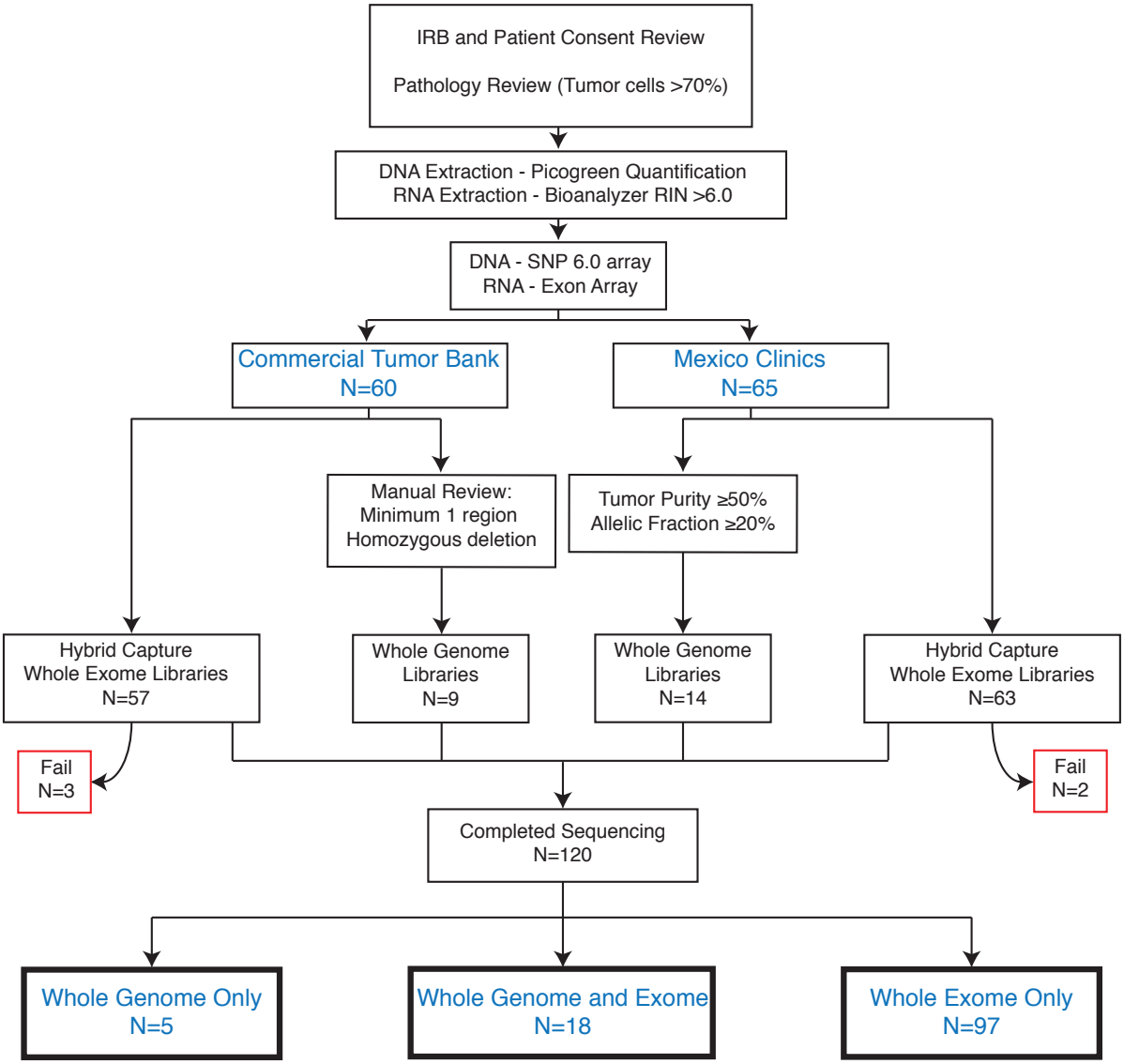
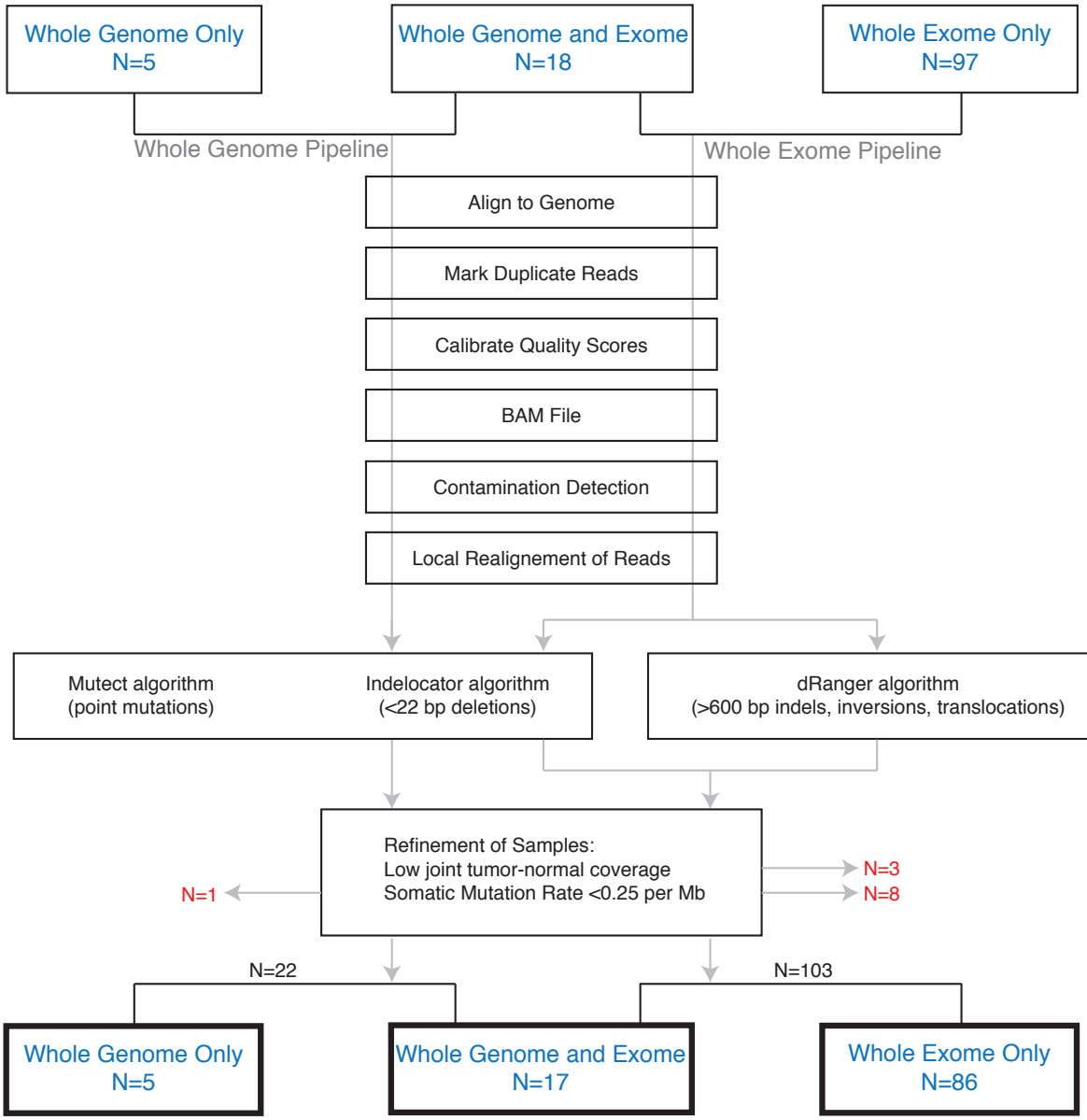
doi:10.1038/nature11154

Supplementary Figures and Legends

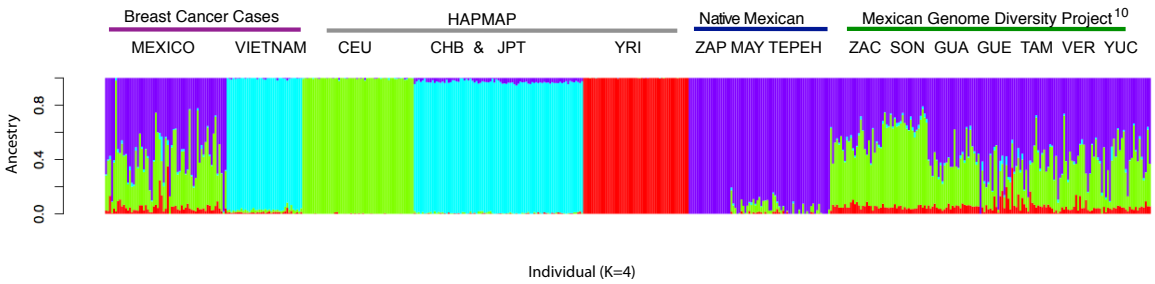


Supplementary Figure 1 - Sample processing pipeline. Vietnamese samples for whole-genome sequencing were selected based on manual review showing at least one region of homozygous deletion while Mexican samples were selected based on tumor purity ≥50% and allelic fraction ≥20%.



Supplementary Figure 2 - Sample analysis pipeline for determining somatic mutations and rearrangements.

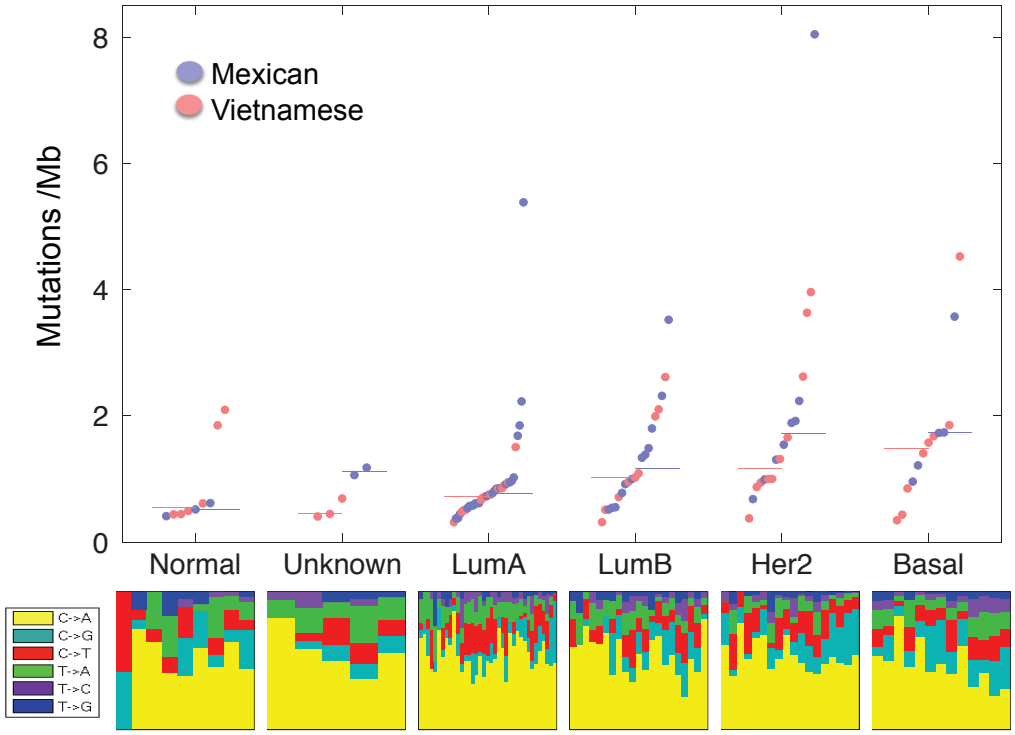
A



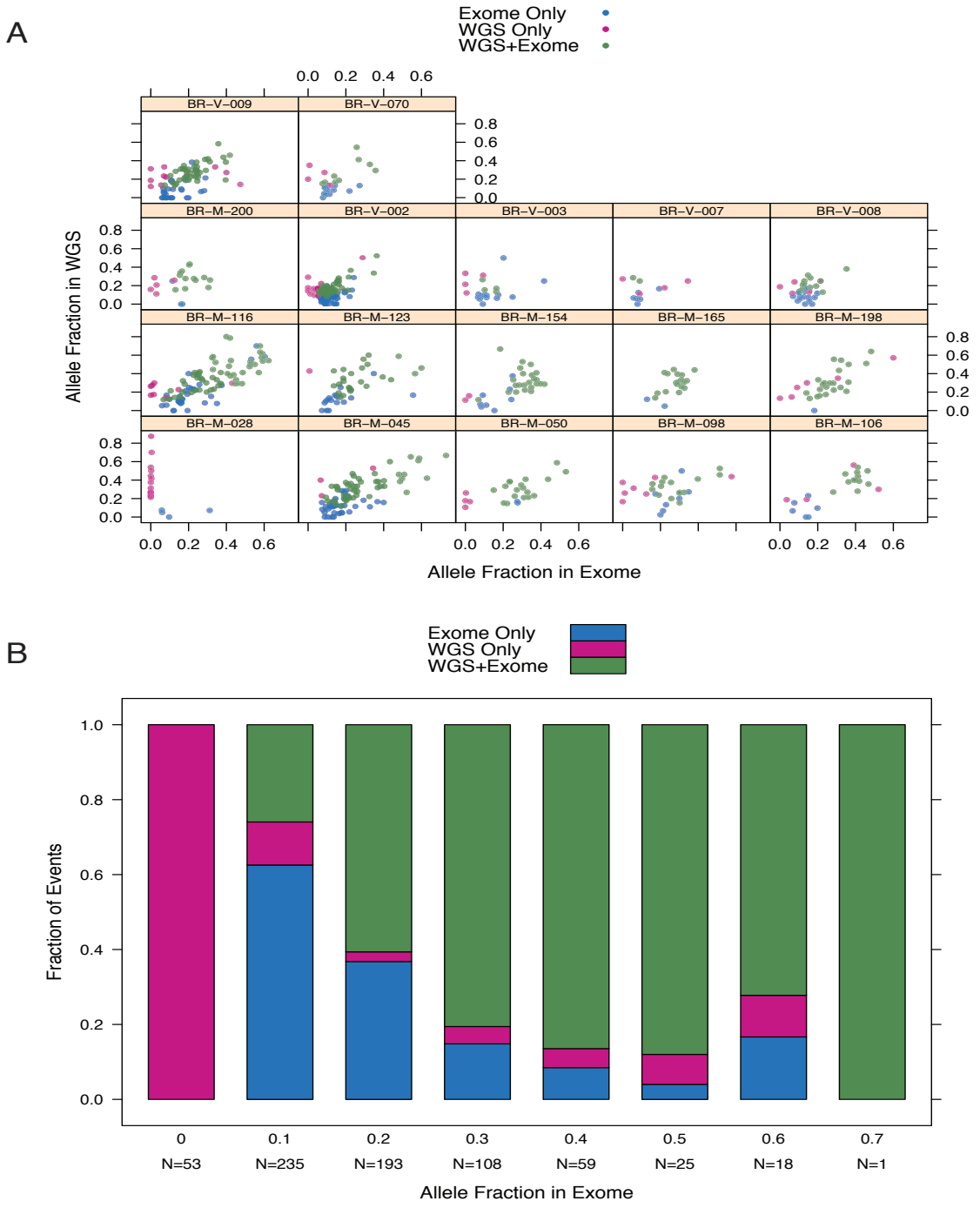
B

	Population	Samples	Details
HapMap		196	
	CEU (56), CHB+JPT (87) and YRI (53)		
MGDP[7]		161	From seven states of Mexico
	GUA	26	Guanajuato
	GUE	27	Guerrero
	SON	24	Sonora
	TAM	9	Tamaulipas
	VER	26	Veracruz
	YUC	24	Yucatan
	ZAC	25	Zacatecas
Native Mexican		71	
	ZAP	21	Oaxaca (Center)
	MAY	27	Campeche (Southeast)
	TEP	23	Durango (North)

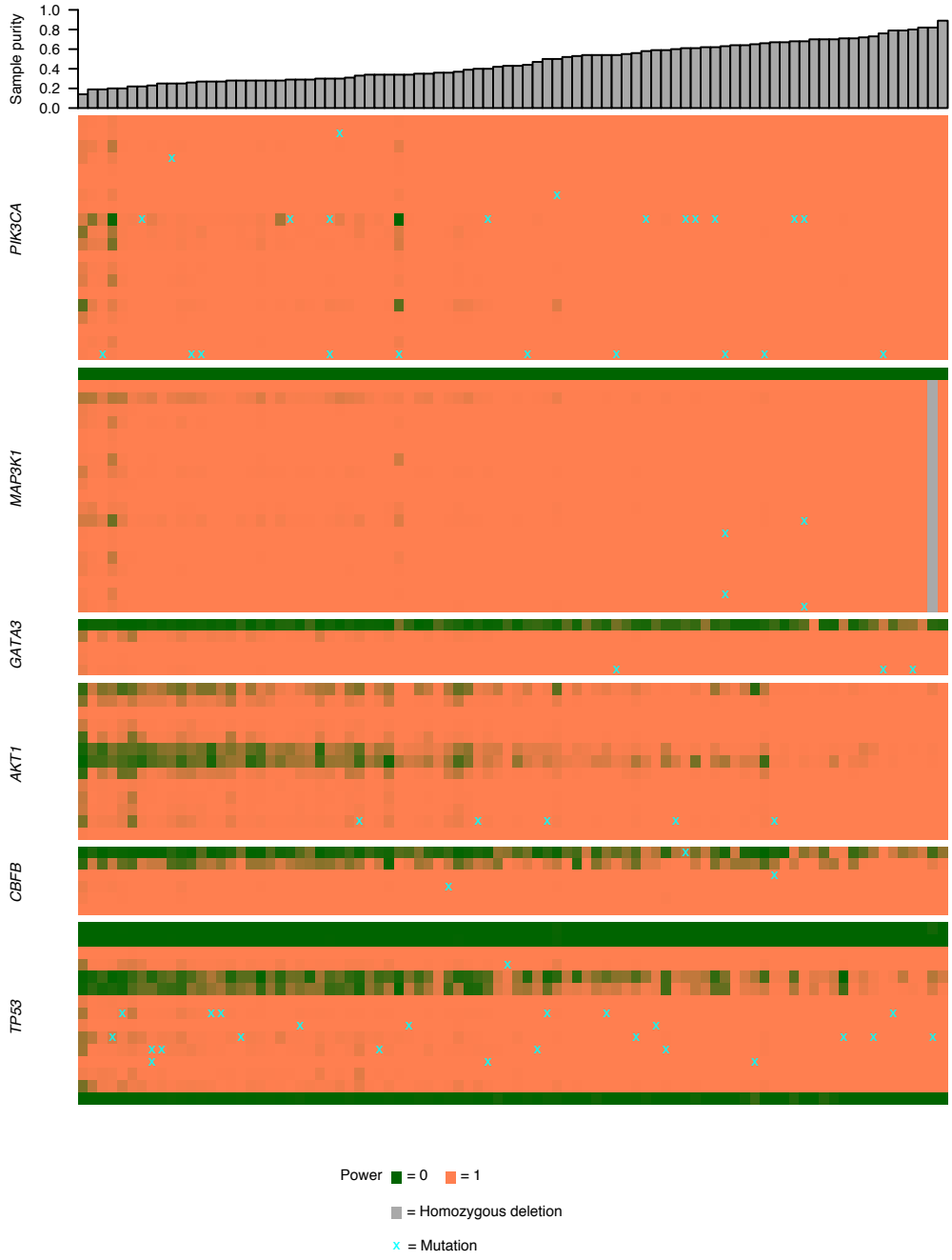
Supplementary Figure 3 - Ancestry analysis. A. Individual ancestry proportions. Four parental groups K=4: Asia (CHB+JPT), Europe (CEU), Africa (YRI) and Native Mexican - Zapoteca, Maya, and Tepehuano (ZAP, MAY, TEPEH) and Breast Cancer cases. B. Populations used for ancestry assessment[1].



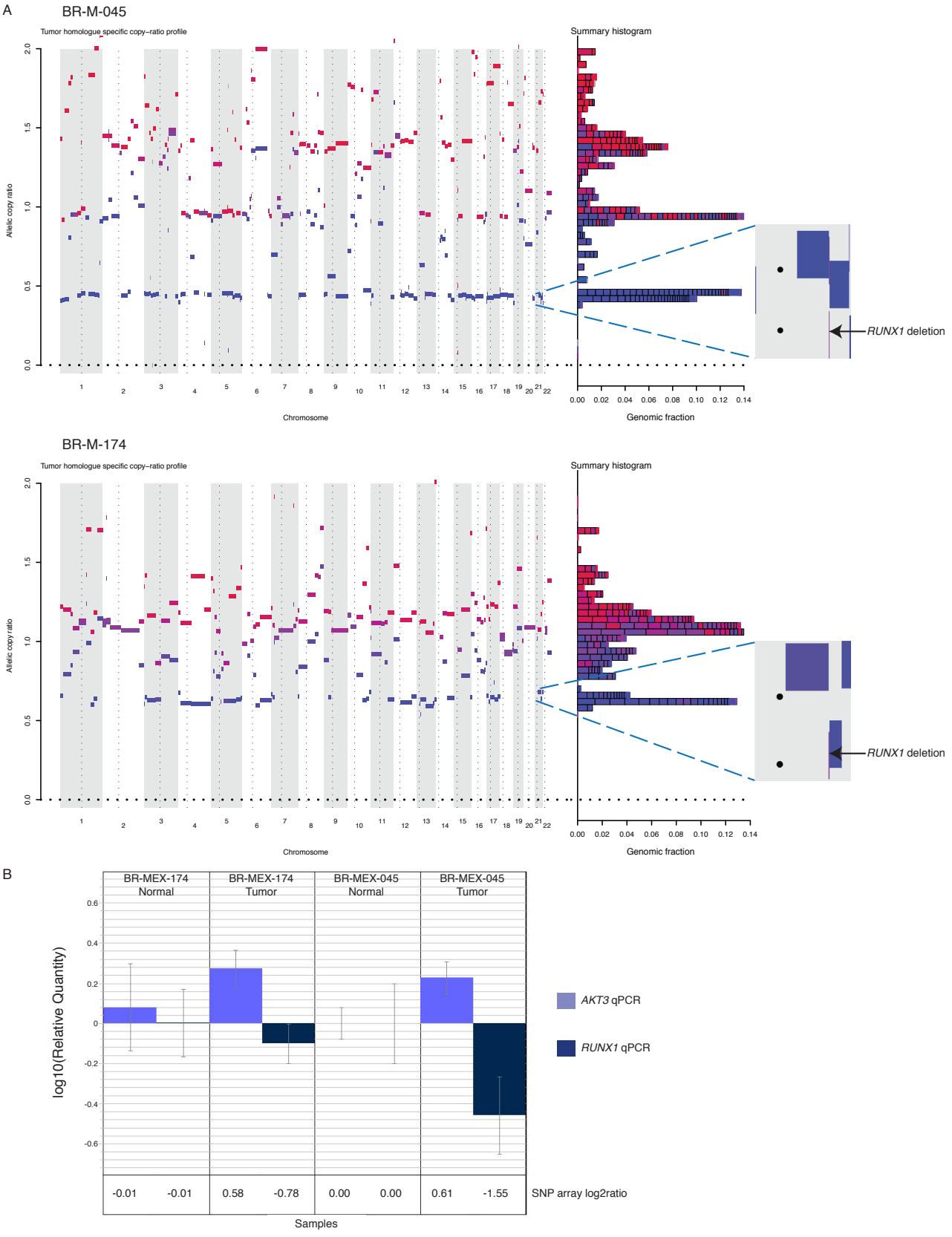
Supplementary Figure 4 - Mutation rate by expression subtype. Plot: Overall mutation rate of samples plotted according to breast expression subtype as determined by PAM50 classification. Histogram: Breakdown of mutation spectra across expression subtypes and samples.



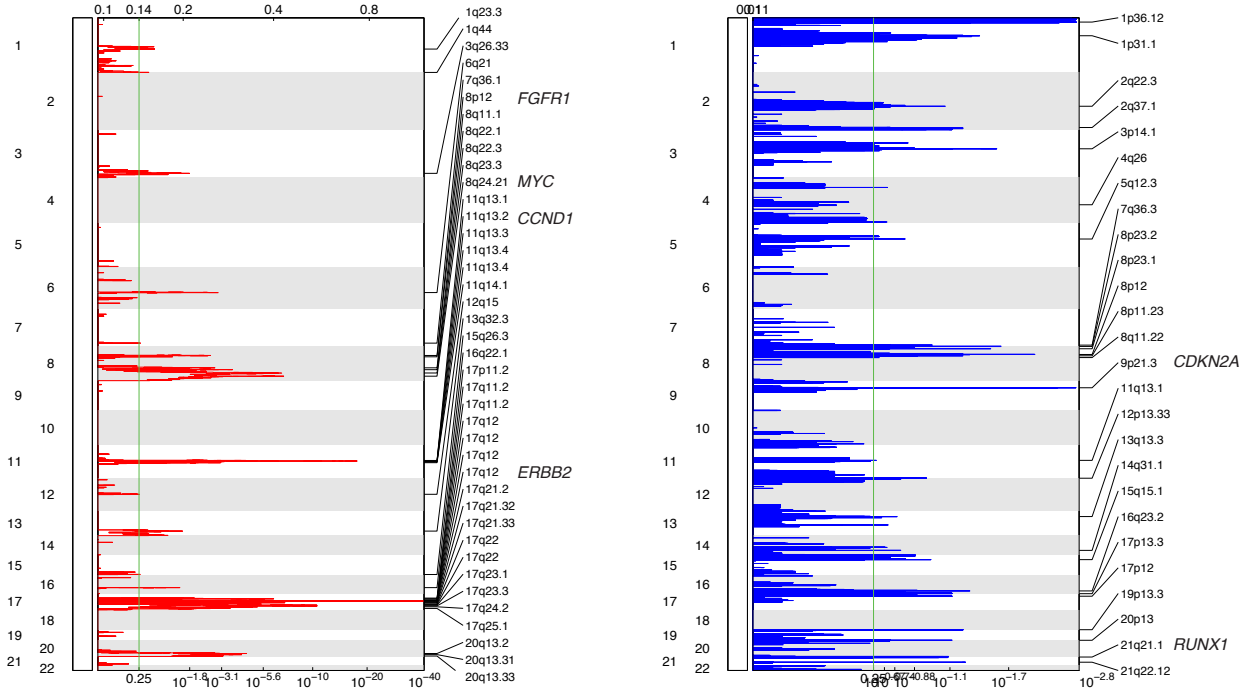
Supplementary Figure 5 - Comparison of somatic mutations identified via whole-exome and whole-genome sequencing. A. Plot of somatic mutations in genomes and exomes according to allelic fraction of event by each method. Mutations found by both methods shown in green. B. Concordance of somatic mutations called by both methods binned by allelic fraction. Whole-exome sequencing overall is able to identify mutations at a lower allelic fraction. Mutations found only in whole-genomes likely represent false-positive somatic mutations due to lower depth of sequence coverage.



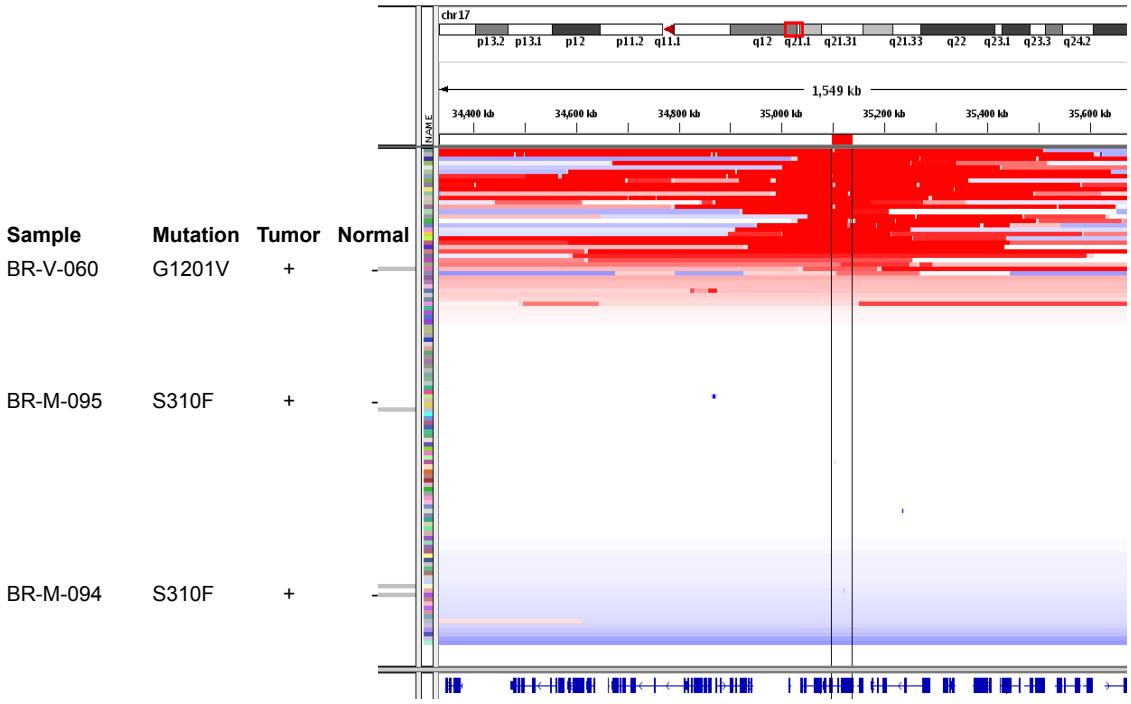
Supplementary Figure 6 - Power to detect mutations in significantly mutated genes as determined by ABSOLUTE. Samples along x-axis arranged from least to greatest tumor cell purity. Genes represented along y-axis with each row representing individual exons within gene. Dark green squares represent exons with zero power while salmon squares represent exons with power = 1. Grey squares represent regions of homozygous deletion. Mutations shown with cyan "x".


A
BR-M-045
Tumor homologue specific copy−ratio profile
Summary histogram
Allelic copy ratio
Chromosome
Genomic fraction
RUNX1 deletion
BR-M-174
Tumor homologue specific copy−ratio profile
Summary histogram
Allelic copy ratio
Chromosome
Genomic fraction
RUNX1 deletion
B
BR-MEX-174 Normal
BR-MEX-174 Tumor
BR-MEX-045 Normal
BR-MEX-045 Tumor
log10(Relative Quantity)
AKT3 qPCR
RUNX1 qPCR
-0.01 -0.01 0.58 -0.78 0.00 0.00 0.61 -1.55
SNP array log2ratio
Samples

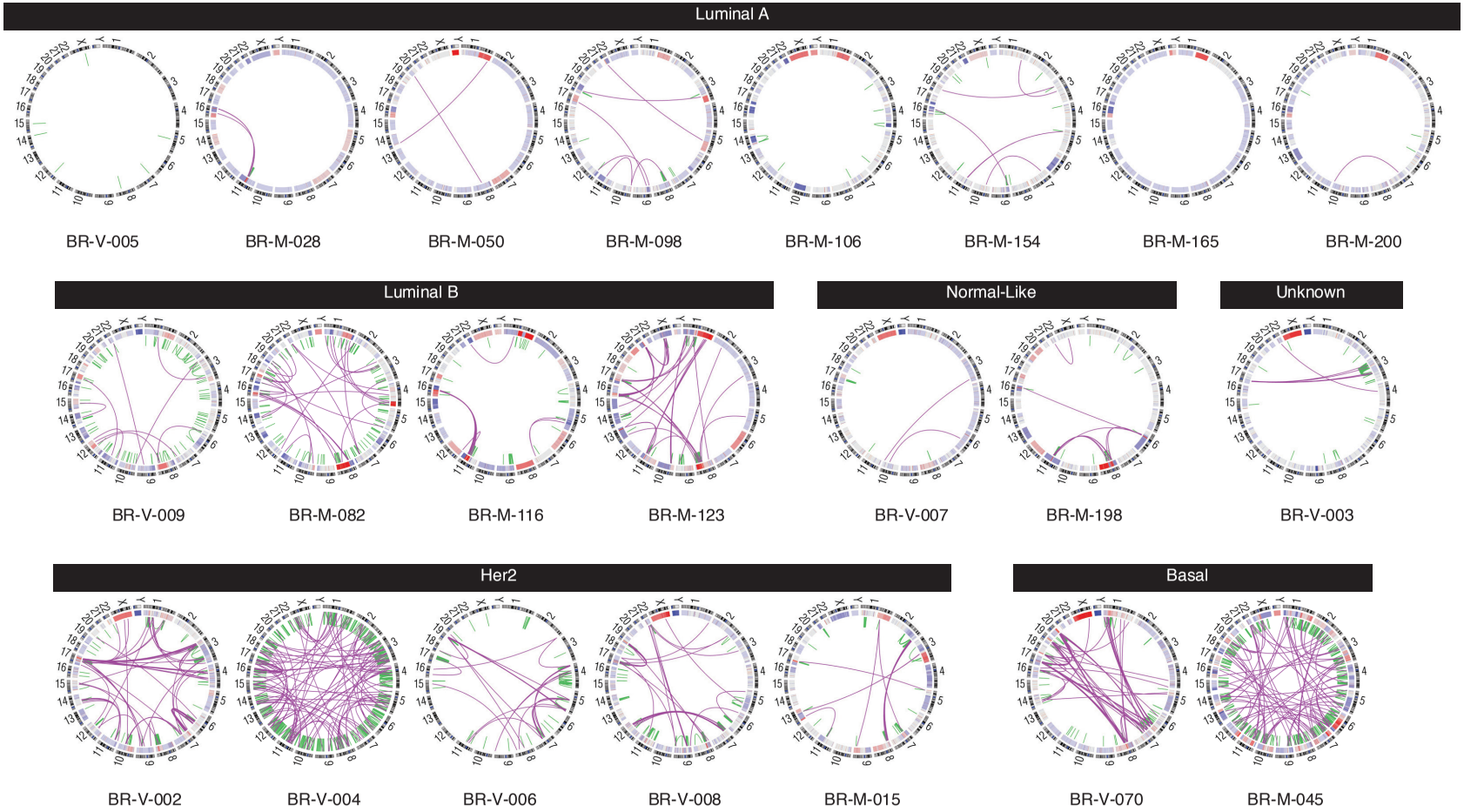
Supplementary Figure 7 - Representative ABSOLUTE plots of samples harboring *RUNX1* deletions. A. For each sample, plot on left demonstrates genome-wide view of copy ratios for both homologous chromosomes. The copy ratios are shown for each genomic segment with locally constant copy number. Color-axis indicates distance between low (blue) and high (red) homologue concentration; segments where these are similar (homologous-allele balance) are purple. Inset zoomed on Chromosome 21 with homozygous *RUNX1* deletion shown by arrow. Plot on right shows homologous copy-ratio histogram obtained by binning at 0.04 resolution (*y*-axis); the length of each block corresponds to the (haploid) genomic fraction (*x*-axis) of each corresponding segment. B. Plot of relative copy number of *RUNX1* and *AKT3* in tumor and normal DNA from samples suspected to harbor *RUNX1* homozygous deletions. Relative quantities are normalized to *CDH7*, used as a diploid internal control. Findings are consistent with homozygous *RUNX1* deletion in both tumor samples, with lower purity of BR-M-174 tumor DNA.



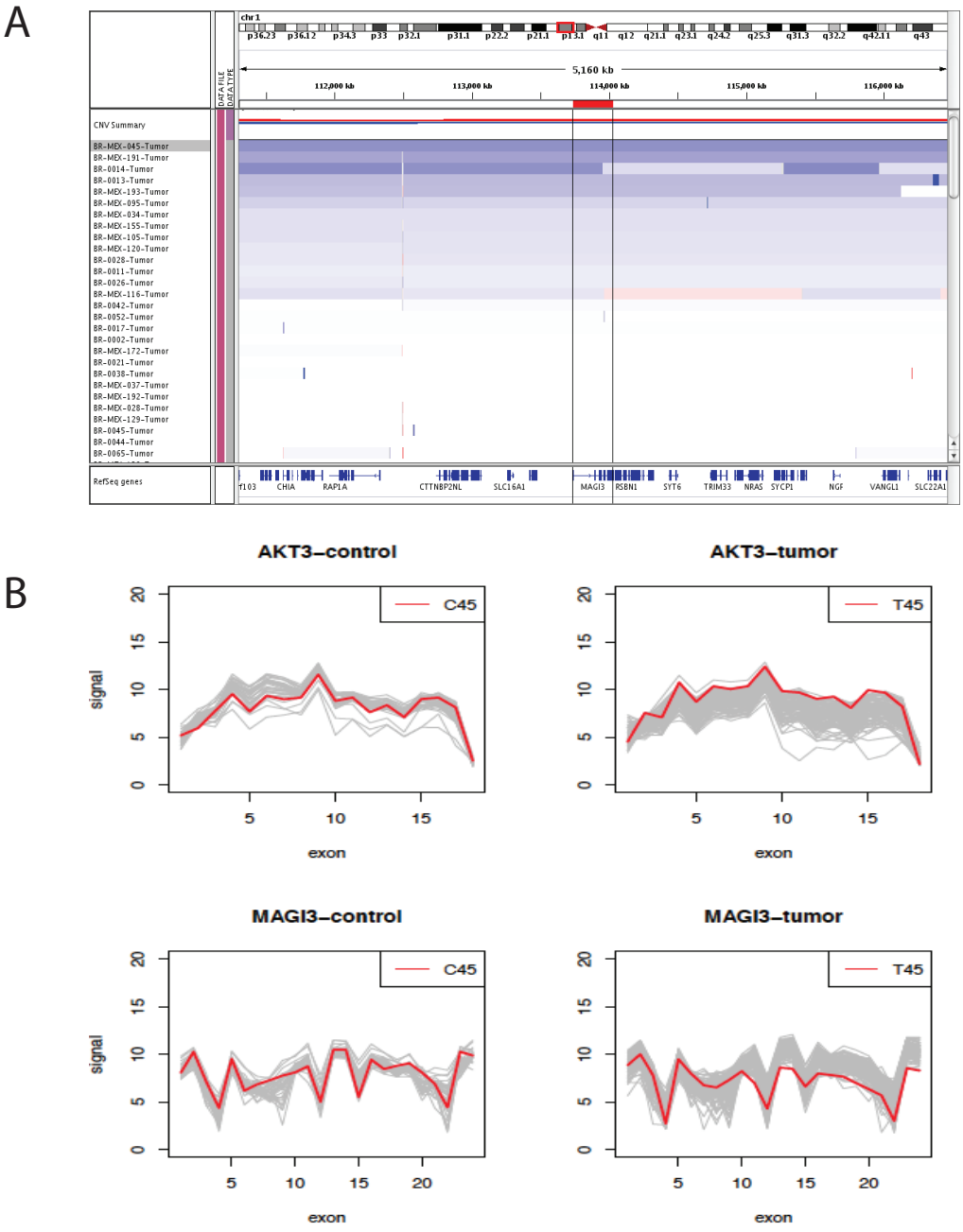
Supplementary Figure 8 - Significant GISTIC amplification and deletion peaks in our collection. Amplification in red and deletions in blue. Green line indicates FDR q-value=0.25. Chromosomal position indicated to right of plot with focus of amplification and deletion as labeled.



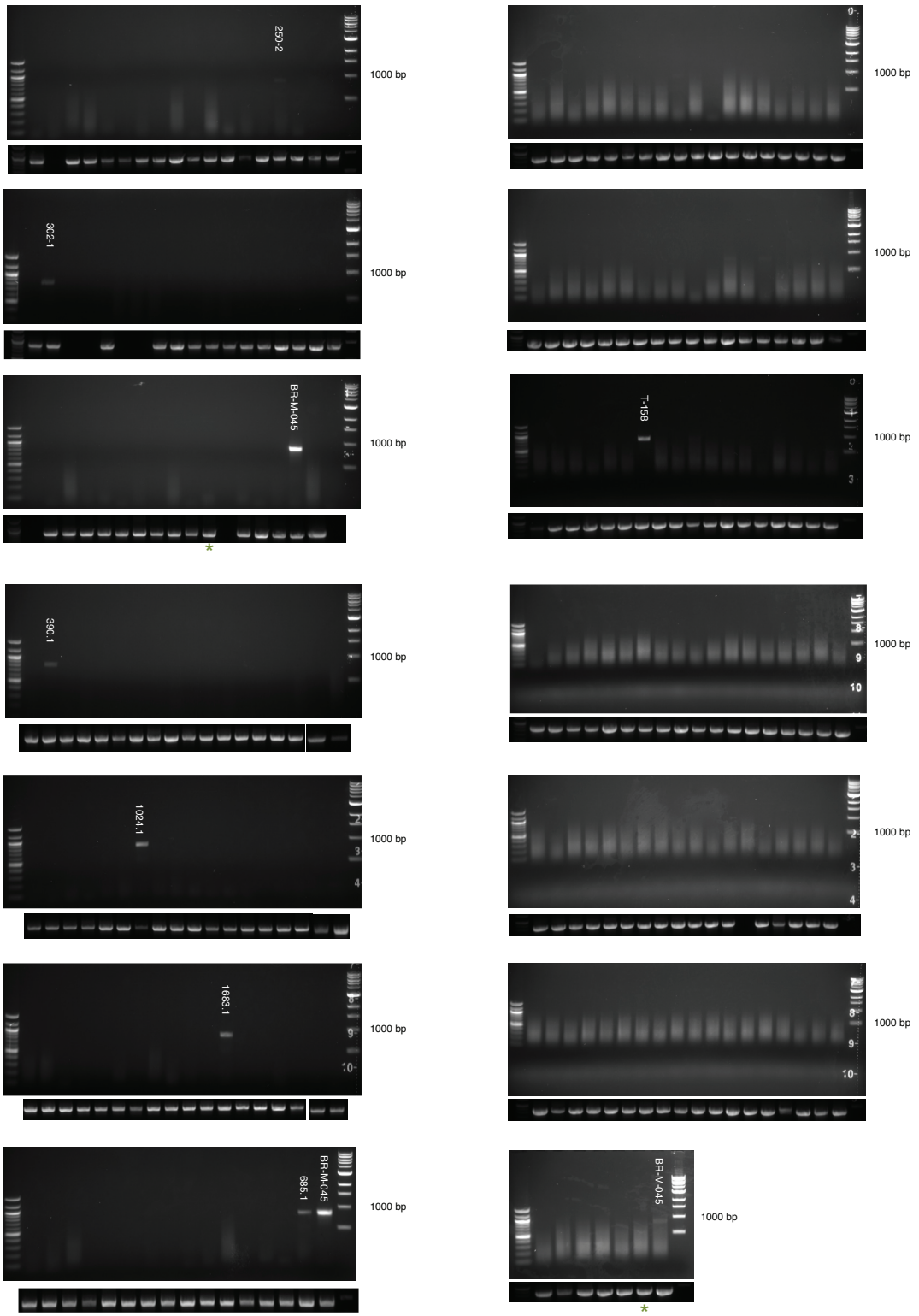
Supplementary Figure 9 - *ERBB2* copy number and mutation status across samples. DNA amplification shown in red. Samples with indicated *ERBB2* somatic mutations shown on left.


Luminal A
BR-V-005
BR-M-028
BR-M-050
BR-M-098
BR-M-106
BR-M-154
BR-M-165
BR-M-200
Luminal B
BR-V-009
BR-M-082
BR-M-116
BR-M-123
Normal-Like
BR-V-007
BR-M-198
Unknown
BR-V-003
Her2
BR-V-002
BR-V-004
BR-V-006
BR-V-008
BR-M-015
Basal
BR-V-070
BR-M-045

Supplementary Figure 10 - Somatic rearrangements observed in 22 whole-genomes. Genome-wide Circos plots shown organized by breast expression subtype. Chromosomal position shown in outer ring, copy number shown in inner ring. Inter-chromosomal rearrangements - red lines; intra-chromosomal rearrangements - green lines.



Supplementary Figure 11 - *MAGI3* copy number status and *MAGI3-AKT3* expression. A. Copy number of *MAGI3* as determined by SNP array. BR-M-045 sample with fusion gene shown at top. B. Expression levels of *MAGI3* and *AKT3* in tumor and normal control samples as determined by exon arrays. Red line indicates exon expression profile of *MAGI3* and *AKT3* in BR-M-045 sample.



Supplementary Figure 12 - Recurrence of *MAGI3-AKT3* fusion across breast cancer samples. Sample identity indicated for lanes with positive band. BR-M-045 used as positive control and appears multiple times. Green asterisk indicates repeated sample.

Supplementary Methods

A. Sample and Collection Attributes, DNA/RNA Collection, and Quality Control

Clinical cohorts

Mexican samples were collected under an IRB approved protocol during the 2008-2010 period at the Instituto de Enfermedades de la Mama – FUCAM A. C. hospital. Tumor, normal adjacent tissue, and peripheral blood were obtained from each patient after informed consent during surgery by S.R.C. for tumor resection. After macroscopic inspection by a pathologist, sections of tumor and normal tissues were immediately frozen in liquid nitrogen and stored at -80°C until further processing. A section of these tissues were formalin fixed, paraffin embedded (FFPE) and 5-micron sections were stained using hematoxylin and eosin (H&E) for confirmation of diagnosis, assessing grade, and tumor cell content evaluation. Estrogen and progesterone receptors as well as HER2 expression were evaluated using the ER/PR pharmDx and HercepTest, respectively (Dako, Denmark).

Fresh frozen Vietnamese samples were acquired from the BioServe commercial tissue repository (www.bioserve.com) following careful review of IRB and informed consent documents applicable to each sample. According to their guidelines, a board certified pathologist reviewed all samples to confirm diagnosis, assess grade, and evaluate tumor cell content. H&E slides were provided for each sample to confirm of diagnosis. ER/PR/HER2 expression status by immunohistochemistry was available only if provided by the original hospital responsible for specimen collection.

DNA/RNA extraction

Mexican Samples: After tumor cell content confirmation, DNA and RNA were extracted from the frozen tissues and peripheral blood lymphocytes using the AllPrep DNA/RNA mini kit (Qiagen, Valencia, CA) according to manufacturers instructions. DNA integrity was evaluated by 1% agarose gel electrophoresis and RNA integrity by capillary electrophoresis using the Bioanalyzer system (Agilent, Santa Clara, CA). Only samples with RNA integrity number (RIN) greater than 6.0, were used for expression microarray analysis.

Vietnamese Samples: DNA extraction was performed on fresh frozen tumor and adjacent normal tissue using DNAQuik reagents developed by BioServe. DNA was run on 1% agarose gels to assess structural integrity. RNA extraction was performed using Trizol (Qiagen) and the quality was determined using the Bioanalyzer system. RNA with a RIN score >6.0 was stored at -80°C until use.

DNA quality control

We used standard Broad Institute protocols as recently described[2-4]. Tumor and normal DNA concentration was measured using PicoGreen® dsDNA Quantitation Reagent (Invitrogen, Carlsbad, CA). A minimum DNA concentration of 60 ng/µl was required for sequencing. In select cases where concentration was <60 ng/µl, ethanol precipitation and re-suspension was required to increase concentration. Gel electrophoresis confirmed that the large majority of DNA was high molecular weight. We prepared reserve stocks of each sample using whole genome amplification (WGA) for use in subsequent validation efforts. All Illumina sequencing libraries were created with the native DNA. The identities of all tumor and normal DNA samples (native and WGA

product) were confirmed by mass spectrometric fingerprint genotyping of 24 common SNPs (Sequenom, San Diego, CA).

B. cDNA Microarrays and Expression Subtype Determination

Expression Microarrays

cDNA generated from RNA was hybridized on human whole-transcript microarrays (Human Gene ST 1.0, Affymetrix, Santa Clara CA), according to manufacturer's instructions. Samples classified as 141 Mexican samples included 35 normal and 106 tumors that were processed at the Affymetrix Unit of the Instituto Nacional de Medicina Genomica (INMEGEN) in Mexico City. cDNA from tumor for all Vietnamese samples was processed at the Genetic Analysis Platform (GAP) at the Broad Institute in Cambridge, MA.

Breast expression subtyping

Raw gene expression profiles from all 201 samples were obtained after low-level analysis and quality assessment for the two sets separately since no comparison between the two populations was planned. Probe level data on each set were log2 transformed, background corrected using RMA[5] and normalized using quantile normalization[6]. These algorithms are coded in the "oligo" package in Bioconductor. Gene expression data was further processed to determine breast cancer molecular subtypes according to the expression profiles classification of PAM50[7]. The PAM50 gene expression test aims at classifying breast cancer tumors into 5 known intrinsic subtypes: Luminal A, Luminal B, Her2, Basal-like, and Normal-like and also provides a continuous risk of recurrence (ROR) score based on the similarity of an individual sample to the prototypic subtypes.

C. Single-Nucleotide Polymorphism (SNP) Array Based Analysis

Single-nucleotide polymorphism arrays

Non-WGA genomic DNA from tumor and paired normal samples was processed using Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Inc.) according to manufacturer's protocols. DNA was digested with NspI and StyI enzymes (New England Biolabs), ligated to the respective Affymetrix adapters using T4 DNA ligase (New England Biolabs), amplified (Clontech), purified using magnetic beads (Agencourt), labeled, fragmented, and hybridized to the arrays. Following hybridization, the arrays were washed and stained with streptavidin-phycoerythrin (Invitrogen). Array preparation and scanning was performed at the genotyping core laboratory of INMEGEN and GAP at the Broad Institute for the Mexican and Vietnamese samples respectively.

Copy-number assessment

Data preprocessing was performed using Affymetrix Power Tools. Copy number data was evaluated after segmenting the log 2 ratios between tumor and paired normal levels on a sample basis. Quality control, data integrity, segmentation and copy number analysis were performed as previously described[8]. Segmented copy number data was visualized with the Integrative Genomics Viewer (IGV)[9]. The Genomic Identification of Significant Targets in Cancer (GISTIC) algorithm was used to identify broad and focal regions of copy number alterations in individual samples as described[10,11].

Purity, ploidy, and allele-specific copy number analysis

SNP Array data was analyzed using the HAPSEG[12] and ABSOLUTE[13] algorithms to infer the tumor purity, average ploidy, and allele-specific copy number levels. Allelic fraction for each tumor was calculated, indicative of the fraction of sequence reads expected to harbor the non-reference allele at a locus with a somatic mutation existing at a single copy per nucleus.

Ancestry Analysis

A total of 140 sample SNP arrays were used for the ancestry analysis: 100 samples from Mexico and 40 from Vietnam. Genotypes of 301,219 common SNPs in three genotyping platforms (SNP 6.0, Affymetrix 500K, and Illumina 1M) were used. 196 HapMap samples (CEU: northern European ancestry; YRI, Africans from Nigeria and CHB+JPT, east Asian population) and 71 native Mexican samples were included as parental populations for the analysis, while 161 Mexican mestizo samples from the Mexican Genome Diversity Project (MGDP)[1] were included to evaluate ancestry proportions in the general Mexican population. Four major quality control tests were performed: 1) Missing rate per person excluding individuals with more than 5% missing genotypes, 2) Missing rate per SNP: only SNPs with a 95% genotyping rate were included, 3) Exclusion of markers that failed the Hardy-Weinberg Equilibrium test at 0.00001 significance threshold, 4) Identity-by-descent (IBD) test to assess quality on the full set of samples. Quality of all samples was good and no familial relationships were found. Principal components analysis (PCA) was used to detect population substructure using genome-wide data using EIGENSOFT 3.0[14]. Individual average ancestral proportions were determined using ADMIXTURE 1.1 software[15]. Based on the origin of the samples, the value of K was chosen to be 4 meaning that four parental groups were considered to quantify ancestral contribution: CEU, YRI, CHB+JPT and NATMEX and to explain the major substructure in this set of 140 individuals.

D. Sequence Data Generation

A total of 125 samples were initially sequenced with 120 successfully completed (Supplementary Figure 1). Ninety-seven underwent whole exome sequencing only and 5 samples whole genome sequencing only. Additional 18 samples were sequenced with both methods. Tumor and normal samples were sequenced according to the manufacturer's protocols (Illumina, San Diego, CA) as previously described[2-4] with a brief summary provided below.

Whole Genome Sequencing Library Construction

We followed established protocols at the Broad Institute as previously described[4]. A total of 1 μg of genomic DNA was sheared to a range of 100-700 bp. Each of the resulting WGS libraries was sequenced on an average of 13 flow cells lanes of the Illumina GA-II or HiSeq sequencers. Using 101 bp paired-end reads, we aimed to reach 30X average genomic coverage for each of the tumor and normal genomes. The mean coverage achieved was 36x in tumors and 38x in normals.

Whole Exome Sequencing Library Construction

We follow the procedure described by Gnirke *et al.*[16] adapted for production-scale exome capture libraries as described[4]. Resulting exome sequencing libraries were sequenced on 3 lanes of an Illumina GA-II sequencer, using 76 bp paired-end reads. The mean coverage achieved was 141x in the tumors and 133x in the normals.

Illumina sequencing

Libraries were quantified using a SYBR Green qPCR protocol with specific probes for the ends of the adapters[4]. Libraries were normalized to 2nM and then denatured using 0.1 N NaOH. Cluster amplification of denatured templates occurred according to manufacturer's protocol (Illumina) using V2 Chemistry and V2 Flowcells (1.4mm channel width). SYBR Green dye was added to all flowcell lanes to provide a quality control checkpoint after cluster amplification to ensure optimal cluster densities on the flowcells. Flowcells were paired-end sequenced on Genome Analyzer II or HiSeq machines, using V3 Sequencing-by-Synthesis kits and analyzed with the Illumina v1.3.4 pipeline. Standard quality control metrics including error rates, % passing filter reads, and total Gb produced were used to characterize process performance prior to downstream analysis. The Illumina pipeline generates data files that contain the reads and qualities.

E. Sequence Data Processing

Sequencing data were processed using two consecutive pipelines[2-4]:

(1) Sequencing data-processing pipeline – "Picard" - uses the reads and qualities produced by the Illumina software for all lanes and libraries generated for a single sample (either tumor or normal) and produces a single BAM file (http://samtools.sourceforge.net/SAM1.pdf) representing the sample. The final BAM file stores all reads and calibrated qualities along with their alignments to the genome.

(2) Cancer genome analysis pipeline – "Firehose" – takes the BAM files for the tumor and patient-matched normal samples and performs analyses including quality control, local-realignment, mutation calling, small insertion and deletion identification, rearrangement detection, coverage calculations and others as described briefly below and more extensively in Stransky et al[4].

The Cancer Genome Analysis Pipeline ("Firehose")

The pipeline represents a set of tools for analyzing massively parallel sequencing data for both tumor DNA samples and their patient-matched normal DNA samples. Firehose uses GenePattern[17] as its execution engine for pipelines and modules based on input files specified by Firehose. The pipeline contains the following steps (described in detail in Chapman et al.[3] and Stransky et al.[4]) (Supplementary Figure 2):

1. Quality control – confirms identity if individual tumor and normal to avoid mix-ups between tumor and normal data for the same individual.
2. Local-realignment of reads – realigns sites potentially harboring small insertions or deletions in either the tumor or the matched normal to decrease the number of false positive single nucleotide variations caused by misaligned reads.
3. Identification of somatic single nucleotide variations (SSNVs) – MuTect algorithm – candidate SSNVs were detected using a statistical analysis of the bases and qualities in the tumor and normal BAMs (described in detail below).
4. Identification of somatic small insertions and deletions – Indelocator algorithm – putative somatic events were first identified within the tumor BAM file an then filtered out using the corresponding normal data.
5. Identification of inter-chromosomal and large intra-chromosomal structural

rearrangements – dRanger algorithm – candidate rearrangements were identified as groups of paired-end reads which connected genomic regions with an unexpected orientation and/or distance. When possible, breakpoints are mapped to basepair resolution using BreakPointer[2].

6. Mutation rate calculation – we calculated base mutation rates using the detected mutations (SSNVs and indels) and the coverage statistics.
7. Identification of significantly mutated genes – MutSig algorithm – genes that harbored a greater number of mutations than expected by chance were detected by comparing the observed number of mutations across the samples to the expected number based on the background mutation rates and the covered bases in all samples. Genes list in Figure 1 of main manuscript were selected after filtering that included: eliminating gene with q value of >0.1 after correction for multiple hypothesis testing, manual review of reads, and fewer than 2 mutations per sample. Subsequent input of samples that failed orthogonal validation were used to correct the background mutation rate.
8. Mutation annotation - Detected point mutation and indels calls were annotated with the annotation pipeline Oncotator.

Candidate SSNV detection (MuTect)

Detection of somatic mutations, from paired tumor/normal next-generation sequencing data BAM files[18], was performed using *muTect* (Cibulskis K. *et al*., in preparation), available at http://www.broadinstitute.org/cancer/cga/mutect. In brief, muTect consists of three steps.

1. Preprocessing

For all regions where there is at least one aligned read in the tumor data, or a subset of those regions that are defined by the analyst, aligned reads in the tumor and normal sequencing data are processed as follows. To remove low-quality sequence data, any read where the observed Phred-like base quality score[19] at the candidate locus is <= 5 is removed from mutation analysis. In addition, to decrease the number of poorly aligned reads, those reads where the sum of the Phred-like base quality scores for all non-reference bases within that read is ≥ 100, are also removed from mutation analysis.

2. Core Statistical Test to Identify Candidate Somatic Mutations

A statistical analysis that identifies sites that are likely to carry somatic mutations with high confidence. The statistical analysis predicts a somatic mutation by using two Bayesian classifiers – the first aims to detect whether the tumor DNA sequence is non-reference at a given site and, for those sites that are found as non-reference, the second classifier assesses whether the normal DNA sequence from the same individual does not carry the variant allele. In practice the classification is performed by calculating a LOD score (log odds ratio) and comparing it to a cutoff determined by the log ratio of prior probabilities of the considered events. We define this LOD score as follows:

For each site we denote the reference allele as $r \in \{A, C, G, T\}$ and denote by b_i and the called base of the i-th read (i=1…d) that covers the site and by e_i the probability of error of that base call based on the Phred base quality score, q_i, at base I, where $e_i = 10^{-q_i/10}$. To call a variant in the tumor we compare two models: a model M_f^m in which we explain all the observed sequence data at each base with a variant m having an allele fraction f in

addition to sequencing noise; and a model M_0 in which there is no variant (i.e. f=0) and we explain all the observed sequence data at each base as a result of sequencing noise.

The likelihood of the model M_f^m is given by

$$L\left[M_f^m\right] = P(\{b_i\}\,|\{e_i\}, r, m, f) = \prod_{i=1}^{d} P(\,b_i|e_i, r, m, f)$$

assuming the errors are statistically independent, where

$$P(\,b_i|e_i, r, m, f) = \left\{ \begin{array}{cc} f\,{}^{e_i}/_3 + (1-f)(1-e_i) & if\ b_i = r \\ f(1-e_i) + (1-f)\,{}^{e_i}/_3 & if\ b_i = m \\ {}^{e_i}/_3 & otherwise \end{array} \right\}$$

assuming all machine errors are equally likely. The classification is then performed by calculating a LOD score (log odds). We calculate

$$LOD_T = -\log_{10}\left(\frac{L\left[M_f^m\right]}{L[M_0])}\right)$$

For the normal we the calculate LOD_N using the same formula, but f is fixed to be 0.5 as the expected case for a heterozygous variant.

Since we expect somatic mutations to occur at a rate of ~1 in a Mb in the average tumor type, we require $\mathrm{LOD}_T > \log_{10}\left(0.5\times10^{-6}\right) \approx 6.3$ which guarantees that our false positive rate, due to noise in the tumor, is less than 0.5 per Mb of sequenced territory. In the normal, not at dbSNP sites, we require $\mathrm{LOD}_N > \log_{10}\left(0.5\times10^{-2}\right) \approx 2.3$ since non-dbSNP germline variants occur roughly at a rate of 100 per Mb[20]. This cutoff guarantees that the false positive somatic call rate, due to missing the variant in the normal, is also less than half the somatic mutation rate if all mutation candidates were exactly at the LOD thresholds and no further filtering was applied. Candidates having $\mathrm{LOD_T}$ and $\mathrm{LOD_N}$ scores higher than the above-described thresholds are then passed through a set of post-processing filters.

3. Post Processing of Candidate Mutations

Post-processing of candidate somatic mutations is performed to eliminate artifacts of next-generation sequencing, short read alignment and hybrid capture which often exhibit correlated noise and therefore are not rejected by the core detection statistic. In order to eliminate these false positives, we reject somatic mutation candidates if are ≥ 3 reads with insertions within an 11bp window centered on the candidate mutation OR if there are ≥ 3 reads with deletions within the same 11bp window. We reject candidates where ≥ 50% of

the reads at the locus have a mapping quality of zero. We also reject candidates where the normal sequence is heterozygous. We reject candidates with ≥ 2 observations of the alternate allele in the normal where the sum of their quality scores ≥ 20. Finally, we apply a Fisher's exact test with $P < 0.05$ to test whether the direction of reads supporting the alternate are biased compared to the direction of reads supporting the reference in the tumor and reject the candidate mutation if there is evidence for directional bias.

F. High-Throughput Experimental Validation of Point Mutations, Indels, and Rearrangements

Somatic Mutations

We obtained independent validation for 494 candidate mutations using mass spectrometric genotyping (Sequenom) of the tumor and normal DNA, or alternative next-generation sequencing of tumor DNA using 454 pyrosequencing, Pacific Biosciences SMRT cell targeted sequencing (for *PIK3CA* and *TP53* mutations), and Illumina exome sequencing of frozen tumor-matched formalin-fixed paraffin embedded tissues. Whole-genome amplified DNA was used for all validation experiments except Illumina sequencing of FFPE tissue where non-amplified DNA was used. Mutations selected for validation included all candidate protein-coding mutations in significantly-mutated genes in the Illumina whole-exome data with q-value < 0.2, and all genes in significantly-mutated genesets with q-value < 0.1.

Somatic Rearrangements

We attempted PCR validation of all dRanger identified rearrangements across the 22 genomes that met the criteria: 1. dRanger score >= 4; 2. event results in duplication/deletion of entire exons, or results in *in frame* protein/transcript fusion. Two unique primer pairs were designed for each event.

G. Other Methods

Germline mutation calling

Mutation calling was performed using Unified Genotyper as previously described[21]. Called germline variations were compared to a list of functionally annotated variations to assess for pathogenic significance[22].

qPCR to confirm absolute copy number

To verify whether two samples, MEX-BR-45 and MEX-BR 174, had homozygous deleted *RUNX1*, qPCR was run following the manufacture's protocol for the Brilliant II SYBR® Green qPCR Master Mix with a gDNA input of 15ng per reaction in triplicates. The following primers were used.

	Forward	Reverse
RUNX1	GGCTCCACTCAGCATGGCACA	GGCGTACTCGCTGCCCTCA
CDH7	CGCCCCTGAATTTGCCATGGACT	AGACTCTTACCTTGCCCCGGCT
AKT3	TCCTGTCCCTGCTGTTTACCCTGC	CGCTCCTCAGAGAACACCCGC

CDH7 was used as the diploid endogenous control, and AKT3 as a representative amplified region in these two tumors. Tumor normal pairs were run on an ABI Prism 7900HT with an annealing temperature of 61°C at one minute and with the extension time at 45 seconds. Ct values, ΔΔCt

values, and relative quantification for each target gene was determined by RQ Manager 1.2 software (Applied Biosystems).

Independent validation of balanced translocation involving MAGI3 and AKT3 in tumor and normal genomic DNA from case BR-M-045

Aliquots of tumor and blood normal DNA was obtained for case BR-M-045. PCR amplification was performed using AccuPrime Taq DNA Polymerase (Invitrogen) with the following primer pairs.

	Forward	Reverse
AKT3	TGCTGCCAATCATGTGCCCGACA	TCCTCTGACCCCCAGGGCCA
MAGI3	ATGAGGGTTCCCTTTTCACC	AAATGGGCAAAACATTGGAA
5'-AKT3, 3'-MAGI3	GTGCTGCTGCTTCACTTCGGGTG	TGCCTTCTGGCCTTTCGGCACACCA
5'-MAGI3, 3'-AKT3	ATGAGGGTTCCCTTTTCACC	TGCTGCCAATCATGTGCCCGACA

PCR amplification of MAGI3-AKT3 fusion gene from patient cDNA

Total RNA was obtained from tumor for case BR-M-045. Double stranded cDNA was made from 200 ng of RNA using the Superscript III cDNA Synthesis kit (Invitrogen) with and without the inclusion of RNA polymerase for first strand synthesis. PCR amplification was performed using the AccuPrime Taq DNA Polymerase on double strand cDNA using forward primer (5'-AAGCCCCTGAAGACTGTGAA-3') in *MAGI3* and reverse primer (5'-ACTTGCCTTCTCTCGAACCA-3') in *AKT3*. 35 PCR cycles were performed as follows: 95°C for 2 min, 95°C for 30 sec, 57.2°C for 30 sec, 72°C for 100 sec, 72°C for 5 min, 4°C hold.

Assessment of expressed MAGI3-AKT3 fusion prevalence across breast cancer samples

Tumor RNA was obtained from the Massachusetts General Hospital via D.C.S., the Susan F. Smith Women's Cancers Tissue Repository at the Brigham and Women's Hospital via A.L.R., and from the Instituto de Enfermedades de la Mama FUCAM via S.R.C and A.H.M. First strand cDNA was made from 50 ng total RNA using the Superscript III First Strand cDNA Synthesis System (Invitrogen) and purified using the MinElute PCR Purification Kit (Qiagen). PCR amplification was performed on the purified first strand cDNA using primers spanning the *MAGI3-AKT3* breakpoint. The forward primer is located on exon six of *MAGI3* (5'-AAGCCCCTGAAGACTGTGAA-3') and the reverse primer is located on exon five of *AKT3* (5'-ACTTGCCTTCTCTCGAACCA-3'). Forty PCR cycles were performed as follows: 95°C for 2 min, 95°C for 30 sec, 57.2°C for 30 sec, 72°C for 100 sec, 72°C for 5 min, 4°C hold. An 833-bp PCR product was expected for the fusion. Bands were excised and purified using the QIAquick Gel Extraction Kit (Qiagen). TOPO TA Cloning (Invitrogen) was performed prior to Sanger sequencing.

Gateway Cloning of MAGI3-AKT3 fusion for validation experiments

Double stranded cDNA was generated from total RNA from case BR-M-045 as described above, and purified using the MinElute PCR Purification Kit (Qiagen). Two PCR products were generated with overlapping sequence using Gateway® cloning compatible primers and Taq DNA polymerase HiFi (Invitrogen). Fragment1: attb1 (5'-GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCGAAGACGCTGAAG-3') and *AKT3* reverse primer (5'-ACTTGCCTTCTCTCGAACCA-3'), Fragment 2: attb2 (5'-GGGGACCACTTTGTACAAGAAAGCTGGGTCTTATTCTCGTCCACTTGCAGA-3') and the *MAGI3* forward primer (5'-AAGCCCCTGAAGACTGTGAA-3').

The 5' and 3' overlapping PCR products were added to the BP reaction with pDONR221 plasmid. In vitro recombination resulted in insertion of the full-length fusion as tested by restriction digest and PCR across the overlapping region. The fusion was subcloned into the pBabe-Puro and pLX304-Blast destination vectors.

Cell culture
ZR75 cells were maintained in RPMI-1640 media (Cellgro; Manassas, VA) supplemented with 10% fetal bovine serum (FBS) (GIBCO; Carlsbad, CA). Rat-1 fibroblasts and HEK-293T cells were maintained in DMEM (Cellgro) supplemented with 10% FBS (GIBCO).

Plasmids and lentivirus production
Plasmids (pLX304 and pLX304-MAGI3-Akt3): pLX304 plasmid constructs were co-transfected in HEK-293T cells with the packaging vectors pCMV-VSVG and psPAX2 using polyethylenimine (PEI). Lentiviral supernatents were harvested 48 hr after transfection, passed through a 0.45 μm filter and used to infect target cells.
HA-Akt1 Glu17Lys (E17K) was constructed by site-directed mutagenesis. Cells were transfected with pcDNA3-Akt1-E17K using Lipofectamine 2000 (Invitrogen; Carlsbad, CA) according to the manufacturer's protocol.

Growth factors and inhibitors
Serum-starved cells were stimulated with recombinant human IGF-1 (R&D Systems; Minneapolis, MN) at a final concentration of 100 ng/mL for 20 min. Serum-starved cells were exposed to the pan-Akt inhibitors MK-2206 (Active Biochem; Wanchai, China) and GSK-690693 (SynKinase; Shanghai, China) at final concentrations of 1 μM for 20 min.

Immunoblotting
ZR75 cells were infected with viral supernatent and 5 μg/mL polybrene (Millipore; Billerica, MA) or transfected with Akt1-E17K for 48 hr prior to serum-starvation for an additional 16 hr. Cells were exposed to IGF-1 or inhibitors, washed with ice-cold PBS and lysed in ice-cold lysis buffer (1% NP-40, 150 mM NaCl, 10 mM KCl, 20 mM Tris-HCl [pH 7.5], 0.1% SDC, 0.1% SDS, protease inhibitor cocktail [Sigma-Aldrich; St. Louis, MO], 50 nM calyculin A [Sigma-Aldrich], 1 mM sodium pyrophosphate, 20 mM sodium fluoride) for 20 min on ice. Cell extracts were cleared by centrifugation at 13,000 rpm for 10 min at 4°C and protein concentration was measured with the Bio-Rad protein assay reagent (Bio-Rad; Hercules, CA). Lysates were resolved by SDS-PAGE and transferred to nitrocellulose membrane (Bio-Rad). Membranes were blocked in TBST buffer (10 mM Tris-HCl [pH 8], 150 mM NaCl, 0.2% Tween 20) containing 5% (w/v) non-fat dry milk and then incubated with primary antibodies diluted in TBST buffer containing 2% (w/v) non-fat dry milk at 4°C overnight. Membranes were washed in TBST and incubated with horseradish-peroxidase-conjugated secondary antibodies for 1 hr at room temperature. Membranes were washed in TBST and developed using a chemiluminescent substrate (Millipore).

Antibodies

Anti-Akt, anti-phospho Akt (Ser473), anti-GSK3β and anti-phospho GSK3β (Ser9) antibodies were obtained from Cell Signaling Technology (Danvers, MA). Anti-β-actin antibody was purchased from Sigma-Aldrich. Horseradish peroxidase-conjugated anti-mouse and anti-rabbit immunoglobulin (IgG) antibodies were purchased from Millipore.

Focus formation assay

Rat-1 cells were infected with viral supernatent and 5 μg/mL polybrene (Millipore). 48 hours after infection, cells were split into 100 mm dishes for focus formation. 8 days later, cells were fixed with ice-cold methanol and stained with crystal violet (0.5% crystal violet, 25% methanol). Images of cells and foci were acquired using an inverted microscope (Eclipse T*i*; Nikon, Melville, NY).

Supplementary Tables

	Expression Subtype¶						
				Mexico			
	Luminal A	Luminal B	Her2	Basal	Normal-Like	Unknown	Sum
Histology							
Ductal	19	11	7	4	3	2	**46**
Lobular	2	1	1	0	0	0	**4**
Tubular	1	0	0	0	0	0	**1**
Medullary	0	0	0	1	0	0	**1**
Mucinous	1	0	0	0	0	0	**1**
Mixed	1	1	1	0	0	0	**3**
Sum	**24**	**13**	**9**	**5**	**3**	**2**	**56**

				Vietnam			
	Luminal A	Luminal B	Her2	Basal	Normal-Like	Unknown	Sum
Histology							
Ductal	10	7	10	7	5	2	**41**
DCIS	4	1	2	1	1	0	**9**
Mucinous	0	1	0	0	0	1	**2**
Mixed	0	0	0	0	0	0	**0**
Sum	**14**	**9**	**12**	**8**	**6**	**3**	**52**

¶ Based on PAM-50 classification of exon array data
DCIS = Ductal carcinoma in situ

Supplementary Table 1 - Sequenced samples by histology and expression subtype.

Sequencing ID	Subtype¶	Age	Country	Gender	HER 2 Positive*	Her2 Amplified§	Menopausal Status	Stage	ER Positive*	Grade	Histology	PR Positive*	Purity#	Ploidy#	Silent Rate$	Non-silent rate
BR-V-002	HER2	46	Vietnam	Female	-	Yes	Premenopausal	III	Negative	III	Infiltrating Ductal Carcinoma	Negative	0.27	1.95	0.72	2.62
BR-V-003	-	60	Vietnam	Female	-	No	Post Menopausal	II	-	III	Infiltrating Ductal Carcinoma	-	-	-	0.21	0.44
BR-V-007	Normal	38	Vietnam	Female	-	No	Premenopausal	III	-	III	Infiltrating Ductal Carcinoma	-	0.23	2.16	0.17	0.61
BR-V-008	HER2	52	Vietnam	Female	-	Yes	Perimenopausal	III	-	III	Infiltrating Ductal Carcinoma	-	0.27	2.01	0.14	1
BR-V-009	Luminal B	53	Vietnam	Female	-	No	Post Menopausal	III	Positive	III	Infiltrating Ductal Carcinoma	Positive	0.52	1.99	0.54	2.07
BR-V-011	Luminal A	69	Vietnam	Female	-	No	Post Menopausal	II	-	-	Carcinoma, Ductal In Situ (DCIS)	-	0.7	1.88	0.17	0.72
BR-V-012	HER2	42	Vietnam	Female	-	Yes	Premenopausal	II	Negative	-	Carcinoma, Ductal In Situ (DCIS)	Negative	0.28	3.24	0.13	0.87
BR-V-013	Luminal A	38	Vietnam	Female	-	No	Premenopausal	II	Positive	-	Carcinoma, Ductal In Situ (DCIS)	Positive	0.43	3.87	0.24	0.68
BR-V-014	HER2	45	Vietnam	Female	-	Yes	Premenopausal	II	Negative	-	Infiltrating Ductal Carcinoma	Negative	0.73	3.98	0.68	1.56
BR-V-015	HER2	43	Vietnam	Female	-	Yes	Premenopausal	II	-	-	Carcinoma, Ductal In Situ (DCIS)	-	0.22	2.04	0.1	0.37
BR-V-016	Luminal B	46	Vietnam	Female	-	No	Premenopausal	II	Negative	-	Infiltrating Ductal Carcinoma	Negative	0.6	1.97	0.1	0.51
BR-V-017	Luminal A	43	Vietnam	Female	-	No	Premenopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.33	3.83	0.035	0.67
BR-V-018	Normal	34	Vietnam	Female	-	Yes	Premenopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.36	2.12	0.067	0.44
BR-V-019	Luminal B	59	Vietnam	Female	-	No	-	II	-	-	Carcinoma, Mucinous	-	0.8	1.97	0.1	1.05
BR-V-020	Luminal A	39	Vietnam	Female	-	No	Premenopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.76	2.04	0.1	0.54
BR-V-021	Luminal B	48	Vietnam	Female	-	No	Perimenopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.61	2.28	0.1	0.51
BR-V-022	Basal	40	Vietnam	Female	-	No	Premenopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.43	1.75	0.34	1.64
BR-V-023	Luminal A	40	Vietnam	Female	-	No	Premenopausal	II	-	-	Carcinoma, Ductal In Situ (DCIS)	-	0.28	1.92	0.38	1.47
BR-V-024	HER2	49	Vietnam	Female	-	Yes	Premenopausal	II	Negative	-	Infiltrating Ductal Carcinoma	Negative	-	-	0.14	1
BR-V-026	Luminal B	43	Vietnam	Female	-	No	Premenopausal	II	Positive	-	Carcinoma, Ductal In Situ (DCIS)	Positive	0.72	3.59	0.28	0.35
BR-V-027	HER2	41	Vietnam	Female	-	Yes	Premenopausal	II	Negative	-	Infiltrating Ductal Carcinoma	Negative	0.29	1.92	1.09	4
BR-V-028	HER2	52	Vietnam	Female	-	No	Perimenopausal	III	-	-	Infiltrating Ductal Carcinoma	-	0.3	3.25	0.43	1.32
BR-V-030	Luminal B	58	Vietnam	Female	-	Yes	Post Menopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.82	1.92	0.34	0.85
BR-V-031	Luminal A	81	Vietnam	Female	-	No	Post Menopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.28	2.09	0.24	0.62
BR-V-032	Luminal A	39	Vietnam	Female	-	No	Premenopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.28	4.05	0.34	0.72
BR-V-033	Normal	53	Vietnam	Female	-	Yes	Perimenopausal	III	-	-	Infiltrating Ductal Carcinoma	-	0.2	3.21	0.76	2.24
BR-V-034	Basal	48	Vietnam	Female	-	No	Premenopausal	II	-	-	Infiltrating Ductal Carcinoma	-	-	-	0.59	1.54
BR-V-036	Luminal B	42	Vietnam	Female	-	No	Premenopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.34	3.81	0.38	2.02
BR-V-037	Luminal B	56	Vietnam	Female	-	No	Post Menopausal	III	-	-	Infiltrating Ductal Carcinoma	-	0.79	2.08	0.73	2.61
BR-V-038	Normal	55	Vietnam	Female	-	No	Post Menopausal	I	Negative	-	Infiltrating Ductal Carcinoma	Negative	0.79	2.28	0.034	0.44
BR-V-039	Luminal A	38	Vietnam	Female	-	No	Premenopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.54	2.03	0.14	0.55
BR-V-040	-	48	Vietnam	Female	-	No	Premenopausal	I	-	-	Carcinoma, Mucinous	-	0.28	3.96	0.21	0.72
BR-V-042	HER2	39	Vietnam	Female	-	Yes	Premenopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.22	1.66	0.31	0.94
BR-V-043	Basal	70	Vietnam	Female	-	No	Post Menopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.71	1.85	1.15	4.46
BR-V-044	Luminal A	50	Vietnam	Female	-	No	Perimenopausal	II	Positive	-	Carcinoma, Ductal In Situ (DCIS)	Positive	-	-	0.27	0.79
BR-V-045	Normal	54	Vietnam	Female	-	Yes	Post Menopausal	II	Negative	-	Carcinoma, Ductal In Situ (DCIS)	Negative	0.25	2.1	0.14	0.49
BR-V-047	Luminal A	38	Vietnam	Female	Negative	Yes	Premenopausal	I	Negative	-	Infiltrating Ductal Carcinoma	Positive	-	-	0.1	0.47
BR-V-048	Luminal A	49	Vietnam	Female	-	No	Perimenopausal	II	Positive	-	Infiltrating Ductal Carcinoma	Positive	0.19	2.13	0.34	0.98
BR-V-049	-	47	Vietnam	Female	-	No	Premenopausal	II	Negative	-	Infiltrating Ductal Carcinoma	Negative	0.34	2.21	0.14	0.44
BR-V-050	Basal	54	Vietnam	Female	-	Yes	Post Menopausal	II	-	-	Carcinoma, Ductal In Situ (DCIS)	-	0.47	2.05	0.34	1.85
BR-V-051	Basal	37	Vietnam	Female	-	No	Premenopausal	II	-	-	Infiltrating Ductal Carcinoma	-	0.62	1.63	0.17	0.34
BR-V-052	Luminal B	31	Vietnam	Female	-	No	Premenopausal	III	-	III	Infiltrating Ductal Carcinoma	-	0.19	2.08	0.38	0.96
BR-V-054	Luminal A	40	Vietnam	Female	-	No	Premenopausal	II	-	III	Infiltrating Ductal Carcinoma	-	0.67	1.94	0.31	0.31
BR-V-060	Normal	57	Vietnam	Female	-	Yes	Post Menopausal	II	-	III	Infiltrating Ductal Carcinoma	-	0.34	2.27	0.82	1.64
BR-V-064	Luminal A	41	Vietnam	Female	-	No	Premenopausal	II	-	III	Infiltrating Ductal Carcinoma	-	0.34	2.07	0.07	0.67
BR-V-067	HER2	55	Vietnam	Female	-	Yes	Post Menopausal	II	-	III	Infiltrating Ductal Carcinoma	-	0.59	3.25	1.41	3.63
BR-V-069	Basal	42	Vietnam	Female	-	No	Premenopausal	II	Negative	III	Infiltrating Ductal Carcinoma	Negative	-	-	0.11	0.43
BR-V-070	Basal	35	Vietnam	Female	-	Yes	Premenopausal	II	Negative	III	Infiltrating Ductal Carcinoma	Negative	0.28	2.31	0.17	0.82
BR-V-071	Basal	45	Vietnam	Female	-	No	Premenopausal	II	Negative	III	Infiltrating Ductal Carcinoma	Negative	0.3	1.99	0.34	1.41
BR-M-005	HER2	66	Mexico	Female	Positive	Yes	-	II	Negative	II	Infiltrating Ductal Carcinoma	Negative	0.59	2.08	0.17	0.68
BR-M-026	Luminal A	54	Mexico	Female	Negative	No	-	III	Positive	I	Infiltrating Tubular Carcinoma	Positive	0.44	2.01	0.51	0.95
BR-M-027	Luminal A	51	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Negative	0.31	3.12	0.45	1.85
BR-M-028	Luminal A	79	Mexico	Female	Negative	No	-	II	Positive	I	Mixed carcinoma	Positive	0.58	2.18	0.2	0.85
BR-M-030	Luminal A	89	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Negative	0.5	2.04	0.24	0.8
BR-M-034	Luminal A	66	Mexico	Female	Negative	No	-	II	Positive	I	Infiltrating Ductal Carcinoma	Positive	-	-	0.34	0.62
BR-M-036	Luminal A	62	Mexico	Female	Negative	No	-	II	Positive	I	Infiltrating Ductal Carcinoma	Positive	0.26	2.06	0.1	0.41
BR-M-037	Luminal A	58	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	-	-	1.85	5.25
BR-M-038	Luminal B	41	Mexico	Female	Negative	No	-	II	Negative	II	Infiltrating Ductal Carcinoma	Positive	0.42	1.9	0.21	0.51
BR-M-041	Luminal A	50	Mexico	Female	Negative	No	-	I	Negative	II	Infiltrating Ductal Carcinoma	Positive	0.34	2.03	0.66	2.19
BR-M-045	Basal	37	Mexico	Female	Negative	No	-	II	Negative	III	Infiltrating Ductal Carcinoma	Negative	0.56	2.67	1.27	3.57
BR-M-047	HER2	78	Mexico	Female	Positive	Yes	-	II	Negative	II	Infiltrating Ductal Carcinoma	Negative	0.71	1.91	0.48	2.27
BR-M-048	Normal	37	Mexico	Female	Positive	Yes	-	II	Positive	II	Infiltrating Ductal Carcinoma	Negative	0.14	2.03	0.068	0.41
BR-M-050	Luminal A	47	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.63	2.13	0.24	0.85
BR-M-055	Luminal B	52	Mexico	Female	Positive	Yes	-	I	Positive	III	Infiltrating Ductal Carcinoma	Negative	0.62	3.44	0.9	2.35
BR-M-059	Luminal A	54	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.68	1.98	0.34	0.58
BR-M-073	Luminal B	65	Mexico	Female	Negative	Yes	-	III	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.82	3.44	0.24	1.52
BR-M-074	Luminal A	47	Mexico	Female	Negative	No	-	II	Negative	-	Carcinoma, Mucinous	Positive	0.25	3.88	0.34	1.06
BR-M-076	HER2	44	Mexico	Female	Positive	Yes	-	II	Negative	III	Infiltrating Ductal Carcinoma	Negative	0.35	1.79	0.41	0.99
BR-M-079	Luminal B	56	Mexico	Female	Negative	No	-	III	Positive	II	Infiltrating Ductal Carcinoma	Negative	0.29	2.12	0.41	1.26
BR-M-080	Normal	38	Mexico	Female	Negative	No	-	II	Negative	III	Infiltrating Ductal Carcinoma	Negative	-	-	0.17	0.52
BR-M-083	Luminal A	59	Mexico	Female	Negative	No	-	II	Positive	-	Infiltrating Lobular Carcinoma	Positive	0.3	2	0.2	0.65
BR-M-085	HER2	47	Mexico	Female	Positive	Yes	-	II	Negative	II	Infiltrating Ductal Carcinoma	Positive	0.2	1.75	0.034	0.75
BR-M-094	Luminal B	47	Mexico	Female	Negative	No	-	III	Negative	-	Mixed carcinoma	Negative	0.35	1.89	0.14	0.65
BR-M-095	HER2	52	Mexico	Female	Negative	No	-	I	Positive	-	Infiltrating Lobular Carcinoma	Negative	0.54	3.08	0.62	1.89
BR-M-098	Luminal A	65	Mexico	Female	Negative	Yes	-	I	Positive	II	Infiltrating Ductal Carcinoma	Negative	0.54	2.05	0.2	0.94
BR-M-105	Luminal B	39	Mexico	Female	Negative	Yes	-	III	Positive	III	Infiltrating Ductal Carcinoma	Positive	0.37	3.78	0.34	0.78
BR-M-106	Luminal A	42	Mexico	Female	Negative	No	-	I	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.89	1.94	0.2	0.84
BR-M-110	Luminal A	39	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Negative	0.39	3.59	0.62	1.62
BR-M-116	Luminal B	92	Mexico	Female	Negative	No	-	III	Positive	III	Infiltrating Ductal Carcinoma	Negative	0.66	4.17	1.47	3.45
BR-M-120	Basal	69	Mexico	Female	Negative	No	-	I	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.61	3.75	0.27	0.96
BR-M-121	-	48	Mexico	Female	Negative	Yes	-	II	Positive	III	Infiltrating Ductal Carcinoma	Positive	0.27	1.94	0.31	0.38
BR-M-122	Basal	51	Mexico	Female	Negative	No	-	II	Negative	III	Infiltrating Ductal Carcinoma	Negative	-	-	0.39	1.22
BR-M-123	Luminal B	71	Mexico	Female	Negative	Yes	-	II	Positive	III	Infiltrating Ductal Carcinoma	Positive	0.65	3.53	0.57	1.42
BR-M-126	Luminal B	63	Mexico	Female	Negative	No	-	II	Positive	-	Infiltrating Lobular Carcinoma	Positive	0.64	1.96	0.14	1
BR-M-129	Luminal A	53	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.4	1.01	0.14	0.54
BR-M-150	Luminal B	49	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	-	-	0.58	0.92
BR-M-154	Luminal A	52	Mexico	Female	Negative	No	-	I	Positive	III	Infiltrating Ductal Carcinoma	Positive	0.7	1.92	0.3	0.78
BR-M-155	-	61	Mexico	Female	Negative	No	-	II	Positive	III	Infiltrating Ductal Carcinoma	Positive	0.67	2.13	0.24	1.28
BR-M-158	Luminal B	56	Mexico	Female	Negative	-	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	-	-	0.3	0.54
BR-M-165	Luminal A	42	Mexico	Female	Negative	No	-	III	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.68	2.06	0.17	0.6
BR-M-166	HER2	53	Mexico	Female	Negative	No	-	III	Positive	-	Mixed carcinoma	Positive	0.4	3.04	0.24	1.41
BR-M-167	Luminal A	48	Mexico	Female	Negative	-	-	II	Positive	-	Infiltrating Lobular Carcinoma	Positive	-	-	0.2	0.61
BR-M-169	Luminal A	52	Mexico	Female	Negative	No	-	I	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.25	2	0.068	0.61
BR-M-172	Luminal A	47	Mexico	Female	Negative	Yes	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.36	3.56	0.2	0.71
BR-M-174	Basal	57	Mexico	Female	Negative	No	-	II	Negative	III	Infiltrating Ductal Carcinoma	Negative	-	-	0.72	1.74
BR-M-184	Luminal A	44	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	-	-	0.24	0.38
BR-M-186	Luminal B	55	Mexico	Female	Negative	Yes	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.54	2.1	0.069	0.55
BR-M-189	Basal	65	Mexico	Female	Negative	No	-	III	Negative	-	Medullary Breast Carcinoma	Negative	0.29	2.93	0.58	1.73
BR-M-191	HER2	55	Mexico	Female	Positive	Yes	-	II	Negative	II	Infiltrating Ductal Carcinoma	Negative	0.7	3.69	2.43	8.05
BR-M-192	Luminal A	59	Mexico	Female	Negative	No	-	III	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.5	3.79	0.2	0.91
BR-M-193	HER2	43	Mexico	Female	Positive	Yes	-	III	Negative	III	Infiltrating Ductal Carcinoma	Negative	0.53	1.86	0.68	1.92
BR-M-198	Normal	44	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.64	2.12	0.24	0.65
BR-M-200	Luminal A	42	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.55	2.04	0.1	0.51

¶ Based on PAM-50 classification of exon array data
* Based on immunohistochemistry
§ Based on SNP array
# Calculated by ABSOLUTE
$ Calculated by MutSig
HER2 = Human epidermal growth factor receptor 2; ER = estrogen receptor; PR = progesterone receptor
- Not Available

Supplementary Table 2 - Cases for whole-exome sequencing.

Sequencing ID	Subtype¶	WGS Only	Age	Country	Gender	HER 2 Positive*	Her2 Amplified§	Menopausal Status	Stage	ER Positive*	Grade	Histology	PR Positive*	Purity#	Ploidy#	High Confidence Rearrangements$	Total Rearrangements$
BR-V-002	HER2		46	Vietnam	Female	-	Yes	Premenopausal	III	Negative	III	Infiltrating Ductal Carcinoma	Negative	0.27	1.95	275	447
BR-V-003	-		60	Vietnam	Female	-	No	Post Menopausal	II	-	III	Infiltrating Ductal Carcinoma	-	-	-	63	86
BR-V-004	HER2	Yes	54	Vietnam	Female	-	-	Post Menopausal	III	-	III	Infiltrating Ductal Carcinoma	-	0.333	2.943	523	1776
BR-V-005	Luminal A	Yes	52	Vietnam	Female	-	-	Perimenopausal	II	-	III	Infiltrating Ductal Carcinoma	-	-	-	2	14
BR-V-006	HER2	Yes	51	Vietnam	Female	-	-	Premenopausal	II	-	III	Infiltrating Ductal Carcinoma	-	0.396	2.083	145	204
BR-V-007	Normal		38	Vietnam	Female	-	No	Premenopausal	III	-	III	Infiltrating Ductal Carcinoma	-	0.23	2.16	4	14
BR-V-008	HER2		52	Vietnam	Female	-	Yes	Perimenopausal	III	-	III	Infiltrating Ductal Carcinoma	-	0.27	2.01	166	306
BR-V-009	Luminal B		53	Vietnam	Female	-	No	Post Menopausal	III	Positive	III	Infiltrating Ductal Carcinoma	Positive	0.52	1.99	74	148
BR-V-070	Basal		35	Vietnam	Female	-	-	Premenopausal	II	Negative	III	Infiltrating Ductal Carcinoma	Negative	0.3	1.99	199	348
BR-M-015	HER2	Yes	45	Mexico	Female	Positive	-	-	III	Negative	II	Infiltrating Ductal Carcinoma	Positive	0.22	2.04	121	757
BR-M-028	Luminal A		79	Mexico	Female	Positive	No	-	II	Positive	I	Mixed carcinoma	Positive	0.3	3.25	13	25
BR-M-045	Basal		37	Mexico	Female	Negative	No	-	II	Negative	III	Infiltrating Ductal Carcinoma	Negative	0.25	2.1	473	649
BR-M-050	Luminal A		47	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.47	2.05	1	725
BR-M-082	Luminal B	Yes	59	Mexico	Female	Negative	-	-	II	Positive	II	Infiltrating Ductal Carcinoma	Negative	0.74	2.071	121	250
BR-M-098	Luminal A		65	Mexico	Female	Negative	Yes	-	I	Positive	II	Infiltrating Ductal Carcinoma	Negative	0.54	2.05	27	486
BR-M-106	Luminal A		42	Mexico	Female	Negative	No	-	I	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.89	1.94	8	49
BR-M-116	Luminal B		92	Mexico	Female	Negative	No	-	III	Positive	III	Infiltrating Ductal Carcinoma	Negative	0.66	4.17	59	79
BR-M-123	Luminal B		71	Mexico	Female	Negative	Yes	-	II	Positive	III	Infiltrating Ductal Carcinoma	Positive	0.65	3.53	139	182
BR-M-154	Luminal A		52	Mexico	Female	Negative	No	-	I	Positive	III	Infiltrating Ductal Carcinoma	Positive	0.7	1.92	16	29
BR-M-165	Luminal A		42	Mexico	Female	Negative	No	-	III	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.68	2.06	0	9
BR-M-198	Normal		44	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.64	2.12	44	41
BR-M-200	Luminal A		42	Mexico	Female	Negative	No	-	II	Positive	II	Infiltrating Ductal Carcinoma	Positive	0.55	2.04	3	9

¶ Based on PAM-50 classification of exon array data
* Based on immunohistochemistry
§ Based on SNP array
# Calculated by ABSOLUTE
$ Calculated by dRanger
WGS = Whole genome sequencing
HER2 = Human epidermal growth factor receptor 2; ER = estrogen receptor; PR = progesterone receptor
- Not available

Supplementary Table 3 - Cases for whole-genome sequencing.

N=103 **Mutation Rate**	**Per Megabase**
Total	1.66
Non-silent	1.27

Type of Somatic Mutation	**Number**
De novo Start In Frame	1
De novo Start Out of Frame	20
Frame Shift Deletion	140
Frame Shift Insertion	96
In Frame Deletion	54
In Frame Insertion	14
Missense	3153
Nonsense	242
Read-through	11
Splice Site Deletion	9
Splice Site Insertion	4
Splice Site SNP	84
Synonymous	1157
Total Mutations	4985

SNP = single nucleotide polymorphism

Supplementary Table 5 - Breakdown of mutations by type - exomes.

N=103

rank	gene	description	n	cos	n_cos	N_cos	cos_ev	p	q
1	PIK3CA	phosphoinositide-3-kinase, catalytic, alpha polypeptide	30	193	30	19,879	19476	5.75E-13	2.08E-09
2	TP53	tumor protein p53	29	312	27	32,136	5340	9.16E-13	2.08E-09
3	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	3	46	2	4,738	4	1.8E-05	0.027

n = number of mutations in this gene in the individual set

cos = number of unique mutated sites in this gene in COSMIC

n_cos = overlap between **n** and **cos**

N_cos = number of individuals × **cos**

cos_ev = total evidence: number of reports in COSMIC for mutations seen in this gene

p = p-value for seeing the observed amount of overlap in this gene

q = q-value, False Discovery Rate (Benjamini-Hochberg procedure)

Supplementary Table 7 - Signficantly mutated genes restricted to COSMIC territory only as determined by MutSig algorithm

Supplementary Discussion

We compared the relative utility of mutation detection using whole-genome and whole-exome sequencing approaches. There was high concordance between whole-genome and whole-exome sequencing for mutation detection at higher allelic fraction (Supplementary Figure 5A). Whole-exome sequencing is more sensitive at detecting mutations at lower allelic fraction (Supplementary Figure 5B). Mutations at a very low allelic fraction were detected only by whole-genome sequencing and likely represent mutation calling artefacts in regions of minimal sequence coverage.

As described in Carter et al.[13] the power to detect a variant depends on the allelic fraction and local depth of coverage. For each exon of the significantly mutated genes in each sample, we calculated the allelic fraction assuming a single mutated copy taking into account the local copy number of the exon and the purity of the sample. The average local depth of coverage was computed directly for each sample-exon. Using this allelic fraction and average local depth, we calculated the power to have observed a clonal mutation in a single copy (Supplementary Figure 6). Power was not uniform across samples and genomic regions. Some genomics regions have suboptimal coverage often due to failed hybrid-capture, GC-bias in sequencing, or lack of unique alignment to the genome. These regions are usually located at the 5'- and 3'- ends of genes. In our 6 significantly mutated genes, the power to detect mutations was not affected by the tumor purity in regions with adequate sequencing coverage (Supplementary Figure 6). In regions with intermediate coverage, power is reduced in samples with lower purity. Therefore our observed frequency of mutations represents a lower-bound of the true mutation frequency.

References

1 Silva-Zolezzi, I. *et al.* Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proc Natl Acad Sci USA* **106**, 8611-8616, doi:10.1073/pnas.0903045106 (2009).

2 Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-220, doi:10.1038/nature09744 (2011).

3 Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467, doi:doi:10.1038/nature09837 (2011).

4 Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-1160, doi:10.1126/science.1208130 (2011).

5 Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264, doi:10.1093/biostatistics/4.2.249 (2003).

6 Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).

7 Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160-1167, doi:10.1200/JCO.2008.18.1370 (2009).

8 Network, C. G. A. R. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068, doi:10.1038/nature07385 (2008).

9 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

10 Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899-905, doi:10.1038/nature08822 (2010).

11 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology* **12**, R41, doi:10.1186/gb-2011-12-4-r41 (2011).

12 Carter, S. L., Meyerson, M. & Getz, G. Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping. *Nature Precedings* (2011). <http://hdl.handle.net/10101/npre.2011.6494.1%3E.

13 Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology*, doi:10.1038/nbt.2203 (2012).

14 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664, doi:10.1101/gr.094052.109 (2009).

15 Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**, 459-463, doi:10.1038/nrg2813 (2010).

16 Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-189, doi:10.1038/nbt.1523 (2009).

17 Reich, M. *et al.* GenePattern 2.0. *Nat Genet* **38**, 500-501, doi:10.1038/ng0506-500 (2006).

18 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

19 Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research* **8**, 175-185 (1998).

20 A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

21 Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

22 Osborne, R. H. *et al.* kConFab: a research resource of Australasian breast cancer families. Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer. *Med J Aust* **172**, 463-464 (2000).