## Title
The clonal and mutational evolution spectrum of primary triple-negative breast cancers



## Journal
Nature, 486(7403)

## ISSN
0028-0836

## Authors
Shah, Sohrab P
Roth, Andrew
Goya, Rodrigo
et al.

## Publication Date
2012-06-01

## DOI
10.1038/nature10933

Peer reviewed

LETTER

doi:10.1038/nature10933

# The clonal and mutational evolution spectrum of primary triple-negative breast cancers

Sohrab P. Shah[1,2], Andrew Roth[1,2]*, Rodrigo Goya[3]*, Arusha Oloumi[1,2]*, Gavin Ha[1,2]*, Yongjun Zhao[3]*, Gulisa Turashvili[1,2]*, Jiarui Ding[1,2]*, Kane Tse[3]*, Gholamreza Haffari[1,2]*, Ali Bashashati[1,2]*, Leah M. Prentice[1,2], Jaswinder Khattra[1,2], Angela Burleigh[1,2], Damian Yap[1,2], Virginie Bernard[4], Andrew McPherson[1,2], Karey Shumansky[1,2], Anamaria Crisan[1,2], Ryan Giuliany[1,2], Alireza Heravi-Moussavi[1,2], Jamie Rosner[1,2], Daniel Lai[1,2], Inanc Birol[3], Richard Varhol[3], Angela Tam[3], Noreen Dhalla[3], Thomas Zeng[3], Kevin Ma[3], Simon K. Chan[3], Malachi Griffith[3], Annie Moradian[3], S.-W. Grace Cheng[3], Gregg B. Morin[3,5], Peter Watson[1,6], Karen Gelmon[6], Stephen Chia[6], Suet-Feung Chin[7,8], Christina Curtis[7,8,9], Oscar M. Rueda[7,8], Paul D. Pharoah[7], Sambasivarao Damaraju[10], John Mackey[10], Kelly Hoon[11], Timothy Harkins[11], Vasisht Tadigotla[11], Mahvash Sigaroudinia[12], Philippe Gascard[12], Thea Tlsty[12], Joseph F. Costello[13], Irmtraud M. Meyer[5,14,15], Connie J. Eaves[16], Wyeth W. Wasserman[4,5], Steven Jones[3,5,17], David Huntsman[1,2,18], Martin Hirst[3,15,19], Carlos Caldas[7,8,20,21], Marco A. Marra[3,5] & Samuel Aparicio[1,2]

**Primary triple-negative breast cancers (TNBCs), a tumour type defined by lack of oestrogen receptor, progesterone receptor and *ERBB2* gene amplification, represent approximately 16% of all breast cancers[1]. Here we show in 104 TNBC cases that at the time of diagnosis these cancers exhibit a wide and continuous spectrum of genomic evolution, with some having only a handful of coding somatic aberrations in a few pathways, whereas others contain hundreds of coding somatic mutations. High-throughput RNA sequencing (RNA-seq) revealed that only approximately 36% of mutations are expressed. Using deep re-sequencing measurements of allelic abundance for 2,414 somatic mutations, we determine for the first time—to our knowledge—in an epithelial tumour subtype, the relative abundance of clonal frequencies among cases representative of the population. We show that TNBCs vary widely in their clonal frequencies at the time of diagnosis, with the basal subtype of TNBC[2,3] showing more variation than non-basal TNBC. Although *p53* (also known as *TP53*), *PIK3CA* and *PTEN* somatic mutations seem to be clonally dominant compared to other genes, in some tumours their clonal frequencies are incompatible with founder status. Mutations in cytoskeletal, cell shape and motility proteins occurred at lower clonal frequencies, suggesting that they occurred later during tumour progression. Taken together, our results show that understanding the biology and therapeutic responses of patients with TNBC will require the determination of individual tumour clonal genotypes.**

To understand the patterns of somatic mutation in TNBC, we enumerated genome aberrations at all scales from 104 cases of primary TNBC (Affymetrix SNP6.0, 104 cases; RNA-seq, 80 cases; genome/exome sequencing, 65 cases: 54 exomes, 15 genomes with 4 overlapping) (Supplementary Table 1 and Supplementary Fig. 1), annotated with clinical information (Supplementary Table 2). We revalidated 2,414 somatic single nucleotide variants[4,5] (SNVs) (Supplementary Table 3) with targeted deep sequencing to a median of 20,000× coverage, including 43 non-coding splice site dinucleotide mutations (Supplementary Table 4) and 104 genes with 107 indels (Supplementary Table 5 and Supplementary Methods). Notably, the distribution of somatic mutation abundance varies in a continuous distribution among tumours (Fig. 1a) and seems to be unrelated to the proportion of the genome altered by copy number alterations (CNAs) (Fig. 1b) or tumour cellularity (Supplementary Fig. 2b). Although this distribution could be partially explained by a false-negative rate in mutation discovery, others have noted similar distributions in epithelial cancers[6], suggesting that the total mutation content of individual tumours may be shaped by biological processes or differential exposure to mutagenic influences in the population.

The overall pattern (Supplementary Fig. 3a, b) of CNA abundance appears similar (Supplementary Fig. 4) to that seen in a larger, independent series of ~2,000 SNP6.0 profiled breast tumours[7]. Among the most frequently observed CNA events (Supplementary Table 6) are the tumour suppressor and oncogenes *PARK2* (6%), *RB1* (5%), *PTEN* (3%) and *EGFR* (5%). Here we report intragenic deletions (Supplementary Fig. 5) in the *PARK2* tumour suppressor[8,9], specifically linking *PARK2* with TNBC for the first time. Consistent with previous reports in breast cancer[10], we did not observe frequent recurrent structural rearrangements (Supplementary Fig. 3d and Supplementary Table 7), although we revalidated many individual fusion events involving known oncogenes or tumour suppressors (for example, *KRAS*, *RB1*, *IDH1*, *ETV6*) (Supplementary Tables 8–10).

A comparison of RNA-seq data with genomes/exomes data revealed that only 36% of validated somatic SNVs were observed in the transcriptome sequence (Supplementary Table 3 and Supplementary Fig. 2b). In a recent lymphoma study, similar proportions were observed (137 of 329 somatic mutations expressed in RNA-seq)[11]. As expected, the proportion of low-abundance somatic SNVs observed in RNA is reflected in the distribution of wild-type, heterozygous and homozygous expressed mutations (Supplementary Fig. 2b), consistent with the notion that

[1]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada. [2]Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, British Columbia V5Z 1L3, Canada. [3]Canada's Michael Smith Genome Sciences Centre, Vancouver, British Columbia V5Z 1L3, Canada. [4]Centre for Molecular Medicine and Therapeutics, 950 West 28th Avenue, Vancouver, British Columbia V5Z 4H4, Canada. [5]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada. [6]British Columbia Cancer Agency, 600 West 10th Avenue, Vancouver, British Columbia V5Z 4E6, Canada. [7]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [8]Department of Oncology, University of Cambridge, Hills Road, Cambridge CB2 2XZ, UK. [9]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA. [10]Departments of Oncology and Laboratory Medicine and Pathology, University of Alberta, 11560 University Avenue, Cross Cancer Institute, Edmonton, Alberta T6G 1Z2, Canada. [11]Life Technologies, 101 Lincoln Centre Dr., Foster City, California 94404, USA. [12]Department of Pathology and Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, California 94143, USA. [13]Brain Tumor Research Center, Department of Neurosurgery, Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, California 94143, USA. [14]Department of Computer Science, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. [15]Centre for High-Throughput Biology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. [16]Terry Fox Laboratory, BC Cancer Agency, 675 W 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. [17]Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Dr., Burnaby, British Columbia V5A1S6, Canada. [18]Centre for Translational and Applied Genomics, BC Cancer Agency, 600 West 10th Ave, Vancouver, British Columbia V5Z 4E6, Canada. [19]Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada. [20]Cambridge Breast Unit, Addenbrookes Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK. [21]Cambridge Experimental Cancer Medicine Centre (ECMC), Cambridge CB2 0RE, UK.

*These authors contributed equally to this work.

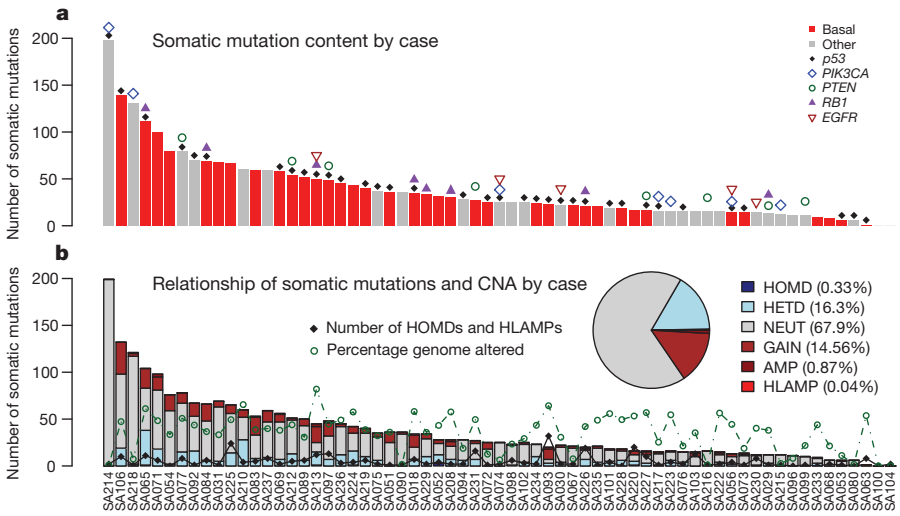


**Figure 1 | Distribution of number of validated somatic mutations by case over 65 cases.** **a**, Mutation frequency (basal, red; other, grey). Patients harbouring known driver gene mutations are indicated. **b**, Case-specific and overall (inset) distributions of mutations in CNA classes. AMP, amplification; GAIN, single copy gain; HETD, hemizygous deletion; HLAMP, high-level amplification; HOMD, homozygous deletion; NEUT, no copy number change. The number of (HOMD, HLAMP) CNAs (black diamonds) and percentage genome altered (green circles) are indicated.

low-abundance alleles may represent rarer clones in the primary tumour. We found 43 splice junction mutations with evidence for an impact on splicing patterns (Supplementary Table 4), encompassing several known tumour suppressors (*p53*, *PIK3R1*; Supplementary Fig. 6) as well as many genes not yet implicated in carcinogenesis. Analysis of 72 somatic mutations in the non-coding space of experimentally determined human regulatory regions[12] showed (Supplementary Table 11) a significant overrepresentation (31.9% versus expected 2.5%, Fisher exact test $P = 2 \times 10^{19}$) of mutations within retinoblastoma-associated protein (RB)-binding sites. Six mutations were predicted to be damaging to RB binding (Supplementary Methods and Supplementary Fig. 7). This is consistent with observations of frequent functional disruption of the RB-regulated cell cycle network[13] in TNBC.

We next searched for mutation enrichment patterns in three ways: by single gene mutation frequency over multiple cases; by the mutation frequency over multiple members of a gene family; and by correlating mutation status with expression networks. First, similar to other studies[14,15], *p53* is the most frequently mutated gene (Supplementary Table 12) with 62% of basal TNBC (determined by gene expression classification with PAM50 (ref. 16) analysis on RNA-seq expression profiles) and 43% of non-basal TNBC cases harbouring a validated somatic mutation. We also observed frequent mutations in *PIK3CA* at 10.2% (7/65), *USH2A* (Usher syndrome gene, implicated in actin cytoskeletal functions) at 9.2% (6/65), *MYO3A* at 9.2%, *PTEN* and *RB1* at 7.7% (5/65) and a further eight genes (including *ATR*, *UBR5* (also known as *EDD1*), *COL6A3*) at 6.2% (4/65) of cases in the cohort (Fig. 2a). Considering background mutation rates[17], *p53*, *PIK3CA*, *RB1*, *PTEN*, *MYO3A* and *GH1* showed evidence of single gene selection ($q < 0.1$) (Supplementary Table 13). Additional recurrent mutations of note occurred in the synuclein genes (*SYNE1* and *SYNE2*, 9.2% 6/65, recently implicated in squamous head and neck cancers[18,19]), *BRCA2* (three cases), and several other well known oncogenes (*BRAF*, *NRAS*, *ERBB2* and *ERBB3*) with mutations in two cases each. Approximately 20% of cases contained examples of potentially 'clinically actionable' somatic aberrations, including *BRAF* V600E, high-level *EGFR* amplifications and *ERBB2* and *ERBB3* mutations.

In the second approach we searched for statistically overrepresented gene families and protein functions using the Reactome functional protein interaction database[20] (Supplementary Methods). This analysis quantifies gene family involvement through sparse mutation patterns in functionally connected genes, which would be statistically underrepresented by single gene recurrent mutation analysis. The overrepresented pathways (false discovery rate (FDR) $< 0.001$) included *p53*-related pathways along with chromatin remodelling, PIK3 signalling, ERBB signalling, integrin signalling and focal adhesion, WNT/cadherin signalling, growth hormone and nuclear receptor co-activators, and ATM/RB-related pathways (Fig. 3a and Supplementary Table 14). We note that the candidate 'driver' *MYO3A*, a cytoskeleton motor protein involved in cell shape and motility, relates to several pathways upstream and downstream of integrin signalling. The mutated genes include extracellular matrix (ECM) interactions (laminins, collagens), ECM receptors (integrins), several proteins

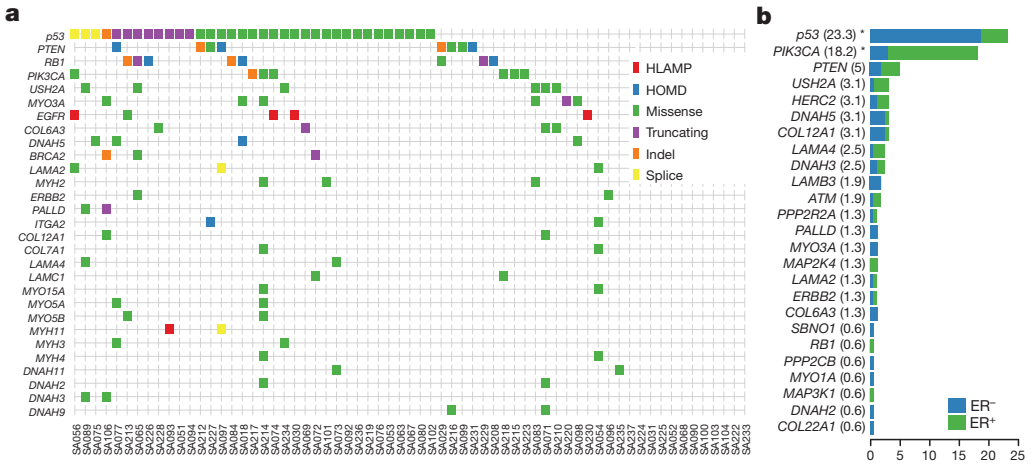


**Figure 2 | Population patterns of co-occurrence and mutual exclusion of genomic aberrations in TNBC.** **a**, Case-specific mutations in known driver genes, plus genes from integrin signalling and ECM-related proteins (laminins, collagens, integrins, myosins and dynein) derived from all aberration types: high-level amplifications (HLAMP), homozygous deletions (HOMD), missense, truncating, splice site and indel somatic mutations are depicted in genes with at least two aberrations in the population. **b**, Distribution of somatic mutations in 25 genes across all exons of 159 additional breast cancers (relative proportion of $ER^+$ cases in green, and $ER^-$ in blue), shown as a percentage of cases (in parentheses) with one or more mutations. $^*P < 0.05$.

regulating actin cytoskeleton dynamics (usherin, palladin, multiple myosins) and microtubule motor proteins (kinesins) (Fig. 2a). All of these contribute to cellular processes that have been functionally implicated in cancer progression; however, a signature of somatic mutation associated with these proteins has not been previously noted in TNBC.

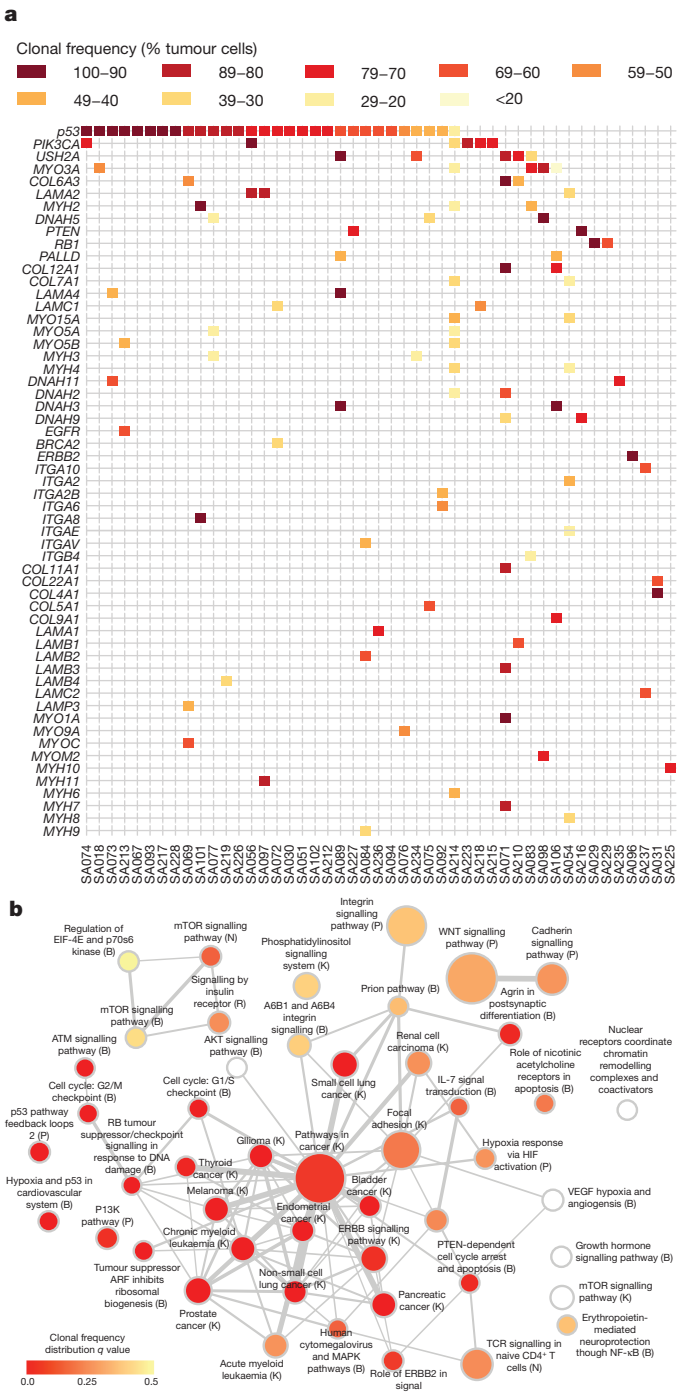


**Figure 3 | Network analysis of 254 recurrently mutated genes by somatic point mutations and indels.** **a**, Case-specific mutations shaded according to clonal frequencies in known driver genes, plus genes from integrin signalling and ECM-related proteins (laminins, collagens, integrins, myosins and dyneins). **b**, Significantly overrepresented pathways (FDR < 0.001) from recurrently mutated genes (see Supplementary Methods). Node shading encodes the adjusted *P* value (*q* value) of the comparison of the distribution of clonal frequencies of mutations in a given pathway to the overall distribution of clonal frequencies. A spectrum of higher (red) and lower (yellow) clonal frequencies is evident. Letters in parentheses indicate database sources.

To confirm the mutational spectrum in the general breast cancer population we re-sequenced all exons of 29 genes in an additional 159 breast cancers (82 oestrogen receptor $(ER)^+$ and 77 $ER^-$, tumour and matched normal) (Fig. 2b), and confirmed that many of the genes found in the discovery cohort were recurrently mutated in an additional population. Whether this pattern of mutation represents the occurrence of disease-modifying mutations, or possibly selection from other processes (for example, transcription-related hypermutation) is unknown. Interestingly, the enrichment of cytoskeletal functions in the somatic aberration landscape is also evident from the copy number and alternative splicing landscapes (Supplementary Fig. 8).

Third, we integrated both the CNA and mutation data with expression data to reveal genomic events associated with extreme changes in the transcription of interacting genes[20] (Table 1), using a bipartite graph-based method (driverNet; Supplementary Methods). The somatic aberrations showing statistically significant association with extreme expression in this analysis ($P < 0.05$) (Table 1 and Supplementary Table 15) implicate well known oncogenes and tumour suppressors (*TP53*, *PIK3CA*, *NRAS*, *EGFR*, *RB1*, *ATM*) and suggest several new genes of interest, including *PRPS2* (a nucleotide biosynthesis enzyme, rank 7), harbouring homozygous deletions in three cases, *NRC31* (a glucocorticoid receptor, rank 10) with SNVs in three cases, four PKC-related genes, *PRKCZ*, *PRKCQ*, *PRKG1* and *PRKCE*. The gene networks show a partial overlap with driverNet applied to the TCGA ovarian high-grade serous data[21] (Supplementary Table 16).

Having identified candidate driver genes and significantly overrepresented pathways, we asked how these are distributed among individual tumours by clustering a pathway–patient-mutation matrix (Supplementary Fig. 9). The abundance of implicated pathways can be

**Table 1 | Analysis of the top somatically aberrated genes influencing expression**

| Rank | Gene | gband | SNV or indel | HLAMP | HOMD | Events | *P* value |
|---|---|---|---|---|---|---|---|
| 1 | *TP53* | 17p13.1 | 35 | 0 | 0 | 2242 | 0 |
| 2 | *PIK3CA* | 3q26.32 | 7 | 0 | 0 | 441 | $1 \times 10^{-4}$ |
| 3 | *NRAS* | 1p13.2 | 2 | 0 | 0 | 271 | $4 \times 10^{-4}$ |
| 4 | *EGFR* | 7p11.2 | 1 | 5 | 0 | 220 | $4 \times 10^{-4}$ |
| 5 | *RB1* | 13q14.2 | 5 | 0 | 5 | 184 | $5 \times 10^{-4}$ |
| 6 | *PGM2* | 4p14 | 1 | 0 | 1 | 172 | $5 \times 10^{-4}$ |
| 7 | *PRPS2* | 23p22.2 | 0 | 0 | 3 | 171 | $5 \times 10^{-4}$ |
| 8 | *PTEN* | 10q23.31 | 5 | 0 | 3 | 150 | $5 \times 10^{-4}$ |
| 9 | *PRKCE* | 2p21 | 0 | 0 | 1 | 136 | $7 \times 10^{-4}$ |
| 10 | *NR3C1* | 5q31.3 | 3 | 0 | 0 | 130 | $7 \times 10^{-4}$ |
| 11 | *CREBBP* | 16p13.3 | 1 | 0 | 1 | 119 | $8 \times 10^{-4}$ |
| 12 | *CS* | 12q13.2 | 1 | 0 | 0 | 108 | 0.0011 |
| 13 | *MAN2A2* | 15q26.1 | 2 | 0 | 1 | 104 | 0.0012 |
| 14 | *HMGCS2* | 1p12 | 1 | 2 | 0 | 100 | 0.0013 |
| 15 | *HEXA* | 15q24.1 | 2 | 1 | 0 | 97 | 0.0013 |
| 16 | *ADCY9* | 16p13.3 | 2 | 1 | 0 | 91 | 0.0017 |
| 17 | *OR4N4* | 15q11.2 | 0 | 0 | 5 | 90 | 0.0017 |
| 18 | *LCLAT1* | 2p23.1 | 0 | 0 | 1 | 85 | 0.002 |
| 19 | *DGKI* | 7q33 | 2 | 0 | 0 | 82 | 0.0022 |
| 20 | *CYP2A6* | 19q13.2 | 1 | 0 | 0 | 80 | 0.0024 |
| 21 | *JAK1* | 1p31.3 | 1 | 0 | 0 | 78 | 0.0026 |
| 22 | *POLR1A* | 2p11.2 | 2 | 0 | 0 | 78 | 0.0026 |
| 23 | *PLD1* | 3q26.31 | 1 | 0 | 0 | 69 | 0.0038 |
| 24 | *IDH3B* | 20p13 | 1 | 0 | 1 | 68 | 0.004 |
| 25 | *PAPSS2* | 10q23.2 | 0 | 0 | 3 | 67 | 0.0041 |
| 26 | *PRKX* | 23p22.33 | 0 | 0 | 2 | 65 | 0.0046 |
| 27 | *TPH2* | 12q21.1 | 1 | 0 | 0 | 65 | 0.0046 |
| 28 | *UGT2B17* | 4q13.2 | 0 | 0 | 1 | 63 | 0.0053 |
| 29 | *RRM2* | 2p25.1 | 1 | 0 | 0 | 57 | 0.0072 |
| 30 | *ATM* | 11q22.3 | 1 | 0 | 0 | 55 | 0.0084 |
| 31 | *CLCA1* | 1p22.3 | 2 | 0 | 0 | 54 | 0.009 |
| 32 | *PRKCZ* | 1p36.33 | 1 | 0 | 0 | 53 | 0.0095 |

Rank, derived by the driverNet algorithm (see Supplementary Methods); gene, somatically aberrated gene; gband, chromosomal band containing gene; SNV or indel, the number of cases harbouring an SNV or indel in the gene; HLAMP, the number of cases harbouring a predicted high-level amplification; HOMD, the number of cases harbouring a predicted homozygous deletion; events, number of gene expression outliers (see Supplementary Methods) coincident with a genomic aberration and where the outlying gene is connected to the aberrated gene; *P* value, statistical significance based on a randomly generated background distribution (Supplementary Methods).

seen to be only partially related to the total number of mutations in a case, groups 1 and 2 having on average fewer mutations per case. The frequent involvement of pathways with *p53*, *PTEN* and *PIK3CA* as members, is noted (Supplementary Fig. 9); however, the case groupings also vary by the progressive inclusion of additional pathways (for example, WNT signalling, integrin signalling, ERBB signalling, hypoxia and PI3K). More than two thirds of cases contained one or more mutations in the actin/cytoskeletal functions group of genes (Supplementary Fig. 9). Some 12% of cases did not contain somatic aberrations in any of the frequent drivers or cytoskeletal genes (Supplementary Table 12). This suggests that primary TNBCs are mutationally heterogenous from the outset, with some patients' tumours having a small number of implicated pathways and few mutations, whereas other patients present with tumours containing extensive mutation burdens and multiple pathway involvement.

Motivated by the observation that early primary TNBCs show a wide variation of mutation content, we asked whether the clonal composition of these primary cancers is similarly varied. We and others have shown[22,23] how deep-frequency measurements of allelic abundance can be used to study tumour clonal evolution. Clonal mutation frequency, a compound measure of clonal complexity, (Fig. 4a) can be estimated from allele abundance, once the influence of copy number states, regional loss of heterozygosity (LOH state) and tumour cellularity have been considered (although we note that approximately 68% of SNVs in this study are in diploid, neutral regions). To extend allelic abundance measurements to estimation of clonal frequencies, we implemented a Dirichlet process clustering model (pyclone; Supplementary Methods and Supplementary Fig. 10) that simultaneously estimates the genotype and clonal frequency given a list of deeply sequenced mutations and their local copy number and heterozygosity contexts.

Using the set of deeply sequenced (median 20,000×), validated SNVs, our analysis revealed (Fig. 4b) that groups of mutations within individual cases have different clonal frequencies, indicative of distinct clonal genotypes. Remarkably, the tumours exhibit a wide spectrum of modes over clonal frequencies (Fig. 4b and Supplementary Fig. 11), with some cases showing only one or two frequency modes (Fig. 4b), indicating a smaller number of clonal genotypes, whereas other tumours exhibit multiple clonal frequency modes, indicating more extensive clonal evolution. Consistent with early 'driver gene' status, mutations in known tumour suppressors such as p53 tend to occur in the highest clonal frequency group in most tumours. However, in some cases (for example, SA219, SA236; Fig. 4b, Supplementary Fig. 11) p53 resides in lower-abundance clonal frequency groups (Supplementary Fig. 12 and Fig. 3a), suggesting that it was not the founding event. Although the number of clonal frequency modes tends to increase with the number of mutations, the relationship is not strictly linear (Fig. 4c). To determine whether basal and non-basal cancers differ in their clonality, we compared the distribution of clonal modes (clusters) by case and as an overall distribution, and note that basal TNBCs have more clonal frequency modes than non-basal TNBCs (Fig. 4c). Both of these distributions emphasize a key observation; namely, that at the time of diagnosis TNBCs already display a widely varying clonal evolution that mirrors the variation in mutational evolution.

Finally, we asked where key pathways appear in the distribution of clonal frequency groups. We examined the clonal frequency of genes in each pathway and ascertained if there was a deviation away from the distribution of clonal frequency for all mutations. As expected, pathways involving p53 and *PIK3CA* showed significantly skewed distributions (Wilcoxon, $q < 0.01$; Fig. 3b and Supplementary Fig. 12) towards higher clonal frequencies, consistent with their roles in early tumorigenesis (Fig. 3a and Supplementary Table 17). Intriguingly, pathways with cytoskeletal genes such as myosins, laminins, collagens and integrins tend to have lower median clonal frequencies, suggesting that somatic mutations in these genes are acquired much later (Fig. 3b). Notably, the median clonal frequency for Reactome pathway 'p53

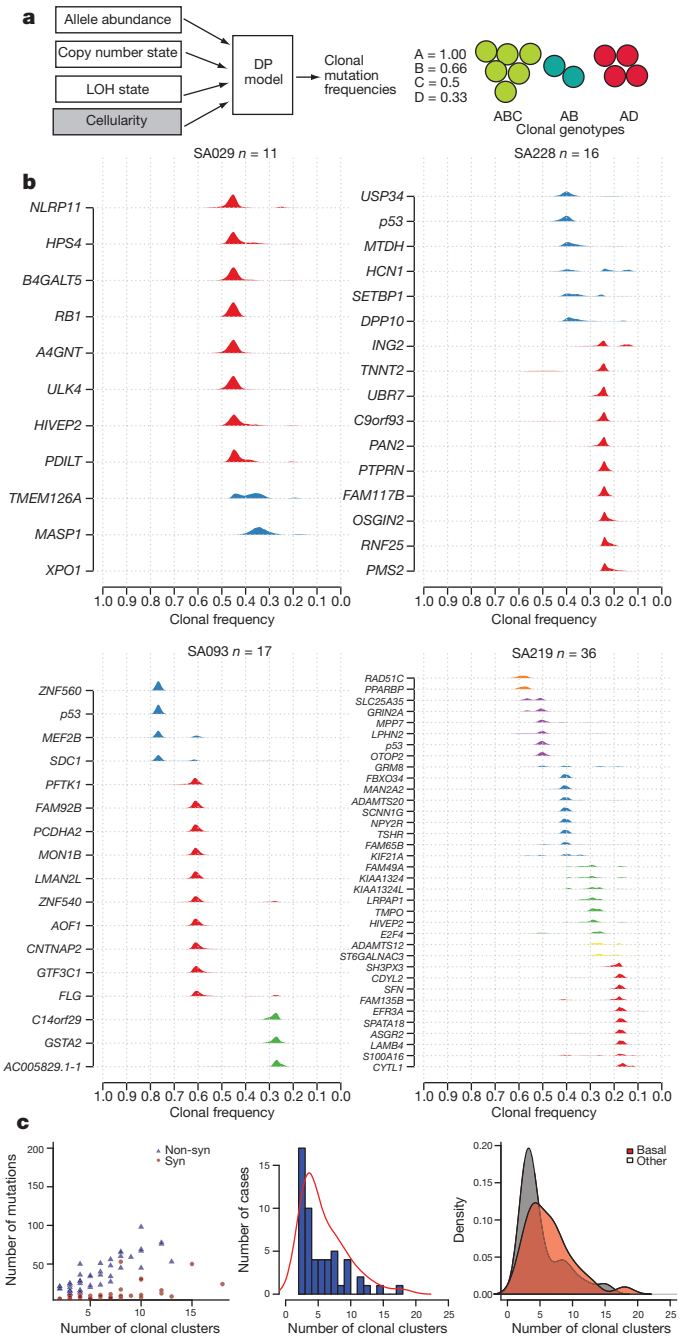


**Figure 4 | Clonal evolution in TNBC.** **a**, Schematic representation of integration of CNA, LOH, allelic abundance measurements and normal cell contamination for clonal frequency estimation using a Dirichlet process (DP) model (left). Example of a mixture of three clonal genotypes composed of four mutations (A, B, C, D) and their resulting clonal frequencies. **b**, Estimated clonal frequencies for four cases are shown as the distribution of posterior probabilities from the pyclone model (Supplementary Methods). Clonal frequency distributions are coloured by their frequency group membership. **c**, Left, relationship of mutation abundance (synonymous (Syn) and non-synonymous (Non-syn)) and the inferred number of clonal clusters. Middle, distribution and kernel density (red line) of the number of inferred clonal clusters over 54 TNBCs. Right, kernel density distribution of clonal clusters for basal (red) and non-basal (grey) tumours.

pathway feedback loops', including 46 mutations in *ATM*, *ATR*, *NRAS*, *PIK3CA*, *PTEN*, *SIAH1* and *p53*,was 73% (Wilcoxon, $q = 0.0007$), whereas 'integrin cell surface interactions', including 23 mutations in integrin, laminin and collagen genes, had a median clonal frequency of 42% (Wilcoxon, $q = 0.9569$).

Primary TNBCs are still treated as if they were a single disease entity, yet it is clear they do not behave as a single entity in response to current therapies. Here we show for the first time, using next-generation sequencing mutational profiling methods, that treatment-naive TNBCs display a complete spectrum of mutational and clonal evolution, with some patients' tumours showing only a few somatic coding sequence point mutations with a limited number of molecular pathways implicated, whereas other patients' tumours exhibit considerable additional mutational involvement. Moreover, the clonal heterogeneity of these cancers is also a continuum, with some patients presenting with low-clonality cancers and other cases exhibiting more extensive clonal evolution at diagnosis. In this respect, the basal expression subtype of TNBCs also tends to show higher clonality at diagnosis, although the relationship is not exact.

In clonally evolving tumours, identification of genes by single gene mutation frequency measurements will probably favour early driver genes, because the subsequent involvement of multiple additional pathways during tumour progression is unlikely to be observed as a frequent single gene mutation. The clonality analysis emphasizes this point: known drivers such as *p53*, *PIK3CA* and *PTEN* have among the highest clonal frequencies, whereas mutations in cell shape/motility and ECM-signalling genes appear in the lower clonal frequency groups, distributed over many genes. Although p53 somatic mutations are clearly early events, the clonal frequencies observed in some TNBC suggest that they are not always the first event, raising a question about what drives early clonal expansion in some of these cancers. Our findings suggest that each TNBC at the time of primary diagnosis may be at a very different phase of molecular progression, with possible implications for approaches to the biology of low clonality versus high clonality primary tumours.

## METHODS SUMMARY

The genomes and transcriptomes of 104 TNBCs were profiled with Affymetrix SNP6.0 arrays (all cases), RNA-seq (80 cases; Illumina GAII), and whole exome/genome sequencing (65 cases; tumour and normal DNA). Exomes were obtained using Agilent's Human All Exon SureSelect Target Enrichment System v.1 followed by Illumina GAII sequencing, and whole genomes were sequenced using Life Technologies SOLiD system. Data were analysed using computational approaches to detect somatic SNVs[4,5], indels, copy number alterations, gene fusions and gene expression patterns. Predictions were then validated using orthogonal experimental assays, including targeted ultra-deep amplicon sequencing of SNVs to ~20,000× redundancy. We determined single genes under selection using a statistical approach that considers patient-specific background mutation and transition/transversion rates. Mutations predicted to alter transcriptional profiles were determined using an integrated bipartite graph-based method (driverNet) that associates genomic aberrations with outlying expression patterns informed by predefined pathway gene sets. Disrupted pathways were determined using the Reactome FI Cytoscape plugin. Clonal analysis was performed (cases with >10 mutations) using a Dirichlet process statistical model that simultaneously estimates clonal frequencies and mutation genotype given deeply sequenced somatic SNVs and copy number estimates. Experimental assays and analytical methodology are detailed in the Supplementary Information.

**Received 16 June 2011; accepted 15 February 2012.**
**Published online 4 April 2012.**

1. Blows, F. M. *et al.* Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS Med.* **7,** e1000279 (2010).
2. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).
3. Sørlie, T. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98,** 10869–10874 (2001).
4. Roth, A. *et al.* JointSNVmix: A probabilistic model for accurate detection of somatic mutations in normal/tumour paired next generation sequencing data. *Bioinformatics* http://dx.doi.org/10.1093/bioinformatics/bts053 (27 January 2012).
5. Ding, J. *et al.* Feature-based classifiers for somatic mutation detection in tumour—normal paired sequencing data. *Bioinformatics* **28,** 167–175 (2012).
6. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).
7. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* http://dx.doi.org/10.1038/nature10983 (this issue).
8. Poulogiannis, G. *et al. PARK2* deletions occur frequently in sporadic colorectal cancer and accelerate adenoma development in *Apc* mutant mice. *Proc. Natl Acad. Sci. USA* **107,** 15145–15150 (2010).
9. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463,** 893–898 (2010).
10. Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462,** 1005–1010 (2009).
11. Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476,** 298–303 (2011).
12. Chicas, A. *et al.* Dissecting the unique role of the retinoblastoma tumor suppressor during cellular senescence. *Cancer Cell* **17,** 376–387 (2010).
13. Herschkowitz, J. I., He, X., Fan, C. & Perou, C. M. The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. *Breast Cancer Res.* **10,** R75 (2008).
14. Langerød, A. *et al.* TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast Cancer Res.* **9,** R30 (2007).
15. Børresen-Dale, A.-L. TP53 and breast cancer. *Hum. Mutat.* **21,** 292–300 (2003).
16. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).
17. Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27,** 175–181 (2011).
18. Agrawal, N. *et al.* Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* **333,** 1154–1157 (2011).
19. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333,** 1157–1160 (2011).
20. Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.* **11,** R53 (2010).
21. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
22. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461,** 809–813 (2009).
23. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).

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

**Acknowledgements** The support of the BC Cancer Agency Tumour Bank, CBCF Breast Tumour Bank Alberta and the Addenbrookes Tumour bank (supported by NIHR and ECMC) is acknowledged. Technical support is acknowledged from the Centre for Translational Genomics, the Michael Smith Genome Sciences Centre technical group, the BCCA Flow Cytometry Core Facility in the Terry Fox Laboratory and the Cancer Research UK Cambridge Research Institute. Supported by the BC Cancer Foundation, US Department of Defense CDMRP program, Canadian Breast Cancer Foundation (BC Yukon) (to S.A. and S.S.), Michael Smith Foundation for Health Research (to S.S.), US National Institutes of Health (NIH) Roadmap Epigenomics Program, NIH grant 5U01ES017154-02 (to M.H., M.A.M., J.C. and T.T.), Cancer Research UK (to C. Caldas and P.D.P.) and the National Institute of General Medical Sciences (R01GM084875 to W.W.W.), the Canadian Breast Cancer Research Alliance and the Canadian Cancer Society (to S.A. and C.E.). We thank B. Reva, Y. Antipin and C. Sander (Memorial Sloan Kettering Cancer Center) for assistance with MutationAssessor, and G. Wu (Ontario Institute for Cancer Research) for assistance with Reactome.

**Author Contributions** S.A., S.P.S., C. Caldas and M.A.M. designed and implemented the research plan and wrote the manuscript. S.P.S., A.R., R. Goya, G. Ha, J.D., G. Haffari, A. Bashashati, A. McPherson, K.S., A.C., R. Giuliany, A.H.-M., J.R., D.L., I.B., R.V., S.W.C., M.G., I.M.M., S.J., C. Curtis, O.M.R., P.D.P., V.B. and W.W.W. conducted bioinformatic analyses of the data and/or gave advice on analytic methodology. G.T. conducted histopathological review and immunohistohemistry. A.O., Y.Z., G.T., K.T., L.M.P., J.K., A.B., D.Y., A.T., N.D., T.Z., S.-F.C., K.M. and M.H. conducted sequencing or experimental validation of somatic aberrations. D.Y., A. Moradian, S.-W.G.C. and G.B.M. conducted proteome validation of splicing. P.W., K.G., S.C., S.-F.C., G.T., J.M., C. Caldas, P.D.P. and D.H. collected and interpreted clinical data. S.D., J.F.C., T.T., M.S., P.G. and C.J.E. contributed materials or reagents. K.H., V.T., T.H., M.H. and M.A.M. generated sequence data.

**Author Information** Aligned exome/genome sequence data, RNA-seq data and Affymetrix SNP6.0 data sets are available at the European Genome-phenome Archive (http://www.ebi.ac.uk/ega/) under study accession number EGAS00001000132. Normal reference RNA-seq datasets are available at the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces) under study accession number SRP000930. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.A. (saparicio@bccrc.ca), C. Caldas (carlos.caldas@cancer.org.uk), S.P.S. (sshah@bccrc.ca) or M.M. (mmarra@bcgsc.ca).