Prospecting whole cancer genomes

A new suite of studies from the Pan Cancer Analysis of Whole Genomes (PCAWG) Consortium provides the most detailed resolution of cancer genomes to date, extending our knowledge of driver genes, mutational features, structural alterations and more. Kreisberg, Ideker, Mills and Meric-Bernstam discuss the foundational and translational insights gained from this project.

Jason F. Kreisberg, Trey Ideker, Funda Meric-Bernstam and Gordon Mills

From the bench: Jason F. Kreisberg and Trey Ideker

Cancer is driven by mutations acquired throughout the genome. Previous large-scale sequencing efforts by The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) sequenced protein-coding regions in patient-matched pairs of tumor and healthy tissue[1,2]. Protein-coding regions, however, account for only about 1% of the entire human genome. In a collaboration spanning both TCGA and ICGC, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium has recently performed a comprehensive meta-analysis of whole genome sequences from matching tumor and healthy samples from over 2,600 patients spanning 38 different types of cancer[3]. Extending earlier non-coding analysis of cancer genomes[4,5], this recent work aggregates the raw sequencing data from many groups and then uses a common analytical pipeline, which results in a set of high-quality somatic-mutation calls.

One of the main motivations for these large-scale genome-sequencing studies is to identify cancer driver genes—genes that when mutated provide cells with a growth advantage. After taking into account background mutation rates and other factors, many frequently mutated genes can be shown to function as cancer drivers. On the other hand, there are also many well-known cancer drivers that are not frequently mutated in tumor genomes, at least not above a level that can be detected as statistically significant[6]. Identifying driver genes from this so-called 'long tail' of rare mutations is a major bioinformatics challenge, with network-based approaches providing potential solutions. In addition, whole-genome studies can reveal new ways in which cancer genes and pathways are dysregulated by somatic mutations in noncoding regions such as promoters and enhancers—regions overlooked by earlier studies focused on protein-coding regions. Presented with this new dataset, collaborators in the PCAWG project tackled the challenge of extending our understanding of cancer driver genes from multiple analytical perspectives.

Raphael and colleagues used an ensemble approach to identify genes and non-coding elements as cancer drivers if they scored as significant in at least four of seven methods that leverage information from molecular pathways or protein-interaction networks[7]. This study identified 87 driver genes with coding variants, including 31 not identified when single-gene approaches were applied to this dataset[8]; this highlights the value of using outside biological knowledge to investigate cancer mutations. In addition, 93 driver genes were identified on the basis of nearby non-coding variants; only three of these genes had been previously identified. Although some pathways were altered mostly by coding mutations, other pathways were impacted by mutations in both coding and non-coding DNA regions. RNA-splicing pathways in particular were found to be impacted mainly by non-coding mutations.

Reimand and colleagues focused on a single approach, ActivePathways, to identify cancer driver genes and pathways altered by either coding or non-coding mutations[9]. ActivePathways uses data-fusion techniques to integrate the separate *P* values calculated by the PCAWG Consortium for implicating coding sequences, untranslated regions, promoters and enhancers in cancer. Since these separate values are combined into a single per-gene score, this approach can elevate multiple borderline signals into a highly significant hit or, conversely, can penalize a single strong value for a gene that is not backstopped by the other layers of data. Subsequently, single-gene scores were used to identify enriched gene sets, as defined by resources such as biological processes from the Gene Ontology and molecular pathways from the Reactome database. Applying ActivePathways to the adenocarcinoma samples in the PCAWG dataset, the authors identified 333 pathway-associated candidate genes, including 60 of the 64 driver genes identified by the PCAWG Consortium[8] and 47 genes not reported by the other Consortium papers but listed in the COSMIC Cancer Gene Census (CGC) Database.

Both approaches above use *P* values calculated by the PCAWG Consortium as the starting point from which to identify new driver genes. In parallel, Stein and colleagues sought to identify cancer drivers using a novel approach that takes into account both mutational burden and functional impacts[10]. This method, DriverPower, builds a global model of the background mutation rate from more than a thousand genomic features. DriverPower identified 271 coding and 95 non-coding driver variants in the PCAWG dataset, most with support from PCAWG's companion studies or the CGC. For the 11 coding and 17 non-coding driver candidates without PCAWG or CGC support, the group was able to identify literature or outside experimental evidence in support of 8 of these candidates. Compared with other algorithms for identifying driver genes, DriverPower had the best balance of recall and precision.

Stein and colleagues also built a deep-learning classifier to identify tissue of origin using features derived from PCAWG whole-genome sequences[11]. In multiple independent cohorts, the classifier achieved an accuracy of 88% in primary samples and of 83% in metastatic samples. Notably, features from driver genes and pathways decreased the accuracy of the classifier, which suggests that the tissue-classification decision is coming predominantly from background genome sequence. One application for this model is to help classify the 3–5% of metastatic disease in which the primary site is unknown, which could then help guide treatment.

Importantly, the PCAWG project generated an enormous dataset of 725 terabytes, which cannot easily or legally be moved from one data center to another. To provide international collaborators access to this massive compendium of data,

Korbel and colleagues developed Butler, a computational tool to manage workflows using cloud-based computation[12]. Butler was deployed to the EMBL-EBI Embassy Cloud, which contains 1,500 computational cores, 5.5 terabytes of RAM, 40 terabytes of local solid-state drive storage and 1 petabyte of shared storage, all accessible over a 10-gigabit network. To help prevent the job failures that are all too common when biological data of variable quality and from various sources are handled, Butler contains anomaly-detection modules that can automatically detect and resolve critical issues.

Exciting as these and other studies have been, these datasets are often missing crucial information about clinical presentation, treatment and outcome. Recognizing this shortcoming, the ICGC recently started the Accelerate Research in Genomic Oncology (ARGO) project, which seeks to analyze specimens from patients with cancer through the use of high-quality clinical data (www.icgc-argo.org), ensuring that the next iterations of sequencing initiatives will extend the essential investigations into the genomic basis of cancer laid out by the TCGA and ICGC to date, by integrating this knowledge with clinical information in the future.

From the bedside: Funda Meric-Bernstam and Gordon Mills

The PCAWG project represents the most comprehensive genomics analysis to date, with 2,658 whole genomes across 38 tumor types, and 1,188 transcriptomes from 27 tumor types[3]. This tremendous team-science initiative covers multiple different aspects of genomics, including identification of non-coding mutational drivers, repertoires of mutational signatures, patterns of somatic structural variations, reconstruction of the evolution of mutational processes, genomic alterations underlying transcriptional changes, and new approaches to identify drivers.

In recent decades, the role of genomic information in oncology—through identifying individuals with a genetic risk for cancer, precision early diagnosis and prognosis, and designing therapies targeted at key alterations that drive disease—has been intensely studied and in some cases validated with patient benefit. As with previous sequencing initiatives, although data from the PCAWG project deepens our understanding of the complexity and heterogeneity of cancer, the feasibility of translating these new insights to the clinic remains an enduring challenge.

For example, we still have an incomplete understanding of genetic variation and heritable risk in cancer. Whole-genome resolution yielded insights into how germline variants, including those in *BRCA1* and *BRCA2*, contribute to the selection of subsequent somatic events[3]. Similarly, genomic reconstruction of evolutionary histories of cancer revealed that the latency between the first genomic event and diagnosis varies greatly between tumor lineages[13], which suggests that early detection approaches will need to be conditioned on the variation between the first event and diagnosis in different tumor lineages. The earliest somatic mutations encompass a discrete set of potential drivers[13]; thus, prevention or management of early disease could be less fraught with complexity arising from the mutational heterogeneity that plagues targeted therapy of advanced disease.

The large-scale genomic data brought about new approaches to identify aberrations that may drive tumorigenesis. However, although the compendium of potential drivers has been extended, looking forward, we are presented with the non-trivial task of determining the contribution of these driver events to clinical management. For instance, combining coding and non-coding elements, genomes contained on average four to five independent putative driver aberrations[3]. This provides insight into why our current single-coding-region alteration-matched therapy approaches have had a more limited impact on cancer outcomes than expected. Interestingly, the key drivers do not appear to change substantially during tumor evolution in as many as 60% of tumors, indicative of potential routes to effective therapy[3]. Despite the deep PCAWG analysis, 5% of tumors lacked detectable genomic drivers. These cases were enriched in certain tumor types, reinforcing previous experience that clinical utility of genomic sequencing varies markedly by disease.

Exome sequencing is increasingly being used to guide diagnostics and selection of treatment. Of the 5,913 potential driver point mutations, 785 (13%) localize to non-coding regions, and 25% of tumors bear at least one putative non-coding driver mutation that would be missed by exome sequencing[8]. This raises the question of whether this additional information would refine clinical decision-making to a degree that would warrant routine whole-genome sequencing of tumors. Although this possibility may seem appealing, in practical terms it is likely premature. These new findings emphasize the incredible heterogeneity of cancer, not only with variation between tumor types, but also between patients within the same tumor lineage. Furthermore, non-coding driver mutations—in particular, those that are not in proximity to coding genes—are uncommon and rarely recurrent, and few, if any, can be targeted by our current catalog of anti-cancer drugs. Translating many of these findings into routine practice will require further understanding of whether the aberrations detected by whole-genome sequencing have therapeutic potential, while we concurrently build much more sophisticated point-of-care analytics and decision support systems.

Transcriptomic profiling identified several categories of RNA alterations associated with germline and somatic DNA alterations[14]. Copy-number changes were major determinants of gene expression. Many somatic mutations affected RNA splicing and likely protein function or expression. 731 genes were recurrently and heterogeneously altered at the RNA level, including splicing, alternative promoter usage, single-nucleotide variants, RNA editing and fusions. Importantly, as fusion genes, including *NTRK1* or *BCR-ABL1*, have exemplified successful targets for therapy, the identification of 2,372 new cancer-specific fusions provides an unexplored suite of potential targets, albeit with few being highly recurrent[14]. *CDK12* was one of the most frequently altered genes, which is noteworthy as *CDK12* aberrations are being explored as biomarkers for response to inhibitors of poly(ADP-ribose) polymerase and immune checkpoints, and a number of CDK12 inhibitors are in preclinical development[15]. Comprehensive analysis of *CDK12* DNA and RNA aberrations could facilitate the identification of patients likely to benefit from CDK12 inhibitors alone or in combination. Implementation of RNA-sequencing approaches into clinical use has the potential to rapidly identify fusion genes as well aberrations in RNA processing and activity that alter function. However, substantial optimization of protocols to analyze, report and interpret results is still required before these techniques can fulfill their clinical promise.

New technologies and collaborative efforts continue to increase our resolution and understanding of the cancer genome. However, we contend that the community still has much to discover about each cancer type and alteration. The tremendous effort by the PCAWG consortium to catalog coding and noncoding genomic drivers with a link to transcriptomics is to be lauded. However, cell behavior is mediated by protein expression and function, and the majority of modern cancer drugs target protein function, which is poorly reflected by genomics and transcriptomics. Extensive investigation of the functional consequences

of the genomic aberrations will be required to realize the clinical impact of the analysis. Additionally, even a dataset of this scale has limitations. As an example, some tumor types had too few cases to identify all key drivers. Most PCAWG samples were primary tumors, but analysis of metastatic tumors or those obtained after intervening therapy is likely to identify additional clinically relevant drivers. Importantly, future genomic profiling studies will have greater clinical impact if linked to high-quality clinical annotation, so that findings can be rapidly translated into biomarkers for early detection, prognosis and therapeutic sensitivity or resistance. Encouragingly, the ICGC-ARGO initiative and the broader cancer community are working to develop such resources now.

A salient observation of genes uncovered as top cancer drivers in this new dataset is their familiarity. While we may now understand new ways these genes are deregulated, the most frequently aberrant drivers of disease were already well known. Frustratingly, most of these genes, such as *TP53*, still are not 'actionable'—that is to say, not targetable directly or indirectly with approved or investigational agents—and for the few that are, we often do not have therapies that achieve durable control. At ClinicalTrials.gov, there are more trials ongoing with 'next-in-class' therapies than with novel strategies that address untargeted known common drivers. In that light, the body of work that the PCAWG effort represents is inspiring. Realization of the investigatory and interventional potential afforded us by the cancer sequencing data now at our disposal will require a coordinated community effort to translate this information into better patient outcomes. ❐

Jason F. Kreisberg[1], Trey Ideker[1]✉, Funda Meric-Bernstam[2]✉ and Gordon Mills[3]✉

[1]Department of Medicine, University of California San Diego, La Jolla, CA, USA. [2]UT MD Anderson Cancer Center, Houston, TX, USA. [3]Knight Cancer Institute Oregon Health and Sciences University, Portland, OR, USA.

✉*e-mail: tideker@ucsd.edu; fmeric@mdanderson.org; millsg@ohsu.edu*

Published online: 20 March 2020
https://doi.org/10.1038/s43018-020-0045-3

References

1. Hoadley, K. A. et al. *Cell* **173**, 291–304.e6 (2018).
2. International Cancer Genome Consortium. et al. *Nature* **464**, 993–998 (2010).
3. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. *Nature* **578**, 82–93 (2020).
4. Zhang, W. et al. *Nat. Genet.* **50**, 613–620 (2018).
5. Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. *Nat. Genet.* **47**, 710–716 (2015).
6. Hofree, M. et al. *Nat. Commun.* **7**, 12096 (2016).
7. Reyna, M. A. et al. *Nat. Commun.* **11**, 729 (2020).
8. Rheinbay, E. et al. *Nature* **578**, 102–111 (2020).
9. Paczkowska, M. et al. *Nat. Commun.* **11**, 735 (2020).
10. Shuai, S., Gallinger, S. & Stein, L. *Nat. Commun.* **11**, 734 (2020).
11. Jiao, W. et al. *Nat. Commun.* **11**, 728 (2020).
12. Yakneen, S. et al. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-019-0360-3 (2020).
13. Gerstung, M. et al. *Nature* **578**, 122–128 (2020).
14. PCAWG Transcriptome Core Group. *Nature* **578**, 129–136 (2020).
15. Lui, G. Y. L., Grandori, C. & Kemp, C. J. *J. Clin. Pathol.* **71**, 957–962 (2018).

Competing interests

G.B.M. is on the scientific advisory board or is a consultant for AstraZeneca, Chrysallis Biotechnology, ImmunoMET, Ionis, Lilly, PDX Pharmaceuticals, Signalchem Lifesciences, Symphogen and Tarveda; has stock, options or other financial interests in Catena Pharmaceuticals, ImmunoMet, SignalChem and Tarveda; has a licensed technology HRD assay to Myriad Genetics, and DSP patents with Nanostring; has research sponsored by Nanostring Center of Excellence and Ionis (provision of tool compounds only). F.M.-B is on the scientific advisory board or is a consultant for Immunomedics, Inflection Biosciences, Mersana Therapeutics, Puma Biotechnology, Seattle Genetics, Silverback Therapeutics, Spectrum Pharmaceuticals, Aduro BioTech, DebioPharm, eFFECTOR Therapeutics, F. Hoffman-La Roche, Genentech, IBM Watson, Jackson Laboratory, Kolon Life Science, OrigiMed, PACT Pharma, Parexel International, Pfizer, Samsung Bioepis, Seattle Genetics, Tyra Biosciences, Xencor and Zymeworks; has research sponsored by Aileron Therapeutics, AstraZeneca, Bayer Healthcare Pharmaceutical, Calithera Biosciences, Curis Inc., CytomX Therapeutics, Daiichi Sankyo, Debiopharm International, eFFECTOR Therapeutics, Genentech, Guardant Health, Millennium Pharmaceuticals, Novartis, Puma Biotechnology and Taiho Pharmaceutical; and receives honoraria from Chugai Biopharmaceuticals, Mayo Clinic, Rutgers Cancer Institute of New Jersey and UT Health San Antonio. T.I. is co-founder of and has an equity interest in Data4Cure; and has an equity interest in Ideaya BioSciences; the terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.