News & views

Tumour genetics

Global cancer genomics project comes to fruition

Marcin Cieslik & Arul M. Chinnaiyan

A massive international effort has yielded multifaceted studies of more than 2,600 tumours from 38 tissues, generating a wealth of insights into the genetic basis of cancer.

See p.82, p.94, p.102, p.112, p.122 & p.129

Comprehensive genomic characterization of tumours became a major goal of cancer researchers as soon as the first human genome had been sequenced in 2001. Since then, advances in sequencing technology and analytical tools have allowed this research field to flourish. In six papers[1–6] in this issue of *Nature*, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium presents the most comprehensive and ambitious meta-analysis of cancer genomes so far. Unlike previous efforts that focused largely on protein-coding regions of the cancer genome, PCAWG analyses whole genomes. Each article scrutinizes an important aspect of cancer genetics — together, their findings will be key to understanding the full genetic complexity of cancer.

Before discussing the impact of these analyses, it is crucial to highlight the massive amount of data and the complex organizational framework that underpin the PCAWG endeavour. The project involved an interdisciplinary group of scientists from 4 continents, with 744 affiliations between them, who had to overcome major technical, legal and ethical challenges to carry out distributed analyses while protecting patient data. Researchers were divided into 16 working groups, each focused on distinct facets of cancer genomics — assessing the recurrence of mutations, for instance, or inferring tumour evolution.

Altogether, the consortium performed integrative analyses of 38 tumour types. The group sequenced 2,658 whole-cancer genomes (Fig. 1), alongside matched samples of non-cancerous cells from the same individuals. These data were complemented by 1,188 transcriptomes — the sequences and abundances of RNA transcripts in a tumour.

These efforts involved extensive quality control and coordinated data processing, as well as massive, systematic experimental validation of the computational pipelines used to detect mutations. Many computational algorithms and pipelines were used and compared in concert. This required hundreds of terabytes of data, spread across multiple data centres, and probably millions of processing hours — all facilitated by cloud computing. Notably, the PCAWG efforts provide a prime example of how cloud computing can make international collaboration possible and help to advance data-intensive fields.

The first of the current papers[1] (page 82) gives an overview of the breadth and depth of the PCAWG data set. The consortium reports that, on average, each cancer genome carries four or five driver mutations, which provide cancer cells with a selective advantage. Only 5% of tumours studied had no identified driver aberrations. By contrast, many cancers exhibited hallmarks of genomic catastrophes called chromoplexy (17.8% of tumours) and chromothripsis (22.3%), which result in major structural changes to the genome.

The other five papers each delve into a different aspect of the data set in more detail. For instance, in the second paper, Rheinbay *et al.*[2] (page 102) set out to identify genetic drivers in non-coding DNA. This is an ambitious undertaking, because it is substantially more difficult to accurately detect mutations in non-coding regions than in coding regions, or to assess their recurrence. The authors used careful modelling to exclude artefacts and systematically identify non-coding driver mutations.

Their results call into question previously reported non-coding drivers, such as the long non-coding RNAs *NEAT1* and *MALAT1*, but also reveal new ones. For example, the authors report a recurrent mutation in a non-coding region of the key tumour-suppressor gene *TP53*. They also found relatively frequent mutations in non-coding regions of the telomerase gene *TERT* that result in overexpression of the telomerase enzyme (which

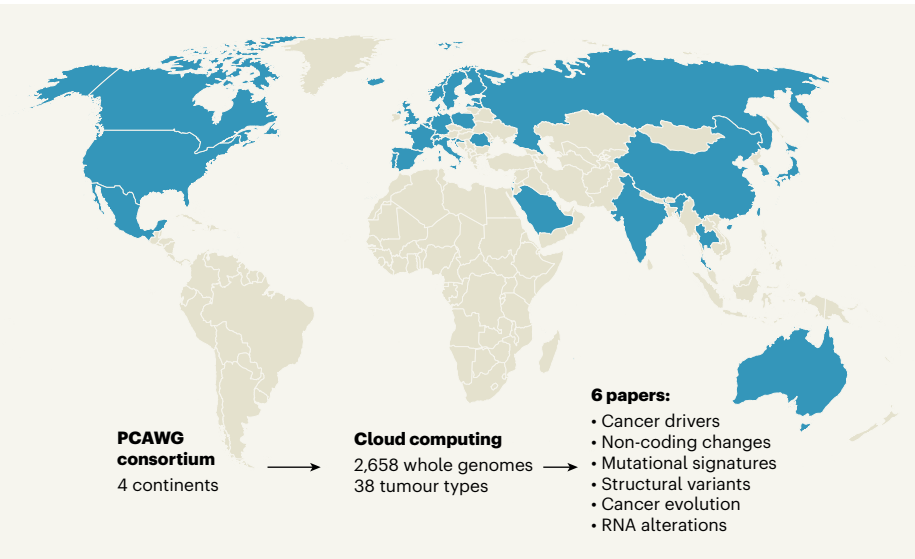

Figure 1 | A worldwide effort to tackle cancer. The Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium is a group of cancer researchers from four continents (blue). The group sequenced and analysed 2,658 whole-cancer genomes from 38 types of tumour. The huge amount of data involved in the effort required sophisticated cloud-computing approaches. Six papers[1–6] from the PCAWG now describe different aspects of the analyses performed. (Nature publications remain neutral with regard to contested jurisdictional claims in published maps.)

helps tumour cells to divide uncontrollably), mirroring the high (12%) prevalence of telomerase mutations found in a previous pan-cancer study of more-advanced (metastatic) tumours[7]. Although the study could not rule out the existence of other non-coding drivers, it decisively shows that this type of mutation is not common.

In the third and fourth papers, Alexandrov *et al.*[3] (page 94) and Li *et al.*[4] (page 112) focus on genomic aberrations called signatures. Different processes, such as defective DNA-repair mechanisms or exposure to environmental mutagens, produce these characteristic patterns of DNA aberrations. Large genomic data sets are crucial if we are to refine known mutational signatures and discover new ones. Impressively, between them, Alexandrov *et al.* and Li *et al.* identify 97 signatures. This expansion on previous work encompasses not only conventional single-nucleotide signatures, but also signatures involving multi-nucleotide variants and small insertions or deletions of DNA.

Notably, Li and colleagues are among the first to uncover reproducible signatures involving structural variants (SVs) — rearrangements of large portions of the genome. The process was much more intricate than that for identifying mutational signatures because of the diversity and complexity of SVs.

Through a series of mutation-subgrouping steps, the researchers identified 16 SV signatures, revealing, for example, a putative mechanistic link between two SVs, deletions and reciprocal inversions (the last of which involves a reversal of the orientation of a segment of DNA). They also gained insights into the roles of all 16 signatures in cancer. Mutations in certain DNA-repair genes were shown to associate with characteristic cancer signatures. For instance, the consortium found that mutations in the gene *CDK12* associate with tandem stretches of duplicated DNA, and that truncated variants of the DNA-repair enzyme MBD4 co-occur with a distinct mutational signature involving DNA sequences called CpG sites. Altogether, these new signatures lay the foundation for understanding mechanisms of cancer development, and the role of mutagenic exposures in this process.

The idea that cancer develops through an evolutionary process was first presented in 1976 (ref. 8). Since then, cancer evolution has been characterized in terms of random mutations and natural selection. A cancer cell harbouring a mutation that confers high fitness proliferates rapidly, becoming the most prominent cell clone in the population. This phenomenon, called a clonal sweep, occurs in recurring cycles during cancer growth. Cancer evolution is most effectively studied by sequencing multiple regions of a tumour over time, but it can also be reconstructed from a single biopsy — the approach taken by Gerstung *et al.*[5] (page 122) in the fifth paper.

The authors introduce the concept of 'molecular time' to classify clonal and subclonal mutations. They reasoned that subclonal mutations, which are present in only a subset of a tumour's cells, must have arisen late in the cancer's evolution. They classify clonal mutations, which are present in all of a tumour's cells, as early or late, depending on whether the mutations arose before or after the clone underwent copy-number gains (an increase in the number of copies of a gene or chromosomal region). The researchers aggregated evolutionary data from multiple tumours, allowing them to identify common mutational trajectories such as *APC*–*KRAS*–*TP53* progression[9], which describes the typical sequence in which mutations arise in colorectal cancer.

Gerstung *et al.* found that the driver mutations that most commonly occur in a given cancer also tend to occur the earliest. Similarly, if copy-number gains are highly recurrent in a particular cancer type, they tend to occur early. For example, a copy-number gain in part of chromosome 5 is common in clear cell kidney cancer, and tends to arise early in the disease's evolution. Conversely, whole-genome duplication is a relatively late event in this cancer. Finally, the researchers found that mutational signatures change over time in at least 40% of tumours. These changes reflect a decreasing role for environmental exposures in disease progression and an increase in the frequency and severity of DNA-repair defects. Overall, the group's findings suggest that driver mutations can occur years before cancer is diagnosed, which has implications for early detection and biomarker development.

"The broad availability and quality of the new data set will almost certainly spur a wave of biological insights."

In the final paper (page 129), the PCAWG Transcriptome Core Group and their colleagues[6] made use of the 1,188 PCAWG samples that had matched transcriptome data, to functionally link DNA and RNA alterations. The group found associations between hundreds of single-nucleotide DNA mutations and the expression of nearby genes. However, larger copy-number alterations were the main drivers of gene-expression changes in cancer cells. Mutations were also associated with changes in transcript structure, such as the formation of a new protein-coding region (an exon) within a non-coding region (an intron).

The authors also characterized the frequency of bridged fusions — a phenomenon in which two genes become fused owing to a third, intervening fragment of DNA. Finally, although 87 of the 1,188 samples analysed did not have a driver alteration at the DNA level, the group showed that every one of these had an RNA-level alteration. Together, these insights illustrate the power of integrated RNA- and DNA-sequencing analysis for cancer studies[10].

These six papers, together with companion papers being co-published elsewhere (see go.nature.com/3boajsm), represent a milestone in cancer and cloud genomics. By focusing on inferences, the PCAWG successfully expands on a decade of cancer sequencing studies that were rooted largely in observations. It is worth noting that, although inferential analyses offer a deeper look at cancer than do descriptive studies, their results are also associated with a higher degree of uncertainty.

The broad availability and quality of the PCAWG data set will almost certainly spur a wave of biological insights and methodological developments. Integration with other functional genomic data sets, for example probing the 3D organization of the genome, will also undoubtedly provide further understanding of the causes and consequences of genetic aberrations.

The biggest limitation of the current studies is the lack of clinical data concerning patient outcomes and treatments. Such data would allow researchers to identify the genetic changes that can predict clinical outcomes. Fortunately, a project called the International Cancer Genome Consortium–Accelerate Research in Genomic Oncology (ICGC–ARGO) is under way that will create such a resource for more than 100,000 people with cancer.

Ultimately, the PCAWG brought together thousands of scientists, working together to achieve its aims. The long-term impact of these efforts will not be limited to the findings published today, but will also come from the collaborations that have formed and the knowledge exchanges that have taken place between members of this global consortium of researchers.

Marcin Cieslik and **Arul M. Chinnaiyan** are in the Michigan Center for Translational Pathology, Rogel Cancer Center, University of Michigan, Ann Arbor, Michigan 48109, USA. **M.C.** is also in the Department of Computational Medicine and Bioinformatics, University of Michigan. **A.M.C.** is also at the Howard Hughes Medical Institute.
e-mails: mcieslik@med.umich.edu; arul@med.umich.edu

1. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. *Nature* **578**, 82–93 (2020).
2. Rheinbay, E. *et al. Nature* **578**, 102–111 (2020).
3. Alexandrov, L. B. *et al. Nature* **578**, 94–101 (2020).
4. Li, Y. *et al. Nature* **578**, 112–121 (2020).
5. Gerstung, M. *et al. Nature* **578**, 122–128 (2020).
6. PCAWG Transcriptome Core Group *et al. Nature* **578**, 129–136 (2020).
7. Priestley, P. *et al. Nature* **575**, 210–216 (2019).
8. Nowell, P. C. *Science* **194**, 23–28 (1976).
9. Fearon, E. R. & Vogelstein, B. *Cell* **61**, 759–767 (1990).
10. Robinson, D. R. *et al. Nature* **548**, 297–303 (2017).