LETTERS

Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution

Sohrab P. Shah[1,2]*, Ryan D. Morin[3]*, Jaswinder Khattra[1], Leah Prentice[1], Trevor Pugh[3], Angela Burleigh[1], Allen Delaney[3], Karen Gelmon[4], Ryan Guliany[1], Janine Senz[2], Christian Steidl[2,5], Robert A. Holt[3], Steven Jones[3], Mark Sun[1], Gillian Leung[1], Richard Moore[3], Tesa Severson[3], Greg A. Taylor[3], Andrew E. Teschendorff[6], Kane Tse[1], Gulisa Turashvili[1], Richard Varhol[3], René L. Warren[3], Peter Watson[7], Yongjun Zhao[3], Carlos Caldas[6], David Huntsman[2,5], Martin Hirst[3], Marco A. Marra[3] & Samuel Aparicio[1,2,5]

Recent advances in next generation sequencing[1–4] have made it possible to precisely characterize all somatic coding mutations that occur during the development and progression of individual cancers. Here we used these approaches to sequence the genomes (>43-fold coverage) and transcriptomes of an oestrogen-receptor-α-positive metastatic lobular breast cancer at depth. We found 32 somatic non-synonymous coding mutations present in the metastasis, and measured the frequency of these somatic mutations in DNA from the primary tumour of the same patient, which arose 9 years earlier. Five of the 32 mutations (in *ABCB11*, *HAUS3*, *SLC24A4*, *SNX4* and *PALB2*) were prevalent in the DNA of the primary tumour removed at diagnosis 9 years earlier, six (in *KIF1C*, *USP28*, *MYH8*, *MORC1*, *KIAA1468* and *RNASEH2A*) were present at lower frequencies (1–13%), 19 were not detected in the primary tumour, and two were undetermined. The combined analysis of genome and transcriptome data revealed two new RNA-editing events that recode the amino acid sequence of *SRP9* and *COG3*. Taken together, our data show that single nucleotide mutational heterogeneity can be a property of low or intermediate grade primary breast cancers and that significant evolution can occur with disease progression.

Lobular breast cancer is an oestrogen-receptor-positive (ER^+, also known as $ESR1^+$) subtype of breast cancer (approximately 15% of all breast cancers). It is usually of low-intermediate histological grade and can recur many years after initial diagnosis. To interrogate the genomic landscape of this class of tumour, we re-sequenced[1–4] the DNA from a metastatic lobular breast cancer specimen (89% tumour cellularity; Supplementary Fig. 1) at approximately 43.1-fold aligned, haploid reference genome coverage (120.7 gigabases (Gb) aligned paired-end sequence; Supplementary Fig. 2, Table 1 and Supplementary Methods). Deep high-throughput transcriptome sequencing (RNA-seq)[5] performed on the same sample generated 160.9-million reads that could be aligned (Supplementary Table 1, see also Supplementary Fig. 2 and Supplementary Methods). The saturation of the genome (Table 1) and RNA-seq (Supplementary Table 1) libraries for single nucleotide variant (SNV) detection is discussed in Supplementary Information. The aligned (hg18) reads were used to identify (Supplementary Fig. 2) the presence of genomic aberrations, including SNVs (Supplementary Table 2), insertions/deletions (indels), gene fusions, translocations, inversions and copy number alterations (Supplementary Methods). We examined predicted coding indels and predicted inversions (coding or non-coding; Supplementary Methods); however, all of the events that were validated by Sanger re-sequencing were also present in the germ line (Supplementary Tables 3 and 4). None of the 12 predicted gene fusions revalidated. We also computed the segmental copy number (Supplementary Methods and Supplementary Table 5a) from aligned reads, and revalidated high level amplicons by fluorescence *in situ* hybridization (FISH) (Supplementary Table 5b), revealing the presence of a new low-level amplicon in the *INSR* locus (Supplementary Fig. 3).

We identified coding SNVs from aligned reads, using a Binomial mixture model, SNVMix (Supplementary Table 2, Methods and Supplementary Appendix 1). From the RNA-seq (WTSS-PE) and genome (WGSS-PE) libraries we predicted 1,456 new coding non-synonymous SNVMix variants (Supplementary Table 2). After the removal of pseudogene and HLA sequences (1,178 positions remaining) and after primer design, we re-sequenced (Sanger amplicons) 1,120 non-synonymous coding SNV positions in the tumour DNA and normal lymphocyte DNA. Some 437 positions (268 unique to WGSS-PE, 15 unique to WTSS-PE, and 154 in common) were confirmed as non-synonymous coding variants. Of these, 405 were new

Table 1 | Summary of sequence library coverage

	WGSS-PE	WTSS-PE
Total number of reads	2,922,713,774	182,532,650
Total nucleotides (Gb)	140.991	7.108
Number of aligned reads	2,502,465,226	160,919,484
Aligned nucleotides (Gb)	120.718	6.266
Estimated error rate	0.021	0.013
Estimated depth (non-gap regions)	43.114	NA
Canonically aligned reads	2,294,067,534	109,093,616
Exons covered	93.5 at >10 reads; 95.7 at >5 reads	82,200 at 10 reads (see also Supplementary Table 1)
Reads aligned canonically (%)	78.49	67.79
Unaligned reads	420,248,548	21,613,166
Mean read length (bp)	48.24	38.94

The WGSS-PE column shows the genome paired-end read coverage for DNA from the metastatic pleural effusion sample. The WTSS-PE column shows coverage for the complementary DNA reads from the matched transcriptome libraries of the metastatic pleural effusion. Coverage of exon bases in the reference genome (hg18) is shown at 5 or more reads per position, and 10 or more reads per position for the metastatic genome. bp, base pairs; NA, not applicable.

[1]Molecular Oncology, [2]Centre for Translational and Applied Genomics, [3]Michael Smith Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver V5Z 1L3, Canada. [4]Medical Oncology, BC Cancer Agency, 600 West 10th Avenue, Vancouver V5Z 1L3, Canada. [5]Department of Pathology, University of British Columbia, G227-2211 Wesbrook Mall, British Columbia, Vancouver V6T 2B5, Canada. [6]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [7]Deeley Research Centre, BC Cancer Agency, Victoria V8R 6V5, Canada.
*These authors contributed equally to this work.

Table 2 | Somatic coding sequence SNVs validated by Sanger sequencing

Gene	Description	Position	Source	Allele change	Amino acid change	Protein domain affected	Expression (sequenced bases per exonic base)	Allelic expression bias (R, NR allele)	Copy number classification (HMM state)
ABCB11	Bile salt export pump (ATP-binding cassette sub-family B member 11)	2:169497197	WGSS	C>T	R>H	Transmembrane helix 3	0.3	1, 1	Amplification (4)
HAUS3	HAUS3 coiled-coil protein (C4orf15)	4:2203607	WGSS, WTSS	C>T	V>M	Unknown	14.1	4, 23	Neutral (2)
CDC6	Cell division control protein 6 homologue	17:35701114	WGSS, WTSS	G>A	E>K	N-terminal, unknown	2.7	3, 3	Amplification (4)
CHD3	Chromodomain-helicase-DNA-binding protein 3	17:7751231	WGSS	G>A	E>K	Unknown, C-terminal	3.9	41, 11 ($Q < 0.01$)	Neutral (2)
DLG4	Disks large homologue 4	17:7052251	WGSS	G>A	P>L	Unknown, N-terminal	5.5	7, 1	Neutral (2)
ERBB2	Receptor tyrosine-protein kinase erb-b2	17:35133783	WGSS, WTSS	C>G	I>M	Kinase domain	67.1	62, 35	Amplification (4)
FGA	Fibrinogen alpha chain	4:155726802	WGSS	C>T	W>stop	Fibrinogen a/b/c domain	0.01	NA	Gain (3)
*GOLGA4**	Golgin subfamily A member 4	3:37267947	WGSS, WTSS	G>C	E>Q	Unknown, N-terminal	111.8	37, 12	Gain (3)
GSTCD	Glutathione *S*-transferase C-terminal domain-containing protein	4:106982671	WGSS, WTSS	G>C	E>Q	Unknown, C-terminal	23.2	23, 8	Neutral (2)
*KIAA1468**	LisH domain and HEAT repeat-containing protein	18:58076768	WGSS, WTSS	G>C	R>T	ARM type fold	36.1	23, 11	Neutral (2)
KIF1C	Kinesin-like protein KIF1C	17:4848025	WGSS, WTSS	G>C	K>N	Kinesin motor domain	28.5	16, 13	Neutral (2)
KLHL4	Kelch-like protein 4	X:86659878	WGSS	C>T	S>L	Unknown, N-terminal	1.7	1, 0	Neutral (2)
MYH8	Myosin 8 (myosin heavy chain 8)	17:10248420	WGSS	C>G	M>I	Actin-interacting protein domain	0	NA	Neutral (2)
PALB2	Partner and localizer of BRCA2	16:23559936	WGSS	T>G	E>A	N-terminal prefolding	13.0	NA	Amplification (4)
PKDREJ	Polycystic kidney disease and receptor for egg-jelly-related protein	22:45035285	WGSS	C>G	E>Q	Unknown	0.1	NA	Gain (3)
RASEF	RAS and EF-hand domain-containing protein	9:84867250	WTSS	G>A	S>L	EF-hand Ca^{2+}-binding motif	65.0	3, 2	Gain (3)
RNASEH2A	Ribonuclease H2 subunit A (EC 3.1.26.4)	19:12785252	WGSS, WTSS	G>A	R>H	Unknown, C-terminal	5.3	2, 2	Neutral (2)
RNF220	RING finger protein C1orf164	1:44650831	WGSS	G>A	D>N	Unknown, N-terminal	16.1	NA	Neutral (2)
SP1	Transcription factor Sp1	12:52063157	WGSS	G>C	E>Q	Glu-rich N-terminal domain	57.3	40, 10 ($Q < 0.01$)	Amplification (4)
USP28	Ubiquitin carboxyl-terminal hydrolase 28	11:113185109	WGSS, WTSS	C>T	D>N	Unknown	12.5	3, 7	Gain (3)
C11orf10	UPF0197 transmembrane protein C11orf10	11:61313958	WGSS	G>A	T>I	Transmembrane domain	28.9	13, 3	Amplification (4)
THRSP	Thyroid hormone-inducible hepatic protein	11:77452594	WGSS	C>T	R>C	Unknown	0.3	NA	Gain (3)
SCEL	Sciellin	13:77076497	WGSS	A>G	K>R	Unknown	0.3	1, 0	Gain (3)
SLC24A4	$Na^+/K^+/Ca^{2+}$-exchange protein 4	14:92018836	WGSS	G>A	V>I	Transmembrane domain	1.2	1, 0	Amplification (4)
COL1A1	Collagen alpha-1(I) chain precursor	17:45625043	WGSS	C>T	G>D	Pro-rich domain	80.0	24, 0 ($Q < 0.01$)	Amplification (4)
KIAA1772	GREB1-like protein	18:17278222	WGSS	A>G	D>G	Unknown	2.8	4, 1	Neutral (2)
CCDC117	Coiled-coil domain-containing protein 117	22:27506951	WGSS	G>C	K>N	Unknown	12.9	2, 0	Neutral (2)
RP1-32I10.10	Novel protein	22:43140252	WGSS	G>C	E>Q	Unknown	0	NA	Gain (3)
MORC1	MORC family CW-type zinc finger protein 1	3:110271286	WGSS	G>A	A>V	Coiled-coil	0.1	NA	Gain (3)
SNX4	Sorting nexin 4	3:126721688	WGSS	C>T	D>N	Unknown, N-terminal	43.4	NA	Gain (3)
LEPREL1	Prolyl 3-hydroxylase 2 precursor (EC 1.14.11.7)	3:191172415	WGSS	T>C	E>G	Hydroxylase domain	1.1	NA	Gain (3)
*WDR59**	WD repeat-containing protein 59	16:73500342	WTSS	C>T	M>I	Unknown, C-terminal	17.3	6, 5	Neutral (2)

Omnibus table showing the features associated with the 32 Sanger amplicon-validated non-synonymous somatic mutations from the WGSS-PE and WTSS-PE libraries. Mutation positions are on the basis of reference genome hg18. The nucleotide substitutions are shown as reference>variant. The amino acid change is shown as reference>variant amino acid. If the mutation occurs in a recognized protein domain or motif this is shown. The transcript expression level in WTSS-PE reads is shown as the mean number of reads supporting each position in the transcript. The allelic expression bias column shows the number of reference (R), non-reference (NR) reads in the WTSS-PE library at the mutated position. Three transcripts (*CHD3*, *SP1* and *COL1A1*) show significant expression bias (annotated with $Q < 0.01$, Supplementary Methods) in favour of the reference allele; however, none of the heterozygous somatic mutations were biased in favour of the non-reference allele. The expression of *HAUS3* is predominantly non-reference as expected for a homozygous allele. The HMM state classifier of copy number for the genomic region encompassing each mutation position is shown in the last column, as state (state number). C-terminal, carboxy-terminal; N-terminal, amino-terminal.
*Genes showing alternative splicing.

germline alleles and 32 were revealed as non-synonymous coding somatic point mutations (Table 2). Of the 32 somatic mutations, 30 were present in WGSS-PE and/or WTSS-PE, whereas two were detected from the WTSS library sequence alone (Table 2). None of the 32 genes were found in common with the CAN breast genes[6], which were discovered from ER$^-$ cell lines. Eleven genes appear in the current release of COSMIC[7] (*CHD3*, *SP1*, *PALB2*, *ERBB2*, *USP28*, *KLHL4*, *CDC6*, *KIAA1468*, *RNF220*, *COL1A1* and *SNX4*) but with mutations at different positions. We examined the population frequency of the somatic mutation positions for *PALB2*, *ERBB2*, *USP28*, *CDC6*, *CHD3*, *HAUS3* (previously known as *C4orf15*), *SP1*, *KIAA1468* and *DLG4* in a further 192 breast cancers (Supplementary Methods; 112 lobular, 80 ductal). None of these 192 breast cancers showed identical mutations to those described here; however, 3 out of 192 cases (2 lobular, 1 ductal) contained neighbouring non-synonymous variants/deletions affecting the *ERBB2* kinase domain (Supplementary Fig. 4). Interestingly, 2 out of 192 cases (both lobular) contained two different heterozygous truncating variants in *HAUS3*: chr4:2203685 G>T on minus strand, GAG>TAG (Glu>stop), and chr4:2203483 C>G on minus strand, TCA>TGA (Ser>stop) (Supplementary Fig. 5). Notably, *HAUS3* is a member of the recently described[8–10] multiprotein augmin complex, the function of which is required for genome stability mediated by appropriate kinetochore attachment and centrosome morphogenesis.

To determine how many of the somatic non-synonymous coding sequence mutations were already present at diagnosis 9 years earlier, we next examined genomic DNA from the primary tumour directly, by a single molecule frequency counting experiment (Supplementary Methods)[4]. Twenty-eight of the 32 mutations yielded amplicons compatible with Illumina sequencing (Supplementary Methods), and two extra mutations were sampled by Sanger sequencing (Supplementary Fig. 5). As controls we selected 36 heterozygous germline SNVs at random. The PCR amplicons for known germline and somatic mutations were sequenced on an Illumina device. After alignment, the observed counts of reference and non-reference bases at the target position were compared using the Binomial exact test. To calibrate the expected mean of the Binomial distribution, we used the non-reference allele frequency from positions −5 to +5 surrounding (but not including) the target position (Supplementary Table 6a, b), where only reference bases should be called. Unequal segmental amplification/deletion in the genome may contribute to a departure from the theoretical ratio of 0.5 for a heterozygous allele. As a result, amplicons from heterozygous germline alleles showed occasional measured frequencies of between 0.2 and 0.8 in both the primary and metastatic tumour DNA (Table 3 and Supplementary Table 7), but with a modal frequency around 0.5, as expected. In the metastatic genomic DNA the somatic mutations showed frequencies of between 0.2 and 0.79 (Table 3). Notably, the somatic coding mutation positions examined in the primary tumour showed three patterns of abundance: prevalent, rare and undetectable (Table 3). Mutations in *ABCB11*, *PALB2* and *SLC24A4* were detected at prevalent frequencies for heterozygous mutations (≥0.2, the lowest value seen for known germline alleles) given a 73% tumour content. The frequency of the mutation in *HAUS3* was 0.79, consistent with it being a prevalent homozygous mutation, also confirmed by Sanger sequencing (Supplementary Fig. 5). Sanger amplicon sequencing showed that the *SNX4* somatic mutation was also present in the primary tumour, whereas the *KIAA1772* (also known as *GREB1L*) mutation was not. Six mutations (*KIF1C*, *USP28*, *MORC1*, *MYH8*, *KIAA1468* and *RNASEH2A*) showed statistically significant ($P < 0.01$, Binomial exact test) intermediate frequencies of between 1% and 13% (Table 3), suggesting that these mutations were

Table 3 | Frequency of germline and somatic alleles in the metastatic and primary genomes

Position	Locus	R	NR	Primary depth	Primary NR ratio	Primary *P* value	Primary status	Metastasis depth	Metastasis NR ratio	M	Copy number classification (HMM state)
4:2203607	*HAUS3*	C	T	5700	0.5472	0.0000	Dominant	762	0.7874	S	Neutral (2)
16:23559936	*PALB2*	T	G	115	0.4957	0.0000	Dominant	669	0.4350	S	Amplification (4)
2:169497197	*ABCB11*	C	T	506	0.3261	0.0000	Dominant	959	0.3691	S	Amplification (4)
14:92018836	*SLC24A4*	G	A	13347	0.2341	0.0000	Dominant	13670	0.3518	S	Amplification (4)
17:10248420	*MYH8*	C	G	10657	0.1353	0.0000	Subdominant	1797	0.5932	S	Neutral (2)
3:110271286	*MORC1*	G	A	24572	0.0468	0.0000	Subdominant	32273	0.4107	S	Gain (3)
17:4848025	*KIF1C*	G	A	8587	0.0107	0.0000	Subdominant	2272	0.3077	S	Neutral (2)
11:113185109	*USP28*	C	T	6654	0.0095	0.0000	Subdominant	1387	0.4484	S	Gain (3)
18:58076768	*KIAA1468*	G	A	719	0.0083	0.0020	Subdominant	1056	0.3059	S	Neutral (2)
19:12785252	*RNASEH2A*	G	A	6537	0.0029	0.0276	Subdominant	1497	0.2806	S	Neutral (2)
4:106982671	*GSTCD*	G	T	7273	0.0008	0.9885	Absent	2208	0.2174	S	Neutral (2)
17:35701114	*CDC6*	G	T	4894	0.0008	0.9733	Absent	4208	0.3577	S	Amplification (4)
17:7751231	*CHD3*	G	A	9665	0.0007	0.9981	Absent	1737	0.2671	S	Neutral (2)
4:155726802	*FGA*	C	T	5756	0.0007	0.9911	Absent	2287	0.2755	S	Gain (3)
17:7052251	*DLG4*	G	A	4383	0.0007	0.9835	Absent	706	0.3272	S	Neutral (2)
3:37267947	*GOLGA4*	G	T	13051	0.0006	0.9999	Absent	3262	0.2235	S	Gain (3)
9:84867250	*RASEF*	G	T	1690	0.0006	0.9500	Absent	796	0.3656	S	Gain (3)
17:35133783	*ERBB2*	C	A	3736	0.0005	0.9899	Absent	1722	0.3612	S	Amplification (4)
X:86659878	*KLHL4*	C	T	6561	0.0005	0.9993	Absent	977	0.3153	S	Neutral (2)
3:191172415	*LPREL1*	T	C	11963	0.0004	1.0000	Absent	8381	0.2148	S	Gain (3)
16:73500342	*WDR59*	C	T	4846	0.0004	0.9982	Absent	1396	0.2629	S	Neutral (2)
1:44650831	*RNF220*	G	A	8160	0.0004	0.9999	Absent	967	0.2203	S	Neutral (2)
22:45035285	*PKDREJ*	C	T	6674	0.0003	0.9999	Absent	1230	0.3366	S	Gain (3)
11:61313958	*C11ORF10*	G	A	116705	0.0003	1.0000	Absent	14354	0.4651	S	Amplification (4)
12:52063157	*SP1*	G	T	7732	0.0003	1.0000	Absent	2011	0.2193	S	Amplification (4)
11:77452594	*THRSP*	C	T	24219	0.0002	1.0000	Absent	40652	0.4750	S	Gain (3)
17:45625043	*COL1A1*	C	A	26343	0.0001	1.0000	Absent	32259	0.2543	S	Amplification (4)
13:77076497	*SCEL*	A	G	49	0.0000	1.0000	Absent	187	0.5722	S	Gain (3)
19:9314428	–	A	G	176	0.5057	0.0000	Present	321	0.4953	G	Neutral (2)
4:130144460	–	A	T	2020	0.2188	0.0000	Present	2081	0.3099	G	Neutral (2)
8:27835012	–	G	A	13587	0.8602	0.0000	Present	10781	0.6667	G	Deletion (1)
6:32908543	–	C	T	4718	0.7484	0.0000	Present	16370	0.4897	G	Amplification (4)
20:43363061	–	G	A	5950	0.5249	0.0000	Present	5540	0.5049	G	Amplification (4)
4:8672089	–	G	A	381	1.0000	0.0000	Present	2850	0.8032	G	Gain (3)
16:1331138	–	C	T	677	0.4963	0.0000	Present	554	0.6245	G	High-level amplicon (5)

Only 7 germline alleles are shown, the full list is in Supplementary Table 7. The genome positions are shown as chr:coordinate. The primary read depth represents the number of reads. Binomial exact *P* values were calculated using a Binomial exact test. R, reference base; NR, non-reference base. Primary status indicates whether the variant was present, subdominant or absent in the primary tumour. Column M denotes somatic (S) or germline (G) single nucleotide variants in the metastasis. HMM state refers to the metastasis.

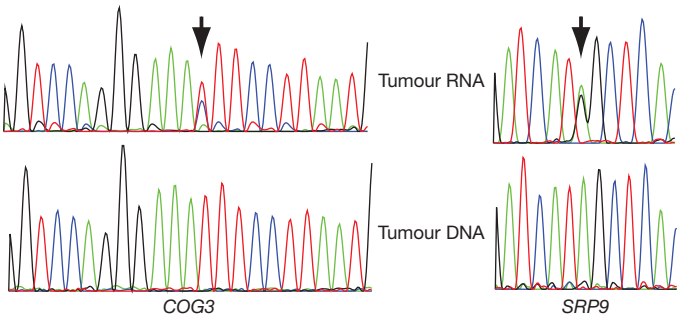

Figure 1 | RNA editing in *COG3* and *SRP9*. Sanger sequence traces from the non-synonymous editing positions in *COG3* and *SRP9*. The editing position is arrowed. Top trace is tumour RNA, bottom trace tumour DNA. The editing positions were confirmed with reverse strand reads (not shown).

restricted to minor subclones of tumour cells. The remaining 19 out of 30 of the somatic coding mutations were not detected in the primary tumour DNA. Thus, significant heterogeneity in tumour somatic mutation content existed in the primary tumour at diagnosis. In contrast with the recently reported sequence of cytogenetically normal acute myeloid leukaemia (AML) tumour[4], significant evolution of coding mutational content occurred between primary and metastasis. It is unknown whether the 19 mutations present in the metastasis, but not detected in the primary, were a consequence of radiation therapy or innate tumour progression.

We also examined how the transfer of information from the nuclear genome to proteins was modified by alternative splicing (Supplementary Table 8 and Supplementary Fig. 6), biased allelic expression (Supplementary Table 9) and RNA editing. At the single nucleotide level, RNA-editing enzymes (which can be regulated by oestrogens[11]) may also recode transcripts resulting in a proteome divergent from the genome[12–15]. Interestingly, the ADAR enzyme—one of the principal RNA-editing enzymes that mediates A→I(G) edits—was one of the top 5% of genes expressed (145.6 reads per base, Supplementary Table 10), and the only editing enzyme expressed at a high level. We searched for potential editing events (Methods) and found 3,122 candidate edits in 1,637 gene loci (Supplementary Table 11). Some 526 out of 3,122 candidate edits are non-synonymous changes and 232 are synonymous changes (with the remainder affecting untranslated regions). We revalidated independently (Supplementary Methods) by Sanger sequencing 75 editing events in 12 gene loci from the lobular metastasis (Supplementary Table 12 and see trace data at http://molonc.bccrc.ca/). Two genes, *COG3* and *SRP9* (Fig. 1), showed confirmed high frequency non-synonymous transcript editing, resulting in variant protein sequences. These observations emphasize the importance of integrating RNA-seq data with tumour genomes in assessing protein variation.

The coding mutation landscape of breast cancers has, so far, been mostly determined from ER^- metastatic cell lines/samples[6,16], and has suggested the presence of large numbers of passenger events as well as drivers. Our results show the importance of sequencing samples of tumour cell populations early as well as late in the evolution of tumours, and of estimating allele frequency in tumour genomes. Our observations suggest that the sequencing of primary breast cancers and pre-invasive malignancy may reveal significantly fewer candidates for tumour initiating mutations.

METHODS SUMMARY

Paired-end reads were assigned quality scores and aligned to the reference genome (hg18) using Maq[17] (Supplementary Methods and Supplementary Fig. 2). For identification of SNVs we used a simple Binomial mixture model, SNVMix (Supplementary Appendix 1), which assigns a probability to each base position as homozygous reference (aa), heterozygous non-reference (ab) and homozygous non-reference (bb), based on the occurrence of reference (hg18) and non-reference bases at each aligned position. This model was calibrated initially, using high confidence allele calls from Affymetrix SNP6.0 hybridization of tumour and normal DNA. We estimated the receiver operating characteristic (ROC) performance (Supplementary Fig. 8) and determined that an SNVMix threshold of $P = 0.77$ for (ab) or (bb) for a non-reference call would yield a false discovery rate (FDR) of 1%. For the RNA-seq library, a threshold of $P = 0.53$ was used (Supplementary Fig. 8; FDR = 0.01) to call non-reference positions. Non-reference positions were then filtered for known variants against the sources of germline variation, the single nucleotide polymorphism database (dbSNP) and the completed individual genomes[18,19] (Supplementary Table 2). Saturation of the libraries for SNV discovery was determined by random re-sampling (Supplementary Fig. 9 and Supplementary Methods). Segmental copy number was inferred with a hidden Markov model (HMM) method (Supplementary Table 4a, b and Supplementary Methods).

We searched for RNA-editing events by examining all very high confidence ($P(ab) + P(bb) > 0.9$) SNVMix predictions from the RNA-seq library of the metastatic tumour, that were not found with extreme confidence ($P(aa) > 0.99$, derived from the SNVMix receiver operating curve at FDR = 0.01) at the same positions in the metastatic tumour genome library.

Received 4 September; accepted 10 September 2009.

1. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genet.* **40,** 722–729 (2008).
2. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
3. Morin, R. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45,** 81–94 (2008).
4. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
5. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcripomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
6. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113 (2007).
7. Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protoc. Hum. Genet.* **Unit 10.11,** doi:10.1002/0471142905.hg1011s57 (2008).
8. Goshima, G., Mayer, M., Zhang, N., Stuurman, N. & Vale, R. D. Augmin: a protein complex required for centrosome-independent microtubule generation within the spindle. *J. Cell Biol.* **181,** 421–429 (2008).
9. Meireles, A. M., Fisher, K. H., Colombie, N., Wakefield, J. G. & Ohkura, H. Wac: a new Augmin subunit required for chromosome alignment but not for acentrosomal microtubule assembly in female meiosis. *J. Cell Biol.* **184,** 777–784 (2009).
10. Lawo, S. *et al.* HAUS, the 8-subunit human Augmin complex, regulates centrosome and spindle integrity. *Curr. Biol.* **19,** 816–826 (2009).
11. Pauklin, S., Sernandez, I. V., Bachmann, G., Ramiro, A. R. & Petersen-Mahrt, S. K. Estrogen directly activates AID transcription and function. *J. Exp. Med.* **206,** 99–111 (2009).
12. Blow, M., Futreal, P. A., Wooster, R. & Stratton, M. R. A survey of RNA editing in human brain. *Genome Res.* **14,** 2379–2387 (2004).
13. Athanasiadis, A., Rich, A. & Maas, S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.* **2,** e391 (2004).
14. Maas, S., Kawahara, Y., Tamburro, K. M. & Nishikura, K. A-to-I RNA editing and human disease. *RNA Biol.* **3,** 1–9 (2006).
15. Li, J. B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324,** 1210–1213 (2009).
16. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
17. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–1858 (2008).
18. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60–65 (2008).
19. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Eaves and M. Pollak for comments on earlier versions of the manuscript. We thank and acknowledge the patients of the BC Cancer Agency for donations of tumour tissues to the TTR-BREAST tumour banking program. S.A. is supported by a Canada Research Chair in Molecular Oncology, S.P.S., J.K., L.P., A.B. and T.P. are supported by Michael Smith Foundation for Health Research awards. R.D.M. is a Vanier scholar (CIHR). A.B. is also supported by an NSERC award, and L.P. by a CIHR award. We are grateful for platform support from CIHR, Genome Canada, Genome BC, Canada Foundation for Innovation and the Michael Smith Foundation for Health Research. The work was funded by the BC Cancer Foundation and the CBCF BC/Yukon chapter.

Author Contributions S.P.S. and R.D.M.: led the data analysis and wrote the manuscript. M.H.: oversaw the sequencing efforts. J.K., L.P., T.P., J.S., C.S., A.B.,

R.M. and T.S.: validation of variants. A.D.: primer design. K.G. and P.W.: establishment of TTR-BREAST tumour bank. K.T., R.G., R.A.H., S.J., M.S., G.L., A.E.T., R.V., G.A.T. and R.L.W.: bioinformatic analysis. G.T., D.H. and P.W.: sample selection and histological grading. Y.Z.: Illumina sequencing library preparation. C.C. and D.H.: data analysis and interpretation. S.A. and M.A.M.: conceived and oversaw the study and wrote the manuscript.

Author Information Genome sequence data have been deposited at the European Genotype Phenotype Archive (http://www.ebi.ac.uk/ega), which is hosted by the EBI, under accession number EGAS00000000054. Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.A.M. (mmarra@bcgsc.ca) or S.A. (saparicio@bccrc.ca).