Viewpoint
A quantum leap in our knowledge of breast cancer mutations

Carlos Caldas[1] and Paul AW Edwards[2]

[1]Cancer Research UK Cambridge Research Institute and Department of Oncology, University of Cambridge, Hills Road, Cambridge CB2 0RE, UK
[2]Hutchison–MRC Research Centre and Department of Pathology, University of Cambridge, Hills Road, Cambridge CB2 2XZ, UK

Corresponding author: Carlos Caldas, cc234@cam.ac.uk

Last month our knowledge of genes mutated in breast cancer took a quantum leap, with the publication in the journal *Science* of the first results of a screen for mutations in the coding sequences of 13,000 genes in human breast cancer samples [1].

Although we believe that breast cancer is caused by alterations to genes, we have very limited knowledge of what genes get altered. It only takes a moment's thought to realise how big a limitation this is: with a catalogue of gene alterations we could classify breast tumours, probably predict their response to therapy, and design new targeted drugs. We would also begin to understand the cell biology of breast cancer – what distinguishes a benign from a malignant breast tumour, which signalling pathways are disturbed, and so on.

The screen
Sjöblom and colleagues [1] determined the sequence of all genes present on the consensus coding sequences database (CCDS) in two common human cancers, breast and colorectal. The work was divided into a discovery screen and a validation screen. In the discovery screen the sequences of exons and exon/intron boundaries of 14,661 transcripts corresponding to 13,023 genes in CCDS were determined in 11 breast cancer cell lines, 11 colorectal cancer xenografts and 2 normal samples. This represented a total of 456 megabases of sequence data, corresponding to 91% of targeted bases in CCDS. Sorting out this massive amount of data to discover 'real' somatic mutations and discard the remaining changes (which, besides some 'noise', represent a treasure trove of valuable data, for example of germline genetic variation) was no simple task, both experimentally and computationally. In total 816,986 putative nucleotide changes were found, which then had to be filtered; 557,029 of these were non-synonymous and were taken forward. The exclusion of false positive calls and changes present in either of the two normal controls or in single-nucleotide polymorphism databases

removed a further 96%. Resequencing removed 9,295 more. The 19,986 'real' nucleotide changes were sequenced in matched normal DNA from the patients, showing that 18,414 were present in the germline as unknown polymorphisms and leaving (after a further filtering step) a final tally of 1,307 confirmed somatic mutations in 1,149 genes. In the validation screen, the 1,149 genes were sequenced in a further 24 breast and 24 colorectal cancers with matched normal DNA, representing an additional 77 megabases of sequencing. With the use of similar criteria to those of the discovery screen, of 133,693 putative changes only 365 somatic mutations were confirmed in 236 genes. In summary, 9% (1,149) of the 13,023 genes sequenced in the discovery screen had somatic mutations; among these, further mutations were identified in 236 (21%) in the validation screen.

As in all such screens, a major issue is whether the mutations observed are selected for in the cancer or are merely accidental 'passenger' mutations. Different groups who have screened for mutations have approached (or ignored) this problem in different ways. The approach used by Sjöblom and colleagues [1] was to estimate the mutation frequency across the genome, allowing for gene size and different frequencies of different base changes, and so estimate the probability that a given gene is mutated more than the expected frequency. The result is a set of genes estimated to have more than 90% probability of undergoing selected mutation, and called candidate cancer genes ('CAN' genes) by the authors – 122 in the case of breast cancer.

Another way of working out whether mutations are passengers is to compare the observed and expected proportions of mutations that change protein sequence [2,3]. This strategy was not possible for the data set of Sjöblom and colleagues because the non-synonymous sequence changes found were not tested for in matching normal DNA. Such data are available on a screen of kinases in breast

---

CAN = candidate cancer; CCDS = consensus coding sequences database; ER = oestrogen receptor.

cancer, where the estimate was that about one-third of somatic mutations are selected [2]. This suggests that Sjöblom and colleagues were conservative in their calculations (122 mutations in CAN genes out of 921 somatic mutations), and that among the mutations in non-CAN genes there were a significant number of selected mutations.

The meaning of what they found
The screen seems to tell us several profoundly important things, of which we highlight the following:
1. A large number of mutated genes had not previously been implicated in breast or other cancers. None of these were mutated in a very high proportion of cases, but a number are estimated to be mutated in around 10 to 20% of (oestrogen receptor (ER)-negative) breast cancers.
2. A large number of mutations per tumour seem to be selected for. Extrapolating to all of the coding sequences in the human genome, the authors estimate an average of 20 CAN genes mutated per breast cancer. Although this will not be a surprise to everyone, some textbooks have got stuck with a much lower estimate of perhaps five to seven. This number is often said to be supported by age–incidence curves, but even Armitage and Doll [4], who pioneered this kind of analysis, pointed out that the five to seven estimate was valid only under rather restrictive assumptions, for example that the rate of the various events was constant over time.
3. There was a great deal of variation in the genes mutated between cases.
4. There was almost no overlap between genes mutated in breast and colon cancers, *OBSCN* (Obscurin) and *TP53/p53* being the only genes appearing in both lists of CAN genes, although some other genes, such as *AKAP6*, had mutations in both cancers but did not meet the criteria for CAN genes.
5. Remarkably few of the mutations were scored as homozygous. This may have reflected some technical bias, but note that in the colorectal samples all except 3 out of 24 samples with *APC* mutations were scored as homozygous or had two heterozygous mutations (presumably bialleleic). This suggests that there are many more dominant mutations than recessive (classic tumour suppressor) mutations, and that mutation plus loss of heterozygosity of tumour suppressor genes makes a very small contribution to the development of breast and colorectal cancer.
6. There was a very different mutation spectrum in colon and breast, suggesting that the mutagenic processes in the two tissues are different.

Putting findings in context
This marks a major step forward in the analysis of the gene changes that drive breast cancer, because it made no assumptions about what kinds of gene might be mutated. This moves us forward dramatically from earlier targeted screens.

What are its limitations?
Only 90% of about 13,000 genes have been examined so far; the authors estimate that this is roughly two-thirds of the whole human coding sequence (one obvious example of a gene missed is *PIK3CA*) and does not include non-coding regions and RNA transcripts including microRNAs.

The discovery screen examined only 11 breast cancer cell lines: all were ER negative, so if there are genes that are specifically mutated in ER-positive tumours – as seems probable – these will have been missed in this screen. Given the lack of overlap between breast and colon genes, this may well be the biggest limitation of the screen.

The way in which the screen was analysed may well have underestimated the number of 'real' mutations; that is, those that are selected for or that are pathogenic. The authors note that their discovery screen will have missed the rarer mutations – only 50% of genes mutated in 6% of tumours would have been found – and this is illustrated by the absence of mutations in *EP300*, for example [5]. They used the validation screen to provide evidence that a gene should be on the significant list of 236 genes, but even a cursory review of the genes found mutated in the discovery screen suggests that more 'real' cancer genes are present in the list of 1,149. For example, both *AKAP6* and *AKAP9* had mutations in both two breast and two colon cases, but did not meet the CAN gene criteria. *Baz1A* and *Baz1B*, two proteins with a bromodomain adjacent to a zinc-finger domain, were found mutated in the discovery screen, suggesting that this gene family is a target of cancer mutations. We also noted additional members of gene families with a member included as a breast CAN gene were found mutated in the discovery screen, for example *CENTD3*, *DNAH9*, *ITGB2*, *PRPF39*, *LRRC4* and *LRRC7*, *SEMA7A* and several solute carriers, suggesting that these genes might also be 'real' breast cancer genes mutated at low frequency.

Point mutation screens like this are also only part of what lies ahead. As the authors point out, the point-mutation screens cannot 'see' DNA rearrangements. Given the finding that 6% breast cancers have translocations of the *NRG1/heregulin* genes [6] and more than half of prostate cancers have inversions, deletions or translocations that fuse members of the *ETS* transcription factors to an androgen-sensitive transcript [7], this may be an important gap. And of course the search for, and understanding of, epigenetic change is only in its infancy.

What next?
We look forward to the screening of the remainder of the transcriptome. Clearly, we also need ER-positive breast cancer cases to be screened. We need more cases screened to assess the frequency of the mutations picked up so far, to improve the discrimination between passenger and

cancer-relevant mutations. Then we need to know whether these genes or others in the same pathways are targets for other kinds of change – epigenetic changes, deletions, amplifications, chromosome translocations and so on.

Further work is now necessary with the breast CAN genes already identified: for example, in most cases it is not clear what the mutations do to the genes' function. The genes can also be scanned for germline variation to determine their role in cancer predisposition and in prognostication.

Perhaps the most exciting immediate prospect is a correlation of mutation of these genes with molecular subtypes of breast cancer that have been defined by expression analysis.

Conclusion

This screen represents a great leap forward in our knowledge of breast cancer. It provides a long list of new cancer-relevant genes and provocative information that forces us to re-examine some of the assumptions we make about the numbers and types of mutation in breast cancer. It vindicates the large-scale unbiased screen as an approach to cancer. However, it is only the first step towards an understanding of the genome changes that drive breast tumorigenesis.

Competing interests

The authors declare that they have no competing interests.

References

1. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, *et al.*: **The consensus coding sequences of human breast and colorectal cancers.** *Science* 2006, **314:**268-274.
2. Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF: **Statistical analysis of pathogenicity of somatic mutations in cancer.** *Genetics* 2006, **173:**2187-2198.
3. Stephens P, Edkins S, Davies H, Greenman C, Cox C, Hunter C, Bignell G, Teague J, Smith R, Stevens C, *et al.*: **A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer.** *Nat Genet* 2005, **37:**590-592.
4. Armitage P, Doll R: **The age distribution of cancer and a multistage theory of carcinogenesis.** *Br J Cancer* 1954, **8:**1-12.
5. Gayther SA, Batly SJ, Linger L, Bannister A, Thorpe K, Chin SF, Daigo Y, Russell P, Wilson A, Sowter HM, *et al.*: **Mutations truncating the EP300 acetylase in human cancers.** *Nat Genet* 2000, **24:**300-303.
6. Huang HE, Chin SF, Ginestier C, Bardou VJ, Adelaide J, Iyer NG, Garcia MJ, Pole JC, Callgy GM, Hewitt SM, *et al.*: **A recurrent chromosome breakpoint in breast cancer at the NRG1/ neuregulin 1/heregulin gene.** *Cancer Res* 2004, **64:**6840-6844.
7. Tomlins SA, Rhodes DR, Perner S, Dhanasekran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, *et al.*: **Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.** *Science* 2005, **310:**644-648.