

The path to routine use of genomic biomarkers in the cancer clinic

Paul C. Boutros

Informatics and Biocomputing Program, Ontario Institute for Cancer Research, Toronto M5G 0A3, Ontario, Canada; Department of Medical Biophysics, University of Toronto, Toronto M5S 1A8, Ontario, Canada; Department of Pharmacology and Toxicology, University of Toronto, Toronto M5S 1A8, Ontario, Canada

It has been almost 15 years since the first microarray-based studies creating multigene biomarkers to subtype and predict survival of cancer patients. This Perspective looks at why only a handful of genomic biomarkers have reached clinical application and what advances are needed over the next 15 years to grow this number. I discuss challenges in creating biomarkers and reproducing them at the genomic and computational levels, including the problem of spatio-genomic heterogeneity in an individual cancer. I then outline the challenges in translating newly discovered genome-wide or regional events, like trinucleotide mutation signatures, kataegis, and chromothripsis, into biomarkers, as well as the importance of incorporating prior biological knowledge. Lastly, I outline the practical problems of pharmaco-economics and adoption: Are new biomarkers viewed as economically rational by potential funders? And if they are, how can their results be communicated effectively to patients and their clinicians? Genomic-based diagnostics have immense potential for transforming the management of cancer. The next 15 years will see a surge of research into the topics here that, when combined with a stream of new targeted therapies being developed, will personalize the cancer clinic.

The potential of clinical cancer genomics

Cancer is, at its heart, a disease of the genome. Individual tumors harbor from hundreds to hundreds of thousands of point mutations (Lawrence et al. 2013). They can have global ploidy changes or local chromosomal abnormalities that alter as much as 50% of the genome (Zack et al. 2013). They can have dozens of genomic rearrangements of various types (Yang et al. 2013). Large-scale sequencing projects like the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) have been sequencing hundreds of tumors of different subtypes to try to create catalogs of those that are recurrent in any given cancer type (The Cancer Genome Atlas Research Network 2008; Hudson et al. 2010). These studies have led to the discovery of fundamental new properties of cancer genomes, such as mutational signatures (Alexandrov et al. 2013a,b), focal genomic abnormalities like kataegis and chromothripsis (Stephens et al. 2011), robust estimates of the distribution and number of driver genes (Lawrence et al. 2014), and classification of many tumor types into distinct subtypes (The Cancer Genome Atlas Network 2012a,b, 2015).

However, these discoveries by themselves are not sufficient to impact patient care. Rather, bringing cancer genomics into the clinic requires two separate and generally orthogonal arms. First, genomic profiles need to be mined to identify candidate genes that can be targeted by novel drugs. By targeting vulnerabilities present in a tumor and not in normal cells, it is believed that drugs can be developed with greater specificity and sensitivity. Second, genomic profiles need to be used to create novel biomarkers that can be used to diagnose disease, to predict patient survival, to predict response to treatment (e.g., companion diagnostics to novel or existing therapies), and to monitor disease relapse. This review will focus on the second problem, discussing the barriers that are lim-

iting the routine use of genomic assays in shaping and improving the care of cancer patients.

Barriers to adoption

Stable biomarkers

In many fields, a very large number of biomarkers have been developed. For example, there are at least 106 separate biomarkers to prognose localized breast cancer, and these differ significantly in accuracy, error profiles (including sensitivity-specificity trade-offs and biases in errors towards specific clinical or molecular characteristics), number of genes, ease of interpretation, and biological origin (Tofigh et al. 2014). Most of these have not moved beyond the research setting, but which might best benefit from additional validation? Which have the most potential for clinical use? Even developing the answer to one of these questions is extremely challenging, but the continual rapid development of new (and sometimes only modestly improved) biomarkers poses several major challenges.

First, because there are no gold standards used in the field for validation, it is difficult or even impossible to assess which markers are performing best. Often, validation cohorts are insufficiently independent and poorly powered, although with some noteworthy exceptions (Kratz et al. 2012). Although challenge-based assessments are just now starting to create such data sets (Margolin et al. 2013; Boutros et al. 2014b), they still remain both statistically underpowered and underrepresentative of the broad diversity of human cancer and of genomic technologies. Second, the sheer number of biomarkers being developed in some fields can directly hinder clinical application both by creating confusion and by fostering an attitude that the field is too dynamic for practical application—that waiting for future research and technological advances is the best decision. Third, commercialization of

Corresponding author: Paul.Boutros@oicr.on.ca

Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.191114.115>. Freely available online through the *Genome Research* Open Access option.

© 2015 Boutros This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

biomarkers is increasingly difficult as more biomarkers are created in a field, because it reduces barriers to entry and limits the ability of an individual biomarker to gain significant market share. Indeed, as more biomarkers are created in a field, the development and advancement of new and improved approaches can be hindered both by intellectual property restrictions from prior art and by reluctance of funders and commercialization offices to support validation studies. Fourth, the ultimate utility of a biomarker is often unknown without long-term clinical follow-up studies in multiple settings, often prospectively, leading to significant development and validation costs.

Reproducibility of analyses

Clinical application of genomic techniques requires that the resulting tests are highly accurate and highly reproducible. Reproducibility can be considered in two different ways. First, there is reproducibility in the actual genomic measurements. Reproducibility of clinical tests is standardized under regulations like CLIA and GLP. Targeted sequencing assays appear to perform very well (Tran et al. 2013), but this result is certainly driven in part by the very high depth of coverage in such assays, and because most mutation-detection algorithms have not been optimized to distinguish low-frequency events from sequencing errors, these panels can yield false negatives. Further, the error rates for sequencing-based discovery of genomic rearrangements (e.g., translocations or inversions) are much less understood than those for single-nucleotide variants. As a result, whole-genome studies—which would be necessary to measure complex phenomena like kataegis or chromothripsis, for example—are likely to be less reproducible, especially given lower coverage levels. There has been a small amount of research into quality control of genomic studies (Daley and Smith 2013; Chong et al. 2014) and almost none into how quality affects final prediction of mutations and other genomic phenomena. An elegant study by the ICGC extracted DNA once from each part of a tumor/normal pair and shipped aliquots of this sample to five large international sequencing centers. Each center sequenced and analyzed the same sample using their own protocols, and the final somatic SNV predictions were compared. Only ~20% of mutations were common to all five centers, while one third were predicted by only a single center (Buchhalter et al. 2014). Clearly, significant work is needed to standardize global analyses of cancer genomes.

There is, similarly, significant diversity in the analysis of cancer genomic data. Even small differences in the way a data set is preprocessed and analyzed can yield massive differences in the predictions of a final biomarker, and it appears that the more complex the biomarker, the more sensitive it is to processing differences, both in terms of computational methodologies (Starmans et al. 2012; Fox et al. 2014) and sample fixation processes (Van Allen et al. 2014). However, analysis methods cannot yet be standardized because there is very little consensus in the field about the best methods for different problems. For example, several studies of microarray processing techniques have yielded discordant results (Shedden et al. 2005; Shi et al. 2005, 2006, 2010; Canales et al. 2006; Zhu et al. 2010). To understand the variability in cancer genome analysis using next-generation sequencing data, the ICGC-TCGA DREAM Somatic Mutation Calling (SMC-DNA) Challenge has been launched (Boutros et al. 2014a). This crowd-sourced challenge, along with efforts by the ICGC Pan-Cancer Project and other groups, will start to create consensus in this area over the next decade. In its first results, the SMC-DNA

Challenge has shown that even on relatively simple tumors (i.e., 100% tumor cellularity, no subclonality, normal ploidy), most groups made a significant number of errors: Across 119 submissions the median *F*-score was 0.88 (Ewing et al. 2015).

However, biomarker reproducibility is not only challenged by the reproducibility of high-throughput assays or their analysis, but also by the inherent biology of a tumor. A series of seminal studies have used high-throughput sequencing to profile the intra-tumoral heterogeneity of kidney (Gerlinger et al. 2012, 2014; Gulati et al. 2014), prostate (Boutros et al. 2015; Cooper et al. 2015; Gundem et al. 2015), breast (Shah et al. 2012; Eirew et al. 2015), lung (de Bruin et al. 2014; Zhang et al. 2014), ovarian (Bashashati et al. 2013; Anglesio et al. 2015), and other tumors. These studies have universally shown that individual tumors are comprised of myriad cell types present at different frequencies in different spatial sites. Importantly, some of these studies have demonstrated that existing biomarkers would give distinct predictions if derived from spatially distinct regions of the tumor. While a few studies have made preliminary estimates of the number of biopsy specimens needed to yield robust conclusions in the face of intra-tumoral heterogeneity (Bachtiary et al. 2006), it remains unclear exactly how biomarkers should be handled in general. For example, should multiple regions be tested and the prediction of the most adverse clinical outcome (e.g., highest drug resistance or shortest survival) used? The average across multiple regions? Should biomarkers focus on clonal driver mutations and, if so, how should variation in the frequencies of truncal mutations be handled (Shah et al. 2009)? Entirely new computational methods may be needed that directly account for intra-tumoral heterogeneity. Indeed, it has been reported that, for poorly understood reasons, some tumors are fundamentally more difficult to develop robust biomarkers for or to make accurate predictions on (Tofigh et al. 2014).

Defining complex phenomena

Some recently uncovered genomic abnormalities are highly complex. For example, several groups have recently shown that genomic instability is a robust biomarker for several tumor types (Vollan et al. 2015). However, there are many potential proxies for genomic instability for use in biomarker studies: number of copy number aberrations (CNAs), the fraction of the genome altered by a CNA, the number of genes showing a CNA, and so forth. Other genomic alterations in cancer are so complex that no real definition exists. Chromothripsis, for example, is generally described as a chromosome “shattering” event where a single chromosome acquires a large number of mutations of different types (Stephens et al. 2011). There is no singular definition of chromothripsis and even only a few operational ways of identifying it (Lapuk et al. 2012; Govind et al. 2014). Similarly, there is not yet a standard library of mutational signatures or standard algorithms to call them uniformly across data sets. The same is true for localized hypermutation at the point-mutation level such as kataegis (Alexandrov et al. 2013a) or for “complex” multichromosomal genomic rearrangements (Berger et al. 2011; Baca et al. 2013). Nevertheless, there is already evidence that global mutation burden can be prognostic in multiple tumor types (Lalonde et al. 2014; Vollan et al. 2015) and that trinucleotide signatures and mutation burden may be predictive of response to targeted therapies (Rizvi et al. 2015), making reproducible measurement critical. This problem is only going to be exacerbated as new methods (Ha et al. 2014; Oesper et al. 2014; Roth et al. 2014; Deshwar et al. 2015) and better understanding of the diversity of cells within a tumor and their evolution (Navin

et al. 2011; Wang et al. 2014; Eirew et al. 2015) start creating population-level features that can be used in biomarker analysis. The next decade will likely see the rise of biomarkers based on nebulous terms such as “subclone number,” “total genetic diversity,” and “tumor heterogeneity index” that will be challenging to define and reproduce, but that will reflect a key aspect of tumor biology with significant predictive potential.

The current round of “compendium” cancer genomic studies thus identify a large number of interesting features that can potentially serve as biomarkers, but these have not yet been defined well enough to serve as components of clinical diagnostics.

Integrating multiple levels of data

Interrogation of any single type of genomic data may provide limited predictive accuracy: Several groups have tested large numbers of random biomarkers to evaluate the probable upper limit of prediction accuracies (Boutros et al. 2009; Starmans et al. 2011; Venet et al. 2011). In several cases, these limits have been surprisingly low. In a recent study, KRAS mutation status could only be predicted, at most, with ~75% accuracy from mRNA abundances (Starmans et al. 2015). Thus, while almost all well-validated genomic tests exploit data of a single class (e.g., copy number aberrations, mRNA abundances, etc.), it is hypothesized that incorporating multiple types of genomic data will improve biomarker accuracy. For example, in the same status of KRAS, it was shown that different prognostic mRNA signatures were optimal in KRAS mutant and KRAS wild-type lung cancers, highlighting the synergy of combining somatic SNV and tumor mRNA abundance information into a composite biomarker (Starmans et al. 2015).

There are not yet any examples of biomarkers that predict clinically relevant endpoints based on simultaneous analysis of methylation levels, specific copy number aberrations or point mutations (both germline and somatic), mRNA abundances, and specific splice-isoform presence or absence. The algorithms required to create such complex biomarkers are now in development (Gonzalez-Perez et al. 2013; Creixell et al. 2015) but are necessarily very complex to develop and require harmonized, multimodal data sets with deep clinical information for both training and testing. Such data sets are not yet broadly available, although some groups have sought to mine TCGA data, despite its somewhat limited clinical follow-up (Yuan et al. 2014), and the METABRIC consortium has profiled miRNA, mRNA, germline SNPs, and somatic copy number aberrations on a coherent set of samples (Curtis et al. 2012; Dvinge et al. 2013). There will be an urgent need for standard data sets to be generated and used for groups to test methods for creating multimodal signatures. There will also be significant challenges in bringing such markers to clinical use, because clinical specimens—particularly those derived from patient biopsies—may not yield sufficient quantity or quality of analytes for simultaneous measurement of all desired biomolecule types. As a result, algorithms will need to be capable of handling missing entire data types, such as when high-quality DNA-based measurements are available but RNA-based ones are not.

However, interrogation of multiple levels of data goes beyond different types of -omic data. For example, several groups have shown that there is significant biomarker content present in the stroma surrounding a tumor (Finak et al. 2008; Hoshida et al. 2008). Others have demonstrated synergy between genomic measurements and tumor microenvironmental factors like hypoxia (Lalonde et al. 2014). A major research direction moving forward will be the integration of clinical imaging data with genomic stud-

ies both through the emergent field of “radiomics” (Aerts et al. 2014) and by exploiting standard pathology images (Yuan et al. 2012). These data types may be generally available on a large fraction of patients, but again, algorithms will be required that can handle missing data types.

Pharmaco-economics of genomic tests

A genomic biomarker may have good accuracy and reproducibility across a range of independent validation data sets. To reach routine adoption, however, it must also guide clinical decision making in a way that is demonstrably and economically efficient for the funders of a healthcare system. That is, one needs to determine if applying a biomarker to specific clinical subgroup will be financially efficient. Consider the use of prostate-specific antigen (PSA) as a population-screening tool to diagnose prostate cancer. Although there is some controversy about the statistical modeling, even conservative estimates suggest that >1250 individuals must be screened and >40 treated to save one life (Loeb et al. 2011). Thus there are many biomarkers that are statistically superior to random chance, but may not be beneficial for the health-care system as a whole. There are many ways of assessing the financial efficiency of a biomarker, although the number of quality-adjusted life years saved per dollar spent (QALY/\$) is often used in formal modeling exercises. There are only a limited number of pharmaco-economic studies for genomic biomarkers to date. It is likely, moving forward, that the pharmaco-economic modeling will be built directly into modeling activities: For example, the cost functions in machine-learning exercises can be modeled explicitly based on the financial benefits or costs of different types of errors or successes.

Explainability

Even if a biomarker is demonstrated to be accurate and economic, this is not always sufficient to guarantee its routine use; that requires adoption and interpretation by clinicians and patients. The development of biomarkers from large genomic data can occur in several ways. Many times a specific drug target is its own biomarker, as with levels of ERBB2 (i.e., HER2) predicting a response to Herceptin or presence of BCR-ABL1 predicting sensitivity to Gleevec. In these cases, the same molecule serves as both biomarker and target. However, single-molecule biomarkers are widely used in many clinical contexts outside of predicting response to treatment. For example, single-molecule biomarkers are widely used to predict prognosis or monitor disease relapse, as in the routine measurement of serum levels of PSA in prostate cancer patients.

Single-gene markers have the immense advantage of simplicity, both in terms of genomics interrogation and in terms of data analysis. However, the biology of a tumor can be extremely complex, especially when considering endpoints like prognosis: No single molecule can fully capture all the determinants of the processes of tumor initiation, progression, or metastasis. Indeed, classically 6–10 distinct molecular or biochemical functions have been identified as associated with these processes (Hanahan and Weinberg 2000, 2011). As a result, modern biomarkers are being developed using statistical and machine-learning techniques, often under the rubric of “data science” or “big-data analysis.” These types of analytical approaches can either be agnostic to the underlying biology or can incorporate domain knowledge such as known pathways (Vaske et al. 2010) or protein complexes (Leiserson et al. 2015), types of information flow between biomolecules, or other types of biological information (Wu et al. 2010).

Independent of whether or not domain knowledge is used, these complex models can use tens to thousands of genes, transcripts, or proteins (Monzon et al. 2009). To better reflect the nonlinearities of biological pathways, this large number of genes is often weighted using mathematical models such as support vector machines, random forests, and network models. Despite the potential greater predictive accuracy introduced by the better fit between true biology and these types of mathematical models, another challenge is introduced: that of interpretability. Patients and their caregivers need to be ready to interpret the results of genomic tests. When these tests involve complex multigene models or sophisticated statistical terminology, that communication can be challenging and can limit uptake.

At least four major changes are likely to occur in this area over the next decade. First, new generations of clinicians are much more familiar with and better trained in genomic techniques, which will facilitate interpretation of final models. Second, patients will become more comfortable with genomics and genomic techniques and be more capable of conversing with their clinicians in this area. Third, standardization of genomic approaches across multiple areas of medicine will create more familiarity and consistency. Fourth, ongoing work by many groups in visualization and communication will provide technical solutions.

The path forward

At times it seems inevitable that those doing genomic research that multimodal -omic biomarkers will become prevalent in routine clinical practice over the next 25 years. However, the path to move from current targeted sequencing panels of specific, carefully selected point mutations to genome-wide assays at multiple levels is unclear. It will require significant advances in genomics and computational biology. The seminal paper demonstrating that gene expression can predict outcome in breast cancer was published 13 years ago (van't Veer et al. 2002), and in the intervening time, few other -omic clinical diagnostics have reached routine clinical practice. In part, this is a function of incomplete clinical annotation of many cohorts with genomic data, particularly with regard to long-term outcomes and response to treatment. This will change as the raw data sets underpinning biomarker discovery and application improve, with more consistent genomic data, better access to and sharing of clinical trial-linked data (as proposed in the next iteration of the ICGC), challenge-based methods assessments, and more frequent assessment of spatial heterogeneity within a tumor. These changes in the raw data will be complemented by improvements in data analysis, particularly in handling heterogeneity, incorporating prior biological knowledge, and in scoring large-scale genomic phenomena. Finally, these improvements in genomics and computational biology will reach their full potential as large numbers of new, targeted therapies continue to be developed, providing the clinical need to drive the development and application of genomic biomarkers for the cancer clinic.

Acknowledgments

This study was conducted with the support of the Ontario Institute for Cancer Research to P.C.B. through funding provided by the Government of Ontario. This work was supported by Prostate Cancer Canada and is proudly funded by the Movember Foundation, Grant #RS2014-01. P.C.B. was supported by a Terry Fox Research Institute New Investigator Award and by a CIHR

New Investigator Award. I thank Renasha Small-O'Connor for editing support.

References

- Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Cavalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, et al. 2014. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* **5**: 4006.
- Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al. 2013a. Signatures of mutational processes in human cancer. *Nature* **500**: 415–421.
- Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013b. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**: 246–259.
- Anglesio MS, Bashashati A, Wang YK, Senz J, Ha G, Yang W, Aniba MR, Prentice LM, Farahani H, Li Chang H, et al. 2015. Multifocal endometriotic lesions associated with cancer are clonal and carry a high mutation burden. *J Pathol* **236**: 201–209.
- Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M, et al. 2013. Punctuated evolution of prostate cancer genomes. *Cell* **153**: 666–677.
- Bachtiary B, Boutros PC, Pintilie M, Shi W, Bastianutto C, Li JH, Schwock J, Zhang W, Penn LZ, Jurisica I, et al. 2006. Gene expression profiling in cervical cancer: an exploration of intratumor heterogeneity. *Clin Cancer Res* **12**: 5632–5640.
- Bashashati A, Ha G, Tone A, Ding J, Prentice LM, Roth A, Rosner J, Shumansky K, Kalloger S, Senz J, et al. 2013. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J Pathol* **231**: 21–34.
- Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, et al. 2011. The genomic complexity of primary human prostate cancer. *Nature* **470**: 214–220.
- Boutros PC, Lau SK, Pintilie M, Liu N, Shepherd FA, Der SD, Tsao MS, Penn LZ, Jurisica I. 2009. Prognostic gene signatures for non-small-cell lung cancer. *Proc Natl Acad Sci* **106**: 2824–2828.
- Boutros PC, Ewing AD, Ellrott K, Norman TC, Dang KK, Hu Y, Kellen MR, Suver C, Bare JC, Stein LD, et al. 2014a. Global optimization of somatic variant identification in cancer genomes with a global community challenge. *Nat Genet* **46**: 318–319.
- Boutros PC, Margolin AA, Stuart JM, Califano A, Stolovitzky G. 2014b. Toward better benchmarking: challenge-based methods assessment in cancer genomics. *Genome Biol* **15**: 462.
- Boutros PC, Fraser M, Harding NJ, de Borja R, Trudel D, Lalonde E, Meng A, Hennings-Yeomans PH, McPherson A, Sabelnykova VY, et al. 2015. Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat Genet* **47**: 736–745.
- Buchhalter I, Hutter B, Alioto TS, Beck TA, Boutros PC, Brors B, Butler AP, Chotewutmontri S, Denroche RE, Derdak S, et al. 2014. A comprehensive multicenter comparison of whole genome sequencing pipelines using a uniform tumor-normal sample pair. *bioRxiv* doi: 10.1101/013177.
- Canales RD, Luo Y, Willey JC, Austerhammer B, Barbacioru CC, Boysen C, Hunkapiller K, Jensen RV, Knight CR, Lee KY, et al. 2006. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol* **24**: 1115–1122.
- The Cancer Genome Atlas Network. 2012a. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**: 330–337.
- The Cancer Genome Atlas Network. 2012b. Comprehensive molecular portraits of human breast tumours. *Nature* **490**: 61–70.
- The Cancer Genome Atlas Network. 2015. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**: 576–582.
- The Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**: 1061–1068.
- Chong LC, Albuquerque MA, Harding NJ, Caloian C, Chan-Seng-Yue M, de Borja R, Fraser M, Denroche RE, Beck TA, van der Kwast T, et al. 2014. SeqControl: process control for DNA sequencing. *Nat Methods* **11**: 1071–1075.
- Cooper CS, Eeles R, Wedge DC, Van Loo P, Gundem G, Alexandrov LB, Kremeyer B, Butler A, Lynch AG, Camacho N, et al. 2015. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat Genet* **47**: 367–372.
- Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, Mustonen V, Gonzalez-Perez A, Pearson J, Sander C, et al. 2015. Pathway and network analysis of cancer genomes. *Nat Methods* **12**: 615–621.
- Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. 2012. The genomic and

- transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**: 346–352.
- Daley T, Smith AD. 2013. Predicting the molecular complexity of sequencing libraries. *Nat Methods* **10**: 325–327.
- de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, Jamal-Hanjani M, Shafi S, Murugaesu N, Rowan AJ, et al. 2014. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**: 251–256.
- Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. 2015. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol* **16**: 35.
- Dvinge H, Git A, Graf S, Salmon-Divon M, Curtis C, Sottoriva A, Zhao Y, Hirst M, Armitage J, Miska EA, et al. 2013. The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature* **497**: 378–382.
- Eirew P, Steif A, Khattra J, Ha G, Yap D, Farahani H, Gelmon K, Chia S, Mar C, Wan A, et al. 2015. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518**: 422–426.
- Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, Bare JC, P'ng C, Waggott D, Sabelnykova VY, et al. 2015. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* **12**: 623–630.
- Finak G, Bertos N, Pepin F, Sadokova S, Souleimanova M, Zhao H, Chen H, Omeroglu G, Meterissian S, Omeroglu A, et al. 2008. Stromal gene expression predicts clinical outcome in breast cancer. *Nat Med* **14**: 518–527.
- Fox NS, Starmans MH, Haider S, Lambin P, Boutros PC. 2014. Ensemble analyses improve signatures of tumour hypoxia and reveal inter-platform differences. *BMC Bioinformatics* **15**: 170.
- Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, et al. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**: 883–892.
- Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, Fisher R, McGranahan N, Matthews N, Santos CR, et al. 2014. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multi-region sequencing. *Nat Genet* **46**: 225–233.
- Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson J, et al. 2013. Computational approaches to identify functional genetic variants in cancer genomes. *Nat Methods* **10**: 723–729.
- Govind SK, Zia A, Hennings-Yeomans PH, Watson JD, Fraser M, Anghel C, Wyatt AW, van der Kwast T, Collins CC, McPherson JD, et al. 2014. ShatterProof: operational detection and quantification of chromothripsis. *BMC Bioinformatics* **15**: 78.
- Gulati S, Martinez P, Joshi T, Birkbak NJ, Santos CR, Rowan AJ, Pickering L, Gore M, Larkin J, Szallasi Z, et al. 2014. Systematic evaluation of the prognostic impact and intratumor heterogeneity of clear cell renal cell carcinoma biomarkers. *Eur Urol* **66**: 936–948.
- Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JM, Papaemmanuil E, Brewer DS, Kallio HM, Hognas G, Annala M, et al. 2015. The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**: 353–357.
- Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, Melnyk N, McPherson A, Bashashati A, Laks E, et al. 2014. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res* **24**: 1881–1893.
- Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. *Cell* **100**: 57–70.
- Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. *Cell* **144**: 646–674.
- Hoshida Y, Villanueva A, Kobayashi M, Peix J, Chiang DY, Camargo A, Gupta S, Moore J, Wrobel MJ, Lerner J, et al. 2008. Gene expression in fixed tissues and outcome in hepatocellular carcinoma. *N Engl J Med* **359**: 1995–2004.
- Hudson TJ, Anderson W, Artz A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, et al. 2010. International network of cancer genome projects. *Nature* **464**: 993–998.
- Kratz JR, He J, Van Den Eeden SK, Zhu ZH, Gao W, Pham PT, Mulvihill MS, Ziaei F, Zhang H, Su B, et al. 2012. A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. *Lancet* **379**: 823–832.
- Lalonde E, Ishkanian AS, Sykes J, Fraser M, Ross-Adams H, Erho N, Dunning MJ, Halim S, Lamb AD, Moon NC, et al. 2014. Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. *Lancet Oncol* **15**: 1521–1532.
- Lapuk AV, Wu C, Wyatt AW, McPherson A, McConeghy BJ, Brahmabhatt S, Mo F, Zoubeydi A, Anderson S, Bell RH, et al. 2012. From sequence to molecular pathology, and a mechanism driving the neuroendocrine phenotype in prostate cancer. *J Pathol* **227**: 286–297.
- Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**: 214–218.
- Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**: 495–501.
- Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, et al. 2015. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* **47**: 106–114.
- Loeb S, Vonesh EF, Metter EJ, Carter HB, Gann PH, Catalona WJ. 2011. What is the true number needed to screen and treat to save a life with prostate-specific antigen testing? *J Clin Oncol* **29**: 464–467.
- Margolin AA, Bilal E, Huang E, Norman TC, Ottestad L, Mechem BH, Sauerwine B, Kellen MR, Mangravite LM, Furia MD, et al. 2013. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci Transl Med* **5**: 181re1.
- Monzon FA, Lyons-Weiler M, Buturovic LJ, Rigl CT, Henner WD, Sciulli C, Dumur CI, Medeiros F, Anderson GG. 2009. Multicenter validation of a 1,550-gene expression profile for identification of tumor tissue of origin. *J Clin Oncol* **27**: 2503–2508.
- Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, et al. 2011. Tumour evolution inferred by single-cell sequencing. *Nature* **472**: 90–94.
- Oesper L, Satas G, Raphael BJ. 2014. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics* **30**: 3532–3540.
- Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, et al. 2015. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**: 124–128.
- Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Cote A, Shah SP. 2014. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* **11**: 396–398.
- Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, et al. 2009. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**: 809–813.
- Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G, et al. 2012. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**: 395–399.
- Shedden K, Chen W, Quirk R, Ghosh D, Macdonald J, Cho KR, Giordano TJ, Gruber SB, Fearon ER, Taylor JM, et al. 2005. Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data. *BMC Bioinformatics* **6**: 26.
- Shi L, Tong W, Fang H, Scherf U, Han J, Puri RK, Frueh FW, Goodsaid FM, Guo L, Su Z, et al. 2005. Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics* **6**: pS12.
- Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, et al. 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**: 1151–1161.
- Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu TM, Goodsaid FM, Pusztai L, et al. 2010. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol* **28**: 827–838.
- Starmans MH, Fung G, Steck H, Wouters BG, Lambin P. 2011. A simple but highly effective approach to evaluate the prognostic performance of gene expression signatures. *PLoS One* **6**: e28320.
- Starmans MH, Pintilie M, John T, Der SD, Shepherd FA, Jurisica I, Lambin P, Tsao MS, Boutros PC. 2012. Exploiting the noise: improving biomarkers with ensembles of data analysis methodologies. *Genome Med* **4**: 84.
- Starmans MH, Pintilie M, Chan-Seng-Yue M, Moon NC, Haider S, Nguyen F, Lau SK, Liu N, Kasprzyk A, Wouters BG, et al. 2015. Integrating RAS status into prognostic signatures for adenocarcinomas of the lung. *Clin Cancer Res* **21**: 1477–1486.
- Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, et al. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**: 27–40.
- Tofigh A, Suderman M, Paquet ER, Livingstone J, Bertos N, Saleh SM, Zhao H, Souleimanova M, Cory S, Lesurf R, et al. 2014. The prognostic ease and difficulty of invasive breast carcinoma. *Cell Rep* **9**: 129–142.
- Tran B, Brown AM, Bedard PL, Winquist E, Goss GD, Hotte SJ, Welch SA, Hirte HW, Zhang T, Stein LD, et al. 2013. Feasibility of real time next generation sequencing of cancer genes linked to drug response: results from a clinical trial. *Int J Cancer* **132**: 1547–1555.

- Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, Jane-Valbuena J, Friedrich DC, Kryukov G, Carter SL, et al. 2014. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* **20**: 682–688.
- van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**: 530–536.
- Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**: i237–i245.
- Venet D, Dumont JE, Detours V. 2011. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* **7**: e1002240.
- Vollan HK, Rueda OM, Chin SF, Curtis C, Turashvili G, Shah S, Lingjaerde OC, Yuan Y, Ng CK, Dunning MJ, et al. 2015. A tumor DNA complex aberration index is an independent predictor of survival in breast and ovarian cancer. *Mol Oncol* **9**: 115–127.
- Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H, et al. 2014. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**: 155–160.
- Wu G, Feng X, Stein L. 2010. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol* **11**: R53.
- Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh CH, Zhang C, Ren X, Protopopov A, Chin L, et al. 2013. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**: 919–929.
- Yuan Y, Failmezger H, Rueda OM, Ali HR, Graf S, Chin SF, Schwarz RF, Curtis C, Dunning MJ, Bardwell H, et al. 2012. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci Transl Med* **4**: 157ra143.
- Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, Byers LA, Xu Y, Hess KR, Diao L, et al. 2014. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol* **32**: 644–652.
- Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, et al. 2013. Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**: 1134–1140.
- Zhang J, Fujimoto J, Wedge DC, Song X, Seth S, Chow CW, Cao Y, Gumbs C, Gold KA, Kalhor N, et al. 2014. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346**: 256–259.
- Zhu Q, Miecznikowski JC, Halfon MS. 2010. Preferred analysis methods for Affymetrix GeneChips. II. An expanded, balanced, wholly-defined spike-in dataset. *BMC Bioinformatics* **11**: 285.