Driver Genes in Breast Cancer: Importance, Integration & Identification

Nicola Faramarzi, MSc,
PhD Student
Dept. of Biomedical Sciences
Faculty of Science and Technology
University of Westminster
London, UK.

Corresponding author:
Dr Nadège Presneau, PGCertHE, FHEA, PhD,
Senior Lecturer in Biomedical Sciences
Dept. of Biomedical Sciences
Faculty of Science and Technology
University of Westminster
London, UK
E: N.Presneau@westminster.ac.uk

It is well-known that breast cancer is highly heterogeneous, behaving not as a single disease, but as a collection of tumour subtypes with distinctive aetiologies, origins and genetic signatures [1]. This heterogeneity is mirrored in its complex genomic landscape, and despite advancements in subtype specific therapeutics, many patients have de novo or develop resistance to current therapies due to underlying genetic mutations which require further characterisation.

In the UK, breast cancer is the most common type of cancer, and incidence has increased by approximately 50% since the 1970s. However, due to increasing knowledge of underlying molecular aberrations and the development of subsequent treatments, survival rates have doubled since the 1970s (Figure 1) [2]. On a global scale, approximately 1.68 million women were diagnosed with invasive breast carcinoma worldwide and an estimated 522,000 cases of mortality from the disease in 2012 alone [2]. These statistics demonstrate the need for greater genomic characterisation and research that can be translated to patient-tumour specific molecular diagnosis in the near future.

Until recently, breast cancer was categorised into three subtypes with overlapping clinicopathological parameters – hormone receptor positive (HR+, consisting of estrogen receptors, ER, and progesterone receptors, PR), human epidermal growth factor positive (HER2+) and triple negative breast cancer (TNBC). Although several types of targeted treatments arose to combat these, such as tamoxifen and trastuzumab for HR+ and HER2+ cancers respectively, these were met with unexpected inefficiency in a large number of patients who were thought to possess the corresponding molecular markers to these treatments [3]. This classification has now been superseded by the 'intrinsic subtypes' taxonomy, based more on gene expression and molecular patterns of the tumours [1]. A basic summary of this is shown in Figure 2 [4].

At present, subtypes are diagnosed on the basis of immunohistochemical analysis of a tissue biopsy, or in situ hybridisation to detect single gene amplification [5]. However, it could be argued that this basic diagnosis using a panel of so few histopathological markers is not reflective of many patients' tumour types and treatment requirements. Hence there is a calling for more widely available clinical molecular diagnostic approaches to reveal patient specific genetic signatures that can be more efficiently targeted by treatment.

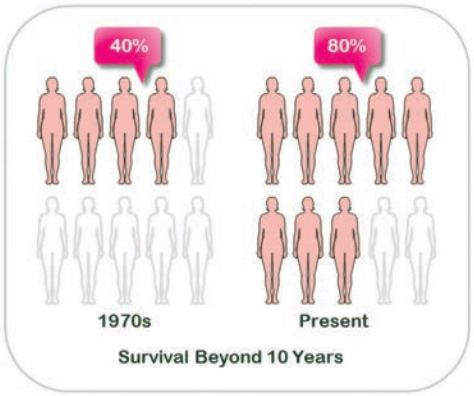

Figure 1: Cancer Research UK statistics illustrating that only 40% of patients survived beyond 10 years in the 1970s compared with 80% of patients today, showing that breast cancer survival beyond 10 years has doubled during this time.

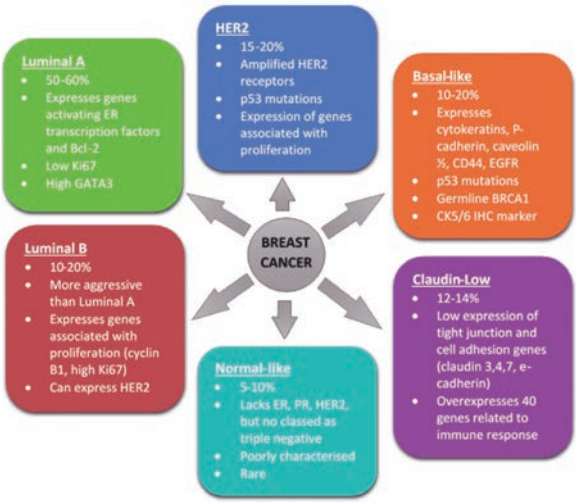

Figure 2: A summary of the intrinsic subtype taxonomy and current molecular markers, adapted from Eroles *et al.*, (2012).

Driver genes in cancer

There has been great progress in identifying many germ-line mutations, such as *BRCA1/2*, which now have the ability to detect susceptibility, predict prognosis and dictate patient stratification. However, the success of these genes is dampened by the fact that hereditary mutations are only responsible for up to approximately 5% of all breast cancers. Therefore,

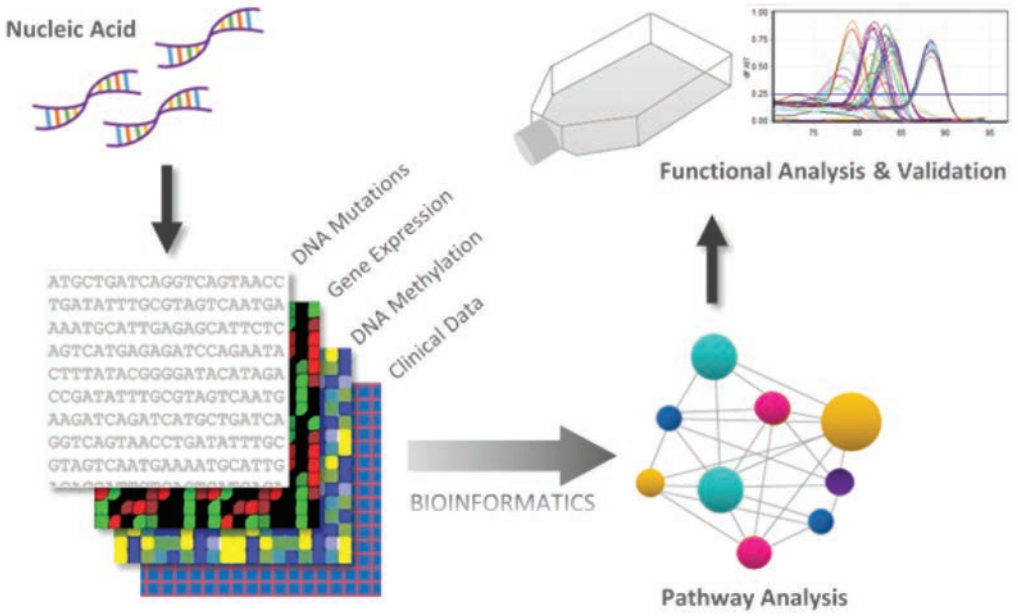

Figure 3: An example of a typical workflow for driver gene identification using an integrated and multi-platform data approach. Bioinformatics analysis can be extremely varied between studies and is dependent on available resources and the expertise of the researcher.

the remaining 95% of breast tumours that develop are instigated by an accumulation of somatic mutations [6].

Somatic mutations occur in all dividing cells due to exogenous (environmental factors such as radiation) or endogenous (faults in DNA replication) mutagens. These types of mutations are acquired and may be classed as a 'driver' or 'passenger' mutation [7]. Solid tumours can typically contain up to thousands of genetic aberrations and alterations, but only a handful of these are considered driver mutations [8]. Driver mutations provide cells a selective growth advantage (Darwinian evolution) and are considered to be positively selected in cancer cells; these alter critical cellular processes leading to the hallmarks of cancer [9]. In contrast, passenger mutations may arise within the cell, but do not give the cell any growth advantage [10] .

Vogelstein *et al.*, [11] have estimated that an average tumour contains two to eight driver gene mutations. These driver mutations are thought to only provide a small growth advantage to cells, which eventually build up over many years and result in billions of additional cells. Hence, it follows that the number of these somatic mutations is correlated to age. In this sense, sequential somatic mutations occurring during tumourigenesis can be thought of as an 'evolutionary clock' [11].

However, despite the exact number of driver gene mutations required for breast tumour initiation and progression being unknown, Tomasetti C et al have shown that for the development of lung and colon adenocarcinomas, only 3 mutations are needed. This has important implications for driver gene identification highlighting that although there is unlikely to be one single gene responsible, there may be only a small handful which can be taken advantage of for therapeutic targeting.

The importance of driver genes

Identifying key driver genes in breast cancer, as well as in other cancers, is pivotal in revealing crucial information regarding tumour biology, such as which pathways are disturbed during tumourigenesis. By identifying the genes responsible for driving and altering oncogenic signalling pathways, these can be further explored and may be used to gather information on individual tumours during diagnosis in order to enhance clinical decisions. Additionally these pathway drivers can be potentially targeted, or used to predict and tailor response to therapy [12].

The important implication that driver genes can be targeted for therapeutic development is supported by a study by Rubio-Perez *et al.* [13] highlighting that of 4000 tumour samples across 28 breast tumour types, only 6% of these were shown to be manageable using currently approved agents [9]. This highlights a need for greater characterisation of driver genes in cancers, particularly for breast cancer patients where current therapies may be ineffective; this is the case for many HER2+ cancers, where nearly half of patients show resistance or inefficiency to trastuzumab [14].

The current genomic landscape of breast cancer

It has been established that there are on average approximately 57 somatic mutations per breast cancer [15]. It is important to note that depending on laboratory methods, sample selection and data analysis, many studies identify different sets or signatures of somatic gene mutations [15]. However, there are still a small set of potential driver genes that are recurrently identified across studies, such as *TP53*, *PIK3CA*, *GATA3* and *MAP3KI*. Studies have shown that ER-positive tumours have less mutations than ER-negative tumours, which primarily affect *PIK3CA*. Of all of the intrinsic subtypes, HER2+ has been shown to have the highest mutation rates, with the most frequently mutated gene in HER2+ and the basal-like subtype being *TP53* [15].
Gatza *et al.*, [12] used gene expression microarray data and a panel of gene expression signatures to examine patterns of pathway activity to identify specific DNA amplifications and genes within these that represent key drivers. This study identified 8 genes (*FGD5*, *METTL6*, *CPT1A*, *DTX3*, *MRPS23*, *EIF2S2*, *EIF6* and *SLC2A10*) amplified only in patients with proliferative luminal breast cancers, a subtype with few therapeutic options. Liu *et al.* [16] have also identified candidate driver mutations in the luminal subtype, revealing mutations in *BRAF, GNAS, IDH1* and *KRAS*, by sequencing hotspot regions from cancer related genes.

A comprehensive genomic,

transcriptomic and proteomic analysis integrated with clinical data, by Michaut *et al.* [17] have confirmed that PI3K pathway mutations and *CDH1* inactivating mutations are most frequently altered in invasive lobular breast carcinoma. Other mutations in *ERBB2*, *MAP3K1*, and *MAP2K4* were revealed at low frequency. As can be seen from these studies, there is seldom complete agreement or overlap of identified driver genes across the different breast cancer subtypes; the variety of subtypes also make it difficult to obtain a generalised picture of the driver genes present in breast cancer. This further demonstrates the complexity of breast tumorigenesis and the challenge of identifying true driver genes, necessitating further investigation and characterisation.

This year, new research from a large team of international researchers led by Mike Stratton at the Sanger Institute in Cambridge has identified 93 protein-coding somatic mutations and potential driver genes, as well as 12 base substitution and 6 base rearrangement driver gene signatures [18]. This study analysed 560 breast cancers using whole genome sequencing and comprehensive molecular and bioinformatics analysis. The research is ground-breaking as evidenced by many news outlets reporting the story. It is the largest study that has worked towards identifying the vast majority of somatic mutations in breast cancer, and it has suggested that each breast tumour genome is individual. Although this leading study has been successful in gaining 'the bigger picture' of the mutations driving breast tumour development it is likely that infrequently mutated genes still require classification, and further analysis must aim to develop this list of mutations by investigating the functionality of these genes [18].

Driver gene identification & data integration

The advent and progression of sequencing technologies over the past few decades have revealed a vast amount of information regarding breast cancer, and provided researchers with an unprecedented ability to identify genetic alterations driving the oncogenic process. In the recent past, gene expression and DNA microarrays have steered the way for understanding the heterogeneity of breast oncogenesis, suggesting that the behaviour of an individual's cancer is based on the tumours' genetic profile and pattern of gene expression [1]. In addition, the emergence of high-throughput, next-generation sequencing (NGS) technology in the past decade has superseded traditional Sanger sequencing, and is evolving rapidly with widespread use across research and clinical laboratories for cancer surveillance, allowing researchers to sequence whole genomes in parallel.

With these technologies being increasingly used for driver gene investigation, a multi-level approach for integrating different types of OMIC data has risen to the forefront of research, with a large number of studies concerned with data-mining from public databases and coupling gene expression with genomic data [19] (Figure 3). An extensive and comprehensive example of a multi-level approach is the use of five data types by The Cancer Genome Atlas Network [20]. This study uses genomic DNA copy number arrays, exome sequencing, DNA methylation, mRNA arrays, microRNA sequencing and reverse phase protein arrays, and integrated data across these platforms to analyse primary breast cancers. This integrated analysis provided confirmation of previously known somatic mutations, as well as novel subtype associated mutations, *GATA3*, *PIK3CA* and *MAP3K1* in the luminal A subtype [20].

Although smaller studies and research groups are often limited in terms of their resources to perform such large scale and varied analysis, the trend of integrating multiple platforms can still be feasible on a smaller scale, for example by using only two or three platforms, smaller sample sizes, or by data mining from free public databases rather than performing all laboratory analysis in-house.

The few driver gene mutations present in cancers in comparison to passenger mutations makes it difficult to investigate the function of all mutations identified by sequencing. In light of this, bioinformatics analysis tools have been developed to predict the most likely driver genes and mutations, which can therefore be preferentially selected for functional analyses. There are two main approaches used in this instance, which either examine the mutation frequencies or aim to predict the functionality of the mutations [21]. A full description of these approaches are beyond the scope of this review, but are described in full by Pon and Marra [21]. Alternatively, systematic approaches can reveal groups of driver genes that are functionally related, or driver genes that are linked by a functional network or significantly enriched signalling pathway (Figure 3) [9].

The challenges of driver gene identification in breast cancer

The ongoing research in this field by academic and industrial laboratories worldwide emulate the challenge of driver gene identification. Despite sequencing and array technologies moving at a phenomenal pace, this is not matched by 'big data' handling techniques or user-friendly bioinformatics analyses. Furthermore, the number of bioinformatics algorithms, possible analysis pipelines and databases can appear very overwhelming for researchers with little bioinformatics experience. Aside from technical challenges, there are many factors that cause difficulties in driver gene identification. One important factor to consider is that somatic mutations rarely occur at greater than 10% prevalence, meaning that most driver genes have a much lower incidence and are mutated much more infrequently. This is due to the fact that there is such an enormous and assorted range of somatic mutations occurring in cancer cells, that the frequency of any identified driving mutations can be extremely low, even if they provide cells with a significant growth advantage [8,22]. This can be seen across all breast cancers where only somatic mutations in *TP53*, *PIK3CA* and *GATA3* appear at >10% incidence [20]. An additional complication is that driver genes can also contain mutations that are not driver mutations [11].

Despite breast tumours originating from the same mammary tissues, the different subtypes can be considered as molecularly different diseases with differing gene expression profiles and therapeutic responses, and it is now accepted that these subtypes do not exhibit identical sets of somatic mutations; it is unlikely that each subtype can be represented by a single driving gene. There is growing evidence suggesting that many primary breast tumours consist of several genetically distinct clones. This inter and intra-tumour heterogeneity has been demonstrated in approximately two-thirds of triple negative breast cancers, particularly in basal-like subtypes, and often means that potential driver somatic mutations are actually only seen in a

minority of tumour cells. This can also affect secondary tumours, with the majority of metastatic lesions varying significantly in their driver mutations compared to primary breast tumours. This presents a challenge to researchers, as driving mutations found in one part of a tumour may not be characteristic of the whole tumour. Although these complex factors pose challenges for driver gene identification in breast cancer, they may be useful for more long term development and implementation of targeted medicine; this would ensure that future therapeutics will be based more upon the molecular biology of individual tumours [15].

Final thoughts & future directions

There is an increasing need for greater characterisation of driver genes in the varying breast cancer subtypes to understand their mechanistic role in tumourigenesis and their respective pathways across the separate subtypes. These may be used in the future for therapeutic exploration or biomarker discovery, or to develop an extensive catalogue of driver genes for which treatment can be built around and used for tailored and personalised therapy; this aims to address the current inefficacy of existing breast cancer therapies.

A multi-platform approach for driver gene identification is a highly strategic, prolific and lucrative method for enhancing our knowledge of the molecular basis underpinning breast carcinogenesis, whilst providing attractive potential pharmaceutical targets. However, in order to exploit this information from sequencing studies, the future must focus on addressing the challenges associated with driver gene identification and 'big data', and developing more robust and user-friendly bioinformatics pipelines for processing large-scale OMIC data.

References

1. Perou, C and Børresen-Dale A. *Systems biology and genomics of breast cancer.* Cold Spring Harbor perspectives in biology, 2011;3(2):a003293.
2. Cancer Research UK (2015). *Breast Cancer Statistics.* Available from: http://publications.cancerresearchuk.org/cancerstats/statsbreast/kfbreast.html [Accessed 14 April 2016].
3. Rexer, B and Arteaga, C. *Optimal targeting of HER2-PI3K signaling in breast cancer: mechanistic insights and clinical implications.* Cancer research 2013;73(13):3817–20.
4. Eroles P et al. *Molecular biology in breast cancer: intrinsic subtypes and signaling pathways.* Cancer treatment reviews 2012;38(6):698–707.
5. Hansen, A and Bedard, P. *Clinical application of high-throughput genomic technologies for treatment selection in breast cancer.* Breast cancer research 2013;15(5):1-10
6. van der Groep P, van der Wall E and van Diest P. *Pathology of hereditary breast cancer.* Cellular Oncology 2011;34(2):71–88.
7. Stratton M, Campbell P, and Futreal P. *The cancer genome.* Nature 2009;458(7239):719-24.
8. Tomasetti C et al. *Only three driver gene mutations are required for the development of lung and colorectal cancers.* Proceedings of the National Academy of Sciences of the United States of America 2015;112(1):118–23.
9. Gonzalez-Perez A. *Circuits of cancer drivers revealed by convergent misregulation of transcription factor targets across tumor types.* Genome medicine 2016;8(1):1-12.
10. Greenman C et al. *Patterns of somatic mutation in human cancer genomes.* Nature 2007;446(7132):153–8.
11. Vogelstein B et al. *Cancer genome landscapes.* Science 2013;339(6127):1546–58.
12. Gatza M et al. *An integrated genomics approach identifies drivers of proliferation in luminal subtype human breast cancer.* Nature genetics 2014; 46(10):1051-9.
13. Rubio-Perez C et al. *In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities.* Cancer Cell 2015;27(3):382–96.
14. Esteva F et al. *PTEN, PIK3CA, p-AKT, and p-p70S6K status: association with trastuzumab response and survival in patients with HER2-positive metastatic breast cancer.* The American journal of pathology 2010;177(4):1647-56.
15. Ng C et al. *Breast cancer genomics from microarrays to massively parallel sequencing: paradigms and new insights.* Journal of the National Cancer Institute 2015;07(5):djv015–13.
16. Liu S et al. *Rapid detection of genetic mutations in individual breast cancer patients by next-generation DNA sequencing.* Human genomics 2015;9:2-11.
17. Michaut M et al. *Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer.* Scientific reports 2016;6:1-13.
18. Nik-Zainal S, et al. *Landscape of somatic mutations in 560 breast cancer whole genome sequences.* Nature 2016, 1-8.
19. Suo C et al. *Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival.* Bioinformatics 2015;31(16):2607–13.
20. The Cancer Genome Atlas Network. *Comprehensive molecular portraits of human breast tumours.* Nature 2012;490(7418):61–70.
21. Pon J and Marra M. *Driver and passenger mutations in cancer.* Annual review of pathology 2015;10:25–50.
22. Liu Y and Hu Z. *Identification of collaborative driver pathways in breast cancer.* BMC genomics 2014;15(1):605-21.