Pan-cancer analysis of whole genomes

Authors: Peter J Campbell (1,2) * ; Gaddy Getz (3,4,5,6) *; Joshua M Stuart (7) *; Jan O Korbel (8,9) *; Lincoln D Stein (10,11) * on behalf of the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network §

(1)Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK
(2)Department of Haematology, University of Cambridge, Cambridge CB2 2XY, UK
(3)The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02124, USA
(4)Massachusetts General Hospital Center for Cancer Research, Charlestown, Massachusetts 02129, USA
(5)Massachusetts General Hospital, Department of Pathology, Boston, Massachusetts 02114, USA
(6)Harvard Medical School, Boston, 02215, USA
(7)Department of Biomolecular Engineering and CBSE, University of California Santa Cruz, Santa Cruz, CA 95064, USA
(8)European Molecular Biology Laboratory (EMBL), Genome Biology Unit, 69117 Heidelberg, Germany
(9)EMBL, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire CB10 1SA, UK
(10) Adaptive Oncology, Ontario Institute for Cancer Research, Toronto, Ontario, Canada, M5G 0A3
(11) Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada, M5S 1A8

* These authors contributed equally to the manuscript
§ Authors listed in appendix
Introduction

Cancer is the second most frequent cause of death worldwide, killing more than 8 million people every year and responsible for 1 in 7 deaths\(^1\). Globally, cancer deaths will increase by more than 50% over the coming decades, attributable to a number of driving forces: an ageing population in high income countries; increased exposure to carcinogens such as tobacco\(^2\), air pollution and asbestos\(^3\) in low- and middle-income countries; declining physical activity with concomitant rise in obesity worldwide; and the continued expansion of the human population. While prevention and treatment of competing causes of mortality, such as cardiovascular disease and infections, have led to major improvements in life expectancy, the gains for cancer mortality have been more modest. For many patients, surgery remains the only curative option, but once the tumour has spread from its original site, cure is often elusive. Nonetheless, in the last 20 years, our deepening understanding of the biology of cancer has enabled development of new therapeutics effective in a handful of cancers\(^4,5\) – it is this success that motivates the desire to systematically characterise cancer biology across all tumour types.

‘Cancer’ is a catch-all term used to denote a set of diseases characterised by autonomous expansion and spread of a somatic clone. To achieve this behavior, the cancer clone must modify multiple cellular pathways that enable it to disregard the normal constraints on cell growth, to modify the local microenvironment favoring its own proliferation, to invade through tissue barriers, to spread to other organs, and to evade immune surveillance\(^6\). No single cellular programme directs these behaviors. Rather there are many different potential abnormalities from which individual cancers draw their own combinations. In that sense, the commonalities of macroscopic features across tumours belie a vastly heterogeneous landscape of cellular abnormalities.

This heterogeneity arises from the fundamentally stochastic nature of Darwinian evolution; a process that operates in somatic cells as much as
species. The preconditions for Darwinian evolution are three: characteristics must vary within a population; this variation must be heritable from parent to offspring; and there must be competition for survival within the population. In the context of somatic cells, heritable variation arises from mutations acquired stochastically throughout life, notwithstanding potential additional contributions from heritable epigenetic variation. A subset of these mutations drive alterations in cellular phenotype, and a small subset of those variants confer an advantage on the clone in its competition to escape the tight physiological controls wired into somatic cells. The mutations conferring selective advantage on the clone we call ‘driver mutations,’ as opposed to the selectively neutral, or possibly slightly deleterious, ‘passenger mutations.’

The discovery that cancers carry recurrent and specific genetic abnormalities in the 1970s\textsuperscript{7} and early 1980s\textsuperscript{8,9} has fuelled four decades of research to define the catalogue of genes and mutations that can drive cancer. This has been accelerated by technological advances in genomic analysis, from gross descriptions of chromosome structure by chromosomal banding\textsuperscript{7} and other cytogenetic techniques, through positional cloning of inherited cancer genes\textsuperscript{10}, low-throughput capillary sequencing\textsuperscript{11} and comparative genomic hybridisation\textsuperscript{12}, to the current era of massively parallel whole genome sequencing\textsuperscript{13–17}. The ever more populous catalogue of cancer genes has opened new therapeutic opportunities, with effective drugs being developed for the \textit{BCR-ABL} fusion gene of chronic myeloid leukaemia, \textit{ERBB2} amplifications of breast cancer and the \textit{BRAF} point mutations of melanoma\textsuperscript{18–20}, amongst others.

**International collaborations to sequence whole cancer genomes**

The advent of massively parallel sequencing promised a future in which the cancer genome was finite and knowable. Early studies showed it was in theory feasible to document every somatic point mutation in a given cancer, every copy number change and every structural variant\textsuperscript{14,15}. In 2008, recognising the opportunity this advance in technology provided,
the global cancer genomics community established The International Cancer Genome Consortium (ICGC) with the goal of systematically documenting the somatic mutations found in 25,000 samples representing all common tumour types\textsuperscript{21}.

The ICGC comprises researchers from The Cancer Genome Atlas (TCGA) in the USA plus those from 17 countries and other jurisdictions in Europe, Asia and the Americas. Each ICGC project is organised around a single tumour type or a set of related types, for which a set of tumour/normal pairs derived from a target of 500 donors were characterised by whole genome sequencing, exome sequencing, transcriptome and/or DNA methylation analysis. The sample size was chosen to provide enough power to detect significantly mutated genes in at least 3\% of patients based on an initial estimate of the background mutation rate.

ICGC samples have been carefully pre-screened by histopathologists and clinicians in order to ensure the accuracy of diagnosis and quality of the sample. Sequencing of both tumour and matched constitutional DNA are required to meet minimum coverage and quality requirements. Following the precepts established in the Human Genome Project, data from ICGC are rapidly released to the wider scientific community under appropriate safeguards to ensure ethical and regulatory compliance\textsuperscript{22}. Since 2008, funding for ICGC projects has amounted to more than USD$900,000,000, with individual funding commitments in some countries being the largest biomedical grants they had ever awarded.

To date, there are 90 ICGC projects, of which 76 have submitted data across 21 primary organ sites and 31 distinct tumour types. At the time of writing, genomic data from 20,343 individual cancer patients were registered in the Data Coordination Center (https://dcc.icgc.org/), of whom 17,570 have molecular data, mostly exomes. Many major breakthroughs in the biology of individual tumour types have emerged from these studies, too numerous to cite exhaustively here, but including discoveries
of new cancer genes and pathways\textsuperscript{23-27}; insights into the underlying mutational processes operative in human cancers\textsuperscript{28-32}; delineation of the patterns of tumour heterogeneity and clonal evolution\textsuperscript{33-36}; and development of genomics approaches to inform cancer prevention\textsuperscript{37} and clinical management of patients with cancer\textsuperscript{38-40}. Many of these discoveries were enabled by novel computational and statistical methods designed to accurately detect various genomic alterations from sequencing data and analyse them across cohorts of patients to extract new biological insights.

**The Pan-Cancer Analysis of Whole Genomes Collaboration**

The early studies from ICGC and TCGA revealed both commonalities and differences of somatic genomic architecture across tumour types. Some cancer genes are mutated in many different tumour types; others are specific to a single histological subtype\textsuperscript{41,42}. All common tumour types are characterised by few frequently mutated genes and many rarely mutated genes; the patterns of co-mutation result in a huge diversity of combinations of driver mutations across individual patients\textsuperscript{43-45}. Some tumours are driven by coding point mutations while others evolve through large-scale restructuring of chromosomes\textsuperscript{46}; some cancer types mutate predominantly tumour suppressor genes\textsuperscript{47} while others have high frequency of driver mutations activating oncogenes\textsuperscript{48}.

Numerous studies point to the relevance of non-coding regions, and projects including ENCODE,\textsuperscript{49} Blueprint\textsuperscript{50} and Epigenome Roadmap\textsuperscript{51} have revealed extensive catalogues of tissue-specific regulatory elements. Transcription factors and other proteins interact with enhancers, silencers, boundary elements, and overall chromatin structure to confer cell-specific regulatory responses, and recent studies have revealed the relevance of this interplay in cancer.\textsuperscript{52-57} Given that cells are pre-wired according to built-in control logics that involve coding and non-coding components, it stands to reason that changes in the DNA that affect these factors may underlie the tissue-specific nature of cancer onset and progression.
Indeed, some evidence points this way, for example there is evidence that epigenetic marks are associated with mutation densities in cancer,\textsuperscript{58,59} whereas cancer-risk associated germline variants typically occur in intergenic regions and show enrichment with enhancers.\textsuperscript{60}

The large number of samples subjected to whole genome sequencing by the ICGC now provides the opportunity to closely examine cancers beyond their protein-coding exomes, which are likely to provide only partial insights into the genomic landscape of cancer. Beyond providing insights into how mutations affect regulatory regions, whole genome sequencing can detail the full repertoire of classes of structural variation in cancers, facilitate resolving mutational processes and signatures acting in these, enable identifying viruses associated with cancers, and allow defining the full repertoire of germline variants in cancer patients. To tackle the various opportunities resulting from numerous cancer whole genome sequencing, 16 thematic Scientific Working Groups were formed and overseen by a Steering Committee for the PCAWG collaboration to pursue a multipronged analysis of the non-coding genome's influence on cancer (Table 1).

The maturing of datasets from individual ICGC and TCGA working groups presented the opportunity to formalise a meta-analysis of whole cancer genomes. However, algorithms for calling somatic mutations were not standardised among the different groups and had evolved considerably in the first few years of the consortium. For cross-tumour comparisons to be meaningful, the core bioinformatic analyses would need to be repeated using gold-standard, benchmarked, version-controlled algorithms. To achieve this, the Pan-Cancer Analysis of Whole Genomes (PCAWG) collaboration was established, comprising about 700 researchers from around the world. A Technical Working Group implemented the core informatics analyses, aggregating the raw sequencing data from the individual tumour type working groups, aligning it to the human genome and delivering a set of high quality somatic mutation calls for downstream
analysis (Figure 1). Scientists from TCGA and ICGC submitted abstracts outlining potential research projects, which were aggregated into 16 thematic Scientific Working Groups. A Steering Committee oversaw the PCAWG collaboration, reporting to the executive committees of ICGC and TCGA.

**Sample collection**

Beginning in early 2015, we inventoried previous submissions of matched tumour/normal whole cancer genomes to the ICGC Data Coordinating Centre and polled ICGC projects for whole genomes that they anticipated completing in the near future. Our PCAWG inclusion criteria for donors included: a matched tumour and normal specimen pair; a minimal set of clinical information including patient age, sex and histopathological diagnosis; and characterisation of tumour and normal whole genomes using Illumina HiSeq platform 100-150bp paired-end sequencing reads. The minimum average depth required was 30 reads per genome base-pair in the tumour sample, and 25 in the normal sample. For the great majority of donors, the paired specimens consisted of a blood sample for the normal specimen, plus a fresh frozen sample of the primary tumour from a resection specimen. In a small number of cases the normal sample originated from tumour-adjacent normal tissue or another non-blood tissue (especially for blood cancers). Most of the tumour samples came from treatment-naïve, primary cancers, but there were a small number of donors with multiple samples of primary, metastatic and/or recurrent tumour. In addition to whole genome sequencing, roughly half of the donors had at least one tumour specimen that had been subjected to whole transcriptome analysis using RNA-sequencing, also centrally collected and re-analysed.

Ultimately, we collected genome data from a total set of 2,834 donors. After an extensive quality assurance process (described below), the data from 176 donors were deemed unusable and were excluded, leaving 2,658 donors, including 2,605 primary tumours and 173 metastases or local
recurrences. Matching normal samples were obtained from blood (2,064 donors), tissue adjacent to the primary (87 donors), or other sites of normal tissue such as bone marrow, lymph node or skin (507 donors). The mean whole genome sequencing coverage in this set was 30 reads per base-pair for normal samples, while tumours had a bimodal coverage distribution with maxima at 38 and 60 reads per base-pair. For 75 donors, QA results were borderline and these donors were flagged in order to caution consortium members that they might be unsuitable for certain types of analysis, leaving a high-quality core of 2,583 donors. RNA-sequencing data was collected on 1222 donors with genome data, including 1178 primary tumours, 67 metastases or local recurrences, and 153 matched normal tissue adjacent to the primary tumour.

Demographically, the cohort included 1469 males (55%) and 1189 females (45%), with a mean age of 56 years (median 60 years; range 1-90 years). By using population ancestry-differentiated single nucleotide polymorphisms (SNPs) derived from the germline calls, we were able to estimate the population ancestry of each donor. The continental ancestry distribution was heavily weighted towards Europeans (77% of total) followed by East Asians (16%), as expected by large contributions from European, North American, and Australian projects (Supplementary Table 1).

**Histopathology harmonisation**

In order to simplify the process of cross-tumour analyses, the PCAWG Pathology and Clinical Correlates Working Group consolidated and harmonised the histopathology descriptions of the tumour samples, using the icd-0-3 tumour site type controlled vocabulary (https://seer.cancer.gov/icc-o-3/) as its basis, in consultation with the leads of each of the contributing projects and a small group of expert anatomic pathologists. We described each tumour type using a four-tier hierarchical system consisting of Embryonic Origin (Mesoderm, Ectoderm or Endoderm), Organ System (such as Breast), Major Histologic Type (for
example, Adenocarcinoma), and Major Histological Subtype (such as Infiltrating duct carcinoma). In addition, each tumour type was assigned a short abbreviation (e.g., Breast-AdenoCa) and a standard colour for use in charts and tables. Overall, we established 39 distinct tumour types in the PCAWG data set (Table 2). The largest tumour type cohorts were hepatocellular carcinoma (Liver-HCC: 318 donors, 327 tumour specimens), pancreatic adenocarcinoma (Panc-AdenoCa: 239 donors, 241 specimens), and prostate cancer (Prost-AdenoCa: 210 donors, 286 specimens). Twelve tumour types had fewer than 20 representatives, including lobular carcinoma of the breast, cervical adenocarcinoma, and benign neoplasms of bone and cartilage. These tumour types, comprising a total of 56 specimens, were excluded from tumour-type specific cohort analyses due to lack of statistical power, but were included in pan-cancer analyses.

**Uniform processing and variant calling**

In order to generate a consistent callset that could be used for cross-tumour type analysis, we analysed all samples using a uniform set of algorithms for alignment, variant calling, and quality control. We used the BWA-Mem algorithm to align each tumour and normal sample to human reference build hs37d50. Somatic mutations were identified in the aligned data using three established pipelines, run independently on each tumour/normal pair. Each of the three pipelines, labeled “Sanger”, “EMBL/DKFZ” and “Broad” after the computational biology groups that created and/or assembled them, consisted of multiple software packages for calling somatic single nucleotide variations (SNVs) and indels, copy number alterations (CNAs), and somatic structural variations (SVs). Each pipeline provided post-processing filters to remove likely false positive variant calls. A final set of filters were also run systematically across the entire set of PCAWG variants.

To assess the quality of the results from these three core pipelines, and to determine whether any other variant calling approaches would add additional value to the call set, we performed a systematic test and
laboratory-based validation of 16 different computational pipelines. After this assessment, described below, we decided to run two additional callers\textsuperscript{63,64} on all samples to improve our ability to detect low-frequency SNVs and indels.

Following execution of each variant-calling pipeline, we merged the pipeline outputs for each variant type separately (SNVs, indels, CNAs, SVs) in order to achieve greater accuracy than provided by individual pipelines. The SNV and indel merge algorithms were designed and tested using the laboratory validation exercise described below as a gold standard.

RNA-Sequencing data were uniformly processed to produce normalised gene-level expression values, splice variant quantifications and measurements of alternative promoter usage, and to identify fusion transcripts, quantify allele-specific expression, and identify RNA edit sites. Calls of common and rare germline variants including single nucleotide variants, indels, SVs and mobile element insertions were generated using previously established principles for population-scale genetic polymorphism detection.\textsuperscript{65,66} The uniform germline data processing workflow comprised variant discovery using six different variant callers, followed by call-set merging, variant genotyping and statistical haplotype-block phasing. Somatic retrotransposition events, including Alu and LINE/L1 insertions,\textsuperscript{67} L1-mediated transductions\textsuperscript{68} and pseudogene formation,\textsuperscript{69} were called using a single, well-validated pipeline.\textsuperscript{63} We removed these retrotransposition events from the SV call-set. Mitochondrial DNA mutations were called using a published algorithm.\textsuperscript{70}

**Core alignment and variant calling by cloud computing**

The requirement to uniformly realign and call variants on more than 6,800 whole genomes presented significant computational challenges. The raw sequencing reads amounted to over 650 terabytes (TB), which corresponds to the size of a high definition movie running continuously for 30 years. If run serially, the execution of the alignment and the three
variant-calling pipelines would have taken roughly 19 days/donor to execute on a single computer, or 145 years to complete the entire project. To accomplish this part of the analysis, we adopted a cloud-compute based architecture\(^1\) in which the alignment and variant calling was spread across 13 data centres distributed across three continents. The data centres represented a mixture of commercial infrastructure-as-a-service cloud compute, academic cloud compute, and traditional academic high-performance computer clusters, together contributing more than 10 million CPU core-hours to the effort. All told, the uniform alignment and variant calling took 23 months to execute – this included the data transfer, software development, and debugging time. On a cloud compute system running a fleet of 200 virtual machines, we estimate that without the overhead of software development and debugging, the project would take eight months to complete if repeated today.

The PCAWG-generated alignments, variant calls, annotations, and derived data sets are available for browsing and download at http://dcc.icgc.org/pcawg/. In addition, for the convenience of researchers who wish to avoid long data transfer times, a large subset of the data is pre-loaded and available for cloud-based computing on various platforms (see https://dcc.icgc.org/icgc-in-the-cloud).

**Quality assessment and control**

Each donor and specimen was subject to a series of quality assessment (QA) and control (QC) steps. At the level of aligned reads, we tested for: minimum overall coverage of aligned reads; coverage across chromosomes; strand bias; insert size distribution; nucleotide content; base mismatch rate; indel rate; the number of unaligned reads; and concordance between the clinical sex of the donor and the sex inferred from the presence of Y chromosome markers and sex chromosome coverage. At the level of tumour/normal pairs and variant calls, we tested for: sharing of germline polymorphisms among the specimens from the same donor to detect sample swaps; the presence of common
polymorphisms from two or more individuals to detect sample contamination; the presence of low-frequency somatic variants in the normal sample to detect tumour-in-normal contamination;\(^72\) and the presence of mutational signatures associated with sequencing artefacts such as oxidative damage. Of the 176 donors excluded on the basis of failing one or more of the QA tests, the most common reason for failure was RNA (cDNA library) contamination of tumour or normal, which manifested as multiple intron-length deletions in a substantial proportion of reads (39 donors). This was followed by lack of required clinical metadata, apparent misdiagnosis, or a disagreement between the clinical and genomic sex (29 donors), and unacceptably high levels of tumour DNA in the normal sample (15 donors). Sample swaps were relatively rare (6 donors), and there were a small number of donors excluded due to unique artefacts including contamination of tumour with a mouse library and the presence of a sibling’s genome in the blood of a leukaemia donor, presumably due to a bone marrow transplant. One cohort (of 33 acute myeloid leukaemias) was removed entirely due to a pervasive sequencing artefacts in SNV calls.

Among the non-excluded specimens, 735 showed signs of oxidative damage, as evidenced by high levels of G>T transversions among the variant calls.\(^73\) These artefactual variants were identified and removed by a purpose-built filter.\(^74\) The 75 donors that were deemed to be borderline following QA were flagged for a variety of reasons including an unexpectedly high fraction (>15%) of paired reads mapping to different chromosomes, an unusual mutational signature that did not correspond to a known biological process or artefact, or a level of tumour-in-normal contamination that approached, but did not exceed, the cut-off level (15%). We consider these suitable for some, but not all, analytic questions and left the choice of whether to use them or not to the downstream analytic groups.

**Validation, benchmarking and merging**
In order to evaluate the performance of each of the mutation-calling pipelines and determine the strategy for integrating them, we performed a large-scale deep sequencing validation experiment. We selected a pilot set of 63 tumour/normal pairs from 23 cancer types across 26 contributing sequencing projects, on which we ran the three main mutation calling pipelines, and an additional 13 tools. The 63 tumours were chosen to have a wide range of somatic mutation frequencies in order to provide accurate representation of sensitivity and specificity estimates across samples. Of the 63 cases, 50 had sufficient DNA in both tumour and normal samples to enable deep sequencing targeting the putative mutated sites through DNA hybridisation capture. We selected ~250,000 SNVs and indels for validation by first stratifying mutations based on the number of methods that called them and then evenly sampling from each of these strata. This enabled us to estimate, for each method, false-positive and false-negative rates, which were used to calculate performance metrics such as precision, sensitivity and a combined (F1) score.

Next, we examined multiple methods for integrating calls made by each of the three pipelines. We evaluated the performance of simple methods (such as taking the union or intersection of the calls) as well as more sophisticated methods that used, beyond the three pipelines, additional parameters (such as coverage, variant allele frequency and nearby sequence context) to predict whether a mutation is real or not. The final consensus calls for SNVs were based on a simple approach that required two or more methods to agree on a call. For indels, because methods were less concordant, we used logistic regression to integrate the calls. The SV merge accepted all calls made by two or more of the four primary SV callers (one pipeline has two SV callers).

Overall, the sensitivity and precision of the consensus calls were 95% (CI90%: 88-98%) and 95% (71-99%) respectively for SNVs. For indels, in keeping with greater challenges in identification accuracy, sensitivity and precision were 60% (34-72%) and 91% (73-96%). Using manual
assessment of called SVs as a gold standard, the false discovery rate of merged calls was estimated to be 2.5%, with 10% of true calls rejected. For all mutation types, accuracy was reduced in repeat-rich regions relative to coding and other unique regions.

**Pan-cancer burden of somatic mutations**

Across the 2,583 donors in the PCAWG dataset, we called 43,778,859 SNVs; 410,123 somatic multi-nucleotide variants; 2,418,247 somatic indels; 288,416 SVs; 21,076 somatic retrotransposition events; and 8,185 \textit{de novo} mitochondrial DNA variants (Supplementary Table 1). There was considerable heterogeneity in the burden of somatic mutations across patients and tumour types (Figure 2). For example, the median number of base substitutions across different tumour types spanned more than two orders of magnitude, from a median of 169/patient in pilocytic astrocytoma to 70,873/patient in melanoma. Similarly, within each tumour type, the burden of somatic substitutions typically varied over 2 orders of magnitude, with the range observed in breast adenocarcinoma being 1,203 in one patient to 65,065 in another. Similar heterogeneity was observed for other classes of somatic variation.

Strikingly, at the level of tumour types, there was a broad correlation in mutation burden among the different classes of somatic variation (Figure 2). Thus, melanomas, squamous cell carcinomas of the lung and oesophageal adenocarcinomas all showed high rates of somatic substitutions, indels, structural variation and retrotransposition. In contrast, the genomes of blood cancers and childhood brain tumours were generally quiet and stable, with relatively few variants of any type. Analysed at a per-patient level, this correlation held (Supplementary Figure 1).

This correlation in burden among different classes of somatic mutation has not been delineated on a pan-cancer basis before, and the underlying causes are unclear. It is likely that age plays some role – we observe a
correlation of most classes of somatic mutation with age at diagnosis (~190 substitutions/year, p=0.02; ~22 indels/year, p=5x10^{-5}; 1.5 SVs/year, p<2x10^{-16}; **Figure 3**). Other factors are also likely to contribute to the correlations among classes of somatic mutation, since there is evidence that some DNA repair defects can cause multiple types of somatic mutation\textsuperscript{75} and a single carcinogen can cause a range of DNA lesions.\textsuperscript{76}
Table 1. PCAWG Working Groups.

<table>
<thead>
<tr>
<th>Working Group Name</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCAWG Technical Working Group</td>
<td>Clinical and molecular data management, execution of uniform alignment and core somatic mutation calling pipelines.</td>
</tr>
<tr>
<td>PCAWG Quality Control Working Group</td>
<td>Quality assurance</td>
</tr>
<tr>
<td>PCAWG Ethics and Legal Working Group</td>
<td>Policy and data access issues relating to distribution of data across compute clouds.</td>
</tr>
<tr>
<td>PCAWG Reference Annotations Working Group</td>
<td>Developed the reference set of genome annotations for use across the project.</td>
</tr>
<tr>
<td>PCAWG SNV Calling Working Group</td>
<td>Benchmarking of somatic single nucleotide variation and indel calling pipelines and development of methods for merging them.</td>
</tr>
<tr>
<td>PCAWG Drivers and Functional Interpretation Working Group</td>
<td>Identification and characterization of coding and non-coding drivers.</td>
</tr>
<tr>
<td>PCAWG Transcriptome Working Group</td>
<td>Exploration of the effect of genomic variation on transcription.</td>
</tr>
<tr>
<td>PCAWG Epigenome Working Group</td>
<td>Exploration of the interactions between genomic variation and the epigenome.</td>
</tr>
<tr>
<td>PCAWG Structural Variation Working Group</td>
<td>Identification and characterization of structural variations in the cancer genome.</td>
</tr>
<tr>
<td>PCAWG Mutational Signatures Working Group</td>
<td>Characterization of exposures and other mutational processes acting on the cancer genome.</td>
</tr>
<tr>
<td>PCAWG Germline Cancer Genome Working Group</td>
<td>Large scale haplotyping of PCAWG donors. Exploration of the interaction between germline and somatic mutations.</td>
</tr>
<tr>
<td>PCAWG Pathology and Clinical Correlates Working Group</td>
<td>Harmonization of tumour types and clinical descriptions. Exploration of clinical significance of somatic and germline variation.</td>
</tr>
<tr>
<td>PCAWG Evolution and Heterogeneity Working Group</td>
<td>Characterisation of evolutionary history of cancer genomes.</td>
</tr>
<tr>
<td>PCAWG Portals and Visualisation Working Group</td>
<td>Developed software systems for community access to PCAWG data set and interpretation.</td>
</tr>
<tr>
<td>PCAWG Mitochondrial Genome and Immunogenomics Working Group</td>
<td>Exploration of variation affecting the mitochondrial genome and the immune system.</td>
</tr>
<tr>
<td>PCAWG Pathogens Working Group</td>
<td>Discovery of the presence of pathogen DNA in tumour samples and interpretation of significance.</td>
</tr>
</tbody>
</table>

Table 2. Overview of tumour types included in PCAWG project.

<table>
<thead>
<tr>
<th>Organ</th>
<th>Abbreviation</th>
<th>Included subtypes</th>
<th>Cases</th>
<th>Sex</th>
<th>Age</th>
<th>10-90th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural crest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skin</td>
<td>Skin-Melanoma</td>
<td>Malignant melanoma</td>
<td>107</td>
<td>38</td>
<td>69</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>CNS-Medullo</td>
<td>Medulloblastoma; Large cell medullo.; Desmoplastic medullo.</td>
<td>141</td>
<td>65</td>
<td>76</td>
<td>9</td>
</tr>
<tr>
<td>CNS</td>
<td>CNS-PiloAstro</td>
<td>Pilocytic astrocytoma</td>
<td>89</td>
<td>47</td>
<td>42</td>
<td>8</td>
</tr>
<tr>
<td>Tissue Type</td>
<td>Suffix</td>
<td>Condition</td>
<td>Stage</td>
<td>Age</td>
<td>Median</td>
<td>Range</td>
</tr>
<tr>
<td>-------------</td>
<td>--------</td>
<td>-----------</td>
<td>-------</td>
<td>-----</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>CNS</td>
<td>CNS-GBM</td>
<td>Glioblastoma</td>
<td>39</td>
<td>12</td>
<td>27</td>
<td>59</td>
</tr>
<tr>
<td>CNS</td>
<td>CNS-Oligo</td>
<td>Oligodendroglioma</td>
<td>18</td>
<td>9</td>
<td>9</td>
<td>40</td>
</tr>
<tr>
<td><strong>Mesoderm</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uterus</td>
<td>Uterus-AdenoCA</td>
<td>Serous cystadenocarcinoma; Endometrioid adenocarcinoma</td>
<td>44</td>
<td>44</td>
<td>0</td>
<td>69</td>
</tr>
<tr>
<td>Ovary</td>
<td>Ovary-AdenoCA</td>
<td>Adenocarcinoma; Serous cystadenocarcinoma Polycystic fibromatosis</td>
<td>110</td>
<td>110</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>Myeloid</td>
<td>Myeloid-MPN</td>
<td>Acute myeloid leukaemia</td>
<td>13</td>
<td>5</td>
<td>8</td>
<td>50</td>
</tr>
<tr>
<td>Myeloid</td>
<td>Myeloid-AML</td>
<td>MDS with ring sideroblasts; Chronic myelomonocytic leukaemia Burkitt; Follicular; Diffuse large B-cell; Marginal zone; Post-transplant</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>76</td>
</tr>
<tr>
<td>Lymphoid</td>
<td>Lymph-BNHL</td>
<td>Chronic lymphocytic leukaemia</td>
<td>107</td>
<td>51</td>
<td>56</td>
<td>57</td>
</tr>
<tr>
<td>Kidney</td>
<td>Kidney-RCC</td>
<td>Clear cell; Papillary</td>
<td>143</td>
<td>53</td>
<td>90</td>
<td>60</td>
</tr>
<tr>
<td>Kidney</td>
<td>Kidney-ChRCC</td>
<td>Chromophobe RCC</td>
<td>43</td>
<td>19</td>
<td>24</td>
<td>47</td>
</tr>
<tr>
<td>Head/Neck</td>
<td>Head-SCC</td>
<td>Squamous cell carcinoma</td>
<td>56</td>
<td>10</td>
<td>46</td>
<td>53</td>
</tr>
<tr>
<td>Cervix</td>
<td>Cervix-SCC</td>
<td>Squamous cell carcinoma</td>
<td>18</td>
<td>18</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>Bone/Soft Tissue</td>
<td>Bone-Cart</td>
<td>Chondromyxoid fibroma; Chondroblastoma</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>Bone/Soft Tissue</td>
<td>Bone-Leiomyo</td>
<td>Leiomyosarcoma</td>
<td>34</td>
<td>15</td>
<td>19</td>
<td>61</td>
</tr>
<tr>
<td>Bone/Soft Tissue</td>
<td>Bone-AdenoCA</td>
<td>Adenocarcinoma</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>Bone/Soft Tissue</td>
<td>Bone-AdenoCA</td>
<td>Chordoma; Adamantinoma</td>
<td>10</td>
<td>4</td>
<td>6</td>
<td>60</td>
</tr>
<tr>
<td>Bone/Soft Tissue</td>
<td>Bone-Osteo</td>
<td>Osteoblastoma</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>Bone/Soft Tissue</td>
<td>Bone-AdenoCA</td>
<td>Adenocarcinoma; Acinar cell Ca.; Mucinous</td>
<td>232</td>
<td>114</td>
<td>118</td>
<td>67</td>
</tr>
<tr>
<td><strong>Endoderm</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thyroid</td>
<td>Thy-AdenoCA</td>
<td>Adenocarcinoma; Follicular adenocarcinoma; Columnar cell adenocarcinoma</td>
<td>48</td>
<td>37</td>
<td>11</td>
<td>50</td>
</tr>
<tr>
<td>Stomach</td>
<td>Stomach-AdenoCA</td>
<td>Adenocarcinoma; Papillary adenocarcinoma; Tubular; Mucinous adenocarcinoma</td>
<td>68</td>
<td>18</td>
<td>50</td>
<td>65</td>
</tr>
<tr>
<td>Prostate</td>
<td>Prost-AdenoCA</td>
<td>Adenocarcinoma</td>
<td>199</td>
<td>0</td>
<td>199</td>
<td>59</td>
</tr>
<tr>
<td>Pancreas</td>
<td>Panc-AdenoCA</td>
<td>Adenocarcinoma; Acinar cell Ca.; Mucinous</td>
<td>232</td>
<td>114</td>
<td>118</td>
<td>67</td>
</tr>
<tr>
<td>Tissue</td>
<td>Subtype</td>
<td>Neoplasms</td>
<td>Median 95% CI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>----------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>--------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pancreas</td>
<td>Panc-Endocrine</td>
<td>Neoroendocrine carcinoma</td>
<td>81 28 53 59 38-74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>Lung-SCC</td>
<td>Squamous cell carcinoma; Basaloid SCC; Adenocarcinoma; Mucinous adeno.; Adenocarcinoma \textit{in situ}</td>
<td>47 10 37 68 55-77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>Lung-AdenoCA</td>
<td>Hepatocellular carcinoma; HCC + cholangio; Fibrolamellar HCC</td>
<td>37 20 17 66 48-77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liver</td>
<td>Liver-HCC</td>
<td></td>
<td>314 88 226 67 50-78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Esophagus</td>
<td>Eso-AdenoCA</td>
<td>Adenocarcinoma</td>
<td>97 14 83 70 56-79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon/Rectum</td>
<td>colorect-AdenoCA</td>
<td>Adenocarcinoma; Mucinous adeno.</td>
<td>52 28 24 68 48-81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bladder</td>
<td>Bladder-TCC</td>
<td>Transitional cell carcinoma; Papillary TCC</td>
<td>23 8 15 65 52-80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biliary</td>
<td>Biliary-AdenoCA</td>
<td>Cholangiocarcinoma; Papillary cholangioca.</td>
<td>34 15 19 64 53-76</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Ectoderm**

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Subtype</th>
<th>Neoplasms</th>
<th>Median 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>Breast-AdenoCA</td>
<td>Infiltrating duct carcinoma; Mucinous adeno.; Medullary Ca.</td>
<td>195 194 1 56 39-76</td>
</tr>
<tr>
<td>Breast</td>
<td>Breast-LobularCA</td>
<td>Lobular carcinoma</td>
<td>13 13 0 52 42-68</td>
</tr>
<tr>
<td>Breast</td>
<td>Breast-DCIS</td>
<td>Ductal carcinoma \textit{in situ}; Duct micropapillary carcinoma</td>
<td>3 3 0 55 43-60</td>
</tr>
</tbody>
</table>

**Total**                                                                 | 258 11 142 60 21-76
FIGURE LEGENDS

**Figure 1.** Flow-chart showing key steps in the analysis of PCAWG genomes

**Figure 2.** Distribution of numbers of somatic mutations of different classes across the different tumour types included in the PCAWG project. The y axis is on a log scale. SNVs, single nucleotide variants (single base substitutions); Indels, insertions or deletions <100 base pairs in size; SVs, structural variants; Retrotranspositions, counts of somatic retrotransposon insertions, transductions and somatic pseudogene insertions.

**Figure 3.** Numbers of somatic mutations by age at diagnosis. Points are coloured by tumour type, using the colour scheme in Figure 3. The y axis is on a log scale for all except mitochondrial DNA mutations. SNVs, single nucleotide variants (single base substitutions); Indels, insertions or deletions <100 base pairs in size; SVs, structural variants; Retrotranspositions, counts of somatic retrotransposon insertions, transductions and somatic pseudogene insertions; MNVs, multinucleotide variants (mostly dinucleotide substitutions); mtDNA mutations, number of somatic mutations in the mitochondrial genome.
SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Pairwise comparison of rates of different classes of somatic mutation. Points are coloured by tumour type, as depicted in the legend. The y axes are on a log scale. SNVs, single nucleotide variants (single base substitutions); Indels, insertions or deletions <100 base pairs in size; SVs, structural variants; Retrotranspositions, counts of somatic retrotransposon insertions, transductions and somatic pseudogene insertions.
References


38. Waddell, N. et al. Whole genomes redefine the mutational landscape


Central data collation

Alignment to human genome

Sanger Inst pipeline

DKFZ/EMBL pipeline

Broad Inst pipeline

Variant call-set merging

Quality assurance & additional filtering

Downstream scientific analyses

Whole genome sequencing data;
Clinical & histological data;
RNA-sequencing data
Methylation data

Genome build hs37d5
Algorithm: BWA-Mem

Call somatic subs, indels, SVs, CNAs
Run on cloud computing network
Docker images and workflow
Available as open source packages

Merging informed by validation data
Indel merging by logistic regression
Subs and SVs used 2 of 4 callers

5-star quality rating for each sample
Remove artefacts from oxidative damage and other effects

16 scientific Working Groups
Thematic analyses

http://dx.doi.org/10.1101/162784
CC-BY-ND 4.0 International license (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.
Figure 2