Tumour Banking Protocol

The tumour specimens analysed in this study were collected via a BCCA prospective tumour banking protocols for breast tumours (BCCA-BREAST-TTR). Patients are approached for consent to donate tissue surplus to diagnostic requirements plus a blood sample, for use in a research ethics board (REB) approved research protocol. All patients are informed at the time of consent about the potential risks of loss of confidentiality arising from research and that none of the research study data will ever form part of the clinical record or be reported back to the care physicians. Specimen materials are released to investigators only with an REB approved study certificate, via a privacy guardian who anonymises samples such that research investigators have no access to patient identifiers or information that could be used to directly identify a subject. Our REB ethics protocol for genome-scale datasets stipulates that primary datasets will not be released into the public domain, but can be made available via a tiered access mechanism to named investigators of institutions agreeing by a materials transfer agreement to honour the same ethical and privacy principles as the BCCA investigators.

Tumour characteristics and clinical information.

We obtained donated lobular breast cancer pleural effusion and primary tumour tissues via the BC Cancer Agency prospective breast tumour banking mechanism (Certificate H06-00289). The tissues obtained for transcriptome and genome sequencing were from a cryopreserved pleural effusion of metastatic lobular breast carcinoma, together with peripheral blood buffy coat DNA, collected under a prospective consent tumour tissue banking protocol and a study ethical approval for genome scale analysis of nucleic acids (Certificate H08-01230). The morphology and immunohistochemical characteristics of the primary tumour were reviewed by two

pathologists (GT, DH), confirming the appearances of an estrogen receptor positive (*ER*+), E-Cadherin negative (*CDH* –ve) intermediate grade lobular breast carcinoma. Unfractionated cells from the frozen pleural effusion sample were used for the isolation of total DNA and total RNA. The fraction of tumour nuclei to total nuclei was estimated as 89%, by counting 10 randomly selected high power fields in a cell block prepared from this specimen (Figure S1b) and the tumour immunophenotype was confirmed as *ER*+/*PR*+/*EGFR*+ (Figure S1c,d,g), *HER2*-ve (Figure 1h, FISH non amplified) and *CDH1*/*CK5/6* –ve (Figure S1e,f) with a low mitotic index (Ki67, Figure S1j).

Surgery (mastectomy) and loco-regional radiotherapy (4000 cGy to chest wall, 3600 cGy to axilla), were used to treat the primary tumour and in the intervening 9 years, the patient was treated with anti-estrogen therapies in the form of tamoxifen, followed by exemestane (an aromatase inhibitor) but received no neo-adjuvant or adjuvant chemotherapy. The patient relapsed with the pleural effusion 9 years after the initial diagnosis. There was no evidence of local/chest wall recurrence at the time of the pleural effusion. We reviewed the primary tumour histological phenotype (Figure S1a) and immunophenotype, both of which were similar to that of the metastatic tumour. Primary tumour cellularity was estimated at 73%, by counting tumour and normal nuclei in 10 randomly selected high power fields. For sequencing library preparation, we extracted DNA from formalin-fixed paraffin embedded blocks containing the primary lobular tumour, using a modified DNA extraction protocol as described below. For sequence identifications of germline/somatic variants, DNA was extracted from a peripheral blood buffy coat fraction from the patient (Supplemental methods), and this DNA used as the reference normal genome.

Saturation of the tumour genome libraries

A definitive method for calculating the depth required to capture the majority of single base mutations in tumours by random sequencing has yet to be established. In a diploid genome, assuming a desired probability of 0.99 to detect all reference and non-reference heterozygous alleles, each with 2 reads (or more) under Poisson sampling, approximately 27-fold haploid coverage might be required. Tumour cellularity, segmental copy number changes and tumour ploidy would alter this estimation variably and unpredictably across the genome. A practical estimation can be made by determining the rate of discovery of new alleles with increasing aligned sequence coverage and sequencing to within some fraction of the predicted asymptote. In the metastatic genome sequence set, this analysis shows that the rate of allele discovery has significantly tailed off by 43 fold aligned coverage (Figure S9). For validated somatic SNVs the first 30Gb of sequence captured 18 somatic variants, whereas the last 30Gb captured only an additional 2 somatic variants and the last 18Gb (~5 fold haploid coverage) of sequence generated no new *validated* somatic mutations. The saturation of RNA-seq is difficult to estimate, because leaky/low level transcription results in an unknown mixture of distributions with respect to the rate of discovery and the slope of the saturation curves reflects this (Figure S9). For example, we observed (Table S1) that although 17,272 genes were sampled at 1.0 sequenced base per position, or greater, almost all genes (22,165 loci) were observed with minimal coverage of 3 or more reads (Table S1) per locus.

Mutation characteristics

This study reports for the first time cancer associated mutations in a subunit of the recently identified augmin complex. The augmin complex was discovered in flies[1,2] and mammalian cells[3] and is required for proper centrosome formation and kinetochore attachment during mitosis. Depletion of any one of the 8 subunits of the human augmin/HAUS complex (HAUS1-8) results in abnormal centrosome formation or spindle attachment[3]. The third member of the HAUS complex HAUS3 was until very recently anonymous, as C4ORF15. From the metastatic lobular genome, the only somatic homozygous coding mutation occurred in *HAUS3(C4ORF15)* and the somatic mutation was one of five present from the outset in the primary tumour. Moreover we discovered two additional breast cancers (192 sequenced), with heterozygous truncating variants in HAUS3. Centrosome dysfunction is thought to be a critical component of genome instability in breast cancer. Based on the nature of the alleles we have found in HAUS3, we suggest the augmin complex members as potential novel players in this respect. We also found a somatic mutation in PALB2, which like HAUS3, was present at dominant frequency in the primary tumour. Germline mutations in PALB2 (Partner and Localiser of BRCA2[4]) have been shown to predispose to heritable breast cancer[5-9] and pancreatic cancer, presumably through BRCA2 related mechanisms. The functions of ABCB11, SLC24A4 are not well characterised but they fall into the class of multi-pass membrane solute transporters. SNX4 is a sorting nexin, containing a phophoinositide binding domain and is thought to be linked with endocytic recycling processes.

SUPPLEMENTAL METHODS

Immunohistochemistry

A cell block from the pleural effusion was prepared in the usual manner, with the cell pellet obtained by spinning ~100 million cells at 1,000 RPM for 5 minutes in 10% neutral buffered formalin solution. After removing formalin and re-suspending cells in 2 drops of 95% ethanol, one drop of the 1:4 albumen solution and 1 ml of 95% ethanol were added, vortexed and centrifuged at 1,500 RPM for 10 minutes. The cell pellet was then incubated in 1 ml of formalin for 3 hours and subsequently embedded in paraffin.

Five-micron paraffin sections were stained with hematoxylin-eosin (H&E). Immunohistochemical studies were performed on Ventana Discovery XT autostainer (Ventana Medical Systems, Tucson, AZ, USA) with appropriate negative and positive controls. The following antibodies were used: estrogen receptor (*ER*, rabbit monoclonal, clone 6F11, prediluted, Ventana), progesterone receptor (*PR*, rabbit monoclonal, clone 1A6, prediluted, Ventana), *HER2* (rabbit monoclonal, clone 4B5, prediluted, Ventana), E-cadherin (mouse monoclonal, clone NCH-38, dilution, 1:250, Dako, Mississauga, ON, Canada), epidermal growth factor receptor (*EGFR*, mouse monoclonal, clone 3C6, prediluted, Ventana), cytokeratin 5/6 (*CK5/6*, mouse monoclonal, clone D5/16B4, prediluted, Ventana), *Ki-67* (mouse monoclonal, clone K-2, prediluted, Ventana), calretinin (rabbit polyclonal, prediluted, Ventana), Insulin-like growth factor-2 (*IGF2*, rabbit polyclonal, dilution 1:100, Abcam, Cambridge, MA, USA), Insulin-like growth factor receptor 1 beta (*IGF-1Rβ*, mouse monoclonal, dilution 1:15, Santa Cruz Biotechnology, Inc. Santa Cruz, CA, USA), Insulin Receptor, β subunit (*IR-β*, rabbit polyclonal, dilution 1:60, Millipore Corp., Billerica, MA, USA), p16 (mouse monoclonal, clone 16P04, dilution 1:20, CellMarque, Rocklin, CA, USA), p53 (mouse monoclonal, clone DO-7,

dilution 1:400, Dako) and *MUC1* (mouse monoclonal, clone 214D4, dilution 1:700, StemCell Technologies, Inc., Vancouver, BC, Canada). Heat-induced antigen retrieval method was performed in Cell Conditioning solution 1 (CC1-Tris based EDTA buffer, pH 8.0, Ventana). Standard avidin-biotin detection kit (Ventana) was used and the colour was developed with DAB for 5 min. Finally, sections were counterstained with hematoxylin and mounted.

Illumina Genome Sequencing Library preparations.

A total of 3µg DNA was used for the metastatic pleural effusion library (WGSS-PE). Genomic DNA for the pleural effusion was prepared by standard phenol chloroform extraction followed by treatment with DNAse free RNAse.

DNA from the FFPE primary block was prepared as follows. Sections were cut (6 microns) from two FFPE blocks from the primary tumour (a T2 lesion) each with ~73% tumour nuclear cellularity. The initial 3 (10µm thick) sections were discarded, and subsequent 40 sections were processed, 4 sections per tube using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Canada) with the following modifications. De-paraffinisation was conducted with xylene at 65°C (3 changes). Sections were treated for 15-minutes in ATL buffer (tissue lysis buffer) at 95°C, prior to proteinase K digestion. Digestion was carried out over three days at 56°C, with daily spike-ins of the same amount of Proteinase K used on the first day. After spin column purification of DNA as per the Qiagen protocol, column eluted nucleic acids were pooled and then microdialysed against low TE (Tris 10mM, pH 8.0, EDTA 0.1mM) using a MF-Millipore membrane filter (cat no. VMWP02500, 0.05 mm, 25 mm, white, plain) for 2 hours. DNA quality was assessed by spectrophotometry (260/280 and 260/230) and gel electrophoresis before library construction.

Genomic DNA was sheared for 10 min using Sonic Dismembrator 550 (cup horn, Fisher Scientific, Canada), and run in 8% PAGE. 190-210bp DNA fraction was excised and eluted from the gel slice overnight at 4 °C in 300 µl of elution buffer (5:1, LoTE buffer (3 mM Tris-HCl, pH 7.5, 0.2 mM EDTA)-7.5 M ammonium acetate), and was purified using a Spin-X Filter Tube (Fisher Scientific Canada), and by ethanol precipitation. The library was prepared by following the pair end library protocol (Illumina Inc., USA). Briefly, the DNA was subject to end-repair, and phosphorylation by T4 DNA polymerase, Klenow DNA Polymerase, and T4 polynucleotide kinase respectively in a single reaction, and then 3' A overhangs generation by Klenow fragment (3' to 5' exo minus), and ligated to Illumina PE adapters, which contain 5' T overhangs. The adapter-ligated products were purified on Qiaquick spin columns (Qiagen), then PCR-amplified with Phusion DNA Polymerase in 10 cycles using Illumina's PE primer set (Illumina). PCR products were purified on Qiaquick MinElute columns (Qiagen) and the DNA quality assessed and quantified using an Agilent DNA 1000 series II assay and Nanodrop 7500 spectrophotometer (Nanodrop, USA) and diluted to 10nM.

RNA-Seq library construction

20µg DNaseI treated RNA was used to purify polyA+RNA fraction using the MACS^{TM} mRNA Isolation Kit (Miltenyi Biotec, Germany). Double-stranded cDNA was synthesized from the purified polyA+RNA using Superscript 2^{TM} Double-Stranded cDNA Synthesis kit (Invitrogen, USA) and random hexamer primers (Invitrogen) at a concentration of 5µM. The cDNA was sonicated and a PET library was prepared by following the pair end library preparation protocol described above. For RNA-edit re-validation in the lobular transcriptome RNA, we used Superscript 3 (Invitrogen).

Mutation revalidation.

Mutations were validated initially by Sanger amplicon sequencing on both strands, in the metastatic tumour and germline DNA. Confirmed somatic mutations were subsequently re-validated in at least 2 additional independent PCR amplicon reactions. For primary tumour, amplicons were sequenced in both directions from at least 2 independent PCR amplicon reactions, from two independent DNA isolations of the primary tumour and in several instances by cloning and sequencing of PCR amplicons. (Figure S7 and supplemental trace data).

Frequency analysis.

In initial experiments we compared the efficiency of longer amplicons ligated and sheared, for library construction, with direct ligation of library adapaters to PCR amplicons and found no substantial difference. Amplicons were therefore designed to span the target mutation positions with short amplicons (~100bp) such that the mutation position could be reached with 75bp Illumina reads. The same amplicon primers and batch were used for all amplifications. To control for amplification from the primary and metastatic samples, 28 randomly selected heterozygous germline amplicons were also amplified and sequenced together with the mutation positions. PCR amplicons were purified followed by library construction from the adapter ligation step. All of the control germline and somatic amplicons were pooled together for each of the DNA sources. Paired end sequences were collected from an Illumina GA2 with a standard protocol for 75bp reads. The primary and metastasis were sequenced in two separate lanes of the same flowcell

INFORMATIC SUPPLEMENTAL METHODS

Illumina read alignments

Short read sequences obtained from the Illumina Genome Analyzer for WGSS-PE and WGSS-PRI were mapped to the reference human genome (NBCI build 36.1, hg18) using Maq in paired end mode. The WTSS-PE reads were mapped to the human reference genome plus a database of known exon junctions (Morin et al Biotechniques (2008)). The exon junction sequences consisted of N-1 nucleotides of the donor exon 'spliced' to N-1 nucleotides of the acceptor exon for every known pair of exons in the human transcriptome, where N is the read length. This was designed to rescue reads that would not map to the human reference genome due to their spanning a splice site. The maximum insert size (-a parameter) to MAQ was 500 for all alignments. In addition we used all reads (including "Chastity filtered" reads) as input to MAQ.

Single nucleotide variants

Single nucleotide variants were predicted using a novel Bayesian mixture model, *SNVmix* (Supplemental appendix 1), whereby aligned reads were processed by counting the number of reference and non-reference bases at every coding nucleotide in the genome that had at least one read mapped to it. Briefly, SNVmix is a probabilistic generative model that assumes the count data were generated by one of three genotype states: *aa* (homozygous for the reference base), *ab* (heterozygous), and *bb* (homozygous for the non-reference base). We used *SNVmix* to determine the probability of each of these states given the model parameters and the counts for each position in the WGSS-PE, WTSS-PE and WGSS-PRI libraries. Only bases with > Q20 base quality were considered in determining the counts in order to minimize errors. Those positions whose probabilities of seeing p(*S*)=p(*ab)* + p(*bb)* exceeded the FPR=0.01 threshold (determined

using high confidence genotype calls from an Affymetrix SNP 6.0 arrays) were considered predicted SNVs. These SNVs were cross-referenced against dbSNP version 129, the genomes of James Watson[13] and Craig Venter[14], the genome of an anonymous Yoruban[15] and Asian individual[16] , and six individuals from the 1000 genomes project (http://www.1000genomes.org) in order to eliminate any previously described germline variants. Any remaining SNVs were carried forward to subsequent analyses.

RNA-seq variant detection, gene and exon quantitation

The same SNP calling approach was used for aligned RNA-SEQ reads with the following additions. Firstly, the reads aligning to exon-exon junction sequences were computationally repositioned to their two corresponding exons. This was achieved by first checking that the aligned junction read was anchored by at least 6 bases in each exon. The bases for each exon were merged to the pileup file to provide enhanced coverage and SNP calling at exon ends. The total coverage of a gene is the sum of the exonic bases of all exonic and junction reads. The corrected coverage of an individual gene is its total coverage corrected for the total number of non-redundant exonic bases. Total and corrected coverage is also computed for each individual exon of a gene, where an exon is defined as a contiguous nonredundant set of exonic bases.

Copy number alterations

We investigated using the read coverage in the WGSS-PE and WGSS-PRI data as a direct measure of copy number in the respective genomes. In order to adjust for variable regions of mappability, we computed mappability windows using a similar approach described Campbell et al[10]. We randomly sampled 200 x 10^6 pairs of reads from the reference human genome

(NCBI36, hg18, ensembl49) using error models trained on the WGSS-PE reads. We mapped the sampled reads back to the genome and filtered out all reads with low mapping quality (< Q=10). We set the boundaries of the mappability windows by delineating non-overlapping segments containing a fixed number of reads.

Formally, let W_i define a unique window i positioned from s_i to t_i on the chromosome. Each W_i contained exactly N=500 simulated reads that aligned with mapping quality > Q=10. The windows cover the genome non-redundantly such that $s_{i+1} - 1 = t_i$. Let C_i be the number of reads from the WGSS-PE (or WGSS-PRI) data that map to each W_i with mapping qual > Q. We then computed an estimate of copy number change as $Y_i = (C_i - \text{median}(C_X))/N$, where M is the total number of windows in the genome and C_X is the counts for windows on chromosome X. The data were median normalized to chr X since we observed through an aneusomy multi-colour probe FISH assay (see Methods in main text) that it had normal copy number of two with little variance over 40 cells (see Figure S5).

We then segmented $Y_{1:M}$ by fitting a robust HMM, described previously in Shah *et al* (2006)[11], but modified here by incorporating a heavy-tailed, Student-t 5-state emission model capable of capturing high-level amplicons. The five states used were loss, neutral, gain, amplification, and high-level amplification. The model was fit to the data using expectation maximization (EM) and the model hyperparameters for the emission model were set using a data driven approach similar to that described in Shah et al (2007)[12]. The most likely sequence of states $Z_{1:M}$, computed using the Viterbi algorithm was then output.

Table M1. Distribution of HMM CNA states

Copy number type	WGSS-PRI number of segments	% genome
loss	90	2.9
neutral	882	27.3
gain	1229	44.5
amplification	835	19.1
high-level amplicon	135	4.2
Total number of segments	3171	

Results

Validated and HMM inferred segmental copy number states are shown in Tables S13 and S14. The computation of mappability windows resulted in *M*= 333236 unique windows with a median length of 8106 bp spanning autosomes 1-22 and chromosome X. In total, 2.997 Gb of the genome was represented. Table M1 shows the summary of how the WGSS-PE reads were segmented by the robust HMM. WGSS-PE was segmented into 1005 segments distributed as 12 loss, 189 neutral, 453 gain, 301 amplifications and 50 high-level amplicons. These comprised 1.4, 22.6, 37.4, 28.1 and 8.6% of the genome respectively.

Unequal Allelic expression

The Binomial exact test was performed using R v2.8.0 with the method call `binom.test(x,p,alternative="two-sided")` where 'x' contained the allelic counts of corresponding WTSS-PE positions to WGSS-PE heterozygous positions (determined by SNVMix concordance with dbSNP v129 or by confirmed novel heterozygous variants), and 'p' was equal to the converged SNVMix Binomial model parameter (μ_{ab}) for the RNA-SEQ

heterozygous genotype (see Supplementary Appendix 1). P-values were adjusted by the Benjamini-Hochberg correction using `p.adjust` in R. All positions with q-values below the 0.01 FDR threshold were deemed as significantly skewed.

Analysis of frequency PCR amplicons

Reads obtained from the primary and metastatic libraries were aligned to the human genome reference (hg18, NCBI 36.1) as outlined above. Allelic counts were computed for the targeted positions as well as 5 positions immediately flanking the target. The allelic counts of the flanking positions were used to compute the population mean representing the expectation of the proportion of reference reads in non-variant positions (the null distribution). A one-tailed Binomial exact test was then used to test the target positions against the null distribution. Positions with a p-value <0.01 were counted as being present. This procedure was applied independently for the metastatic library and primary library. The average observed error rate in the flanking positions for the metastatic library was 0.0005/base, while for the primary (which was obtained from paraffin) it was 0.0012/base.

Identification and PCR/sequence validation of candidate genomic inversions

Each read aligned by Maq is (by default) given a flag indicating its orientation/placement with regards to its mate. The normal pairing situation is termed FR (forward reverse), and represents mates pointing 'towards' each other on the same chromosome. Two classes of pairings indicative of inversions are FF and RR and should be accompanied by a significant increase or decrease in pairing distance between mates (see below). Hence, each edge of an inversion (i.e. breakpoint) should be distinguishable to fairly high resolution using these pairings.

Read pairs with FF and RR flags accompanied by a high (>=40) mapping quality are used for this analysis. Of the remaining read pairs, those with a pairing distance greater than some large cutoff (D) are identified. Amongst these reads, clusters aligning in close proximity (within a window of 500bp) are collected. These clusters should comprise a set of reads that lie either within the inversion (and near the breakpoint) or lie just outside of the inversion. For reads within a cluster, the mean 'mate distance' for all reads is computed, which gives an estimate of the size of the inversion. Finally, each inversion should have two clusters (one for each breakpoint). High-confidence inversions are identified by retaining only pairs of clusters whose mean mate-distance allows them to be joined (i.e. the two clusters have mate distances consistent with each other). This approach assumes that both breakpoints have at least 3x physical coverage (clusters must contain at least 3 reads). It is also assumed that the regions around the breakpoints are mappable. Any inversion having one unmappable breakpoint or less than 3x physical coverage will have an orphaned cluster. Orphaned clusters were not carried forward for validation.

The outer boundaries of each read cluster in a pair provide the best estimate for the breakpoint of the inversion. We prepared primers that would amplify normal (un-inverted) genomic DNA at each breakpoint producing a ~500bp amplicon. Two pairs of primers were designed for each breakpoint. PCR reactions were prepared (with either tumour or germline DNA template) with these primer pairs in either normal or swapped arrangement. The swapped primers would produce an amplicon if the inversion was real (as they flank the breakpoint). Amplicons produced from swapped primer pairs were sequenced, assembled (using Phred/Phrap) and contig sequences aligned by BLAT. Amplicons for which distinct non-overlapping portions

aligned to the two estimated breakpoints were deemed successful. All successful inversion validations are included in the supplemental table.

References.

1 G. Goshima, M. Mayer, N. Zhang et al., *The Journal of cell biology* **181** (3), 421 (2008).
2 A. M. Meireles, K. H. Fisher, N. Colombie et al., *The Journal of cell biology* **184** (6), 777 (2009).
3 S. Lawo, M. Bashkurov, M. Mullin et al., *Curr Biol* **19** (10), 816 (2009).
4 B. Xia, Q. Sheng, K. Nakanishi et al., *Molecular cell* **22** (6), 719 (2006).
5 H. Erkko, B. Xia, J. Nikkila et al., *Nature* **446** (7133), 316 (2007).
6 K. J. Patel, *Nature genetics* **39** (2), 142 (2007).
7 N. Rahman, S. Seal, D. Thompson et al., *Nature genetics* **39** (2), 165 (2007).
8 S. Reid, D. Schindler, H. Hanenberg et al., *Nature genetics* **39** (2), 162 (2007).
9 B. Xia, J. C. Dorsman, N. Ameziane et al., *Nature genetics* **39** (2), 159 (2007).
10. Campbell PJ, Stephens PJ, Pleasance ED, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008 Jun;40(6):722-9.
11. Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, Ng R, Murphy KP. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics. 2006 Jul 15;22(14):e431-9.
12. Shah SP, Lam WL, Ng RT, Murphy KP. Modeling recurrent DNA copy number alterations in array CGH data. Bioinformatics. 2007 Jul 1;23(13):i450-8.
13 Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 2008;452:872-6.
14 Levy S, Sutton G, Ng PC, et al. The diploid genome sequence of an individual human. PLoS Biol 2007;5:e254.
15 Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53-9.
16 Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual. Nature 2008;456:60-5.