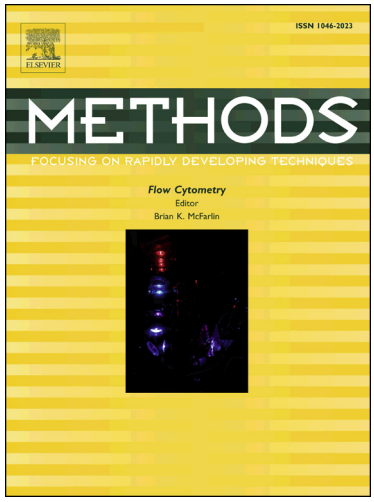


Molecular Signatures in Breast Cancer

Samir Lal, Amy E McCart Reed, Xavier M de Luca, Peter T Simpson

PII: S1046-2023(17)30058-0
DOI: http://dx.doi.org/10.1016/j.ymeth.2017.06.032
Reference: YMETH 4262

To appear in: *Methods*

Received Date: 17 May 2017
Revised Date: 26 June 2017
Accepted Date: 28 June 2017

Please cite this article as: S. Lal, A.E. McCart Reed, X.M. de Luca, P.T. Simpson, Molecular Signatures in Breast Cancer, *Methods* (2017), doi: http://dx.doi.org/10.1016/j.ymeth.2017.06.032

Molecular Signatures in Breast Cancer

Samir Lal[1], Amy E McCart Reed[1], Xavier M de Luca[1] and Peter T Simpson*[1]

The University of Queensland, Centre for Clinical Research, Faculty of Medicine, Herston, QLD 4029, Australia

*Corresponding author - p.simpson@uq.edu.au

Abstract

The use of molecular signatures to add value to standard clinical and pathological parameters has impacted clinical practice in many cancer types, but perhaps most notably in the breast cancer field. This is, in part, due to the considerable complexity of the disease at the clinical, morphological and molecular levels. The adoption of molecular profiling of DNA, RNA and protein continues to reveal important differences in the intrinsic biology between molecular subtypes and has begun to impact the way patients are managed. Several bioinformatic tools have been developed using DNA or RNA-based signatures to stratify the disease into biologically and/or clinically meaningful subgroups. Here, we review the approaches that have been used to develop gene expression signatures into currently available diagnostic assays (*e.g.* OncotypeDX® and Mammaprint®), plus we describe the latest work on genome sequencing, the methodologies used in the discovery process of mutational signatures, and the potential of these signatures to impact the clinic.

Keywords Signature, breast cancer, biomarker, prognostic, genomic test

Introduction

Breast cancer is an extremely diverse and complex disease, and is one of the leading causes of death amongst women. There is marked tumour heterogeneity between patients, with specific breast cancer subtypes associated with differing prognoses. Differentiating breast tumour types is a key component of the clinical management process to ensure patients are given the most appropriate type of therapy. In this review we briefly illustrate the best practices in tumour classification from a pathology context, including currently utilised predictive and prognostic biomarkers. We will then highlight the advances made in the molecular arena, which have shed light onto the differences in intrinsic biology between subtypes of the disease and how these have been developed into molecular signatures with clinical utility.

Pathological Classification of Disease

As part of the diagnostic process, a pathologist examines a tissue biopsy or resection specimen. A diagnosis will be made based on key parameters, which include histological type, tumour grade, and tumour stage using criteria outlined by the World Health Organisation (WHO) [1]. There are at least 20 different histological subtypes of breast cancer, which display differences in morphology and growth pattern. The most common is Invasive Carcinoma of No Special Type (IC-NST; previously called Invasive Ductal Carcinoma (IDC) accounting for 80% of all cases [1]). The remaining are classified as 'special' histological types in that they exhibit unique patterns of growth. Invasive Lobular Carcinoma (ILC) is the most common special type, accounting for between 5-15% of cases, with others including medullary, metaplastic, tubular and mucinous subtypes which all have distinctive growth patterns and variable prognoses.

Several diagnostic systems give insight into the behaviour of a tumour, including tumour grade and stage (Figure 1). *Histological grade* describes how abnormal the tumour appears relative to normal tissue, as a measure of tumour cell differentiation. Grading of breast cancer is performed using the Nottingham grading system [2, 3]. This is a three-tiered scoring system, assessing the number of visible mitoses, the presence of tumour cells creating tubule structures and evidence of nuclear

pleomorphism. The number of mitoses acts as a surrogate for growth rate, while tubule formation is a measure of whether the tumour tissue resembles normal-like ductal structures. Pleomorphism is a measure of the size, shape and variability of tumour nuclei. The prognostic value of the grading system in predicting behaviour and patient outcome has long been established [4, 5]. A histological grade 1, well-differentiated tumour is associated with a significantly better prognosis compared to a grade 3, poorly differentiated tumour.

Tumour Stage is a measure of how far the tumour has spread, and so is also a highly prognostic tool. The American Joint Committee on Cancer (AJCC) TNM staging system is used for most organ systems, including breast. T is a measure of the tumour size (<2cm, between 2-5cm and >5 cm) and whether the tumour has invaded the chest wall; N refers to the number of lymph nodes that show evidence of cancer (0, 1-3, 4-9, >10) and the position of the node in the nodal system; and M is a measure of distant metastasis, *i.e.* if there is a sign of cancer spread beyond the site of the primary tumour.

Clinical biomarkers in breast cancer

Biomarkers play important roles in diagnosis and prediction of prognosis, and may also represent therapeutic targets. Key breast cancer biomarkers include Oestrogen Receptor (ER), Progesterone Receptor (PR) and Human Epidermal Growth factor Receptor 2 (HER2/ERBB2); these markers have been reviewed extensively and their expression correlates with differences in tumour behaviour and patient outcome and the potential response to targeted endocrine therapy or HER2 therapy [6]. The protein expression levels of ER, PR and HER2 are assessed using immunohistochemistry and, in addition, the *ERBB2* gene copy number is also quantified using *in situ* hybridization [7]. If a breast cancer is positive for either ER or PR the breast cancer is termed as Hormone Receptor positive (HR+) and these patients will likely receive endocrine therapy, while patients with HER2+ breast cancers will receive trastuzumab or other HER2 targeted therapies. According to the Surveillance, Epidemiology, and End Results (SEER) survey that covers 28% of the North American population, 72.7% of breast cancer patients had HR+/HER2- disease, 14.9% had HER2+ disease (of which, 10.3% were HR+/HER2+ and 4.6% were HR-/HER2+) and 12.2% had

triple-negative disease [8]. As the name suggests, triple-negative breast cancer (TNBC) encompasses all tumours that are negative for ER, PR and HER2. The majority of TNBC are high-grade and aggressive tumours associated with a poorer outcome than other breast cancer subtypes, despite a good response to chemotherapy [9].

Ki67 is a proliferative marker used to predict tumour growth rate, which has been shown to be a prognostic biomarker with predictive ability in the adjuvant context [10-14]. Ki-67 staining has not yet been fully translated to the clinical setting owing to difficulties in standardising technical aspects of the procedure, including appropriate scoring methods and thresholds [15, 16].

A combination of the 4 IHC based biomarkers (ER, PR, HER2 and Ki67) exists as a protein-based 'signature'. This panel was termed IHC4 and the algorithm has been validated as a predictor of risk of distant recurrence in breast cancer [17, 18]. Although the current well-established clinical variables mentioned above show associations with prognosis and outcome, there are increasing concerns that these variables are limited in their ability to capture the diversity of breast cancer and tailor the therapy to individual patients.

Various methods have been developed to help improve the management of breast cancer patients based on the combination of this clinical, pathological and biomarker data that is collected at diagnosis. Various tools have been developed (*e.g.* Adjuvant! Online, Predict, the Nottingham Prognostic Index (NPI)) to help inform clinicians about their patient's potential prognosis and whether they are likely to benefit from adjuvant therapy following breast cancer surgery. These mathematical algorithms incorporate various clinical and pathological variables described above (patient age, tumour size, grade and lymph node status) together with tumour expression of these molecular biomarkers (ER, HER2 and Ki67) to predict survival with or without adjuvant therapy. The NPI is a highly prognostic tool based on combining pathological parameters of tumour size, tumour grade and the number of axillary lymph nodes that are involved with tumour. The NPI Plus was recently developed to incorporate a panel of 10 immunohistochemical biomarkers of demonstrated prognostic significance [19]. This modification identifies prognostic subgroups within

the well-described molecular subtypes of disease (ER+, HER2+ or triple negative) and so may provide additional benefit to managing the individual patient.

Developing mRNA based gene classifiers

Gene expression profiling of breast cancer has illustrated that the disease exhibits significant molecular heterogeneity, even among tumours with the same morphological features. This has led to the development of classification tools to stratify tumours into diverse molecular subtypes that have clinical relevance. To briefly illustrate the power of this approach, the early analysis of just 84 breast tumours by hierarchical clustering using genes whose expression levels differed from the median revealed several molecular subtypes of disease, including the Luminal A, Luminal B, Her2-enriched and Basal-like, as well as a normal-like group [20]. Each of these subtypes is clinically important as they exhibit differences in incidence, prognosis and response to therapies [21]. For example ER+ breast tumours are a heterogeneous group, which can be stratified, in simple terms, into Luminal A and B subtypes. The luminal A subtype has the best prognosis, whereas the Luminal B subtype collectively describes an aggressive group of tumours with a higher proliferative index indicated by Ki67 or Aurora A kinase (AURKA) staining. The amplification and overexpression of HER2 contributes to this aggressive behaviour in some Luminal B tumours [22] along with the activation of a plethora of other oncogenes found in regions of recurrent gene amplification (e.g. at 8p11-12 and 11q13) [23]. This work has therefore had a significant impact on the way breast cancer is described, researched and managed clinically. The classification of these 'intrinsic' subtypes has evolved since this first description, and they can now be defined by a 50 gene quantitative measurement known as the PAM50 subtype classifier [21], which is commercially packaged as a diagnostic classification tool called Prosigna® [24], for use in clinical practice (see below).

Gene signatures are sets of genes, or meta-genes, that together have predictive power to differentiate tumours based on broader biology; and so in this sense they are a form of biomarker. The process of identifying gene signatures can be summarised into a few key steps, illustrated in Figure 2, and with specific examples given below. Each

step requires a unique set of skills within a multidisciplinary team. The first step is signature discovery and involves acquiring a retrospective tumour cohort reflecting the clinical question in mind, together with detailed clinical history data on the patients. The tumour samples within the cohort would then be subject to molecular profiling by gene expression arrays, historically, or by whole transcriptome sequencing (RNAseq). A class comparison or regression analysis would be performed to identify an initial set of genes that best discriminate between a phenotype or an end point of interest, such as survival. Various filtering steps such as cross-validation would then be applied to arrive at a minimal set of genes predictive of the end-point. A scoring algorithm, which may include clinical features such as the tumour size or the number of lymph nodes involved, would be developed to assign thresholds or scores to classify the tumours into groups. The analysis of a validation cohort, together with permutation-based testing would be used to evaluate the robustness of the gene signature. Such statistical tools are user-friendly and allow the comparison of the performance of the signature against random gene sets and known gene signatures. Individual biomarkers from the signature might be studied in a candidate gene approach, for functional implications in the disease or as a surrogate biomarker in clinical samples. For clinical implementation, a prospective evaluation of the biomarker or gene signature would be required to determine the robustness of the signature in identifying risk groups with distinct clinical outcomes.

A single sample predictor (SSP) is a strategy that has been employed in order to best classify an individual tumour into one of several predefined molecular subtypes, such as those mentioned above. The idea being that an individual tumour can be classified based on how similar its gene expression profile associates with the molecular based centroids of these subtypes, and hence the approach should not need the simultaneous profiling of a group of cases for classification. An SSP would therefore have clinical utility, similar to other simple biomarkers currently used in diagnostic practice, to classify tumours into clinically meaningful groups (*e.g.* with distinct response to therapy, risk of relapse and/or outcome). In practice, SSPs have some limitations, including low concordance rates on sample assignment across several studies [25] [26]. In particular a lack of concordance was observed amongst the assignment of tumours into Luminal A, Luminal B and HER2 subtypes. These findings were further supported by those of a meta-analysis from a diverse set of different microarrays

which showed a limited concordance of SSPs in 22 uniformly pre-processed datasets in 4000 hybridizations [27]. Based on these findings, the reliability of single sample predictors has been brought into question, which could be based on the bioinformatics algorithms used for classifications [28] (Figure 3) and/or be due to the biological nature of disease, which in the case of luminal/HER2 tumours is a spectrum of disease rather than being discrete entities that are easily stratified.

An alternative approach is the development of subtype classification models (SCM), which have an advantage over SSPs in their versatility. An SCM can be applied to different microarray platforms and a variety of normalisation methods [29-32]. To illustrate this concept, a subtype classification model was developed using a mixture of three Gaussian distributions representing three important breast cancer phenotypes (ER, HER2 amplification signalling and cell proliferation) in a two-dimensional space [29, 32] (Figure 3). The model consisted of co-expression gene modules representing these phenotypes, each consisting of genes associated with the key biological process. These ER, HER2 and proliferation co-expression gene modules were derived from a meta-analysis of 2,833 breast tumours, using a multiple regression approach with a Gaussian error model. For example, the ER gene module consisted of a set of 288 genes with correlated expression to *ESR1*; similarly there were 20 genes in the HER2 module and there were 355 genes in the proliferation module [29]. These modules were refined to a single three-gene subtype classification module using the 3 original prototype genes (*ESR1*, *ERBB2* and *AURKA*) in 5715 tumours [32]. In a comparative analysis in this large dataset, three different SCM classifiers, including the three gene SCM, were compared to three SSPs for molecular subtypes classification; the SCMs were shown to be more robust than SSPs at subtype classification and provided similar prognostic value [32]. Strong correlations of the raw prediction scores were observed in tumours profiled with different technologies (*i.e.* Affymetrix microarray and Illumina RNA-seq) [33], confirming that the subtype classifiers (SSPs and SCMs) can be transferred from microarray to RNA-seq technologies.

An analyst wishing to define a prognostic gene signature must also be aware of the fact that genes can be associated with outcome purely by chance. Enlightening research by Venet *et al.* revealed that randomly selected gene sets may be associated with breast cancer outcome, to some degree [34]. It is necessary, therefore, for robust checks to be implemented to ensure that a gene signature is unique in its ability to

prognosticate compared to other randomly chosen genes that are prognostic by chance. The package *Sigcheck* has been developed to deal with this particular issue [35] (Table 2) and compares the performance of an identified gene signature against the performance of known and random signatures through permutation testing.

Commercially available mRNA-based diagnostic tests

There are several commercially available mRNA-based diagnostic tests for prognostication and prediction in breast cancer. These gene expression panel assays are risk predictors and examples include: MapQuantDX™ (Genomic Grade Index, GGI) [36], ProSigna® [24] [37], Mammaprint® [38], OncotypeDX® [39] and EndoPredict® [40] (Table 1). Each predictor was derived in a different manner and using a unique algorithm, but collectively they classify the risk of recurrence of breast cancer, and hence provide insight into whether a patient might require adjuvant chemotherapy or not. From a bioinformatic point of view, the original implementations of the mRNA signatures are available in the R programming environment through the package *genefu* [41]. This is therefore a particularly useful package, the 'swiss army knife' of gene expression signatures and can be added to any bioinformatics analysis pipeline (Table 1) to enable *in silico* comparison between these established signatures and novel metagene sets using publically available gene expression data.

The Genomic Grade Index (GGI; MapQuantDX™) is an RT-PCR-based assay that can be applied to formalin fixed, paraffin embedded (FFPE) tissues and is performed by a centralised laboratory to predict the risk of distant recurrence in ER-positive breast tumours. The GGI was first developed through testing the hypothesis that gene expression profiling could add value to conventional histological grading. Grade 1 and 3 tumours exhibit low and high risks of recurrence, respectively, and hence patients can be managed with some confidence. However, patients diagnosed with grade 2 disease are more challenging to manage and so the intent with the GGI was specifically in stratifying grade 2 tumours into good (grade 1-like; low risk of recurrence) and poor (grade 3-like; high risk of recurrence) prognostic groups [36]. This tool was derived from a total set of 64 ER positive tumour tissue samples and a comparison was performed between grade 1 and grade 3 tumours using a differential

gene expression analysis. The GGI score was determined through the difference of the log-transformed gene expression values of grade 3 and grade 1 associated genes. The tumour cohort was scaled and normalised further using offset parameters finally revealing a total of 97 genes [36], which were validated in an independent set of 125 patients as being able to stratify grade 2 tumours into grade 1-like and grade 3-like. A cut-off point of the GGI score was determined rather than utilising GGI as a continuous score, whereby high risk was defined as (GGI ≥0) and the low-risk group was defined as (GGI<0). The GGI was investigated alongside other signatures in a large cohort of 2833 patients, highlighting the importance of proliferation-associated genes in breast cancer prognostication [29]. Subsequent evaluation in various clinical cohorts demonstrated the utility of this signature, relative to other measures of a tumours proliferative capacity (Ki67 staining, histological grade, mitotic activity) for predicting disease free survival [42-47].

ProSigna® is the commercial iteration of the PAM50 subtyping tool, and generates a risk of recurrence score primarily to predict chemotherapy benefit [21, 24, 37, 48]. ProSigna® is applied using the nanoString nCounter Analysis System, to obtain digitally quantified gene expression levels, even for archival (FFPE) tumour material. A prognostic risk model was developed by applying the PAM50 subtype predictor in combination with various clinical-pathological parameters (tumour size and grade), to a series of published breast cancer datasets, including the 141 cases of node negative, untreated disease from the NKI cohort that was previously used to develop Mammaprint [21, 38]. The risk of recurrence (ROR) score (combination of PAM50 plus clinical parameters) outperformed the use of clinical parameters alone or molecular subtype classification alone for predicting outcome [21]. The PAM50 assay was validated in the ABCSG-8 trial as a prognostic tool for distant recurrence-free survival (DRFS) [49]. This trial consisted of a set of 1478 patients with postmenopausal, early stage ER positive breast cancer who received endocrine therapy [49]. All examples of the 4 intrinsic subtypes were present in this cohort, as were all ProSigna risk groups (*i.e.* low, intermediate and high). Patients had a probability of 10 years DFRS of 96.7% when classified as low risk, 91.3% for intermediate risk and 79.9% for high risk [49], showing the clinical value of this gene signature. Importantly, ProSigna is suitable for use in core needle biopsies in the

neoadjuvant chemotherapy setting [48] and has also revealed an accurate estimation of the risk of distant recurrence in estrogen receptor-positive (ER+), node-negative patients treated with five years of adjuvant tamoxifen, [24].

MammaPrint® is a microarray-based 70-gene assay that is performed by a centralised service on fresh frozen tumour material. The set of 70 genes were originally derived using 98 patients, 34 of which developed metastatic disease and 44 were metastasis free after 5 years follow up [38]. The neat hypothesis underpinning its development was that the expression of a unique set of genes could distinguish between a good and poor prognosis. The transcriptome was profiled and a filtered set of approximately 5000 genes were obtained based on their ≥2.5 fold differential expression between the 'metastasis' and 'no metastasis' groups at a statistical significance of $p<0.01$ [38]. A supervised classification model was then generated after extensive feature selection [38], as outlined. The first step involved correlating gene expression of each gene with each prognostic category (metastasis *vs.* non-metastasis). Significant random gene correlations were accounted for using 10,000 Monte Carlo simulations, identifying 231 genes with a correlation coefficient >0.3 and less than -0.3. "Leave one out" cross validation was applied and the left out sample was correlated with a good or poor prognosis template. This template is the average gene expression value in each prognostic group. Iteratively, five markers would be added to the classifier from the top of the list of 231 candidates (based on correlation coefficient) until 231 markers were used. The optimal performance of the classifier was reached using 70 genes and thus defining the MammaPrint set to predict poor prognosis. The MammaPrint recurrence score is based on the correlation coefficient of each sample with the good prognosis template. The TRANSBIG consortium and member countries of the Breast International Group (BIG) have validated the power of MammaPrint in distinguishing between low-risk patients and patients with a high risk of distant recurrence and survival [50]. The Microarray In Node negative and 1 to 3 positive lymph node Disease may Avoid ChemoTherapy (MINDACT) trial [51] was performed to provide evidence from a large-scale, prospective, randomised controlled phase III international clinical trial for the value of integrating MammaPrint into clinical practice. Interestingly, the study reported that in 6693 patients, 23.2% were classified as high clinical risk and low genomic risk. Based on the clinical risk factors these patients would ordinarily have been given chemotherapy, yet, since patients

were classified as low-risk based on the gene signature, they were not given chemotherapy. The rate of 5-year survival was 94.7% in this clinically and genomic discordant risk group [51], suggesting that using the genomic risk predictor to identify low-risk patients might be clinically useful in recognising a target population that will have a good prognosis without chemotherapy, despite having contrary clinical indicators of high-risk disease.

OncotypeDX® is a 21-gene panel that was derived using a unique approach; a high throughput Real-time PCR (RT-PCR) approach was used to quantify the expression of a targeted panel of 250 candidate genes identified through a combination of literature analysis, genomic-focused databases, pathway analysis and from the analysis of publically available microarray gene expression profiling data [39]. The expression level of these genes was examined in three independent cohorts [52-54] of 447 patients, spanning different tumour and treatment types. This approach was used to refine the 250 genes down to a powerful prognostic metagene, rationalised by considering that genes robustly associated with recurrence would need to be present in all three cohorts. Nine genes displayed highly correlated expression with recurrence across all cohorts ($P<0.05$), five genes displayed correlated expression with outcome in all cohorts ($P<0.01$), and the expression of an additional nine genes correlated with outcome in two studies ($P<0.05$). A final set of 16 genes was selected based on the technical performance of the genes in the RT-PCR assay. This 21-gene assay (16 cancer genes plus 5 reference genes) constitutes the OncotypeDX assay, and encompasses genes involved in oestrogen signalling (*ESR1, PGR, BCL2, SCUBE2*), HER2 (*GRB7* and *ERBB2*) and proliferation (*MKI67, AURKA, BIRC5, CCNB1* and *MYBL2*). OncotypeDX was initially tested in 668 patients from the tamoxifen-treated NSABP B14 trial; where the recurrence score was significantly associated with both relapse free- and overall-survival and performed better than most clinical-pathological parameters at predicting risk of distant recurrence. The clinical utility of OncotypeDX has been evaluated in numerous retrospective clinical trial cohorts [43] and was recently evaluated in the TAILORx (Trial Assigning IndividuaLised Options for tReatment) [55] and RxPONDER (Rx for POsitive NoDe, Endocrine Responsive breast cancer) [56] phase III prospective clinical trials. Both trials involved patients with ER+ HER2- breast cancer and either lymph node negative disease or lymph node positive disease, respectively and investigated whether patients with low-intermediate

risk of recurrence scores by OncotypeDX could be treated with endocrine therapy alone or also required chemotherapy. While data from RxPONDER trail matures, the results from TAILORx demonstrated that patients with a recurrence score between 0 and 10 given endocrine therapy alone had a 5-year disease free survival rate of 93.8% and overall survival at 5 years of 98% [57]. Similar to MammaPrint, OncotypeDX therefore identifies a low-risk group of patients who have favourable outcomes, with little need for chemotherapy.

EndoPredict is another RNA-based multigene panel that was developed from the analysis of ER+ HER2- patients treated with adjuvant endocrine therapy [40]. The development of this signature employed a multistep-step approach beginning with the analysis of a training cohort of 253 primary tumours by Affymetrix gene expression arrays. The gene list was selected by prioritizing probes with a dynamic range (defined as the ratio between the 90^{th} and 10^{th} percentile) of expression of ≥ 2, plus those with published evidence of being prognostic in breast cancer. Cox regression analyses were conducted using the time to distant recurrence as the endpoint. A total of 104 candidate prognostic genes, based on their ranked p-value in the model and various analytical parameters, were tested by RT-PCR, using FFPE tumour tissue. A refined set of 63 genes had high performing primer-probe pairs, and a linear relationship between RT-PCR and logged (base 2) Affymetrix probe hybridisation values. The algorithm was trained using distance recurrence as the end point and a leave one out cross validation approach, which identified 8 genes that constitute the EndoPredict signature: three proliferation associated genes and five genes associated with ER-signalling/differentiation: *BIRC5*, *UBE2C*, *DHCR7*, *RBBP8*, *IL6ST*, *AZGP1*, *MGP*, and *STC2* (plus an additional three reference genes) [40]. From a bioinformatics perspective, the algorithm involved a committee predictor consisting of a fixed number of members, where each member is a linear combination of a few input variables. A committee predictor was used, as these low dimensional linear combinations tend to yield more powerful and robust prediction algorithms. The signature was evaluated in two large randomised phase III trials involving patients with ER+ HER2- disease treated with tamoxifen. Endopredict (EP) generated a risk score that was shown to be an independent predictor of subsequent risk of distant recurrence at 5 or 10 years. The power of this gene set was further improved with the integration of clinical parameters (tumour size and nodal status) into the diagnostic

algorithm (EPClin) [40]. The prognostic value of EP and EPClin were independently verified using the TransATAC trial (the Translational substudy of the Arimidex, Tamoxifen, Alone or in Combination trial (ATAC)), where they outperformed OncotypeDX in predicting late distant recurrence after endocrine therapy, though some of this benefit came from the integration of clinical parameters into the gene signature [58].

This latter point is an important factor in the evaluation of these gene expression-based molecular signatures in the clinic. Although they were not all developed with the same intention, each of the signatures discussed describe a commonality in predicting risk of recurrence in the setting of ER positive disease. They each appear useful tools for identifying a very good (low risk) prognostic group that could be spared the toxicity of chemotherapy. Many of these patients would have been classified low risk based on clinical and pathological features alone. However, one particular benefit of the molecular signatures over standard clinical-pathological parameters appears to be in identifying molecular low risk patients that were classified as high risk based on clinical parameters. These patients will also benefit from being spared the toxicity of chemotherapy, without affecting their prognosis.

From a clinical point of view these are important advances, however given the overlap in clinical utility across the spectrum of signatures, it begs the question of which gene signature is the most appropriate or most robust or gives most benefit? Studies using the TransATAC trial, and others, have made useful progress in retrospectively comparing signatures (including the IHC4, 21-gene Oncotype DX, PAM50/Prosigna, EP/EPclin) within the controlled environment of a clinical trial [43, 58, 59]. The OPTIMA trial (The Optimal Personalised Treatment of early breast cancer using Multiparameter Analysis) was designed to prospectively compare multiple molecular signatures to identify patients who would most benefit from chemotherapy [60]. These comparative studies illustrate that the molecular tests provide broadly comparative prognostic information and hence similar risk stratification at the population level, and that the integration of clinical and pathological parameters into the molecular signature provides increased power in prognostication. However, importantly, at the patient level there remains discrepancy in classifying individual patients into different risk categories depending on the test used, which is presumably related to the different methods and clinical scenarios used

to develop the signatures in the first place, and the different panels of genes being analysed.

DNA Signatures in Breast Cancer

All cancers are characterised by the acquisition of somatic mutations to the tumour genome, including base substitutions, insertions and deletions, copy-number changes and structural rearrangements. The diversity of this "genomic landscape" among individual tumours is considerable. Some alterations directly affect key cancer genes (driver events) that confer selective growth advantage upon tumour cells, and may be useful therapeutic targets (*e.g.* HER2 amplification). Other mutations (passenger events) accumulate over the lifetime of the cell lineage and although individually and collectively they may confer little selective advantage to the tumour cell, their frequency distribution can serve as a tell-tale imprint of the underlying mutational processes that contributed to tumour growth [61, 62]. It is therefore quite rational to consider DNA mutation profiles as a diagnostic tool or signature to identify clinically useful subgroups of disease based on aetiology, behaviour or outcome.

Genome-wide DNA copy number alterations are derived from array-based Comparative Genomic Hybridization (aCGH) or single nucleotide polymorphism (SNP) arrays. Some copy number alterations of key genomic regions are highly prognostic [63], whilst others correlate closely with defects in specific DNA repair pathways, such as deficiency in homologous recombination (HR) DNA repair due to germline/somatic loss of *BRCA1* or *BRCA2* [64-71] see below for clinical utility of these types of signatures).

A range of copy number-based classifiers have been determined for breast cancer [72], including using variants of random forest [73], logistic regression, logistic group lasso[74], fused support vector machine (SVM) [75] and supervised and unsupervised feature clustering (FC) using silhouette methods for estimating clusters [76]. Both unsupervised and supervised FC performed the best at classifying aCGH data from tumour samples, as a result of their ability to remove unwanted correlation bias [72]. Compared to other methods tested this approach allowed more accurate features to be selected and as a result, greater accuracy could be achieved [72]. This highlights the

importance of trialling different models and having key metrics in place to evaluate the model.

Several types of massively parallel sequencing strategies are used to identify nucleotide level mutations in the cancer genome. Targeted gene panel sequencing and whole exome sequencing (WES) have been most widely applied to characterise the mutations in the coding portion of the genome. While whole-genome sequencing (WGS) gives comprehensive, unbiased access to the complete repertoire of somatic alterations in the genome, including driver events, but also those that are considered passenger events. As mentioned above, the pattern and distribution of these passenger mutations (base substitutions) can be used to classify mutational signatures, which relate to the underlying aetiology of the tumour. There are now 30 mutational signatures reported by COSMIC [77]. Whilst many remain of unknown cause, several known carcinogenic or defective cellular processes produce characteristic mutation signatures. For example, those of lung (C:G > A:T transversions) and skin cancer (C:G > T:A and CC:GG > TT:AA transversions) are caused by exposure to exogenous mutagens in tobacco, and by ultraviolet light, respectively. Other processes that lead to the accumulation of characteristic sets of somatic mutations include enzyme modification (*e.g.* enhanced activation of the DNA cytidine deaminase APOBEC3B [43, 78]), infidelity of the DNA replication machinery, or failure of DNA repair pathways (*e.g.* mismatch repair or base excision repair). Mutation signatures associated with the inactivation of *BRCA1* or *BRCA2* are most common in breast cancer (and ovarian and pancreatic cancers) and indicate tumours with a deficiency in HR- double-strand break repair (HR-DSB) [61, 79, 80].

In addition to the substitution signatures, patterns of structural rearrangements can give insight into tumour aetiology. The largest study of breast cancer whole genomes (560 genomes, [79]) defined six re-arrangement signatures based on the type (deletions, tandem duplications, inversions and translocations), and size (1 kilobase to >1 Mb) of re-arrangement, and the extent of the alteration throughout the genome (clustered or dispersed) (Figure 4). Interestingly, three of the signatures were associated with *BRCA1/2* deficiency, suggesting defective HR DNA repair leads to very prominent patterns of DNA mutations, both at the nucleotide and genome structural levels.

Computational methods for classifying mutation signatures

Deciphering the signatures of mutational processes in cancer genomes is a blind source separation problem [81]. This problem involves unravelling hidden signals in a mixture of various signals, without knowing the mixing that was performed. The intrinsic non-negative nature of the blind source separation problem requires an algorithm that assumes non-negativity of the original source of the signal [81]. The primary algorithms for identifying mutational signatures are non-negative matrix factorization (NMF) and variants, for instance BayesNMF. There is a two pronged approached to NMF; the first approach is to determine the signatures that best account for mutational observations. This approach is achieved through mathematical optimization implemented to identify factors at a fixed rank of the actual number of signatures and a chosen norm. The second approach is to determine the actual number of signatures and this is achieved by further factorising the same data and applying a ranking system for different numbers of signatures and the appropriate rank is determined by identifying the clustering properties of the factors obtained from the original algorithm [82]. The NMF signature discovery approach is now available in the R package *somaticsignatures* [83], which includes visualisation tools (Table 2). The Bayesian non-negative matrix (BayesNMF) variant is available here (http://archive.broadinstitute.org/cancer/cga/msp). BayesNMF offers analytical utility, as you do not have to initially define the number of signatures. An optimal number of signatures is determined directly from the data, striking the balance between data fidelity and the model complexity [84-86]. Simple model-based approaches of signature identification and visualisation are also available (probabilistic mutation signature - *pmsignature*), and differ from the usual NMF approach [87] (Table 2). Mutational patterns are often derived from the standard 6 possible substitution patterns (C>A, C>G, C>T, T>A, T>C, T>G), however DNA sequence context of the substitutions is not taken into account. *pmsignature* includes the 5' and 3' flanking bases of the mutated base. This results in 96 patterns, and with strand information (plus or minus) further extends the number of possible patterns to 192. This approach reduces the number of parameters per signature by dividing each mutational pattern into features (*i.e.* substitution type, 3' base, 5' base) [87]. Each mutational signature is characterized by the probability distribution for each feature. The decomposition is

then analysed using a probabilistic model, which assumes each feature is independent and significantly reduces the number of parameters to 18 per signature, unlike the 3071 parameters per signature employed by other methods [87].

Both *somaticsignature* and *pmsignature* have their advantages and disadvantages (Table 2) in yielding mutational signatures that provide further insight into the mechanistic underpinnings of these mutational processes. However, the use of these tools is not yet widespread due to the cost and challenges associated with processing whole genome sequencing data. There is certainly precedence for these tools to be evaluated against each other, and for new tools that are more user-friendly to be developed. An accurate mutational signature relies heavily on the accuracy of the variant callers; consequently, the robustness of these mutational signatures needs to be evaluated further in a hypothesis-driven fashion.

There are several limitations that need to be addressed in order to derive a full spectrum of mutational signatures. These limitations include the sample size, since at least 200 whole genome-sequencing samples were required to determine 20 mutational signatures [81]. The second limitation is sequencing coverage, as exome sequencing covers approximately 1% of the human genome so fewer mutations are identified and effectively a large number of samples are required to determine the majority of mutational processes [88]. In order to address this particular issue *deconstructSigs* was developed [88]. This tool reconstructs the mutational profile within a single tumour sample rather than relying on a set of tumour samples. A multiple linear regression model is implemented using the fraction of mutations found in each of the 96-trinucleotide contexts in each tumour as input. The approach determines the weights that will best recreate the input data. Signatures are excluded if a single trinucleotide context comprises more than 20% of the signature, which is not present in the input data [88]. The reason for this exclusion is that some signatures are characterised largely by specific trinucleotide contexts and if these contexts are absent in the input data then the signature is unlikely to be active [88]. An initial mutational signature is chosen from the remaining signatures that best encompasses the mutational profile in a single tumour. A forward selection process then determines the optimal weight for each signature based on the contribution of the signature to the reconstructed profile. The weight corresponding to that minimizes the sum squared error between the tumour sample and the reconstructed profile is subsequently

selected. The weights are normalized to 0 and 1 and a reconstructed tumour profile is constructed based on the remaining weights [88].

Clinical Utility of Mutation Signatures

The clinical relevance of genome-scale patterns of DNA alterations has been investigated to gain an understanding of the mechanistic underpinnings of tumour aetiology, but also in particular to raise the possibility that genome profiling might be a useful clinical/diagnostic tool to aid patient management.

In addition to the work described above, where substitution signatures of know aetiology can give insight into causes of some cancers (*e.g.* smoking, ultraviolet irradiation, DNA repair deficiency), a recent pioneering study examined the impact of ionizing radiation (IR) as a cancer-causing agent. The investigators identified patients who received radiotherapy in order to treat a cancer (of any type), but who then developed a second tumour within the field of the therapeutic IR. Whole genome sequencing of the second tumours revealed two unique patterns of mutation signature specifically associated with *in vivo* exposure to IR [89]. These mutation patterns involved a high rate of small (1-100 bp) deletions and balanced inversions, distributed throughout the genome, irrespective of the type of tumour analysed. This study gives *in vivo* insight into the precise role of IR in causing mutations to normal cells in the field of IR, which initiated the malignant process.

Patterns of DNA alterations (loss of heterozygosity and copy number alterations) and DNA mutational signatures are also being developed as diagnostic tools to help stratify patients into certain treatment actionable groups. One area of great interest is in identifying tumours with homologous recombination (HR) deficiency. Between 1-5% of breast cancers are attributable to germline mutations in *BRCA1* and *BRCA2*, and as described above these tumours exhibit considerable genome instability indicative of HR deficiency. Importantly, *BRCA1/2* deficient tumours also exhibit enhanced sensitivity to some DNA damaging chemotherapies, particularly platinum-based compounds, or to targeted therapy using inhibitors to poly(ADP-ribose) polymerase (PARP) [90]. HR-deficient tumour cells exhibit considerable sensitivity to PARP inhibitors because PARP is involved in an alternative repair pathway called

base excision repair (BER). Hence, with BER inactivated, single strand breaks are not effectively repaired and they accumulate as double strand breaks during DNA replication, leading to replication fork collapse, enhanced genome instability and cell death. Genome-scale patterns could therefore be a biomarker of HR deficiency and hence sensitivity to PARP inhibitors or certain chemotherapies. Great interest has therefore been applied to develop 'biomarkers' of HR deficiency based on the pattern of somatic alterations identified in the tumour genome. These genome-based 'tools' utilise either aCGH or SNP array based genome data [64-71] or more recently whole genome sequencing data [91].

Several genome stability scores have been developed that use combinations of DNA copy number and/or allelic imbalance information to infer HR deficiency. For example, telomeric allelic imbalance (AI) is a subchromosomal region of AI that extends to the telomeric end of a chromosome. The number of telomeric AIs was associated with sensitivity to platinum-based therapy in triple negative breast cancer [71]. In another study, the total number of breakpoints within a tumour genome had no association with *BRCA* status, however the number of large-scale transitions (LST), defined as chromosomal breaks between adjacent regions of at least 10 Mb correlated with *BRCA1* inactivation [69]. Loss of heterozygosity (LOH) was also strongly associated with *BRCA* deficiency [68]. These three genome-based scores (telomeric AI, LST and LOH) have been assessed as a combined homologous recombination deficiency (HRD) score (HRD Index) in various cohorts of breast cancer samples. The assay, developed using next generation sequencing, was collectively a more sensitive predictor of *BRCA* deficiency than individual parameters [92] and identified patients with increased benefit to neoadjuvant platinum based therapy [66, 93]. Similar findings have been reported reflecting the burden of allelic imbalanced copy number alterations and response to platinum-based chemotherapy [70].

A recent study applied whole genome sequencing data of breast cancers to develop a probabilistic predictor of HR deficiency [91]. Logistic lasso regression modelling was used to incorporate six genomic parameters, including substitution and rearrangement signatures and the HRD Index (described above) into a predictor, termed HRDetect. The predictor identified 22% of all breast tumours as being HR deficient, including tumours with germline or somatic biallelic inactivation of *BRCA1/2*, but also tumours

in which the cause of the HR deficiency was unclear. In a small pilot, the HRDetect score correlated with pathological response to neoadjuvant anthracycline chemotherapy[91]. This collective data indicates the potential for complex genome-scale mutation data to be distilled into a simple diagnostic score that could be used clinically to identify patients who may gain benefit from certain chemotherapies or PARPi.

Intratumour heterogeneity is a consequence of the acquisition of DNA mutations that drive the evolution of tumour cell clones with enhanced 'fitness'. It is important to consider that the selective pressures of therapy may promote the accumulation of favourable mutations that lead to the development of treatment-resistant clones. The APOBEC substitution signature [43, 94] is a marker of APOBEC mutagenesis and has been linked to ensuing tumour evolution during progression and as a mechanism of resistance to treatment [95]. In breast cancer, a study has revealed that APOBEC3B could be a potential biomarker of tamoxifen resistance in ER+ disease [96]. The levels of ABOPEC3B in primary ER+ tumours inversely correlated with clinical benefit from tamoxifen treated metastatic disease. This finding was also validated *in vitro* using functional manipulation of APOBEC3B levels to either enhance tamoxifen response when APOBEC3B was depleted or to enhance resistance when APOBEC3B was overexpressed [96].

Pioneering work using whole genome sequencing data has also identified a novel phenomenon known as kaetaegis [97], which is defined as clusters of mutations or hypermutation in concentrated genomic regions. Recent work has linked a 628 kaetaegis gene expression signature to higher grade breast tumours that are HER2 positive and have prolonged survival [98]. This study derived a signature based on a differential gene expression analysis between kaetaegis breast tumours and 106 normal breast samples. The expression of genes proximal and distal to kataegis loci was reported. From these studies, it is clear that integration of the findings from whole genome sequencing data and transcriptomic data will reveal key signatures that explain breast cancer prognosis and disease aetiology.

Certainly, the area of cancer genomics has advanced significantly in recent years. Some of the findings highlighted here suggest that the unique insights being derived from different types of mutational profiling are revealing important insights into tumour biology, but also that they could make major contributions in the clinical management of patients, for example in the context of novel biomarkers that might predict therapeutic response.

Conclusions

There are increasing numbers of molecular signatures that are already commercially available, or are in pre-clinical development or show promise from the research setting. These advances illustrate the power of molecular approaches to stratify cancer into biologically and clinically meaningful subgroups beyond what is possible with standard clinical and pathological means of classification. Bioinformatics has also played a crucial role in the development and testing of these signatures, yielding numerous computational methods and tools for the use in the research/discovery setting. Hypothesis-driven research is imperative to validate and compare the performance of these gene signatures and bioinformatics tools across independent studies to ensure the findings are robust and reproducible. We have focused on breast cancer and genomic DNA and mRNA-based signatures, yet it is clear that these approaches reach across cancer types and advances in the signature profiling in the areas of epigenetics, miRNA, proteomics, and immune cell infiltrates will no doubt also play important clinical roles in the near future.

Table 1: Commercial mRNA-based gene signatures. The associated assay details, platforms used to perform the diagnostic assay and the most appropriate patient populations are listed for each test.

Test	Assay	Platform	Patient population	Ref
OncotypeDx® **(Genomic Health)**	21 genes FDA approved	RT-PCR	Stage I or II ER+, LN- and postmenopausal, LN+	[39]
MammaPrint® **(Agendia)**	70 genes FDA approved	Microarray	Stage I and II, ER+ or ER-, HER2+ or HER2-, LN-	[38]
MapQuantDx™ (GGI) (Ipsogen/QIAGEN)	97 genes	Microarray	ER+, intermediate grade	[36]
ProSigna® (nanoString)	58 genes (PAM50) + clinical variables	nanoString	Stage I and II ER+, LN-	[21]
EndoPredict **(Sividon/Myriad Genetics)**	12 genes + tumor size and nodal status	RT-PCR	ER+, HER2-, LN 0-3	[40]

ER: oestrogen receptor; FDA: US Food and Drug Administration; LN: lymph nodes; RT-PCR: Reverse Transcription- Polymerase Chain Reaction; -: negative; +: positive.

Table 2: Bioinformatics tools available for mRNA and DNA analyses. The capability, as well as the associated strengths and limitations of each tool are described.

Bioinformatics tools (URL)	Aims	Strengths	Limitations	References
mRNA profiling				
Genefu http://bioconductor.org/packages/release/bioc/html/genefu.html	Provides a compendia of bioinformatics tools and gene signatures for breast cancer subtyping and prognostication	•Allows the quick manipulation of gene expression datasets, and has tools which facilitate gene selection and probe-gene mapping across platforms •Allows the comparison of multiple signatures (existing and novel)	•Lack of documentation and simple tutorials for first-time users	[41]
SigCheck http://bioconductor.org/packages/release/bioc/html/SigCheck.html	Provide methods for the assessment of a gene signature's prognostic performance	• Allows comparison of signatures performance against, random and known gene signatures •Also allows the assessment of the signature's performance on permuted data and or metadata; useful to ascertain whether the signature has detected a real signal in the original data	•Currently, the package only supports the division of samples into 2 to 3 distinct survival groups, based on the use of a threshold value or percentiles, respectively	[35]
DNA signatures				
SomaticSignatures https://bioconductor.org/packages/release/bioc/html/SomaticSignatures.html	Provide methods for the modelling, identification and visualisation of mutation signatures	•Integrates well with Bioconductor tools and data structures for processing and annotating genomic variants •Supports multiple statistical approaches and other user-defined approaches	•Relies on an unconstrained model, therefore limits the domain of signatures considered, as incorporating more distal bases result in a significant increase in the number of parameters, rendering the mutation signature unstable	[83]
Pmsignature https://github.com/friend1ws/pmsignature	Provide methods for the modelling, identification and visualisation of mutation signatures	•Provides intuitive visualisation of mutation signatures •Uses a probabilistic approach that reduces the number of parameters associated with each signature, thus allows the incorporation of additional sequence context (e.g., allows for incorporation of two bases 3' and 5'of the substitution)	Assumes independence between the mutation features (e.g., type of substitution is a feature, and flanking bases are each, individually, another feature)	[87]
deconstructSigs https://cran.r-project.org/web/packages/deconstructSigs/	Provides methods for the modelling, identification and visualisation of mutational signatures	• Provides intuitive visualisation of mutation signatures • Determines mutation signatures in a single sample.	Assumes any coefficient in the multiple linear regression model must be greater than zero as negative contributions make no biological sense.	[88]

References

1. Lakhani, S.R., et al., *WHO Classification of Tumours of the Breast.* 2012.
2. Bloom, H.J.G. and W.W. Richardson, *Histological Grading and Prognosis in Breast Cancer: A Study of 1409 Cases of which 359 have been Followed for 15 Years.* British Journal of Cancer, 1957. **11**(3): p. 359-377.
3. Elston, C.W. and I.O. Ellis, *pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up.* Histopathology, 1991. **19**(5): p. 403-410.
4. Rakha, E.A., et al., *Breast cancer prognostic classification in the molecular era: the role of histological grade.* Breast Cancer Research, 2010. **12**(4): p. 207.
5. Rakha, E.A., et al., *Prognostic Significance of Nottingham Histologic Grade in Invasive Breast Carcinoma.* Journal of Clinical Oncology, 2008. **26**(19): p. 3153-3158.
6. Rakha, E.A., J.S. Reis-Filho, and I.O. Ellis, *Combinatorial biomarker expression in breast cancer.* Breast Cancer Res Treat, 2010. **120**(2): p. 293-308.
7. Rakha, E., J. Reis-Filho, and I. Ellis, *Combinatorial biomarker expression in breast cancer.* Breast Cancer Research and Treatment, 2010. **120**(2): p. 293-308.
8. Howlader, N., et al., *US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status.* J Natl Cancer Inst, 2014. **106**(5).
9. Foulkes, W.D., I.E. Smith, and J.S. Reis-Filho, *Triple-negative breast cancer.* N Engl J Med, 2010. **363**(20): p. 1938-48.
10. Gerdes, J., et al., *Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation.* International Journal of Cancer, 1983. **31**(1): p. 13-20.
11. Viale, G., et al., *Prognostic and Predictive Value of Centrally Reviewed Ki-67 Labeling Index in Postmenopausal Women With Endocrine-Responsive Breast Cancer: Results From Breast International Group Trial 1-98 Comparing Adjuvant Tamoxifen With Letrozole.* Journal of Clinical Oncology, 2008. **26**(34): p. 5569-5575.
12. Yerushalmi, R., et al., *Ki67 in breast cancer: prognostic and predictive potential.* Lancet Oncol, 2010. **11**(2): p. 174-83.
13. Dowsett, M., et al., *Prognostic value of Ki67 expression after short-term presurgical endocrine therapy for primary breast cancer.* J Natl Cancer Inst, 2007. **99**(2): p. 167-70.
14. Dowsett, M., et al., *Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group.* J Natl Cancer Inst, 2011. **103**(22): p. 1656-64.
15. Mengel, M., et al., *Inter-laboratory and inter-observer reproducibility of immunohistochemical assessment of the Ki-67 labelling index in a large multi-centre trial.* J Pathol, 2002. **198**(3): p. 292-9.
16. Polley, M.Y., et al., *An international Ki67 reproducibility study.* J Natl Cancer Inst, 2013. **105**(24): p. 1897-906.
17. Bartlett, J.M.S., et al., *Validation of the IHC4 Breast Cancer Prognostic Algorithm Using Multiple Approaches on the Multinational TEAM Clinical Trial.* Archives of Pathology & Laboratory Medicine, 2016. **140**(1): p. 66-74.
18. Cuzick, J., et al., *Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the Genomic Health recurrence score in early breast cancer.* J Clin Oncol, 2011. **29**(32): p. 4273-8.

19. Rakha, E.A., et al., *Nottingham Prognostic Index Plus (NPI+): a modern clinical decision making tool in breast cancer.* Br J Cancer, 2014. **110**(7): p. 1688-97.
20. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
21. Parker, J.S., et al., *Supervised risk predictor of breast cancer based on intrinsic subtypes.* J Clin Oncol, 2009. **27**(8): p. 1160-7.
22. Goldhirsch, A., et al., *Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013.* Ann Oncol, 2013. **24**(9): p. 2206-23.
23. Cornen, S., et al., *Candidate luminal B breast cancer genes identified by genome, gene expression and DNA methylation profiling.* PLoS One, 2014. **9**(1): p. e81843.
24. Wallden, B., et al., *Development and verification of the PAM50-based Prosigna breast cancer gene signature assay.* BMC Med Genomics, 2015. **8**: p. 54.
25. Mackay, A., et al., *Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement.* J Natl Cancer Inst, 2011. **103**(8): p. 662-73.
26. Weigelt, B., et al., *Breast cancer molecular profiling with single sample predictors: a retrospective analysis.* Lancet Oncol, 2010. **11**(4): p. 339-49.
27. Mj Sontrop, H., M. Jt Reinders, and D.M. P, *Breast cancer subtype predictors revisited: from consensus to concordance?* BMC Med Genomics, 2016. **9**(1): p. 26.
28. Pusztai, L., et al., *Molecular classification of breast cancer: limitations and potential.* Oncologist, 2006. **11**(8): p. 868-77.
29. Wirapati, P., et al., *Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures.* Breast Cancer Res, 2008. **10**(4): p. R65.
30. Haibe-Kains, B., et al., *A fuzzy gene expression-based computational approach improves breast cancer prognostication.* Genome Biol, 2010. **11**(2): p. R18.
31. Desmedt, C., et al., *Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes.* Clin Cancer Res, 2008. **14**(16): p. 5158-65.
32. Haibe-Kains, B., et al., *A three-gene model to robustly identify breast cancer molecular subtypes.* J Natl Cancer Inst, 2012. **104**(4): p. 311-25.
33. Fumagalli, D., et al., *Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-Sequencing technology.* BMC Genomics, 2014. **15**: p. 1008.
34. Venet, D., J.E. Dumont, and V. Detours, *Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome.* Plos Computational Biology, 2011. **7**(10).
35. Stark, R. and J. Norder, *SigCheck: Check a gene signature's prognostic performance against random signatures, known signatures, and permuted data*. 2016.
36. Sotiriou, C., et al., *Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis.* J Natl Cancer Inst, 2006. **98**(4): p. 262-72.
37. Nielsen, T., et al., *Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens.* BMC Cancer, 2014. **14**: p. 177.
38. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-6.
39. Paik, S., et al., *A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.* N Engl J Med, 2004. **351**(27): p. 2817-26.
40. Filipits, M., et al., *A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors.* Clin Cancer Res, 2011. **17**(18): p. 6012-20.

41. Gendoo, D.M., et al., *Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer.* Bioinformatics, 2016. **32**(7): p. 1097-9.
42. Ignatiadis, M., et al., *The Genomic Grade Assay Compared With Ki67 to Determine Risk of Distant Breast Cancer Recurrence.* JAMA Oncol, 2016. **2**(2): p. 217-24.
43. Burns, M.B., et al., *APOBEC3B is an enzymatic source of mutation in breast cancer.* Nature, 2013. **494**(7437): p. 366-70.
44. Metzger-Filho, O., et al., *Genomic Grade Index (GGI): feasibility in routine practice and impact on treatment decisions in early breast cancer.* PLoS One, 2013. **8**(8): p. e66848.
45. Bertucci, F., et al., *Comparison of the prognostic value of genomic grade index, Ki67 expression and mitotic activity index in early node-positive breast cancer patients.* Ann Oncol, 2013. **24**(3): p. 625-32.
46. Metzger Filho, O., M. Ignatiadis, and C. Sotiriou, *Genomic Grade Index: An important tool for assessing breast cancer tumor grade and prognosis.* Crit Rev Oncol Hematol, 2011. **77**(1): p. 20-9.
47. Tobin, N.P., et al., *Multi-level gene expression signatures, but not binary, outperform Ki67 for the long term prognostication of breast cancer patients.* Mol Oncol, 2014. **8**(3): p. 741-52.
48. Prat, A., et al., *Prediction of Response to Neoadjuvant Chemotherapy Using Core Needle Biopsy Samples with the Prosigna Assay.* Clin Cancer Res, 2016. **22**(3): p. 560-6.
49. Gnant, M., et al., *Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone.* Annals of Oncology, 2014. **25**(2): p. 339-345.
50. Buyse, M., et al., *Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer.* J Natl Cancer Inst, 2006. **98**(17): p. 1183-92.
51. Cardoso, F., et al., *70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer.* New England Journal of Medicine, 2016. **375**(8): p. 717-729.
52. Esteban, J., J. Baker, and M. Cronin, *Tumor gene expression and prognosis in breast cancer: multi-gene RT-PCR assay of paraffin-embedded tissue.* Prog Proc Am Soc Clin Oncol, 2003. **98**: p. 10869-10874.
53. Cobleigh, M., P. Bitterman, and J. Baker, *Tumor gene expression predicts distant disease-free survival (DDFS) in breast cancer patients with 10 or more positive nodes: high throughout RT-PCR assay of paraffin-embedded tumor tissues.* Prog Proc Am Soc Clin Oncol 2003. **22**: p. 850-850.
54. Paik, S., S. Shak, and G. Tang, *Multi-gene RT-PCR assay for predicting recurrence in node negative breast cancer patients -- NSABP studies B-20 and B-14.* Breast Cancer Res Treat, 2003. **82**: p. A16-A16.
55. Sparano, J.A. and S. Paik, *Development of the 21-gene assay and its application in clinical practice and clinical trials.* J Clin Oncol, 2008. **26**(5): p. 721-8.
56. Ramsey, S.D., et al., *Integrating comparative effectiveness design elements and endpoints into a phase III, randomized clinical trial (SWOG S1007) evaluating oncotypeDX-guided management for women with breast cancer involving lymph nodes.* Contemporary Clinical Trials, 2013. **34**(1): p. 1-9.
57. Sparano, J.A., et al., *Prospective Validation of a 21-Gene Expression Assay in Breast Cancer.* N Engl J Med, 2015. **373**(21): p. 2005-14.
58. Buus, R., et al., *Comparison of EndoPredict and EPclin With Oncotype DX Recurrence Score for Prediction of Risk of Distant Recurrence After Endocrine Therapy.* J Natl Cancer Inst, 2016. **108**(11).

59. Alvarado, M.D., et al., *A Prospective Comparison of the 21-Gene Recurrence Score and the PAM50-Based Prosigna in Estrogen Receptor-Positive Early-Stage Breast Cancer.* Adv Ther, 2015. **32**(12): p. 1237-47.
60. Bartlett, J.M., et al., *Comparing Breast Cancer Multiparameter Tests in the OPTIMA Prelim Trial: No Test Is More Equal Than the Others.* J Natl Cancer Inst, 2016. **108**(9).
61. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
62. Nik-Zainal, S., et al., *Mutational processes molding the genomes of 21 breast cancers.* Cell, 2012. **149**(5): p. 979-93.
63. Hicks, J., et al., *Novel patterns of genome rearrangement and their association with survival in breast cancer.* Genome Research, 2006. **16**(12): p. 1465-1479.
64. Joosse, S.A., et al., *Prediction of BRCA1-association in hereditary non-BRCA1/2 breast carcinomas with array-CGH.* Breast Cancer Res Treat, 2009. **116**(3): p. 479-89.
65. Joosse, S.A., et al., *Prediction of BRCA2-association in hereditary breast carcinomas using array-CGH.* Breast Cancer Res Treat, 2012. **132**(2): p. 379-89.
66. Telli, M.L., et al., *Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast Cancer.* Clin Cancer Res, 2016. **22**(15): p. 3764-73.
67. Vollebergh, M.A., et al., *Genomic patterns resembling BRCA1- and BRCA2-mutated breast cancers predict benefit of intensified carboplatin-based chemotherapy.* Breast Cancer Res, 2014. **16**(3): p. R47.
68. Abkevich, V., et al., *Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer.* Br J Cancer, 2012. **107**(10): p. 1776-82.
69. Popova, T., et al., *Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation.* Cancer Res, 2012. **72**(21): p. 5454-62.
70. Watkins, J., et al., *Genomic Complexity Profiling Reveals That HORMAD1 Overexpression Contributes to Homologous Recombination Deficiency in Triple-Negative Breast Cancers.* Cancer Discov, 2015. **5**(5): p. 488-505.
71. Birkbak, N.J., et al., *Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents.* Cancer Discov, 2012. **2**(4): p. 366-75.
72. Tolosi, L. and T. Lengauer, *Classification with correlated features: unreliability of feature ranking and solutions.* Bioinformatics, 2011. **27**(14): p. 1986-94.
73. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.
74. Meier, L., S.A. van de Geer, and P. Buhlmann, *The group lasso for logistic regression.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 2008. **70**: p. 53-71.
75. Rapaport, F., E. Barillot, and J.P. Vert, *Classification of arrayCGH data using fused SVM.* Bioinformatics, 2008. **24**(13): p. I375-I382.
76. Park, M.Y., T. Hastie, and R. Tibshirani, *Averaged gene expressions for regression.* Biostatistics, 2007. **8**(2): p. 212-227.
77. Forbes, S.A., et al., *COSMIC: somatic cancer genetics at high-resolution.* Nucleic Acids Res, 2017. **45**(D1): p. D777-D783.
78. Harris, R.S., *Molecular mechanism and clinical impact of APOBEC3B-catalyzed mutagenesis in breast cancer.* Breast Cancer Res, 2015. **17**: p. 8.
79. Nik-Zainal, S., et al., *Landscape of somatic mutations in 560 breast cancer whole-genome sequences.* Nature, 2016. **534**(7605): p. 47-54.
80. Waddell, N., et al., *Whole genomes redefine the mutational landscape of pancreatic cancer.* Nature, 2015. **518**(7540): p. 495-501.
81. Alexandrov, L.B., et al., *Deciphering signatures of mutational processes operative in human cancer.* Cell Rep, 2013. **3**(1): p. 246-59.

82. Brunet, J.P., et al., *Metagenes and molecular pattern discovery using matrix factorization.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(12): p. 4164-4169.
83. Gehring, J.S., et al., *SomaticSignatures: inferring mutational signatures from single-nucleotide variants.* Bioinformatics, 2015. **31**(22): p. 3673-5.
84. Tan, V.Y. and C. Fevotte, *Automatic relevance determination in nonnegative matrix factorization with the beta-divergence.* IEEE Trans Pattern Anal Mach Intell, 2013. **35**(7): p. 1592-605.
85. Kasar, S., et al., *Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution.* Nat Commun, 2015. **6**: p. 8866.
86. Kim, J., et al., *Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors.* Nat Genet, 2016. **48**(6): p. 600-6.
87. Shiraishi, Y., et al., *A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures.* PLoS Genet, 2015. **11**(12): p. e1005657.
88. Rosenthal, R., et al., *DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution.* Genome Biol, 2016. **17**: p. 31.
89. Behjati, S., et al., *Mutational signatures of ionizing radiation in second malignancies.* Nat Commun, 2016. **7**: p. 12605.
90. Lord, C.J. and A. Ashworth, *BRCAness revisited.* Nat Rev Cancer, 2016. **16**(2): p. 110-20.
91. Davies, H., et al., *HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures.* Nat Med, 2017. **23**(4): p. 517-525.
92. Timms, K.M., et al., *Association of BRCA1/2 defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes.* Breast Cancer Res, 2014. **16**(6): p. 475.
93. Telli, M.L., et al., *Phase II Study of Gemcitabine, Carboplatin, and Iniparib As Neoadjuvant Therapy for Triple-Negative and BRCA1/2 Mutation-Associated Breast Cancer With Assessment of a Tumor-Based Measure of Genomic Instability: PrECOG 0105.* J Clin Oncol, 2015. **33**(17): p. 1895-901.
94. Roberts, S.A., et al., *An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers.* Nat Genet, 2013. **45**(9): p. 970-6.
95. Swanton, C., et al., *APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity.* Cancer Discov, 2015. **5**(7): p. 704-12.
96. Law, E.K., et al., *The DNA cytosine deaminase APOBEC3B promotes tamoxifen resistance in ER-positive breast cancer.* Sci Adv, 2016. **2**(10): p. e1601737.
97. Nik-Zainal, S., et al., *The life history of 21 breast cancers.* Cell, 2012. **149**(5): p. 994-1007.
98. D'Antonio, M., et al., *Kataegis Expression Signature in Breast Cancer Is Associated with Late Onset, Better Prognosis, and Higher HER2 Levels.* Cell Reports, 2016. **16**(3): p. 672-683.

Figure Legends

Figure 1: Conceptual overview of tumour grading and staging.

A) Grading is a measure of tumour cell differentiation, relative to normal cells. Representative histological images of tumours of grade 1, 2 and 3 (see text) are shown, as stained with haemotoxylin and eosin. B) A pictorial representation of the tumour staging system. The different components of the TNM staging system are highlighted: T represents tumour size and extent of local invasion; N is a measure of tumour spread to regional lymph nodes (N); and M is a clinical assessment to record the extent of cancer metastasis to distant sites, such as the lung, liver, brain and bone.

Figure 2: General overview of the signature development process. Each stage of this process is conducted within framework of human ethics approval, and should be subjected to quality control to ensure there is consistency and reproducibility.

Figure 3: Conceptual overview of different subtype classification algorithms. This schematic illustrates an overview for the development of a Single Sample Predictor (SSP) or a Subtype Classification Model (SCM). The SSP does not necessarily involve a model and is calculated for each tumour sample, using a centroid computed from an intrinsic gene list. Each calculated centroid is compared with predefined subtypes from a training set and the nearest centroid is used in the classification of a new tumour. SCMs involve gene modules rather than an intrinsic gene list and each new sample has a module score from the gene modules. The training cohort is classified into three distinct subtypes (ER-/HER2-, HER2+, ER+/HER2+) by fitting a 3 component Gaussian mixture model rather than using centroid based classification. Each new sample and module score is classified based on the maximum posterior probability using the model.

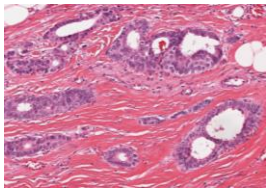
Figure 4: Conceptual overview of DNA mutational signatures. The tumour genome is subjected to whole genome sequencing to identify somatic nucleotide alterations and structural variants. Substitution mutation signatures (left panel) are derived by calculating the frequency of all single base-pair mutations, in the context

of each mutation according to its neighbouring 3' and 5' nucleotides. A 96-trinucleotide matrix, which consists of 96 substitutions by M samples is established and used in non-matrix factorization (NMF; or BayesNMF) to calculate the mutational signatures. The number of signatures (k) must be provided *a priori* in the NMF methodology and this can be calculated using the cophenetic correlation coefficient. Matrix A is split up into matrix W (96 x k matrix) and H (k x M matrix), where M is the number of samples.

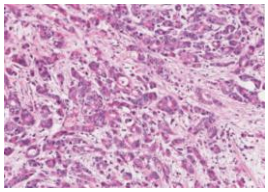
Rearrangement Signatures (right panel) are used to catalogue genomic rearrangements in cancer genomes. A piecewise constant fitting method is applied to discriminate between rearrangements that occurred as focal catastrophic events or focal driver amplicons ("clustered") from genome-wide rearrangement mutagenesis ("non clustered"). The rearrangements are sub-classified based on type and size of the structural variation. This classification produces a 32-structural variant matrix which is decomposed using NMF.

A. Tumour Grade

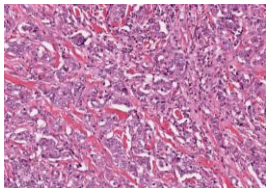
Grade 1



Grade 2

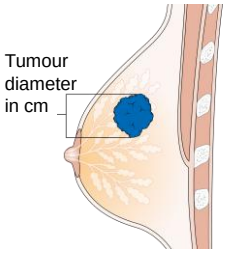


Grade 3

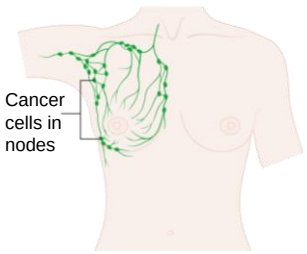


B. Tumour Stage

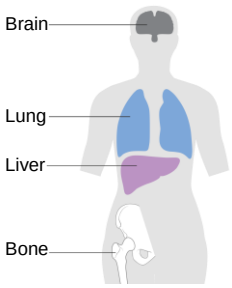
T: Tumour Size

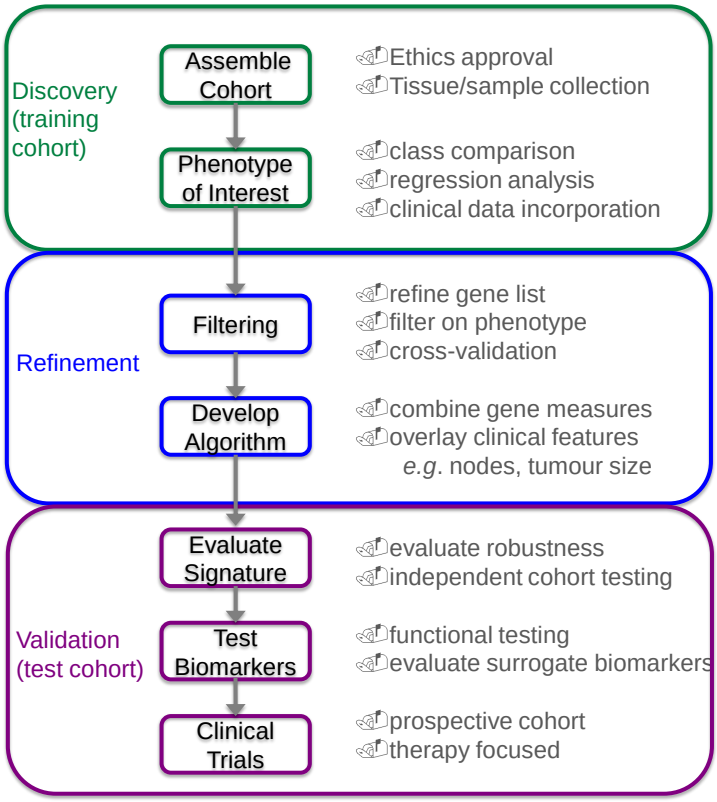


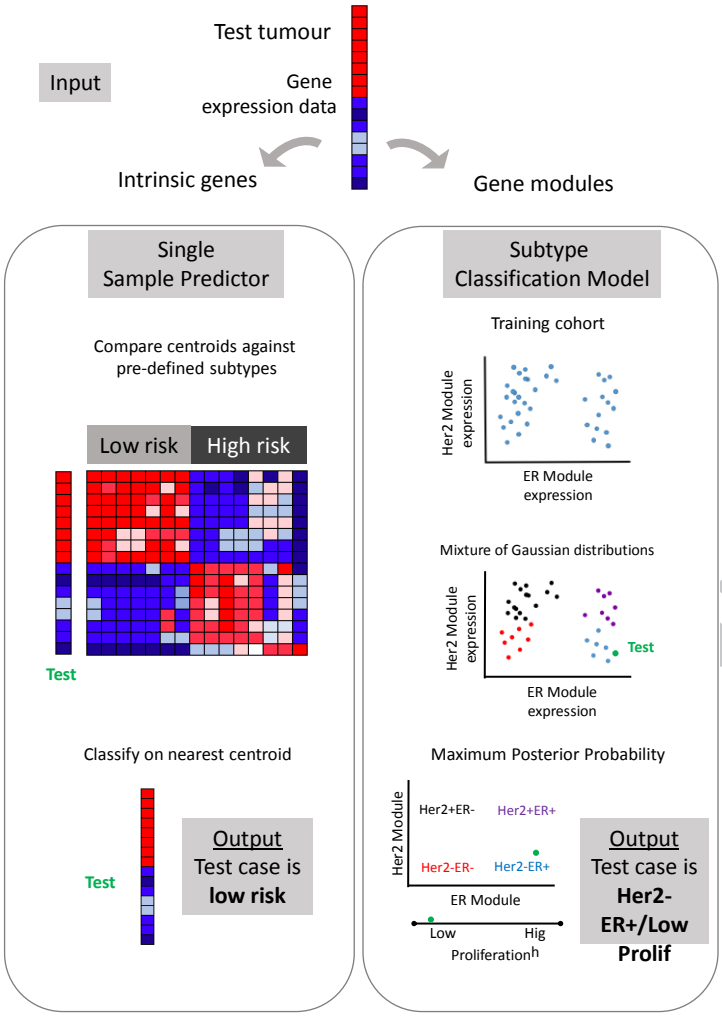
N: Lymph Node Involvement

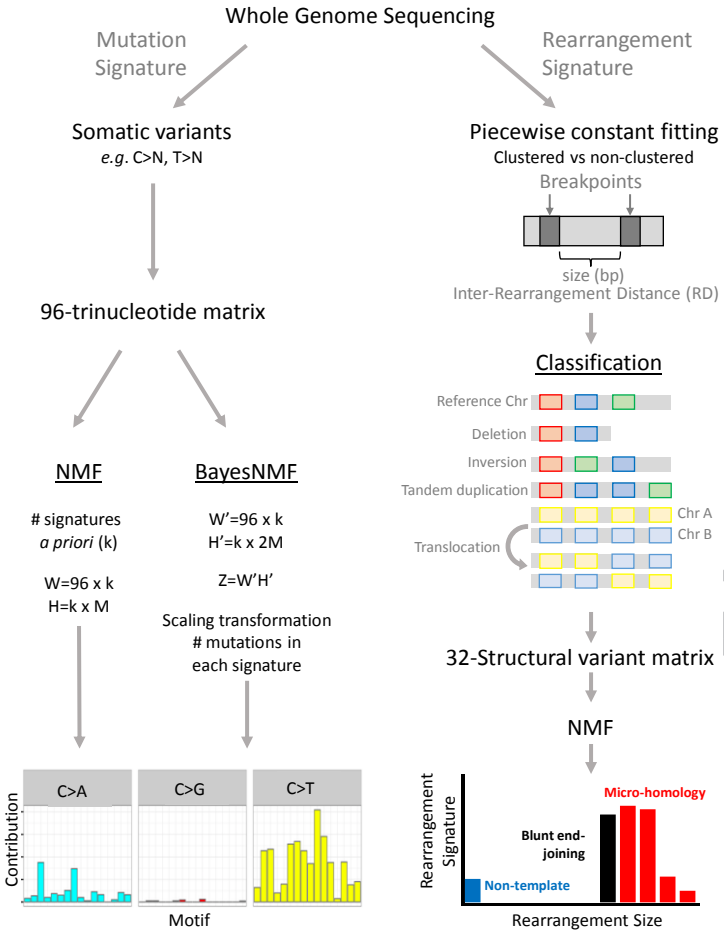


M: Metastasis









Highlights

- An overview of the commercially available prognostic signatures in breast cancer
- Conceptual overview of the molecular signature development process in breast cancer
- Aimed at biologists and computational biologists with an interest in gene signatures.