs consisting of two miniTurbo-eGFP-only samples and two untransduced MCF10A samples. Control runs were not compressed for this analysis. Preys with a Bayesian false discovery rate < 5% were considered high-confidence proximity interactions. gProfiler (https://biit.cs.ut.ee/gprofiler/gost) was used to calculate enrichment of GO cellular component terms.

RNA *in situ* hybridization and immunofluorescence of GCRC1735

FFPE tissue was deparaffinized and underwent heat-mediated antigen retrieval in citrate buffer pH6.0 or EDTA buffer pH9.0. Slides were blocked with Power Block for 5 min at room temperature, and incubated with the primary antibody for 30 min at room temperature followed by washing with TBST (3 x 3min). Slides were incubated with secondary antibody-HRP for 30 min at room temperature, washing with TBST 3x 3min and stained with Opal fluorophore working solution for 10 min. This was followed by heat-mediated antibodies stripping to remove the primary and secondary antibodies in order to repeat additional rounds for labeling with other primary antibodies. The primary antibodies are against Ki67 (Ventana #790-2910) and Pan-Keratin (Cat# 760-2595, Ventana). The antibody specificity and dilution were tested before multiplex assay. Nuclei were stained with 0.5 ng/ml DAPI for 5 minutes at room temperature and counterstaining was done with Harris' hematoxylin. RNA *in situ* hybridization was performed using the RNAscope 2.5 HD Assay (cat#322360. ACD Bio) according to the manufacturer's instruction on FFPE PDX section. The probes used are Hs-RBPJ (cat#448661), the positive control Hs-PPIB housekeeping gene and the negative control dapB. Slides were imaged with an LSM800 confocal microscope (Zeiss). Brightfield slides were scanned using Aperio-XT slide scanner (Aperio). Visual inspection was used to classify cells into RBPJ or Ki67 high and low classes. Due to low number of RBPJ high expression cells a field with equal number of cells of both RBPJ expression classes was used for the quantification.

C4ORF19 immunofluorescence and quantification

C4orf19 immunofluorescence staining: cells were seeded in 24-well plate with coverslips until they reached 80-90% confluence. Then, they were fixed in 4% paraformaldehyde (20 min), permeabilized with 0.2% Triton X-100 (10 min), blocked with 2% BSA (30 min), and then incubated with Anti-C4orf19 primary antibody (1:100, GeneTex, GTX106538) (1 hr). The primary antibody was visualized with a fluorescent secondary antibody conjugated to Alexa Fluor 488 raised in goat (1:1000, Invitrogen Cat#) (1 hr). Nuclei were counterstained with 0.25 ng/ml DAPI (5 min). All steps were performed at room temperature. Images were acquired on the Nikon C2/TIRF confocal laser scanning microscope (Nikon), using a 63X objective.

Fiji (v.2.3, NIH) was used to analyze subcellular localization of C4ORF19 using a custom macro. The DIC image was used to manually outline each cell; DAPI was used to create a nuclear mask and GFP channel was used to quantify C4ORF19 subcellular localization. The macro makes bands of ~3μm from the edge of a cell outline into the middle of the cell, and measures the mean intensity, total amount of signal, the proportion of the cell's total area that is in this band, the proportion of the cell's total signal that is in this band, and the ratio of the signal-to-area. The ratio of the signal-to-area is above 1, if there is a greater proportion of the signal in that band than might be expected based solely upon area. The outer band representing ~ 25% of the cell area was compared to the remainder of the cell to quantify the protein abundance of C4ORF19 in the cell periphery compared to the cytoplasm.

Single cell RNA sequencing

Breast mammoplasty reduction epithelial single-cell suspensions were washed three times in PBS with 0.04% BSA. An aliquot of cells was used for LIVE/DEAD viability testing (Thermo Fisher Scientific). Single-cell libraries were generated using the Chromium Controller and Single Cell 3' Library & Gel Bead Kit v3 and Chip Kit (10x Genomics) according to the manufacturer's protocol. Briefly, cells suspended in reverse transcription reagents, along with gel beads, were segregated into aqueous nanoliter-scale gel bead-in-emulsions (GEMs). The GEMs were then reverse transcribed in a T1000 Thermal cycler (Bio-Rad) programed at 53ºC for 45 min, 85ºC for 5 min, and hold at 4ºC. After reverse transcription, single-cell droplets were broken and the single-strand cDNA was isolated and cleaned with Cleanup Mix containing DynaBeads (Thermo Fisher Scientific). cDNA was then amplified with a T1000 Thermal cycler programed at 98ºC for

3 min, 12 cycles of (98ºC for 15 s, 63ºC for 20 s, 72ºC for 1 min), 72ºC for 1 min, and hold at 4ºC. Subsequently, the amplified cDNA was fragmented, end-repaired, A-tailed and index adaptor ligated, with SPRIselect Reagent Kit (Beckman Coulter) with cleanup in between steps. Post-ligation product was amplified with a T1000 Thermal cycler programed at 98ºC for 45 s, 12 cycles of (98ºC for 20 s, 54ºC for 30 s, 72ºC for 20 s), 72ºC for 1 min, and hold at 4ºC. The sequencing-ready library was cleaned up with SPRIselect and quantified by qPCR (KAPA Biosystems Library Quantification Kit for Illumina platforms). 200 pM of sequencing libraries were loaded on an Illumina HiSeq instruments (see Single-cell RNA sequencing analysis section) and ran using the following parameter: 26 bp Read1, 8 bp I7 Index, 0 bp I5 Index and 98 bp Read2.

Single-cell RNA sequencing analysis

Two samples were sequenced using the Chromium single cell 3' RNA-seq on 0.5 lane the NovaSeq S1 instrument, for a total of 492,626,914 reads, and 327,762 reads per single-cell (saturation 89.2%). Alignment against the human GRCh38 genome was performed using Cell Ranger Pipeline version 3.0.1. The Ensembl annotation for GRCh38 (release 93) was used, keeping only the genes with the biotypes protein_coding, lincRNA and antisense. Empty GEMs containing only background reads were discarded by the pipeline and bar code errors resulting from sequencing were corrected if they contained only one mismatch by assigning them to the closest available bar code, or discarded otherwise, resulting in 1,503 GEMs containing cells. Alignment quality was controlled by assessing the proportion of reads mapping confidently to the transcriptome (52.4%). The total number of genes detected (>1 mapped read) was 20,541, with a median of 880 per cell. The R package Seurat (v3.2.3) was used to analyze the single-cell RNA-seq data (Satija et al., 2015). Cells with over 12% mitochondrial content, over 40,000 UMIs, or less than 500 UMIs were discarded. Gene counts were normalized to a total of 10,000 UMIs for each cell, and transformed to a natural log scale. Counts were then adjusted for library size and mitochondrial proportions. Heat Digestion Stress Response Gene Set as previosuly reported [76] was identified and removed. Cell types were annotated using previously defined markers [76] (Fig. 4S). Cells in the breast epithelial cell cluster were used as a baseline of normal gene expression for inferring copy number described below.

Inferring copy number aberrations from scRNAseq

Copy number profile was inferred from scRNAseq data as previously described [33]. Briefly, the gene expression was normalized to ensure cells are comparable, whereby the Trimmed-Mean M normalization rescale expression in a cell by a factor to match that of a control cell. Consecutive genes were merged to form "bins" with a minimum average expression. Each bin was normalized across cells to produce a Z-score which was computed by subtracting the average expression across all the cells and dividing by the standard deviation. The normalized expression was smoothed using a rolling median approach: the expression in each bin was replaced by the median expression of the surrounding bins in the cell. Specifically, the rolling median was run-in windows of size 5, i.e., the bin of interest and two bins on each side. The smoothed Z-score was winsorized to be within [-3,3] to further reduce the effect of single-gene outliers. After this step, the score is centered on 0 and positive (negative) values support a higher (lower) copy number than in the majority of the cells. A principal component analysis (PCA) was run on the smoothed Z-scores. To minimize the effect of cell cycle, the PCA was run on non-cycling cells and all the cells projected on this principal components (PCs).

Community detection was performed by the Louvain algorithm on a cell network built from the top PCs. A tSNE was run on the top 20 PCs. To build the cell network, we first identify the K-nearest neighbors of each cell based on the Euclidean distance D in the PC space. The K-nearest neighbor cells are linked in the network with a weight defined as 1/(1+D). The Louvain community detection was then run on this cell network. Gamma resolution parameter with a high mean Rand Index across the runs and/or a low Rand Index variance was used. Copy-number aberrations are called at the community level to increase the sensitivity to shorter aberrations. Meta-cells were constructed by combining the expression of randomly selected cells in a community. We created multiple meta-cells for each community and looked for consistent CNA signal in all meta-cells. The expression in each meta-cell was normalized similarly as for the CNA-based community detection: normalization per cell, merging into expressed bins, Z-score computation, and smoothing. The Z-score was computed relative to a specific baseline, e.g., cells identified as normal that were isolated from mammoplasty reduction. CNA were called using an HMM with three states: neutral copy number, loss, and gain. A Gaussian mixture HMM was capable of segmenting together the multiple meta-cells from a community. Short copy-number segments are filtered, for example if spanning less than 5 consecutive bins, as they could result from single genes with strong expression differences. A Wilcoxon test was performed to assess the significance of each loss/gain segment by comparing the expression within the segment with the expression in nearby "neutral" segments.

WGS analysis

Bulk WGS data for PT and PDX including BAM generation, Manta calls for structural variants and Mutect2 calls for somatic mutation variants for were obtained from a previous study [29].

Timing analysis using bulk WGS data for basal breast cancer PT/PDX panel

CNA profiles

CNA profiles were obtained using Battenberg (v2.2.9) [77], integrating SV calls from Manta and correcting logR for both GC content and replication timing. SNP phasing was performed using Beagle 5.1 (18May20.d20). To better capture LOH-related events in PDX samples, because BAF in pure samples tends to be highly squished and might be misinterpreted, purity was artificially decreased by pulling allele counts from both germline and PDX and was set back to 100% when checking sample purity with SNV information from Mutect2. All CNA profiles were manually examined and quality checked (*e.g.* homozygous deletions, superclonal peaks, purity estimates, etc.). Whole-genome doubling (WGD) information was assessed using the relationship between the fraction of the genome with LOH and ploidy, as in PCAWG studies [78].

CNA clustering

Clustering of CNA profiles was performed using MEDICC2 (v0.3) [79]. Genomic regions of more than 500kb and covered in all samples for a given patient were considered, with *cn_a* and *cn_b* defined as the major and the minor allele, respectively. The copy-number state of the most abundant subclone was selected for subclonal CNAs. Since WGDs were clonal events, we defined reference normals with a 1+1 baseline in samples without WGD and 2+2 with WGD so the output was WGD-aware.

Subclone trees

Subclonal compositions were assessed using DPClust (v2.2.8) [77] and SNVs information from Mutect2 calls, leveraging principles of reconstructing subclone trees [80]. Small and noisy clusters (<5% of total SNVs) breaking the pigeonhole principle were discarded from the final trees.

Genomic Event Timing

Clonal copy number gains were timed using an approach similar to that outlined in a previous study [17]. A posterior distribution over SNV multiplicities was measured using the emcee sampler (https://arxiv.org/abs/1202.3665) with a prior that corresponded to a uniform distribution over gain timing. For each segment we ran 30 independent chains for 2000 steps with 1000 burnin steps. The posterior distribution over SNV multiplicities was converted into a distribution over gain timing.

The timing of the WGD for WGD tumors was measured by jointly timing all the gains that resulted in a major copy number state of two. For WGD samples, the timing of gains leading to major copy number three and four states were measured with equal prior probability on a gain occurring before or after the WGD. The relative likelihood of pre- or post-WGD gains was calculated by measuring the similarity between the segment WGD timing distribution measured with the route compared to the sample-wide WGD timing distribution. For gain regions with major copy number four, the timing of the average of the two post-WGD gains was measured as the system is underdetermined. Only gained regions with a major copy number of up to four in WGD tumors and two in non-WGD tumors were timed. The gained regions also needed to have at least 10 SNVs and a minor copy number of no more than two. The timing of SNVs in key genes was measured using MutationTimeR [17].

The PDXs were used to refine our timing of events in the primary tumor. If an event was identified as clonal in the primary tumor, but was not found in the PDX, we reclassified the event as subclonal in the primary as the event was likely not present in the cells from the primary that seeded the PDX.

League Model

A league representing a timeline of genomic events aggregated across tumors were produced from our timing data using a league modeling approach similar to that outlined in previous studies[17,81]. Briefly, the aggregate timing of the events is determined by running a scoring process where the earliest events accumulate the higher score. This is achieved by initialising each genomic event with a score of zero. We then sample the relative timing for each possible pair of genomic events from the subset of individual tumor timelines in our cohort that contain both events. The earlier event has its score increased by one and the later event decreased by one. If the relative timing of the event pair cannot be distinguished, or if no sample has both events, the score for both events is kept the same. After each possible pair of events is considered, the events are ranked according to their score. This process is repeated 100 times to achieve a distribution over the ranks.

Real-Time Timing

A real-time estimate of WGD and the emergence of the MRCA was achieved using the approach outlined in a previous study [17]. Instead of evaluating the timing using both a branching and linear subclonal structure [17] we used the structure inferred from our subclone tree reconstruction.

scDNAseq

Basal breast cancer PDX-derived GCRC1735 single-cell suspensions were washed three times in PBS with 0.04% BSA. An aliquot of cells was used for LIVE/DEAD viability testing (Thermo Fisher Scientific). Single-cell DNA libraries were generated using the Chromium Single Cell DNA Reagent Kit (10X Genomics) according to the manufacturer's protocol. Briefly, an appropriate volume of cell suspension for targeting 500 cells were added to the Single Cell Bead Mix then loaded onto a Chromium Chip C, along with CB polymer. The resultant Cell-Bead was allowed to polymerize overnight shaken at 1000rpm. The encapsulated cells were then lysed and its genomic DNA was NaOH denatured. The Cell-Bead along with a reaction mix and Gel-Bead were loaded onto the Chromium D Chip to generate gel bead-in-emulsions (GEMs) on the Chromium Controller. An ideal GEM will contain reaction mix, one Cell-Bead and one Gel-Bead. The GEMs were incubated in a T1000 Thermal cycler (Bio-Rad) programed at 30ºC for 3hour, 16ºC for 5 hour, 65ºC for 10 min, and hold at 4ºC. Then the GEMs were broken and its amplified DNA were isloated using Dynabeads MyOne Silane beads followed by a SPRIselect cleanup. DNA was quantified on a Caliper Labchip (Beckman Coulter) using High Sensitivity DNA Assays. The DNA was converted to sequence ready library by fragmentation, end-repaired, A-tailing, index adaptor ligated and index PCR with SPRIselect clean ups in between. Four samples were sequenced on 2 lanes the Illumina HiSeqX instrument, for a total of 3,726,525,082 reads, and 505,266 reads per single-cell (saturation 15%).

scDNAseq data processing

Sequencing data was processed by using 10X Cell Ranger DNA pipeline to generate a raw bam for each sample. Briefly, the reads were aligned to the human reference genome build 38 (GRCh38) by using BWA and then converted to sorted BAM. The bam file was demultiplexed into individual bam files by using in-house python script to represent the sequencing reads from each single cell. Poorly mapped reads with mapping quality < 25 were filtered out by using SAMtools. PCR duplicates were removed by using Picard. Noisy cells detected by 10X Cell Ranger DNA pipeline with depth independent MAPD statistically higher than the sample distribution (with p-value < 0.01) and low ploidy confidence were excluded. We also filtered out cell outliers with large Lorenz curve area [82].

scDNAseq SNV analysis

To create a pseudo-bulk sample, single-cell DNA (scDNA) samples were merged while preserving the cell of origin information. This was achieved by incorporating the cell of origin information into the read group field of each read. The pseudo-bulk sample was then processed as a standard whole-genome sequencing (WGS) sample using the tumor_pair pipeline from Genpipes [83]. Somatic variants were generated using Mutect 2 and were utilized for single-nucleotide variant (SNV) fishing in individual cells. To reduce the false positive rate during the fishing process, we excluded indels and retained only high-quality somatic variants (TLOD >= 40). For each cell (represented by each read-group in the pseudo-bulk), we extracted the base distribution at each selected somatic position using BVAtools basefreq (https://bitbucket.org/mugqic/bvatools/src/master/). Cell-specific somatic variants were determined by comparing the extracted base frequency with the expected variant allele detected by Mutect 2. In order to ensure robustness, cell-specific somatic variants were excluded if they were not genotyped in at least 10 different cells.

Building the clone tree and assigning SNVs to clones using heuristics

SNVs were used to build the clone tree for GCRC1735 primary and PDX tumors. SNVs were independently called in four datasets: primary tumour bulk (PT_bulk), PDX bulk (PDX_bulk), single-cell samples 1 and 2 pseudobulk (SCS12) and single-cell samples 3 and 4 pseudobulk (SCS34). After examining the patterns of SNVs presence and absence in the samples, and based on the known ancestral relationships among the sample, the following heuristics were developed to assign mutations to clones (**Fig S3A**). First, if a mutation was present in all datasets, it was deemed a clonal mutation. Second, if a mutation was present in PT_bulk and SCS12 and not present in SCS34 or PDX_bulk, then it was assigned to subclone 1. Third, if a mutation is present in PT_bulk, PDX_bulk and SCS34 but not in SCS12, then it is set to subclone 2. Fourth, if a mutation is present only in PDX_bulk and SCS34 then it is set to subclone 3. Due to the low count of unique mutations to PT_bulk or PDX_bulk, we did not attempt to further define smaller subclones present in these samples.

Inference of haplotype-specific copy number profiles in single cells using CHISEL

CHISEL[84] is a tool that infers haplotype-specific copy number profiles of each cell in a low coverage single-cell sequencing dataset. This was used on the single-cell datasets to determine copy number profiles for subclones 1 and 3. To run CHISEL we first installed the package and dependencies as instructed on their GitHub page. Germline SNPs were called using bcftools' `mpileup` and `call` methods on the normal sample .baf file. These SNPs were phased using the Michigan Imputation Server (MIS). As required for MIS input, the SNP calls were separated by chromosome into separate .bcf files, sorted by genomic position, and uploaded to their server. For phasing on MIS, we used as arguments: reference panel `HRC r1.1 2016 (GRCh37/hg19)`, array build `GRCh38/hg38`, phasing `Eagle v2.4`, and mode `Quality control and phasing only`. The MIS results were downloaded as a set of chromosome-specific .vcf files and merged. The X chromosome was omitted during this merge as, at the time of running, sex chromosomes were not permitted in CHISEL's input. The MIS outputs phased SNP data using hg19 reference genome coordinates. Therefore, these coordinates were lifted over to the hg38 reference genome. This was done using picardtool's `LiftoverVcf` function with the requisite UCSF chain file. CHISEL also requires a .bam file for the single-cell sequencing data. In CHISEL, haplotypes are categorized as maternal or paternal arbitrarily, and so single-cell datasets could not be input to CHISEL separately. Instead, all single-cell data files were merged into a single .bam file and used as input. With these phased germline SNPs and the single-cell .bam file, CHISEL was run with default settings. This process was repeated using a different seed to confirm reproducibility.

Inferring the Most Recent Common Ancestor (MRCA) copy number profile

The results from CHISEL reveal two predominant subclones, each defined by the PDX line the cells are derived from. To estimate the haplotype-specific copy number profile of their MRCA, we developed and applied the following heuristics. First, if an allele's copy number was the same between the two subclones, that allele's copy number was set to the same for the MRCA. Second, if there was a LOH event in one subclone but not the other, the MRCA copy number state was set to contain the lost allele. Third, in other regions of differing copy number, if there was an adjacent region with shared copy number between the two subclones, and that region's copy number matched that of one of the two mismatched copy numbers, then the MRCA copy number for the differing region was set to be the same as the matching neighbour. If none of the previous three heuristics apply, then the MRCA copy number is set to be the minimum of the two subclonal copy numbers.

Supplementary tables

Table S1. GISTIC2.0 Broad deletion analysis in TCGA basal breast cancer cohort.
Table S2. Differential gene expression of chr4p genes in TCGA basal breast cancer cohort with deletion or copy neutral status of chr4p.
Table S3. GISTIC2.0 Broad deletion analysis in TCGA pancancer cohort.
Table S4. Transcriptome-wide differential gene expression in TCGA basal breast cancer cohort with deletion or copy neutral status of chr4p.
Table S5. Aneuploidy in basal breast cancer with different chr4p copy number states. Aneuploidy score as quantified by Chrom.Arm.SCNA.Level median reported by Davoli et al Science 2017.
Table S6. Phylogenetic reconstruction using bulk WGS data of PT/PDX basal breast cancer cohort.
Table S7. Phylogenetic reconstruction using scDNAseq data of PT/PDX GCRC1735 basal breast cancer.
Table S8. Gene expression and cell clusters as identified from scRNAseq of normal breast epithelial tissue.
Table S9. Relationship between transcriptional programs and inferred copy number changes using scRNAseq data of GCRC1735 PDX.
Table S10. miniTurboID screen for C4orf19.
Table S11. Gene overexpression screen for MCF10A and GCRC1335 PDX-derived cell line.

References

1. Hammond, M.E., Hayes, D.F., Dowsett, M., Allred, D.C., Hagerty, K.L., Badve, S., Fitzgibbons, P.L., Francis, G., Goldstein, N.S., Hayes, M., et al. (2010). American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. J Clin Oncol *28*, 2784-2795. 10.1200/JCO.2009.25.6529.
2. Haffty, B.G., Yang, Q., Reiss, M., Kearney, T., Higgins, S.A., Weidhaas, J., Harris, L., Hait, W., and Toppmeyer, D. (2006). Locoregional relapse and distant metastasis in conservatively managed triple negative early-stage breast cancer. J Clin Oncol *24*, 5652-5657. 10.1200/JCO.2006.06.5664.
3. Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., and Pietenpol, J.A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest *121*, 2750-2767. 10.1172/JCI45014.
4. Lehmann, B.D., Jovanovic, B., Chen, X., Estrada, M.V., Johnson, K.N., Shyr, Y., Moses, H.L., Sanders, M.E., and Pietenpol, J.A. (2016). Refinement of Triple-Negative Breast

Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. PLoS One *11*, e0157368. 10.1371/journal.pone.0157368.

5. Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A *98*, 10869-10874. 10.1073/pnas.191367098.
6. Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747-752. 10.1038/35021093.
7. Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol *27*, 1160-1167. 10.1200/JCO.2008.18.1370.
8. Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70. 10.1038/nature11412.
9. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346-352. 10.1038/nature10983.
10. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674. 10.1016/j.cell.2011.02.013.
11. Hanahan, D., and Coussens, L.M. (2012). Accessories to the crime: functions of cells recruited to the tumor microenvironment. Cancer Cell *21*, 309-322. 10.1016/j.ccr.2012.02.022.
12. Sotiriou, C., and Pusztai, L. (2009). Gene-expression signatures in breast cancer. N Engl J Med *360*, 790-800. 10.1056/NEJMra0801289.
13. Russnes, H.G., Navin, N., Hicks, J., and Borresen-Dale, A.L. (2011). Insight into the heterogeneity of breast cancer through next-generation sequencing. J Clin Invest *121*, 3810-3818. 10.1172/JCI57088.
14. Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400-404. 10.1038/nature11017.
15. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhsng, C.Z., Wala, J., Mermel, C.H., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat Genet *45*, 1134-1140. 10.1038/ng.2760.
16. Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899-905. 10.1038/nature08822.
17. Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Rosebrock, D., Mitchell, T.J., Rubanova, Y., Anur, P., Yu, K., et al. (2020). The evolutionary history of 2,658 cancers. Nature *578*, 122-128. 10.1038/s41586-019-1907-7.
18. Minussi, D.C., Nicholson, M.D., Ye, H., Davis, A., Wang, K., Baker, T., Tarabichi, M., Sei, E., Du, H., Rabbani, M., et al. (2021). Breast tumours maintain a reservoir of subclonal diversity during expansion. Nature *592*, 302-308. 10.1038/s41586-021-03357-x.

19. Sack, L.M., Davoli, T., Li, M.Z., Li, Y., Xu, Q., Naxerova, K., Wooten, E.C., Bernardi, R.J., Martin, T.D., Chen, T., et al. (2018). Profound Tissue Specificity in Proliferation Control Underlies Cancer Drivers and Aneuploidy Patterns. Cell. 10.1016/j.cell.2018.02.037.
20. Davoli, T., Uno, H., Wooten, E.C., and Elledge, S.J. (2017). Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science *355*. 10.1126/science.aaf8399.
21. Knight, J.F., Sung, V.Y.C., Kuzmin, E., Couzens, A.L., de Verteuil, D.A., Ratcliffe, C.D.H., Coelho, P.P., Johnson, R.M., Samavarchi-Tehrani, P., Gruosso, T., et al. (2018). KIBRA (WWC1) Is a Metastasis Suppressor Gene Affected by Chromosome 5q Loss in Triple-Negative Breast Cancer. Cell Rep *22*, 3191-3205. 10.1016/j.celrep.2018.02.095.
22. Cai, Y., Crowther, J., Pastor, T., Abbasi Asbagh, L., Baietti, M.F., De Troyer, M., Vazquez, I., Talebi, A., Renzi, F., Dehairs, J., et al. (2016). Loss of Chromosome 8p Governs Tumor Progression and Drug Response by Altering Lipid Metabolism. Cancer Cell *29*, 751-766. 10.1016/j.ccell.2016.04.003.
23. Liu, Y., Chen, C., Xu, Z., Scuoppo, C., Rillahan, C.D., Gao, J., Spitzer, B., Bosbach, B., Kastenhuber, E.R., Baslan, T., et al. (2016). Deletions linked to TP53 loss drive cancer through p53-independent mechanisms. Nature *531*, 471-475. 10.1038/nature17157.
24. Taylor, A.M., Shih, J., Ha, G., Gao, G.F., Zhang, X., Berger, A.C., Schumacher, S.E., Wang, C., Hu, H., Liu, J., et al. (2018). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell *33*, 676-689 e673. 10.1016/j.ccell.2018.03.007.
25. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer *18*, 696-705. 10.1038/s41568-018-0060-1.
26. Li, H., Cao, Y., Ma, J., Luo, L., and Ma, B. (2021). Expression and prognosis analysis of GINS subunits in human breast cancer. Medicine (Baltimore) *100*, e24827. 10.1097/MD.0000000000024827.
27. Zhang, S., Wu, H., Wang, K., and Liu, M. (2019). STK33/ERK2 signal pathway contribute the tumorigenesis of colorectal cancer HCT15 cells. Biosci Rep *39*. 10.1042/BSR20182351.
28. Lee, L.J., Papadopoli, D., Jewer, M., Del Rincon, S., Topisirovic, I., Lawrence, M.G., and Postovit, L.M. (2021). Cancer Plasticity: The Role of mRNA Translation. Trends Cancer *7*, 134-145. 10.1016/j.trecan.2020.09.005.
29. Savage, P., Pacis, A., Kuasne, H., Liu, L., Lai, D., Wan, A., Munoz-Ramos, V., Pilon, V., Monast, A., Zhao, H., et al. (2020). Chemogenomic profiling of breast cancer patient-derived xenografts reveals targetable vulnerabilities for difficult-to-treat tumors. Communications Biology.
30. Karlsson, K., Przybilla, M.J., Kotler, E., Khan, A., Xu, H., Karagyozova, K., Sockell, A., Wong, W.H., Liu, K., Mah, A., et al. (2023). Deterministic evolution and stringent selection during preneoplasia. Nature. 10.1038/s41586-023-06102-8.
31. Scriver, C.R. (2001). Human genetics: lessons from Quebec populations. Annu Rev Genomics Hum Genet *2*, 69-101. 10.1146/annurev.genom.2.1.69.
32. Savage, P., Blanchet-Cohen, A., Revil, T., Badescu, D., Saleh, S.M.I., Wang, Y.C., Zuo, D., Liu, L., Bertos, N.R., Munoz-Ramos, V., et al. (2017). A Targetable EGFR-

Dependent Tumor-Initiating Program in Breast Cancer. Cell Rep *21*, 1140-1149. 10.1016/j.celrep.2017.10.015.

33. Kuzmin, E., Monlong, J., Martinez, C., Kuasne, H., Kleinman, C.L., Ragoussis, J., Bourque, G., and Park, M. (2021). Inferring Copy Number from Triple-Negative Breast Cancer Patient Derived Xenograft scRNAseq Data Using scCNA. Methods Mol Biol *2381*, 285-303. 10.1007/978-1-0716-1740-3_16.
34. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189-196. 10.1126/science.aad0501.
35. Collaboration, O.R. (2016). The ORFeome Collaboration: a genome-scale human ORF-clone resource. Nat Methods *13*, 191-192. 10.1038/nmeth.3776.
36. Shen, L., Pugsley, L., Cencic, R., Wang, H., Robert, F., Naineni, S.K., Sahni, A., Morin, G., Zhang, W., Nijnik, A., et al. (2021). A forward genetic screen identifies modifiers of rocaglate responsiveness. Sci Rep *11*, 18516. 10.1038/s41598-021-97765-8.
37. Alliance of Genome Resources, C. (2022). Harmonizing model organism data in the Alliance of Genome Resources. Genetics *220*. 10.1093/genetics/iyac022.
38. Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., et al. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci *30*, 187-200. 10.1002/pro.3978.
39. Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402-408. 10.1038/s41586-020-2188-x.
40. Huttlin, E.L., Bruckner, R.J., Navarrete-Perea, J., Cannon, J.R., Baltier, K., Gebreab, F., Gygi, M.P., Thornock, A., Zarraga, G., Tam, S., et al. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell *184*, 3022-3040 e3028. 10.1016/j.cell.2021.04.011.
41. Ceccarelli, D.F., Laister, R.C., Mulligan, V.K., Kean, M.J., Goudreault, M., Scott, I.C., Derry, W.B., Chakrabartty, A., Gingras, A.C., and Sicheri, F. (2011). CCM3/PDCD10 heterodimerizes with germinal center kinase III (GCKIII) proteins using a mechanism analogous to CCM3 homodimerization. J Biol Chem *286*, 25056-25064. 10.1074/jbc.M110.213777.
42. Branon, T.C., Bosch, J.A., Sanchez, A.D., Udeshi, N.D., Svinkina, T., Carr, S.A., Feldman, J.L., Perrimon, N., and Ting, A.Y. (2018). Efficient proximity labeling in living cells and organisms with TurboID. Nat Biotechnol *36*, 880-887. 10.1038/nbt.4201.
43. Uhlen, M., Zhang, C., Lee, S., Sjostedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., et al. (2017). A pathology atlas of the human cancer transcriptome. Science *357*. 10.1126/science.aan2507.
44. Thul, P.J., Akesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Bjork, L., Breckels, L.M., et al. (2017). A subcellular map of the human proteome. Science *356*. 10.1126/science.aal3321.
45. Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*, 1515-1526. 10.1016/j.cell.2015.11.015.

46. Pacini, C., Dempster, J.M., Boyle, I., Goncalves, E., Najgebauer, H., Karakoc, E., van der Meer, D., Barthorpe, A., Lightfoot, H., Jaaks, P., et al. (2021). Integrated cross-study datasets of genetic dependencies in cancer. Nat Commun *12*, 1661. 10.1038/s41467-021-21898-7.

47. Wang, Y., Yang, W., Pu, Q., Yang, Y., Ye, S., Ma, Q., Ren, J., Cao, Z., Zhong, G., Zhang, X., et al. (2015). The effects and mechanisms of SLC34A2 in tumorigenesis and progression of human non-small cell lung cancer. J Biomed Sci *22*, 52. 10.1186/s12929-015-0158-7.

48. Lv, Y., Zhang, W., Zhao, J., Sun, B., Qi, Y., Ji, H., Chen, C., Zhang, J., Sheng, J., Wang, T., et al. (2021). SRSF1 inhibits autophagy through regulating Bcl-x splicing and interacting with PIK3C3 in lung cancer. Signal Transduct Target Ther *6*, 108. 10.1038/s41392-021-00495-6.

49. Horiguchi, H., Xu, H., Duvert, B., Ciuculescu, F., Yao, Q., Sinha, A., McGuinness, M., Harris, C., Brendel, C., Troeger, A., et al. (2022). Deletion of murine Rhoh leads to de-repression of Bcl-6 via decreased KAISO levels and accelerates a malignancy phenotype in a murine model of lymphoma. Small GTPases *13*, 267-281. 10.1080/21541248.2021.2019503.

50. Bachetti, T., and Ceccherini, I. (2020). Causative and common PHOX2B variants define a broad phenotypic spectrum. Clin Genet *97*, 103-113. 10.1111/cge.13633.

51. Mondal, M., Conole, D., Nautiyal, J., and Tate, E.W. (2022). UCHL1 as a novel target in breast cancer: emerging insights from cell and chemical biology. Br J Cancer *126*, 24-33. 10.1038/s41416-021-01516-5.

52. Xue, W., Kitzing, T., Roessler, S., Zuber, J., Krasnitz, A., Schultz, N., Revill, K., Weissmueller, S., Rappaport, A.R., Simon, J., et al. (2012). A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. Proc Natl Acad Sci U S A *109*, 8212-8217. 10.1073/pnas.1206062109.

53. Wistuba, II, Behrens, C., Virmani, A.K., Mele, G., Milchgrub, S., Girard, L., Fondon, J.W., 3rd, Garner, H.R., McKay, B., Latif, F., et al. (2000). High resolution chromosome 3p allelotyping of human lung cancer and preneoplastic/preinvasive bronchial epithelium reveals multiple, discontinuous sites of 3p allele loss and three regions of frequent breakpoints. Cancer Res *60*, 1949-1960.

54. International Stem Cell, I., Amps, K., Andrews, P.W., Anyfantis, G., Armstrong, L., Avery, S., Baharvand, H., Baker, J., Baker, D., Munoz, M.B., et al. (2011). Screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage. Nat Biotechnol *29*, 1132-1144. 10.1038/nbt.2051.

55. Mayshar, Y., Ben-David, U., Lavon, N., Biancotti, J.C., Yakir, B., Clark, A.T., Plath, K., Lowry, W.E., and Benvenisty, N. (2010). Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. Cell Stem Cell *7*, 521-531. 10.1016/j.stem.2010.07.017.

56. Gruosso, T., Gigoux, M., Manem, V.S.K., Bertos, N., Zuo, D., Perlitch, I., Saleh, S.M.I., Zhao, H., Souleimanova, M., Johnson, R.M., et al. (2019). Spatially distinct tumor immune microenvironments stratify triple-negative breast cancers. J Clin Invest *129*, 1785-1800. 10.1172/JCI96313.

57. Weigman, V.J., Chao, H.H., Shabalin, A.A., He, X., Parker, J.S., Nordgard, S.H., Grushko, T., Huo, D., Nwachukwu, C., Nobel, A., et al. (2012). Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to

therapy, and patient survival. Breast Cancer Res Treat *133*, 865-880. 10.1007/s10549-011-1846-y.

58. Goudreault, M., D'Ambrosio, L.M., Kean, M.J., Mullin, M.J., Larsen, B.G., Sanchez, A., Chaudhry, S., Chen, G.I., Sicheri, F., Nesvizhskii, A.I., et al. (2009). A PP2A phosphatase high density interaction network identifies a novel striatin-interacting phosphatase and kinase complex linked to the cerebral cavernous malformation 3 (CCM3) protein. Mol Cell Proteomics *8*, 157-171. 10.1074/mcp.M800266-MCP200.

59. Kean, M.J., Ceccarelli, D.F., Goudreault, M., Sanches, M., Tate, S., Larsen, B., Gibson, L.C., Derry, W.B., Scott, I.C., Pelletier, L., et al. (2011). Structure-function analysis of core STRIPAK Proteins: a signaling complex implicated in Golgi polarization. J Biol Chem *286*, 25065-25075. 10.1074/jbc.M110.214486.

60. Valentino, M., Dejana, E., and Malinverno, M. (2021). The multifaceted PDCD10/CCM3 gene. Genes Dis *8*, 798-813. 10.1016/j.gendis.2020.12.008.

61. Zhang, Y., Tang, W., Zhang, H., Niu, X., Xu, Y., Zhang, J., Gao, K., Pan, W., Boggon, T.J., Toomre, D., et al. (2013). A network of interactions enables CCM3 and STK24 to coordinate UNC13D-driven vesicle exocytosis in neutrophils. Dev Cell *27*, 215-226. 10.1016/j.devcel.2013.09.021.

62. Jenny Zhou, H., Qin, L., Zhang, H., Tang, W., Ji, W., He, Y., Liang, X., Wang, Z., Yuan, Q., Vortmeyer, A., et al. (2016). Endothelial exocytosis of angiopoietin-2 resulting from CCM3 deficiency contributes to cerebral cavernous malformation. Nat Med *22*, 1033-1042. 10.1038/nm.4169.

63. Fujii, M., Yan, J., Rolland, W.B., Soejima, Y., Caner, B., and Zhang, J.H. (2013). Early brain injury, an evolving frontier in subarachnoid hemorrhage research. Transl Stroke Res *4*, 432-446. 10.1007/s12975-013-0257-2.

64. Davoli, T., Xu, A.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J., and Elledge, S.J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell *155*, 948-962. 10.1016/j.cell.2013.10.011.

65. Sheltzer, J.M., Blank, H.M., Pfau, S.J., Tange, Y., George, B.M., Humpton, T.J., Brito, I.L., Hiraoka, Y., Niwa, O., and Amon, A. (2011). Aneuploidy drives genomic instability in yeast. Science *333*, 1026-1030. 10.1126/science.1206412.

66. Ben-David, U., and Amon, A. (2020). Context is everything: aneuploidy in cancer. Nat Rev Genet *21*, 44-62. 10.1038/s41576-019-0171-x.

67. Anders, K.R., Kudrna, J.R., Keller, K.E., Kinghorn, B., Miller, E.M., Pauw, D., Peck, A.T., Shellooe, C.E., and Strong, I.J. (2009). A strategy for constructing aneuploid yeast strains by transient nondisjunction of a target chromosome. BMC Genet *10*, 36. 10.1186/1471-2156-10-36.

68. Ravichandran, M.C., Fink, S., Clarke, M.N., Hofer, F.C., and Campbell, C.S. (2018). Genetic interactions between specific chromosome copy number alterations dictate complex aneuploidy patterns. Genes Dev *32*, 1485-1498. 10.1101/gad.319400.118.

69. Liu, X., Krawczyk, E., Suprynowicz, F.A., Palechor-Ceron, N., Yuan, H., Dakic, A., Simic, V., Zheng, Y.L., Sripadhan, P., Chen, C., et al. (2017). Conditional reprogramming and long-term expansion of normal and tumor cells from human biospecimens. Nature protocols *12*, 439-451. 10.1038/nprot.2016.174.

70. Coelho, P.P., Hesketh, G.G., Pedersen, A., Kuzmin, E., Fortier, A.N., Bell, E.S., Ratcliffe, C.D.H., Gingras, A.C., and Park, M. (2022). Endosomal LC3C-pathway

selectively targets plasma membrane cargo for autophagic degradation. Nat Commun *13*, 3812. 10.1038/s41467-022-31465-3.

71. Liu, G., Knight, J.D., Zhang, J.P., Tsou, C.C., Wang, J., Lambert, J.P., Larsen, B., Tyers, M., Raught, B., Bandeira, N., et al. (2016). Data Independent Acquisition analysis in ProHits 4.0. J Proteomics *149*, 64-68. 10.1016/j.jprot.2016.04.042.
72. Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis *20*, 3551-3567. 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.
73. Eng, J.K., Jahan, T.A., and Hoopmann, M.R. (2013). Comet: an open-source MS/MS sequence database search tool. Proteomics *13*, 22-24. 10.1002/pmic.201200439.
74. Shteynberg, D., Deutsch, E.W., Lam, H., Eng, J.K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R.L., Aebersold, R., and Nesvizhskii, A.I. (2011). iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics *10*, M111 007690. 10.1074/mcp.M111.007690.
75. Teo, G., Liu, G., Zhang, J., Nesvizhskii, A.I., Gingras, A.C., and Choi, H. (2014). SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. J Proteomics *100*, 37-43. 10.1016/j.jprot.2013.10.023.
76. O'Flanagan, C.H., Campbell, K.R., Zhang, A.W., Kabeer, F., Lim, J.L.P., Biele, J., Eirew, P., Lai, D., McPherson, A., Kong, E., et al. (2019). Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. Genome Biol *20*, 210. 10.1186/s13059-019-1830-0.
77. Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al. (2012). The life history of 21 breast cancers. Cell *149*, 994-1007. 10.1016/j.cell.2012.04.023.
78. Consortium, I.T.P.-C.A.o.W.G. (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82-93. 10.1038/s41586-020-1969-6.
79. Kaufmann, T.L., Petkovic, M., Watkins, T.B.K., Colliver, E.C., Laskina, S., Thapa, N., Minussi, D.C., Navin, N., Swanton, C., Van Loo, P., et al. (2022). MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution. Genome Biol *23*, 241. 10.1186/s13059-022-02794-9.
80. Dentro, S.C., Wedge, D.C., and Van Loo, P. (2017). Principles of Reconstructing the Subclonal Architecture of Cancers. Cold Spring Harb Perspect Med *7*. 10.1101/cshperspect.a026625.
81. Leshchiner, I., Mroz, E.A., Cha, J., Rosebrock, D., Spiro, O., Bonilla-Velez, J., Faquin, W.C., Lefranc-Torres, A., Lin, D.T., Michaud, W.A., et al. (2023). Inferring early genetic progression in cancers with unobtainable premalignant disease. Nat Cancer *4*, 550-563. 10.1038/s43018-023-00533-y.
82. Wang, R., Lin, D.Y., and Jiang, Y. (2020). SCOPE: A Normalization and Copy-Number Estimation Method for Single-Cell DNA Sequencing. Cell Syst *10*, 445-452 e446. 10.1016/j.cels.2020.03.005.
83. Bourgey, M., Dali, R., Eveleigh, R., Chen, K.C., Letourneau, L., Fillon, J., Michaud, M., Caron, M., Sandoval, J., Lefebvre, F., et al. (2019). GenPipes: an open-source framework

for distributed and scalable genomic analyses. Gigascience *8*. 10.1093/gigascience/giz037.

84. Zaccaria, S., and Raphael, B.J. (2021). Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. Nat Biotechnol *39*, 207-214. 10.1038/s41587-020-0661-6.

Figure 1

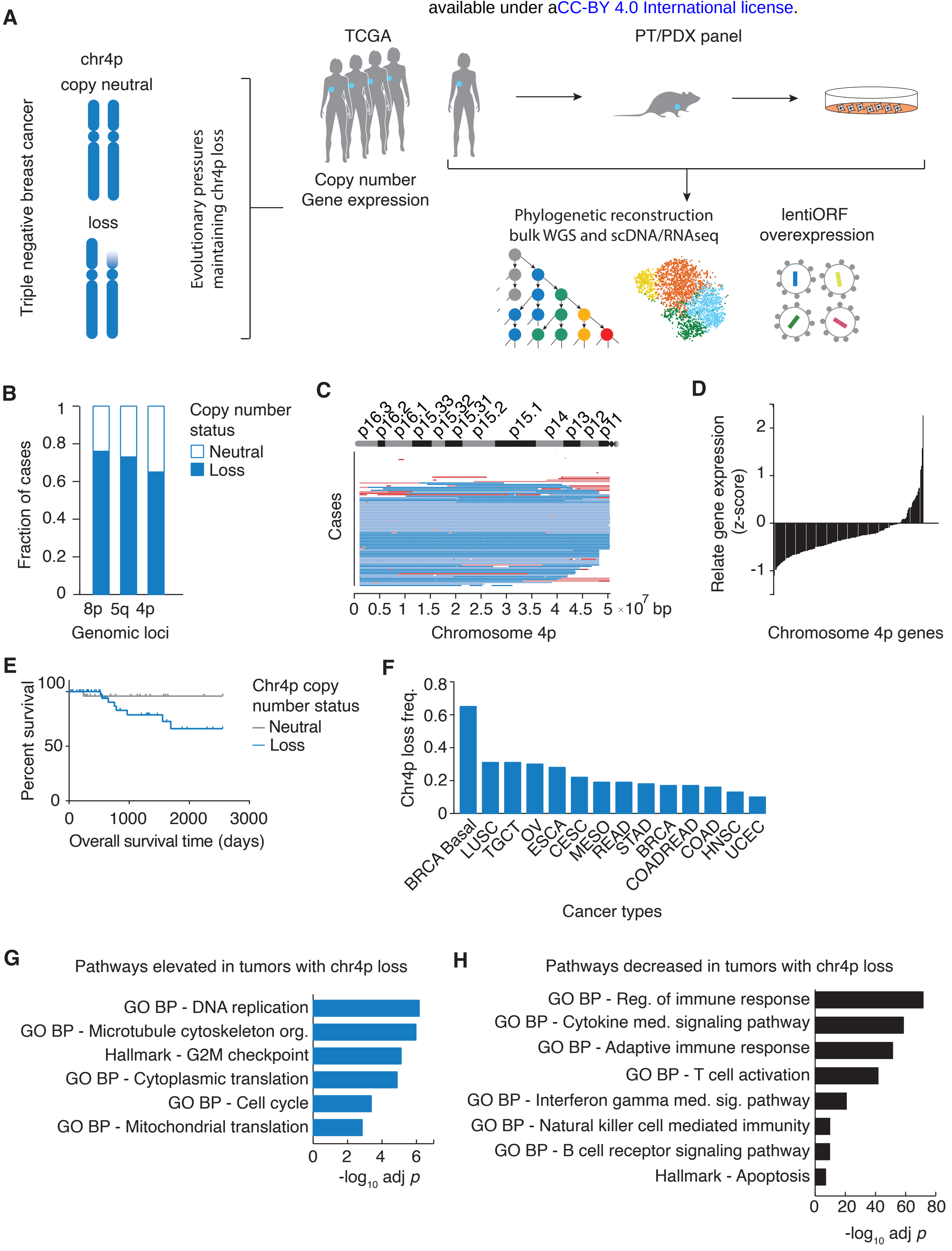
A
Triple negative breast cancer
chr4p
copy neutral
loss
Evolutionary pressures
maintaining chr4p loss
TCGA
Copy number
Gene expression
PT/PDX panel
Phylogenetic reconstruction
bulk WGS and scDNA/RNAseq
lentiORF
overexpression
B
Fraction of cases
1
0.8
0.6
0.4
0.2
0
8p
5q
4p
Genomic loci
Copy number
status
Neutral
Loss
C
p16.3
p16.2
p16.1
p15.33
p15.32
p15.31
p15.2
p15.1
p14
p13
p12
p11
Cases
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
×10^7 bp
Chromosome 4p
D
Relate gene expression
(z-score)
2
1
0
-1
Chromosome 4p genes
E
Percent survival
100
50
0
0
1000
2000
3000
Overall survival time (days)
Chr4p copy
number status
Neutral
Loss
F
Chr4p loss freq.
0.8
0.6
0.4
0.2
0
BRCA Basal
LUSC
TGCT
OV
ESCA
CESC
MESO
READ
STAD
BRCA
COADREAD
COAD
HNSC
UCEC
Cancer types
G
Pathways elevated in tumors with chr4p loss
GO BP - DNA replication
GO BP - Microtubule cytoskeleton org.
Hallmark - G2M checkpoint
GO BP - Cytoplasmic translation
GO BP - Cell cycle
GO BP - Mitochondrial translation
0
2
4
6
-log10 adj p
H
Pathways decreased in tumors with chr4p loss
GO BP - Reg. of immune response
GO BP - Cytokine med. signaling pathway
GO BP - Adaptive immune response
GO BP - T cell activation
GO BP - Interferon gamma med. sig. pathway
GO BP - Natural killer cell mediated immunity
GO BP - B cell receptor signaling pathway
Hallmark - Apoptosis
0
20
40
60
80
-log10 adj p

Figure 2

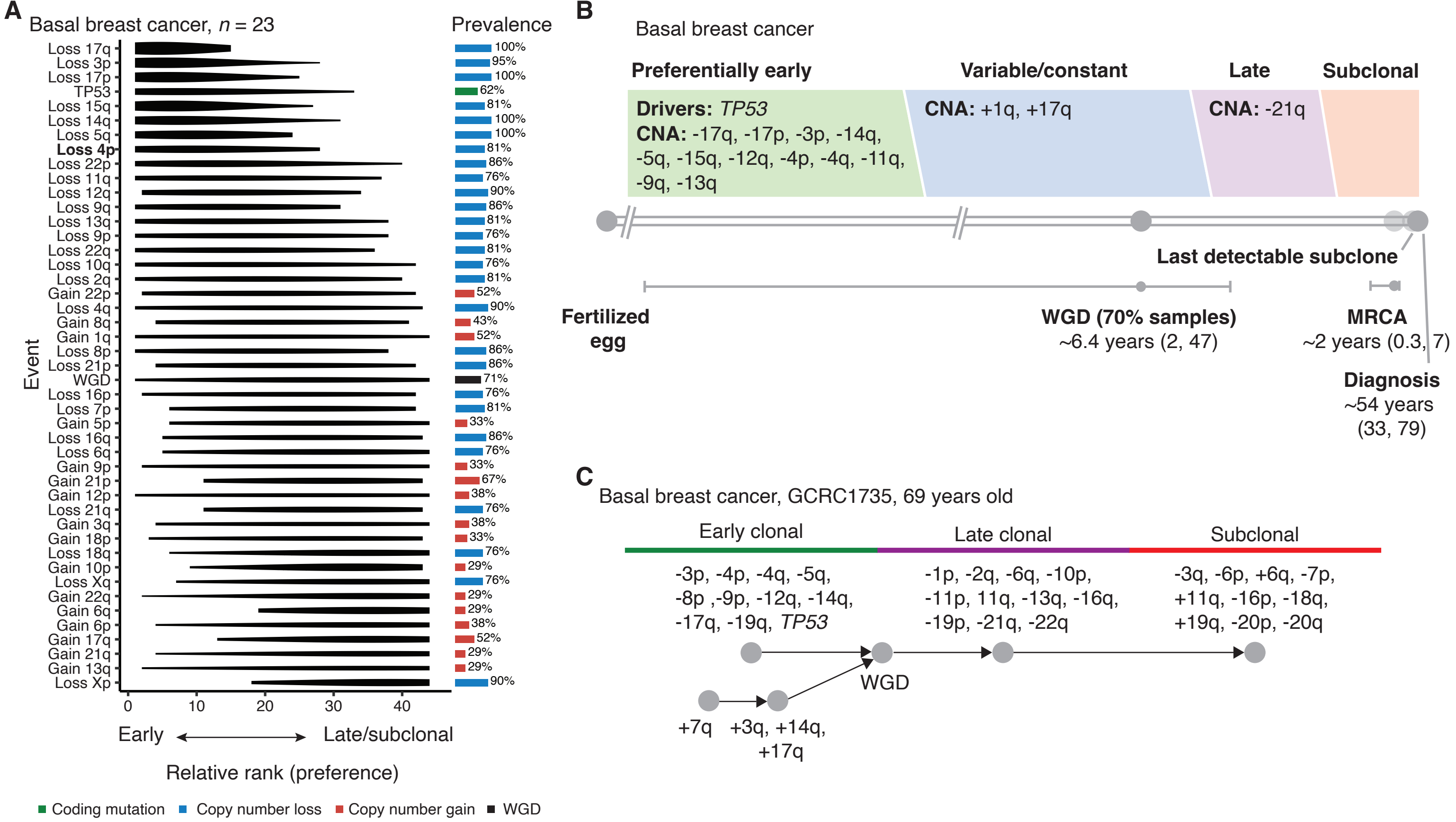
A
Basal breast cancer, n = 23
Prevalence
Loss 17q 100%
Loss 3p 95%
Loss 17p 100%
TP53 62%
Loss 15q 81%
Loss 14q 100%
Loss 5q 100%
Loss 4p 81%
Loss 22p 86%
Loss 11q 76%
Loss 12q 90%
Loss 9q 86%
Loss 13q 81%
Loss 9p 76%
Loss 22q 81%
Loss 10q 76%
Loss 2q 81%
Gain 22p 52%
Loss 4q 90%
Gain 8q 43%
Gain 1q 52%
Loss 8p 86%
Loss 21p 86%
WGD 71%
Loss 16p 76%
Loss 7p 81%
Gain 5p 33%
Loss 16q 86%
Loss 6q 76%
Gain 9p 33%
Gain 21p 67%
Gain 12p 38%
Loss 21q 76%
Gain 3q 38%
Gain 18p 33%
Loss 18q 76%
Gain 10p 29%
Loss Xq 76%
Gain 22q 29%
Gain 6q 29%
Gain 6p 38%
Gain 17q 52%
Gain 21q 29%
Gain 13q 29%
Loss Xp 90%
Event
0
10
20
30
40
Early
Late/subclonal
Relative rank (preference)
🟩 Coding mutation
🟦 Copy number loss
🟥 Copy number gain
⬛ WGD
B
Basal breast cancer
Preferentially early
Variable/constant
Late
Subclonal
Drivers: TP53
CNA: -17q, -17p, -3p, -14q, -5q, -15q, -12q, -4p, -4q, -11q, -9q, -13q
CNA: +1q, +17q
CNA: -21q
Last detectable subclone
Fertilized egg
WGD (70% samples)
~6.4 years (2, 47)
MRCA
~2 years (0.3, 7)
Diagnosis
~54 years
(33, 79)
C
Basal breast cancer, GCRC1735, 69 years old
Early clonal
Late clonal
Subclonal
-3p, -4p, -4q, -5q, -8p ,-9p, -12q, -14q, -17q, -19q, TP53
-1p, -2q, -6q, -10p, -11p, 11q, -13q, -16q, -19p, -21q, -22q
-3q, -6p, +6q, -7p, +11q, -16p, -18q, +19q, -20p, -20q
WGD
+7q
+3q, +14q, +17q

Figure 3

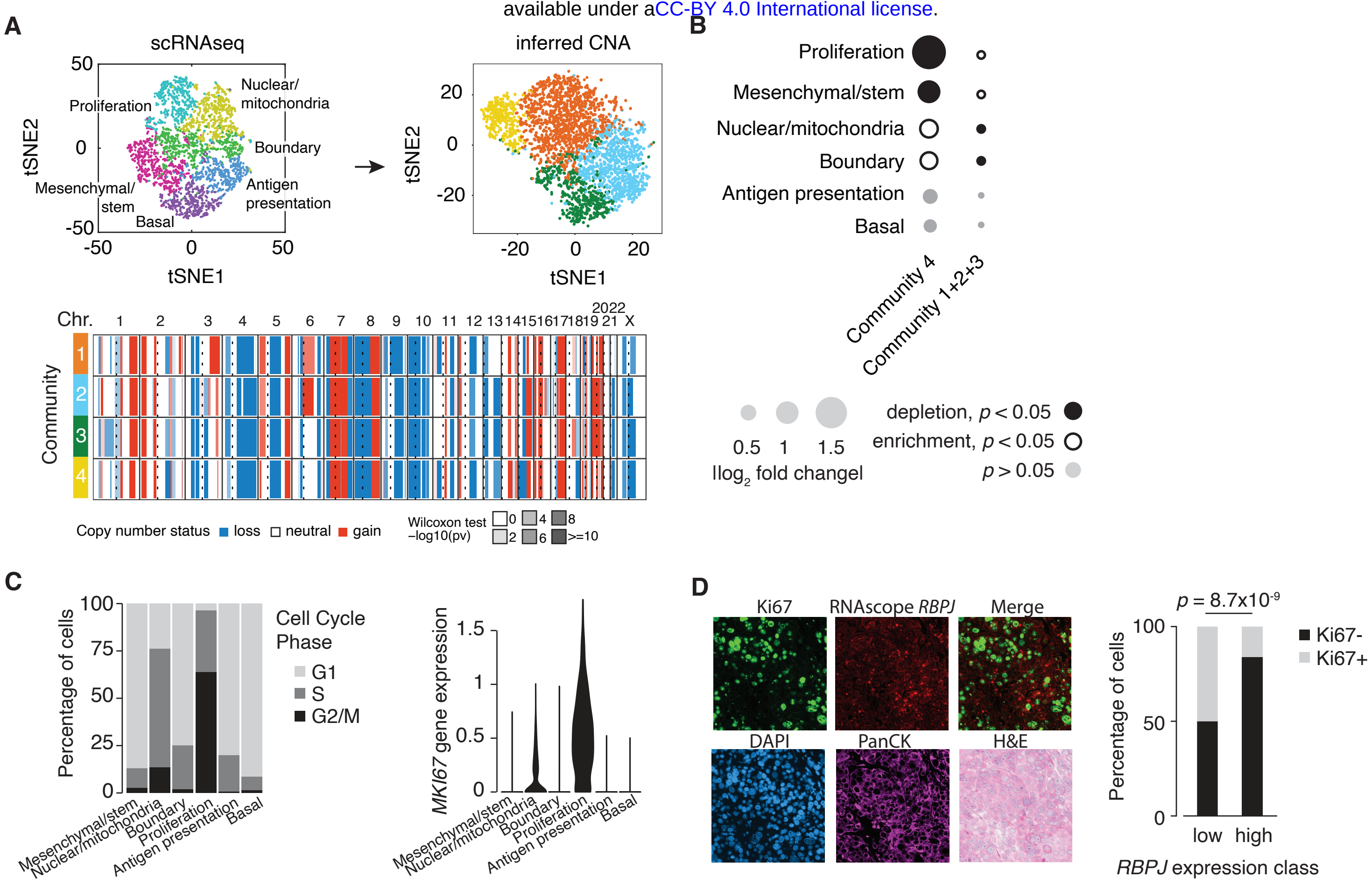
A
scRNAseq
50
0
-50
-50
0
50
tSNE2
tSNE1
Proliferation
Nuclear/
mitochondria
Boundary
Antigen
presentation
Mesenchymal/
stem
Basal
inferred CNA
20
0
-20
-20
0
20
tSNE2
tSNE1
Chr.
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Community
1
2
3
4
Copy number status ■ loss □ neutral ■ gain
Wilcoxon test
−log10(pv)
0 2 4 6 8 >=10
B
Proliferation
Mesenchymal/stem
Nuclear/mitochondria
Boundary
Antigen presentation
Basal
Community 4
Community 1+2+3
0.5 1 1.5
|log$_2$ fold change|
depletion, $p < 0.05$ ●
enrichment, $p < 0.05$ ○
$p > 0.05$ ●
C
Percentage of cells
100
75
50
25
0
Cell Cycle
Phase
G1
S
G2/M
Mesenchymal/stem
Nuclear/mitochondria
Boundary
Proliferation
Antigen presentation
Basal
MKI67 gene expression
1.5
1
0.5
0
Mesenchymal/stem
Nuclear/mitochondria
Boundary
Proliferation
Antigen presentation
Basal
D
Ki67
RNAscope RBPJ
Merge
DAPI
PanCK
H&E
$p = 8.7\text{x}10^{-9}$
Percentage of cells
100
50
0
low
high
RBPJ expression class
Ki67-
Ki67+

Figure 4

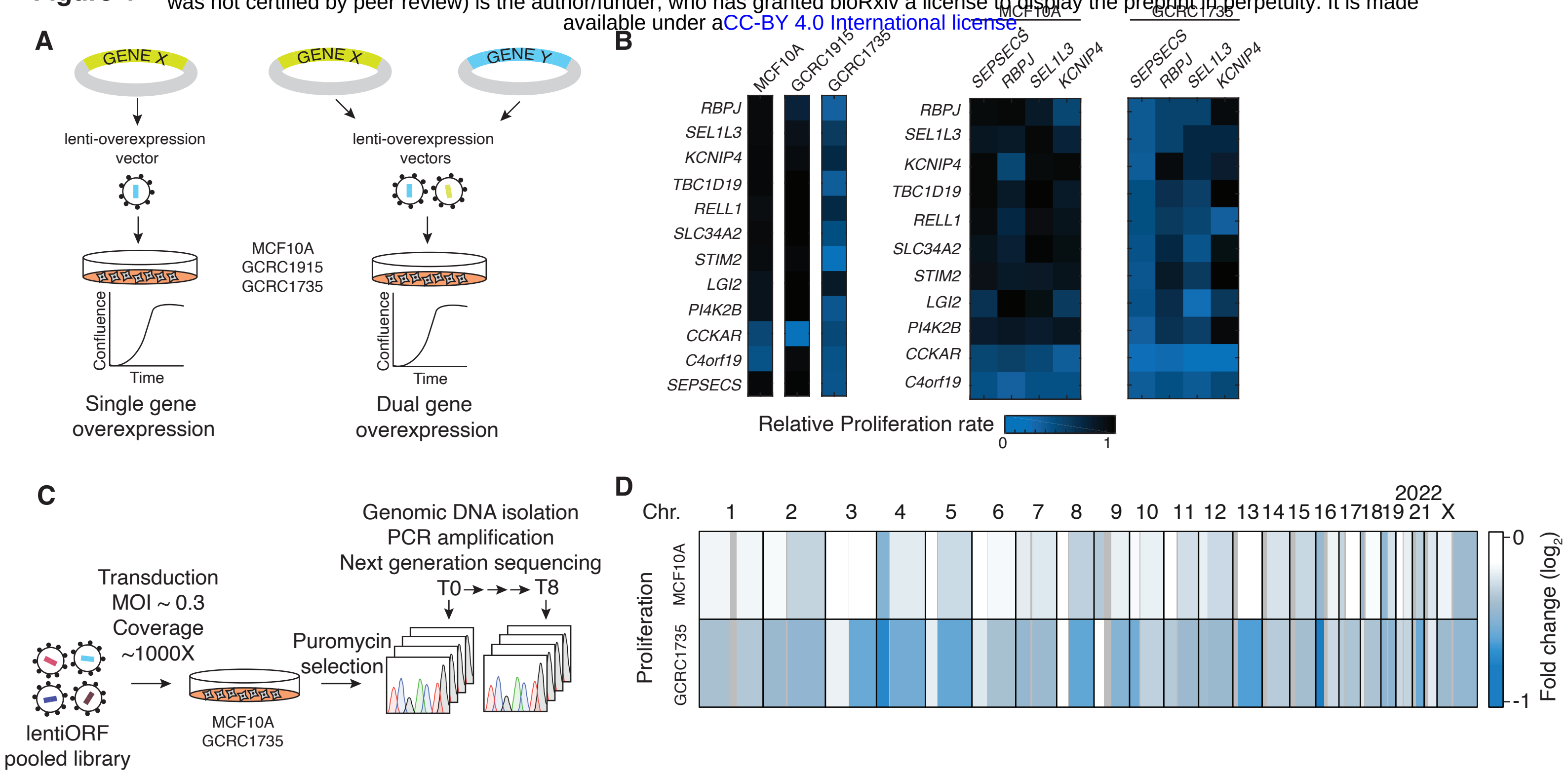
A
GENE X
lenti-overexpression vector
Confluence
Time
Single gene overexpression
MCF10A
GCRC1915
GCRC1735
GENE X
GENE Y
lenti-overexpression vectors
Confluence
Time
Dual gene overexpression
B
MCF10A
GCRC1915
GCRC1735
RBPJ
SEL1L3
KCNIP4
TBC1D19
RELL1
SLC34A2
STIM2
LGI2
PI4K2B
CCKAR
C4orf19
SEPSECS
MCF10A
GCRC1735
SEPSECS
RBPJ
SEL1L3
KCNIP4
Relative Proliferation rate
0
1
C
Transduction
MOI ~ 0.3
Coverage ~1000X
Genomic DNA isolation
PCR amplification
Next generation sequencing
T0
T8
Puromycin selection
lentiORF pooled library
MCF10A
GCRC1735
D
Chr.
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Proliferation
MCF10A
GCRC1735
Fold change (log2)
0
-1

Figure 5

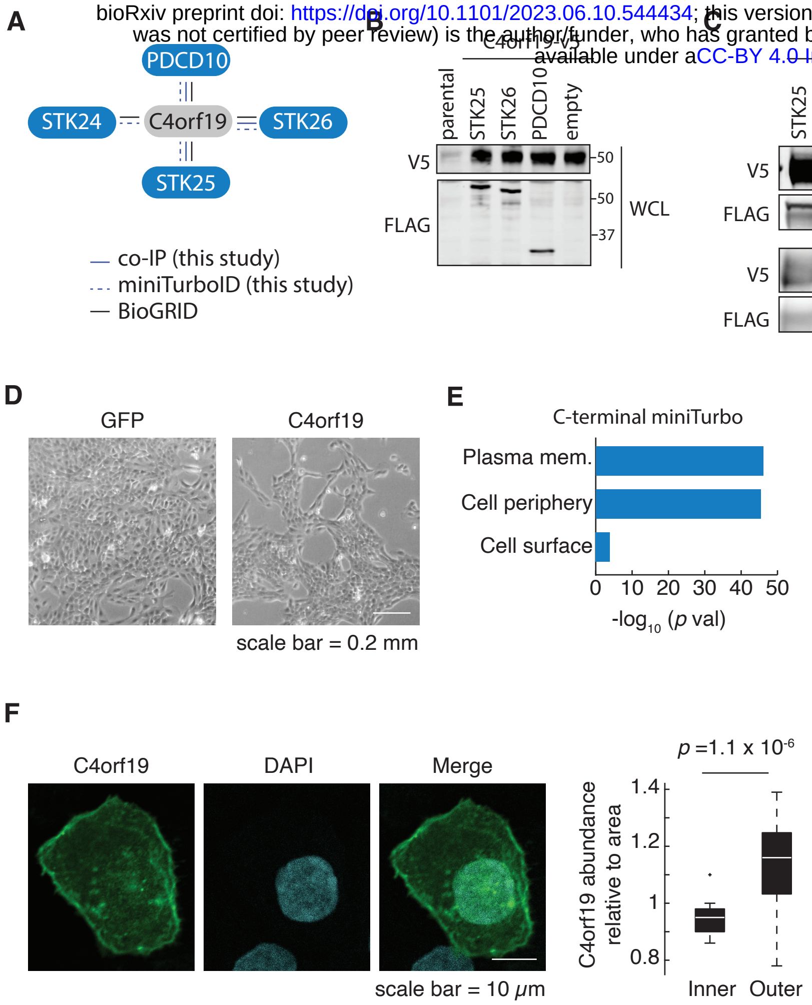
A
PDCD10
STK24
C4orf19
STK26
STK25
co-IP (this study)
miniTurboID (this study)
BioGRID
B
C4orf19-V5
parental
STK25
STK26
PDCD10
empty
V5
FLAG
50
50
37
WCL
C
C4orf19-V5
STK25
STK26
PDCD10
empty
V5
FLAG
V5
FLAG
50
50
50
37
50
50
50
37
IP: V5
IgG[m]
D
GFP
C4orf19
scale bar = 0.2 mm
E
C-terminal miniTurbo
Plasma mem.
Cell periphery
Cell surface
0
10
20
30
40
50
-log10 (p val)
F
C4orf19
DAPI
Merge
scale bar = 10 μm
p =1.1 x 10-6
C4orf19 abundance relative to area
1.4
1.2
1
0.8
Inner
Outer

Figure S1

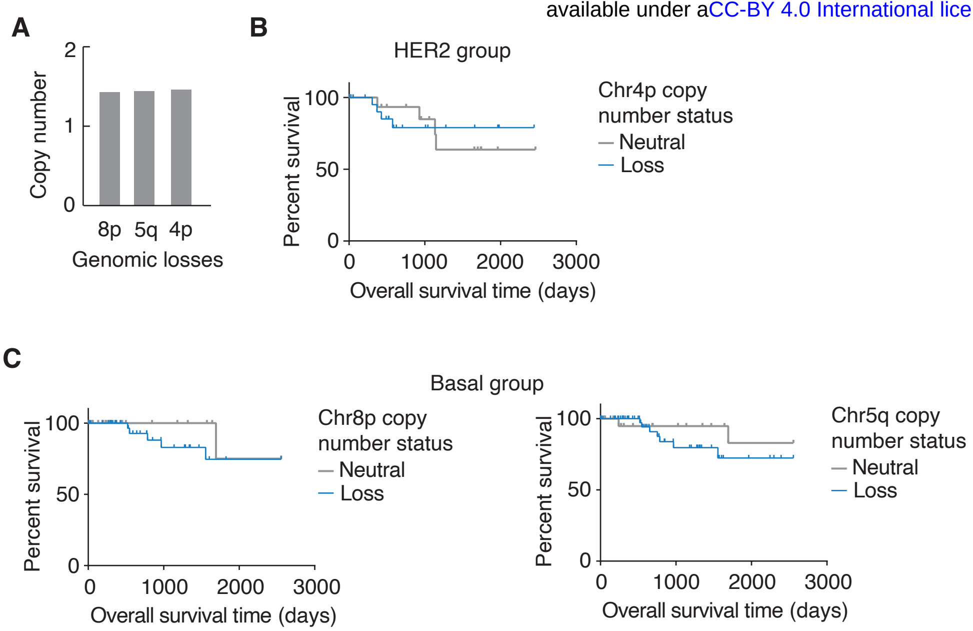

Figure S1. Recurrent large chromosomal deletions in breast cancer. (A) Copy number was obtained from the TCGA segmented mean showing that frequently recurrent large chromosomal deletions in basal breast cancer are hemizygous, n = 91. **(B)** Overall survival of Her2 breast cancer patients with copy neutral and deletion status of chr4p shows no difference between groups, n = 55. **(C)** Overall survival of basal breast cancer patients with copy neutral and deletion status of chr8p and chr5q shows a trends towards a worse survival of patients with these chromosome arm losses (chr 8p loss *p* = 0.5, chr5q loss *p* = 0.4 as assessed by long rank test; n = 91).

Figure S2

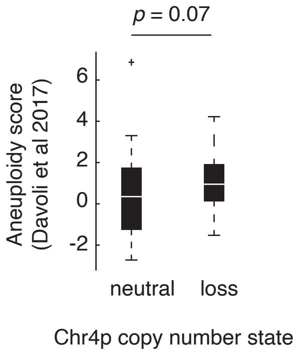

Figure S2. Aneuploidy in basal breast cancer with different chr4p copy number states. Aneuploidy score as quantified by Chrom.Arm.SCNA.Level median reported by Davoli et al Science 2017 shows no statistically significant difference in aneuploidy between chr4p copy neutral vs deletion basal breast cancer samples. Significance was assessed using Wilcoxon rank sum test.

Figure S3

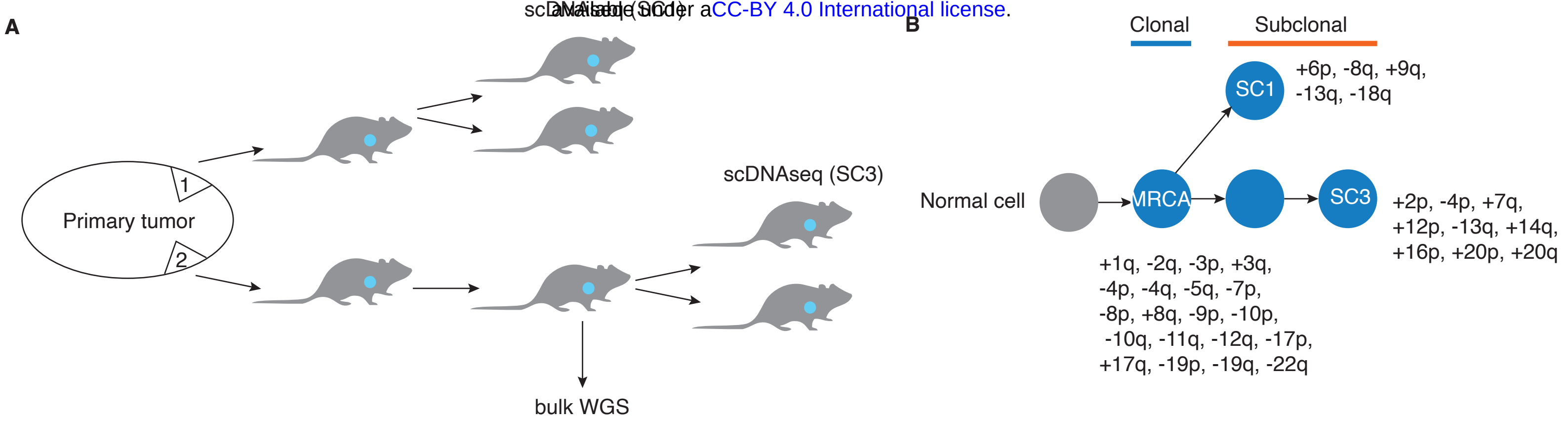

Figure S3. Timing of chromosome arm aberrations of PDX samples using CHISEL. (A) Schematic of PDX generation which was used for bulk WGS and scDNAseq. Single cell DNA sequencing was conducted on four GCRC1735 PDX samples. Two different locations within the primary tumor were biopsied, cryopreserved and propageted in NOD-SCID mice. Two mice were engrafted using a fragment derived from one location from passage 2 in the PDX and two mice were engrafted using a fragment drived from another location from passage 3 in the PDX. **(B)** CHISEL (Zaccaria et al Nat Biotech 2020) was used to generate an evolutionary timeline showing that chr4p loss is an early event in basal breast cancer PDX progression.

Figure S4

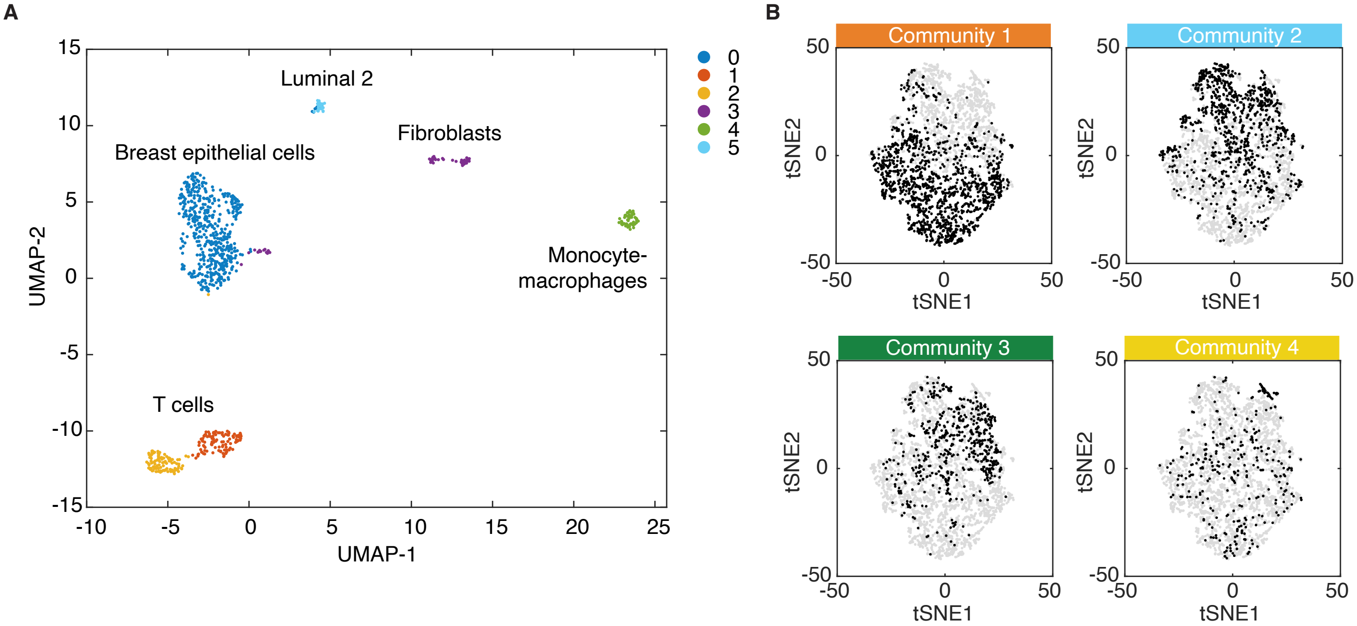

Figure S4. Inferring copy number aberrations from scRNAseq. (A) scRNAseq of normal breast epithelial cells. UMAP plot of single cell RNA seq data from two reduction mammoplasty samples. Clusters were annotated using previously defined cell type markers. **(B)** Overlapping inferred copy number communities with scRNAseq gene expression clusters. scRNAseq map as reported in a previous study is used to annotate single cells with inferred copy number communities as analyzed in this study. Single cells coloured in black belong to the specifiied community.

Figure S5

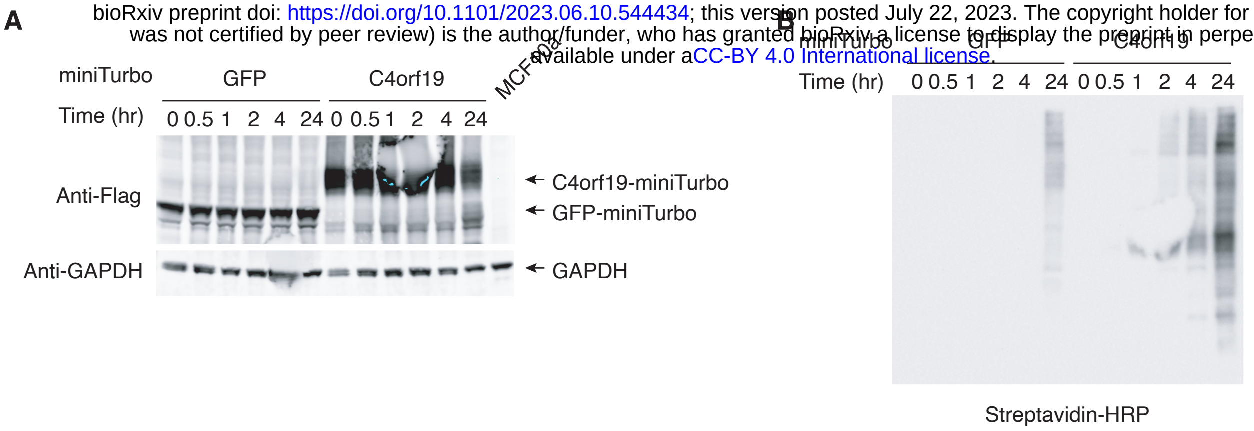

Figure S5. miniTurboID screen bait protein expression and biotinylation level. (A) miniTurboID C4orf19 protein bait expression in MCF10a. MCF10a cells stably expressing miniTurboID bait proteins: C4orf19-3XFLAG-miniTurbo and GFP-3XFLAG-miniTurbo control were induced with 0.5 μg/ml doxycycline for designated time. Protein bait expression was assessed using Anti-Flag antibody. GAPDH protein level served as the loading control. **(B)** Biotinylation level. Biotin labeling was induced with 0.5 μg/ml doxycycline and 40 μM biotin for designated time. Optimal biotinylation was achieved 4 hr post induction.