The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups

Christina Curtis1,2,4*, Sohrab P. Shah3,4*, Suet–Feung Chin3,4*, Gulisa Turashvili3,4*, Oscar M. Rueda1,2, Mark J. Dunning2, Doug Speed3,4,5*, Andy G. Lynch1,2, Shamith Samarajiwa1,2, Yinyin Yuan1,2, Stefan Gräf1,2, Gavin Ha3, Gholamreza Haffari3, Ali Bashashati3, Roslin Russell2, Steven McKinney5,4, METABRIC Group6, Anita Langerød7, Andrew Green7, Elena Provenzano8, Gordon Wishart8, Sarah Pinder9, Peter Watson1,2,10, Florian Markowitz1,2, Leigh Murphy10, Ian Ellis11, Arnie Purushotham9,11, Anne–Lise Børresen–Dale12,12, James D. Brenton2,13, Simon Tavare6,2,1,14, Carlos Caldas3,2,8,13 & Samuel Aparicio7,4

The elucidation of breast cancer subgroups and their molecular drivers requires integrated views of the genome and transcriptome from representative numbers of patients. We present an integrated analysis of copy number and gene expression in a discovery and validation set of 997 and 995 primary breast tumours, respectively, with long–term clinical follow–up. Inherited variants (copy number variants and single nucleotide polymorphisms) and acquired somatic copy number aberrations (CNAs) were associated with expression in ~40% of genes, with the landscape dominated by cis– and trans–acting CNAs. By delineating expression outlier genes driven in cis by CNAs, we identified putative cancer genes, including deletions in PPP2R2A, MTAP and MAP2K4. Unsupervised analysis of paired DNA–RNA profiles revealed novel subgroups with distinct clinical outcomes, which reproduced in the validation cohort. These include a high–risk, oestrogen–receptor–positive 11q13/14 cis–acting subgroup and a favourable prognosis subgroup devoid of CNAs. Trans–acting aberration hotspots were found to modulate subgroup–specific gene networks, including a TCR deletion–mediated adaptive immune response in the ‘CNA–devoid' subgroup and a basal–specific chromosome 5 deletion–associated mitotic network. Our results provide a novel molecular stratification of the breast cancer population, derived from the impact of somatic CNAs on the transcriptome.

Inherited genetic variation and acquired genomic aberrations contribute to breast cancer initiation and progression. Although somatically acquired CNAs are the dominant feature of sporadic breast cancers, the driver events that are selected for during tumorigenesis are difficult to elucidate as they co–occur alongside a much larger landscape of random non–pathogenic passenger alterations1,2 and germline copy number variants (CNVs). Attempts to define subtypes of breast cancer and to discern possible somatic drivers are still in their relative infancy3,4, in part because breast cancer represents multiple diseases, implying that large numbers (many hundreds or thousands) of patients must be studied. Here we describe an integrated genomic/transcriptomic analysis of breast cancers with long–term clinical outcomes composed of a discovery set of 997 primary tumours and a validation set of 995 tumours from METABRIC (Molecular Taxonomy of Breast Cancer International Consortium).

A breast cancer population genomic resource

We assembled a collection of over 2,000 clinically annotated primary fresh–frozen breast cancer specimens from tumour banks in the UK and Canada (Supplementary Tables 1–3). Nearly all oestrogen receptor (ER)–positive and/or lymph node (LN)–negative patients did not receive chemotherapy, whereas ER–negative and LN–positive patients did. Additionally, none of the HER2 patients received trastuzumab. As such, the treatments were homogeneous with respect to clinically relevant groupings. An initial set of 997 tumours was analysed as a discovery group and a further set of 995 tumours, for which complete data later became available, was used to test the reproducibility of the integrative clusters (described below). An overview of the main analytical approaches is provided in Supplementary Fig. 1. Details concerning expression and copy number profiling, including sample assignment to the PAM50 intrinsic subtypes4,5,6 (Supplementary Fig. 2), copy number analysis (Supplementary Tables 4–8) and validation (Supplementary Figs 3 and 4 and Supplementary Tables 9–11), and TP53 mutational profiling (Supplementary Fig. 5) are described in the Supplementary Information.

Genome variation affects tumour expression architecture

Genomic variants are considered to act in cis when a variant at a locus has an impact on its own expression, or in trans when it is associated with a distant gene. The high concordance of CNAs and gene expression (Supplementary Fig. S1A) suggests that these variants contribute to expression variation, in line with the well–established association between CNAs and gene expression variation7,8. For example, deletions in MAP2K4 were associated with decreased expression of a large number of genes (Fig. 1a). It is possible that CNAs directly contribute to expression variation through trans–acting mechanisms, for instance, by modulating transcription factor binding sites or by altering transcripts levels indirectly through regulation of the transcriptome of genes in linked chromatin regions. Genetic variation contributes to variation in gene expression through both cis– and trans–acting mechanisms9,10.

1Department of Oncology, University of Cambridge, Hills Road, Cambridge CB2 1XZ, UK. 2Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. 3Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada. 4Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, British Columbia V5Z 1L3, Canada. 5Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Centre for Mathematical Sciences, Cambridge CB3 0WA, UK. 6Department of Genetics, Institute for Cancer Research, Osaka University Hospital Radiumhaeftefalen, Saint Olaf, 0310 Oslo, Norway. 7Department of Histopathology, School of Molecular Medical Sciences, University of Nottingham, Nottingham NG5 1PB, UK. 8Cambridge Breast Unit, Addenbrooke’s Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge, Cambridge CB2 2QQ, UK. 9King’s College London, Breakthrough Breast Cancer Research Unit, London WC2R 2LS, UK. 10Misho Institute of Cell Biology, University of Manitoba, Manitoba RS3 OY9, Canada. 11NIHR Comprehensive Biomedical Research Centre at Guy’s and St Thomas’ NHS Foundation Trust and King’s College London, London WC2R 2LS, UK. 12Institute for Clinical Medicine, Faculty of Medicine, University of Southern California, Los Angeles, California 90089, USA. 13Present addresses: Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA (Ch.C.). 14University College London, Genetics Institute, WC1E 6BT, UK (D.S.).

*These authors contributed equally to this work.

©2012 Macmillan Publishers Limited. All rights reserved
with genes at other sites in the genome. We generated a map of CNAs, CNVs (Supplementary Fig. 6, Supplementary Tables 12–15) and single nucleotide polymorphisms (SNPs) in the breast cancer genome to distinguish germline from somatic variants (see Methods), and to examine the impact of each of these variants on the expression landscape. Previous studies have shown that most heritable gene expression traits are governed by a combination of cis (proximal) loci, defined here as those within a 3-megabase (Mb) window surrounding the gene of interest, and trans (distal) loci, defined here as those outside that window. We assessed the relative influence of SNPs, CNVs and CNAs on tumour expression architecture, using each of these variants as a predictor (see Methods) to elucidate expression quantitative trait loci (eQTLs) among patients.

Both germline variants and somatic aberrations were found to influence tumour expression architecture, having an impact on >39% (11,198/28,609) of expression probes genome-wide based on analysis of variance (ANOVA; see Methods), with roughly equal numbers of genes associated in cis and trans. CNVs were associated with the greatest number of expression profiles (Fig. 1, Supplementary Figs 7–13 and Supplementary Tables 16–20), but were rivalled by SNPs to explain a greater proportion of expression variation on a per-gene basis genome-wide, whereas the contribution from CNVs to explain a greater proportion of expression variation on a per-gene basis genome-wide, whereas the contribution from CNVs was more moderate (Fig. 1b and Supplementary Table 21). The true ratio of putative trans versus cis eQTLs is hard to estimate; however, the large sample size used here allowed the detection of small effects, with 5,401 and 5,462 CNAs significantly (Sidak adjusted P-value <0.0001) associated in cis or in trans, respectively. Whereas cis-associations tended to be stronger, the trans-acting loci modulated a larger number of messenger RNAs, as described below.

**Expression outliers refine the breast cancer landscape**

As shown above, ~20% of loci exhibit CNA-expression associations in cis (Supplementary Fig. 14). To refine this landscape further and identify the putative driver genes, we used profiles of outlier expression (see Methods and ref. 10) and the high resolution and sensitivity of the Affymetrix SNP 6.0 platform to delineate candidate regions. This approach markedly reduces the complexity of the landscape to 45 regions (frequency > 5, Fig. 2) and narrows the focus, highlighting novel regions that modulate expression. The full enumeration of regions delineated by this approach and their subtype-specific associations (Supplementary Figs 15 and 16 and Supplementary Tables 22–24) includes both known drivers (for example, ZNF703 (ref. 11), PTEN (ref. 12), MYC, CCND1, MDM2, ERBB2, CCNE1 (ref. 13)) and putative driver aberrations (for example, MDM1, MDM4, CDK3, CDK4, CAMK1D, PIK3R, NCOR1).

The deletion landscape of breast cancer has been poorly explored, with the exception of PTEN. We illustrate three additional regions of significance centred on PPP2R2A (8p21, Fig. 2, region 11), MTAP (9p21, Fig. 2, region 15) and MAP2K4 (17p11, Fig. 2, region 33), which exhibit heterozygous and homozygous deletions (Supplementary Figs 15, 17–19 and Supplementary Table 24) that drive expression of these loci. We observe breast cancer subtype-specific (enriched in mitotic ER-positive cancers) loss of transcript expression in PPP2R2A, a B-regulatory subunit of the PP2A mitotic exit holoenzyme complex. Somatic mutations in PPP2R1A have recently been reported in clear cell ovarian cancers and endometrioid cancers, and methylation silencing of PPP2R2B has also been observed in colorectal cancers. Thus, dysregulation of specific PPP2R2A functions in luminal B breast cancers adds a significant pathophysiology to this subtype.

**MTAP** (9p21, a component of methyladenosine salvage) is frequently co-deleted with the CDKN2A and CDKN2B tumour suppressor genes in a variety of cancers as we observe here (Supplementary Figs 17c and 18). The third deletion encompasses MAP2K4 (also called MKK4) (17p11), a p38/Jun dual specificity serine/threonine protein kinase. MAP2K4 has been proposed as a recessive cancer gene, with mutations noted in cell lines. We show, for the first time, the recurrent deletion of MAP2K4 (Supplementary Figs 17d and 19) concomitant with outlying expression (Supplementary Fig. 15) in predominantly ER-positive cases, and verify homozygous deletions (Supplementary Table 9) in primary tumours, strengthening the evidence for MAP2K4 as a tumour suppressor in breast cancer.

**Figure 1 | Germline and somatic variants influence tumour expression architecture.** a, Venn diagrams depict the relative contribution of SNP, CNV and CNA to genome-wide, cis and trans tumour expression variation for significant expression associations (Sidak adjusted P-value \( \leq 0.0001 \)).

b, Histograms illustrate the proportion of variance explained by the most significantly associated predictor for each predictor type, where several of the top associations are indicated.
Frequency of cases exhibiting an outlying expression profile at regions across the genome is indicated by the arrows and numbered regions. The frequency (absolute count) copy number events in the CNA landscape highlights putative driver genes, as drivers.

7p, 8, 11q, 14q, 16, 17q and 20q (Fig. 3a), including both positive and This revealed strong off-diagonal patterns at loci on chromosomes 1q, examining the matrices of CNA–expression associations (see Methods).

We next asked how Trans modules, which highlight known driver loci such as Importantly, these aberration hotspots can be grouped into pathway modules, which highlight known driver loci such as ERBB2 and MYC, as well as novel loci associated with large trans expression modules (Supplementary Tables 25 and 26). The T-cell receptor (TCR) loci on chromosomes 7 (TRG) and 14 (TRA) represent two such hotspots that modulated 381 and 153 unique mRNAs, respectively, as well as 19 dually regulated genes (Supplementary Fig. 20). These cognate mRNAs were highly enriched for T-cell activation and proliferation, dendritic cell presentation, and leukocyte activation, which indicate the induction of an adaptive immune response associated with tumour-infiltrating lymphocytes (Fig. 3b, Supplementary Fig. 20 and Supplementary Tables 27 and 28), as described later.

In a second approach, we examined the genome-wide patterns of linear correlation between copy number and expression features (see Methods), and noted the alignment of several off-diagonal signals, including those on chromosome 1q, 8q, 11q, 14q and 16 (Supplementary Fig. 21). Additionally, a broad signal on chromosome 5 localizing to a deletion event restricted to the basal-like tumours was observed (Supplementary Fig. 21), but was not detected with the eQTL framework, where discrete (as opposed to continuous) copy number values were used. This basal-specific trans module is enriched for transcriptional changes involving cell cycle, DNA damage repair and apoptosis (Supplementary Table 29), reflecting the high mitotic index typically associated with basal-like tumours, described in detail below.

Integrative clustering reveals novel subgroups

Using the discovery set of 997 breast cancers, we next asked whether novel biological subgroups could be found by joint clustering of copy number and gene expression data. On the basis of our finding that cis-acting CNAs dominated the expression landscape, the top 1,000 cis-associated genes across all subtypes (Supplementary Table 30) were used as features for input to a joint latent variable framework for integrative clustering (see Methods). Cluster analysis suggested 10 groups (based on Dunn’s index) (see Methods and Supplementary Figs 22 and 23), but for completeness, this result was compared with the results for alternative numbers of clusters and clustering schemes (see Methods, Supplementary Figs 22–27 and Supplementary Tables 31–33). The 10 integrative clusters (labelled IntClust 1–10) were typified by well-defined copy number aberrations (Fig. 4, Supplementary Figs 22, 28–30 and Supplementary Tables 34–39), and split many of the intrinsic subtypes (Supplementary Figs 31–33). Kaplan–Meier plots of disease-specific survival and Cox proportional hazards models indicate subgroups with distinct clinical outcomes (Fig. 5, Supplementary Figs 34, 35 and Supplementary Tables 40 and 41). To validate these results, we trained a classifier (754 features) for the integrative subtypes in the discovery set using the nearest shrunk centroids approach (see Methods and Supplementary Tables 42 and 43), and then classified the independent validation set of 995 cases into the 10 groups (Supplementary Table 44). The reproducibility of the clusters in the validation set is shown in three ways. First, classification of the validation set resulted in the assignment of a similar proportion of cases to the in-group proportions (IGP) measure.

Among the integrative clusters, we first note an ER-positive subgroup composed of 11q13/14 cis-acting luminal tumours (IntClust 2,
where it is more challenging to distinguish the driver UVRAG–GAB2 (Fig. 2) suggest at least two separate amplicons at 11q13/14, one at 11q13.3, which have been previously linked to breast cancer subgroups (Supplementary Tables 46 and 47) and the expression outlier profiles for this region are enriched for samples belonging to IntClust 2 (Fig. 2, inset region 23) and all 45 members of this subgroup harboured amplifications of these genes, with high frequencies of amplification also observed for CCND1 (n = 39) and EMSY (n = 34). In light of these observations, the 11q13/14 amplicon may be driven by a cassette of genes rather than a single oncogene.

Second, we note the existence of two subgroups marked by a paucity of copy number and cis–acting alterations. These subgroups cannot be explained by low cellularity tumours (see Methods). One subgroup (IntClust3, n = 156) with low genomic instability (Fig. 4 and Supplementary Fig. 22) was composed predominantly of luminal A cases, and was enriched for histotypes that typically have good prognosis, including invasive lobular and tubular carcinomas. The other subgroup (IntClust 4, n = 167) was also composed of favourable outcome cases, but included both ER-positive and ER-negative cases and varied intrinsic subtypes, and had an essentially flat copy number landscape, hence termed the ‘CNA-devoid’ subgroup. A significant proportion of cases within this subgroup exhibit extensive lymphocytic infiltration (Supplementary Table 45).

Third, several intermediate prognosis groups of predominantly ER-positive cancers were identified, including a 17q23/20q cis–acting luminal B subgroup (IntClust 1, n = 76), an 8p12 cis–acting luminal subgroup (IntClust 6, n = 44), as well as an 8q cis–acting/20q-amplified mixed subgroup (IntClust 9, n = 67). Two luminal A subgroups with similar CNA profiles and favourable outcome were noted. One subgroup is characterized by the classical 1q gain/16q loss (IntClust 8, n = 143), which corresponds to a common translocation event25, and the other lacks the 1q alteration, while maintaining the 16p gain/16q loss with higher frequencies of 8q amplification (IntClust 7, n = 109). We also noted that the majority of basal-like tumours formed a stable, mostly high-genomic instability subgroup (IntClust 10, n = 96). This subgroup had relatively good long-term outcomes (after 5 years), consistent with ref. 26, and characteristic cis–acting alterations (5 loss/8q gain/10p gain/12p gain).

The ERBB2-amplified cancers composed of HER2-enriched (ER-negative) cases and luminal (ER-positive) cases appear as IntClust 5 (n = 94), thus refining the ERBB2 intrinsic subtype by grouping additional patients that might benefit from targeted therapy. Patients in this study were enrolled before the general availability of trastuzumab, and as expected this subgroup exhibits the worst disease-specific survival at both 5 and 15 years and elevated hazard ratios (discovery set: 3.899, 95% confidence interval (2.234–6.804); validation set: 4.447, 95% confidence interval (2.284–8.661)).

Pathway deregulation in the integrative subgroups

Finally, we projected the molecular profiles of the integrative subgroups onto pathways to examine possible biological themes among breast cancer subgroups (Supplementary Tables 46 and 47) and the relative impact of cis and trans expression modules on the pathways. The CNA-devoid (IntClust 4) group exhibits a strong immune and inflammation signature involving the antigen presentation pathway, OX40 signalling, and cytotoxic T-lymphocyte-mediated apoptosis (Supplementary Fig. 36). Given that trans–acting deletion hotspots were localized to the TRG and TRA loci and were associated with an adaptive immune response module, we asked whether these deletions contribute to alterations in this pathway. The CNA-devoid subgroup (IntClust 4) was found to exhibit nearly twice as many deletions (typically heterozygous loss) at the TRG and TRA loci (~20% of cases) as compared to the other subtypes (with the exception of IntClust 10), and deletions of both TCR loci were significantly associated with severe lymphocytic infiltration (χ² test, P < 10⁻⁷ and P < 10⁻⁸, respectively). Notably, these trans–associated mRNAs were significantly enriched in the immune response signature of the CNA-devoid subgroup (Supplementary Fig. 36) as well as among genes differentially expressed in CNA-devoid cases with severe lymphocytic infiltration (Supplementary Fig. 37). We conclude that genomic copy number loss
at the TCR loci drives a trans-acting immune response module that associates with lymphocytic infiltration, and characterizes an otherwise genomically quiescent subgroup of ER-positive and ER-negative patients with good prognosis. These observations suggest the presence of mature T lymphocytes (with rearranged TCR loci), which may explain an immunological response to the cancer. In line with these findings, a recent study\(^{27}\) demonstrated the association between CD8\(^+\) lymphocytes and favourable prognosis.

Also among the trans-influenced groups is IntClust 10 (basal-like cancer enriched subgroup), which harbours chromosome 5q deletions (Supplementary Fig. 21). Numerous signalling molecules, transcription factors and cell division genes were associated in trans with this deletion event in the basal cancers, including alterations in AURKB, BCL2, BUB1, CDCA3, CDCA4, CDC20, CDC45, CHEK1, FOXM1, HDAC2, IGF1R, KIF2C, KIFC1, MTHFD1L, RAD51AP1, TTK and UBE2C (Supplementary Fig. 38). Notably, TTK (MPS1), a dual specificity kinase that assists AURKB in chromosome alignment during mitosis, and recently reported to promote aneuploidy in breast cancer\(^{28}\), was upregulated. These results indicate that 5q deletions modulate the coordinate transcriptional control of genomic and chromosomal instability and cell cycle regulation within this subgroup.

In contrast to these subtype-specific trans-associated signatures, the high-risk 11q13/14 subgroup was characterized by strong cis-acting associations. Like the basal cancers, this subgroup also exhibited alterations in key cell-cycle-related genes (Supplementary Fig. 39), which probably have a role in its aggressive pathophysiology, but the nature of the signature differs. In particular, the regulation of the G1/S transition by BTG family proteins, which include CCND1, PPP2R1B and E2F2, was significantly enriched in the 11q13/14 cis-acting subgroup, but not the basal cancers, and this is consistent with CCND1 and the PPP2R subunit representing subtype-specific drivers in these tumours.

**Discussion**

We have generated a robust, population-based molecular subgrouping of breast cancer based on multiple genomic views. The size and nature of this cohort made it amenable to eQTL analyses, which can aid the identification of loci that contribute to the disease phenotype\(^{29}\). CNAs and SNPs influenced expression variation, with CNAs dominating the landscape in cis and trans. The joint clustering of CNAs and gene expression profiles further resolves the considerable heterogeneity of the expression-only subgroups, and highlights a high-risk 11q13/14 cis-acting subgroup as well as several other strong cis-acting clusters and a genomically quiescent group. The reproducibility of subgroups with these molecular and clinical features in a validation cohort of 995 tumours suggests that by integrating multiple genomic
features it may be possible to derive more robust patient classifiers. We show here, for the first time, that subtype-specific trans-acting aberrations modulate concerted transcriptional changes, such as the TCR deletion-mediated adaptive immune response that characterizes the T cell program in the basal cancers.

The integrated CNA-expression landscape highlights a limited number of genomic regions that probably contain driver genes, including ZNF703, which we recently described as a luminal B specific driver, as well as somatic deletion events affecting key subunits of the PP2A holoenzyme complex and MTAP, which have previously been under-explored in breast cancer. The CNA-expression landscape also illuminates rare but potentially significant events, including IGFR1, KRAS and EGFR amplifications and CDKN2B, BRCAl, RB1, ATM, SMAD4, NCOAI, and UTX homozygous deletions. Although some of these events have low overall frequencies (<1% patients) (Figs 2, Supplementary Fig. 15 and Supplementary Tables 22–24), they may have implications for understanding therapeutic responses to targeted agents, particularly those targeting tyrosine kinases or phosphatases.

Finally, because the integrative subgroups occur at different frequencies in the overall population, focusing sequencing efforts on representative numbers from these groups will help to establish a comprehensive breast cancer somatic landscape at sequence-level resolution. For example, a significant number (~17%, n = 167 in the discovery cohort) of breast cancers are devoid of somatic CNAs, and are ripe for mutational profiling. Our work provides a definitive framework for understanding how gene copy number aberrations affect gene expression in breast cancer and reveals novel subgroups that should be the target of future investigation.

**METHODS SUMMARY**

All patient specimens were obtained with appropriate consent from the relevant institutional review board. DNA and RNA were isolated from samples and hybridized to the Affymetrix SNP 6.0 and Illumina HT-12 v3 platforms for genomic and transcriptional profiling, respectively. A detailed description of the experimental assays and analytical methods used to analyse these data are available in the Supplementary Information.

Received 24 April 2011; accepted 22 February 2012.

**Published online 18 April 2012.**

1. Leary, R. J. et al. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. Proc. Natl Acad. Sci. USA 105, 16224–16229 (2008).
2. Bignell, G. R. et al. Signatures of mutation and selection in the cancer genome. Nature 463, 893–898 (2010).
3. Perou, C. M. et al. Molecular portraits of human breast tumours. Nature 406, 747–752 (2000).
4. Sarile, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subtypes with clinical implications. Proc. Natl Acad. Sci. USA 98, 10869–10874 (2001).
5. Chin, K. et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiology. Cancer Cell 10, 529–541 (2006).
6. Chin, S. F. et al. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. Genome Biol. 8, R215 (2007).
7. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. 27, 1160–1167 (2009).
8. Young, J. L. et al. Genome-wide association of gene expression variation in humans. PLoS Genet. 1, e78 (2005).
9. Gilad, Y., Rifkin, S. A. & Pritchard, J. K. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 24, 408–415 (2008).
10. Teschendorff, A. E., Naderi, A., Barbosa-Morais, N. L. & Caldas, C. Profile analysis using clustering and kurtosis to find molecular classifiers in cancer. Bioinformatics 22, 2269–2275 (2006).
11. Holland, D. et al. ZNF703 is a common Luminal B breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium. EMBO Mol. Med. 3, 167–180 (2011).
12. Li, J. et al. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science 275, 1943–1947 (1997).
13. Santarius, T., Shiply, J., Brewer, D., Stratton, M. R. & Cooper, C. S. A census of amplified and overexpressed human cancer genes. Nature Rev. Cancer 10, 59–64 (2010).
14. Jones, S. et al. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. Science 330, 228–231 (2010).
15. McConkey, M. K. et al. Subtype-specific mutation of PPP2R1A in endometrial and ovarian carcinomas. J. Pathol. 223, 567–573 (2011).
16. Tan, J. et al. B55-associated PP2A complex controls PDK1-directed MYC signaling and modulates rapamycin sensitivity in colorectal cancer. Cancer Cell 18, 459–471 (2010).
17. Christopher, S. A., Diegelman, P., Porter, C. W. & Kruger, W. D. Methylthioadenosine phospholase, a gene frequently codelted with p16 (CDKN2A/ARF), acts as a tumor suppressor in a breast cancer cell line. Cancer Res. 62, 6639–6644 (2002).
18. Teng, D. H. et al. Human mitogen-activated protein kinase kinase 4 as a candidate tumor suppressor. Cancer Res. 57, 4177–4182 (1997).
19. Höllestelle, A. et al. Distinct gene mutation profiles among luminal-type and basal-type breast cancer cell lines. Breast Cancer Res. Treat. 121, 53–64 (2010).
20. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 25, 2906–2912 (2009).

21. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl Acad. Sci. USA 99, 6567–6572 (2002).

22. Kapp, A. V. & Tibshirani, R. Are clusters found in one dataset present in another dataset? Biostatistics 8, 3–31 (2007).

23. Hughes-Davies, L. et al. EMSY links the BRCA2 pathway to sporadic breast and ovarian cancer. Cell 115, 523–535 (2003).

24. Brown, L. A. et al. Amplification of 11q13 in ovarian carcinoma. Genes Chromosom. Cancer 47, 481–489 (2008).

25. Russnes, H. G. et al. Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. Sci. Transl. Med. 2, 38ra47 (2010).

26. 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).

27. Mahmoud, S. M. A. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl Acad. Sci. USA 99, 6567–6572 (2002).

28. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl Acad. Sci. USA 99, 6567–6572 (2002).

29. Hughes-Davies, L. et al. EMSY links the BRCA2 pathway to sporadic breast and ovarian cancer. Cell 115, 523–535 (2003).

30. Brown, L. A. et al. Amplification of 11q13 in ovarian carcinoma. Genes Chromosom. Cancer 47, 481–489 (2008).

31. Russnes, H. G. et al. Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. Sci. Transl. Med. 2, 38ra47 (2010).

32. 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).

33. Mahmoud, S. M. A. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl Acad. Sci. USA 99, 6567–6572 (2002).

34. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl Acad. Sci. USA 99, 6567–6572 (2002).

35. Hughes-Davies, L. et al. EMSY links the BRCA2 pathway to sporadic breast and ovarian cancer. Cell 115, 523–535 (2003).

36. Brown, L. A. et al. Amplification of 11q13 in ovarian carcinoma. Genes Chromosom. Cancer 47, 481–489 (2008).

37. Russnes, H. G. et al. Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. Sci. Transl. Med. 2, 38ra47 (2010).

38. 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).

39. Mahmoud, S. M. A. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl Acad. Sci. USA 99, 6567–6572 (2002).

40. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl Acad. Sci. USA 99, 6567–6572 (2002).

41. Hughes-Davies, L. et al. EMSY links the BRCA2 pathway to sporadic breast and ovarian cancer. Cell 115, 523–535 (2003).

42. Brown, L. A. et al. Amplification of 11q13 in ovarian carcinoma. Genes Chromosom. Cancer 47, 481–489 (2008).

43. Russnes, H. G. et al. Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. Sci. Transl. Med. 2, 38ra47 (2010).

44. 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).

45. Mahmoud, S. M. A. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl Acad. Sci. USA 99, 6567–6572 (2002).

46. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl Acad. Sci. USA 99, 6567–6572 (2002).