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[7] and early 1980s[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[7] and other cytogenetic techniques, through positional cloning of inherited cancer genes[10], low-throughput capillary sequencing[11] and comparative genomic hybridisation[12], to the current era of massively parallel whole genome sequencing[13–17]. The ever more populous catalogue of cancer genes has opened new therapeutic opportunities, with effective drugs being developed for the *BCR-ABL* fusion gene of chronic myeloid leukaemia, *ERBB2* amplifications of breast cancer and the *BRAF* point mutations of melanoma[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[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[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[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[23–27]; insights into the underlying mutational processes operative in human cancers[28–32]; delineation of the patterns of tumour heterogeneity and clonal evolution[33–36]; and development of genomics approaches to inform cancer prevention[37] and clinical management of patients with cancer[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[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[43–45]. Some tumours are driven by coding point mutations while others evolve through large-scale restructuring of chromosomes[46]; some cancer types mutate predominantly tumour suppressor genes[47] while others have high frequency of driver mutations activating oncogenes[48].

Numerous studies point to the relevance of non-coding regions, and projects including ENCODE,[49] Blueprint[50] and Epigenome Roadmap[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.[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,[58,59] whereas cancer-risk associated germline variants typically occur in intergenic regions and show enrichment with enhancers.[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 opportunties 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/icd-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[61] to align each tumour and normal sample to human reference build hs37d50.[62] 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[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.[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,[67] L1-mediated transductions[68] and pseudogene formation,[69] were called using a single, well-validated pipeline.[63] We removed these retrotransposition events from the SV call-set. Mitochondrial DNA mutations were called using a published algorithm.[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[71] 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% ($CI_{90\%}$: 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 *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[75] and a single carcinogen can cause a range of DNA lesions.[76]

Table 1. PCAWG Working Groups.

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

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

Organ	Abbreviation	Included subtypes	Cases	Sex		Age	
				F	M	Med	10-90th
Neural crest							
Skin	Skin-Melanoma	Malignant melanoma	107	38	69	57	38-77
CNS	CNS-Medullo	Medulloblastoma; Large cell medullo.; Desmoplastic medullo.	141	65	76	9	3-27
CNS	CNS-PiloAstro	Pilocytic astrocytoma	89	47	42	8	3-16

CNS	CNS-GBM	Glioblastoma	39	12	27	59	43-70
CNS	CNS-Oligo	Oligodendroglioma	18	9	9	40	27-59
Mesoderm							
Uterus	Uterus-AdenoCA	Serous cystadenocarcinoma; Endometrioid adeno.	44	44	0	69	57-80
Ovary	Ovary-AdenoCA	Adenocarcinoma; Serous cystadenocarcinoma	110	110	0	60	48-74
Myeloid	Myeloid-MPN	Polycythaemia vera; Essential thrombocythaemia; Mastocytosis	23	12	11	54	38-72
Myeloid	Myeloid-AML	Acute myeloid leukaemia	13	5	8	50	39-69
Myeloid	Myeloid-MDS	MDS with ring sideroblasts; Chronic myelomonocytic leukaemia	2	1	1	76	74-77
Lymphoid	Lymph-BNHL	Burkitt; Follicular; Diffuse large B-cell; Marginal zone; Post-transplant	107	51	56	57	10-74
Lymphoid	Lymph-CLL	Chronic lymphocytic leukaemia	90	30	60	61	46-78
Kidney	Kidney-RCC	Clear cell; Papillary	143	53	90	60	48-75
Kidney	Kidney-ChRCC	Chromophobe RCC	43	19	24	47	34-67
Head/Neck	Head-SCC	Squamous cell carcinoma	56	10	46	53	34-68
Cervix	Cervix-SCC	Squamous cell carcinoma	18	18	0	39	26-52
Cervix	Cervix-AdenoCA	Adenocarcinoma	2	2	0	39	33-45
Bone/Soft Tissue	Bone-Osteosarc	Osteosarcoma	36	20	16	20	9-58
Bone/Soft Tissue	Bone-Leiomyo	Leiomyosarcoma	34	15	19	61	51-78
Bone/Soft Tissue	Bone-Epith	Chordoma; Adamantinoma	10	4	6	60	37-67
Bone/Soft Tissue	Bone-Cart	Chrondromyxoid fibroma; Chondroblastoma	9	2	7	16	13-48
Bone/Soft Tissue	Bone-Osteoblast	Osteoblastoma	5	2	3	18	12-30
Bone/Soft Tissue	Bone-Benign	Osteofibrous dysplasia	1	1	0	26	26-26
Endoderm							
Thyroid	Thy-AdenoCA	Adenocarcinoma; Follicular adeno.; Columnar cell adeno.	48	37	11	50	30-75
Stomach	Stomach-AdenoCA	Adenocarcinoma; Papillary; Tubular; Mucinous adeno.	68	18	50	65	48-78
Prostate	Prost-AdenoCA	Adenocarcinoma	199	0	199	59	47-71
Pancreas	Panc-AdenoCA	Adenocarcinoma; Acinar cell Ca.; Mucinous;	232	114	118	67	50-79

		Adenosquamous Ca.					
Pancreas	Panc-Endocrine	Neoroendocrine carcinoma	81	28	53	59	38-74
Lung	Lung-SCC	Squamous cell carcinoma; Basaloid SCC	47	10	37	68	55-77
Lung	Lung-AdenoCA	Adenocarcinoma; Mucinous adeno.; Adenocarcinoma *in situ*	37	20	17	66	48-77
Liver	Liver-HCC	Hepatocellular carcinoma; HCC + cholangio; Fibrolamellar HCC	314	88	226	67	50-78
Esophagus	Eso-AdenoCA	Adenocarcinoma	97	14	83	70	56-79
Colon/Rectum	colourect-AdenoCA	Adenocarcinoma; Mucinous adeno.	52	28	24	68	48-81
Bladder	Bladder-TCC	Transitional cell carcinoma; Papillary TCC	23	8	15	65	52-80
Biliary	Biliary-AdenoCA	Cholangiocarcinoma; Papillary cholangioca.	34	15	19	64	53-76
Ectoderm							
Breast	Breast-AdenoCA	Infiltrating duct carcinoma; Mucinous adeno.; Medullary Ca.	195	194	1	56	39-76
Breast	Breast-LobularCA	Lobular carcinoma	13	13	0	52	42-68
Breast	Breast-DCIS	Ductal carcinoma *in situ*; Duct micropapillary carcinoma	3	3	0	55	43-60
Total			**2583**	**1160**	**1423**	**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

1. American Cancer Society. Global Cancer Facts & Figures 3rd Edition. *Am. Cancer Soc.* 1–64 (2015). doi:10.1002/ijc.27711
2. Jha, P. Avoidable global cancer deaths and total deaths from smoking. *Nat Rev Cancer* **9,** 655–664 (2009).
3. Frank, A. L. & Joshi, T. K. The global spread of asbestos. *Ann. Glob. Heal.* **80,** 257–262 (2014).
4. McDermott, U., Downing, J. R. & Stratton, M. R. Genomics and the continuum of cancer care. *N Engl J Med* **364,** 340–350 (2011).
5. Melero, I. *et al.* Evolving synergistic combinations of targeted immunotherapies to combat cancer. *Nat. Rev. Cancer* **15,** 457–472 (2015).
6. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
7. Rowley, J. D. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243,** 290–3 (1973).
8. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300,** 149–52 (1982).
9. Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature* **300,** 143–9 (1982).
10. Friend, S. H. *et al.* A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323,** 643–6 (1986).
11. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colourectal cancers. *Science (80-.).* **314,** 268–274 (2006).
12. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
13. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40,** 722–9 (2008).
14. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
15. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463,** 184–190 (2010).
16. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
17. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science (80-.).* **318,** 420–426 (2007).
18. O'Brien, S. G. *et al.* Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. *N Engl J Med* **348,** 994–1004 (2003).
19. Hudis, C. A. Trastuzumab — Mechanism of Action and Use in Clinical Practice. *N. Engl. J. Med.* **357,** 39–51 (2007).

20. Chapman, P. B. *et al.* Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med* **364,** 2507–2516 (2011).
21. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
22. Joly, Y., Dove, E. S., Knoppers, B. M., Bobrow, M. & Chalmers, D. Data Sharing in the Post-Genomic World: The Experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). *PLoS Comput. Biol.* **8,** e1002549 (2012).
23. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475,** 101–105 (2011).
24. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511,** 428–34 (2014).
25. Papaemmanuil, E. *et al.* Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med* **365,** 1384–1395 (2011).
26. The Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
27. Richter, J. *et al.* Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat. Genet.* **44,** 1316–1320 (2012).
28. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
29. The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).
30. Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148,** 59–71 (2012).
31. Haradhvala, N. J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164,** 538–549 (2016).
32. Totoki, Y. *et al.* Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nat. Genet.* **46,** 1267–73 (2014).
33. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
34. Patch, A.-M. *et al.* Whole–genome characterization of chemoresistant ovarian cancer. *Nature* **521,** 489–494 (2015).
35. Cooper, C. S. *et al.* 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 (2015).
36. Ross-Innes, C. S. *et al.* Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nat. Genet.* **47,** 1038–1046 (2015).
37. Scelo, G. *et al.* Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat. Commun.* **5,** 5135 (2014).
38. Waddell, N. *et al.* Whole genomes redefine the mutational landscape

of pancreatic cancer. *Nature* **518,** 495–501 (2015).
39. Yuan, Y. *et al.* Assessing the clinical utility of cancer genomic and proteomic data across tumour types. *Nat. Biotechnol.* **32,** 644–52 (2014).
40. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23,** 517–525 (2017).
41. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–9 (2013).
42. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).
43. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47,** 106–114 (2014).
44. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).
45. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–7 (2012).
46. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45,** 1127–1133 (2013).
47. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499,** 43–9 (2013).
48. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511,** 543–550 (2014).
49. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57-74 (2012).
50. Stunnenberg HG; International Human Epigenome Consortium, Hirst M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167,** 1145-1149 (2016).
51. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative

analysis of 111 reference human epigenomes. *Nature* **518,** 317-30 (2015).

52. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339,** 957-9 (2013).
53. Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, Gast A, Kadel S, Moll I, Nagore E, Hemminki K, Schadendorf D, Kumar R. TERT promoter mutations in familial and sporadic melanoma. *Science* **339:**959-61 (2013).
54. Northcott PA, Lee C, Zichner T, Stütz AM, Erkek S, Kawauchi D, Shih DJ, Hovestadt V, Zapatka M, Sturm D, Jones DT, Kool M, Remke M, Cavalli FM, Zuyderduyn S, Bader GD, VandenBerg S, Esparza LA, Ryzhova M, Wang W, Wittmann A, Stark S, Sieber L, Seker-Cin H, Linke L, Kratochwil F, Jäger N, Buchhalter I, Imbusch CD, Zipprich G, Raeder B, Schmidt S, Diessl N, Wolf S, Wiemann S, Brors B, Lawerenz C, Eils J, Warnatz HJ, Risch T, Yaspo ML, Weber UD, Bartholomae CC, von Kalle C, Turányi E, Hauser P, Sanden E, Darabi A, Siesjö P, Sterba J, Zitterbart K, Sumerauer D, van Sluis P, Versteeg R, Volckmann R, Koster J, Schuhmann MU, Ebinger M, Grimes HL, Robinson GW, Gajjar A, Mynarek M, von Hoff K, Rutkowski S, Pietsch T, Scheurlen W, Felsberg J, Reifenberger G, Kulozik AE, von Deimling A, Witt O, Eils R, Gilbertson RJ, Korshunov A, Taylor MD, Lichter P, Korbel JO, Wechsler-Reya RJ, Pfister SM. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511,** 428-34 (2014).
55. Hnisz D, Weintraub AS, Day DS, Valton AL, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, Reddy J, Borges-Rivera D, Lee TI, Jaenisch R, Porteus MH, Dekker J, Young RA. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454-8 (2016).
56. Weischenfeldt J, Dubash T, Drainas AP, Mardin BR, Chen Y, Stütz AM, Waszak SM, Bosco G, Halvorsen AR, Raeder B, Efthymiopoulos T, Erkek S, Siegl C, Brenner H, Brustugun OT, Dieter SM, Northcott PA, Petersen I, Pfister SM, Schneider M, Solberg SK, Thunissen E, Weichert W, Zichner T, Thomas R, Peifer M, Helland A, Ball CR, Jechlinger M, Sotillo R, Glimm H, Korbel JO. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49,** 65-74 (2017).
57. Rheinbay E, Parasuraman P, Grimsby J, Tiao G, Engreitz JM, Kim J, Lawrence MS, Taylor-Weiner A, Rodriguez-Cuevas S, Rosenberg M, Hess J, Stewart C, Maruvka YE, Stojanov P, Cortes ML, Seepo S, Cibulskis C, Tracy A, Pugh TJ, Lee J, Zheng Z, Ellisen LW, Iafrate AJ, Boehm JS, Gabriel SB, Meyerson M, Golub TR, Baselga J, Hidalgo-Miranda A, Shioda T, Bernards A, Lander ES, Getz G. Recurrent and functional regulatory mutations in breast cancer. *Nature* **547,** 55-60 (2017).
58. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, Reynolds A, Rynes E, Vlahoviček K, Stamatoyannopoulos JA, Sunyaev SR. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360-4 (2015).

59. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488,** 504-7 (2012).
60. Yao L, Tak YG, Berman BP, Farnham PJ. Functional annotation of colon cancer risk SNPs. *Nat. Commun.* **5,**5114 (2014).
61. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
62. 1000 Genomes Project. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
63. Moncunill, V. *et al.* Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* **32,** 1106–12 (2014).
64. Fan, Y. *et al.* MuSE: accounting for tumour heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17,** 178 (2016).
65. 1000 Genomes Project. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
66. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
67. Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. *Science (80-.).* **337,** 967–971 (2012).
68. Tubio, J. M. C. *et al.* Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science (80-.).* **345,** 1251343–1251343 (2014).
69. Cooke, S. L. *et al.* Processed pseudogenes acquired somatically during cancer development. *Nat. Commun.* **5,** 1–9 (2014).
70. Ju, Y. S. *et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *Elife* **3,** e02935 (2014).
71. Stein, L. D., Knoppers, B. M., Campbell, P., Getz, G. & Korbel, J. O. Data analysis: Create a cloud commons. *Nature* **523,** 149–51 (2015).
72. Amaro Taylor-Weiner*, Chip Stewart*, Thomas Giordano, Mara Rosenberg, Alyssa Macbeth, Niall Lennon, Esther Rheinbay, Dan-Avi Landau, Catherine J. Wu, Gad Getz. DeTiN : Overcoming Tumor in Normal Contamination. Submitted
73. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41,** 1–12 (2013).
74. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).
75. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534,** 47–54 (2016).
76. Meier, B. *et al.* C. elegans whole genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res.* **24,** gr.175547.114- (2014).

Figure 1

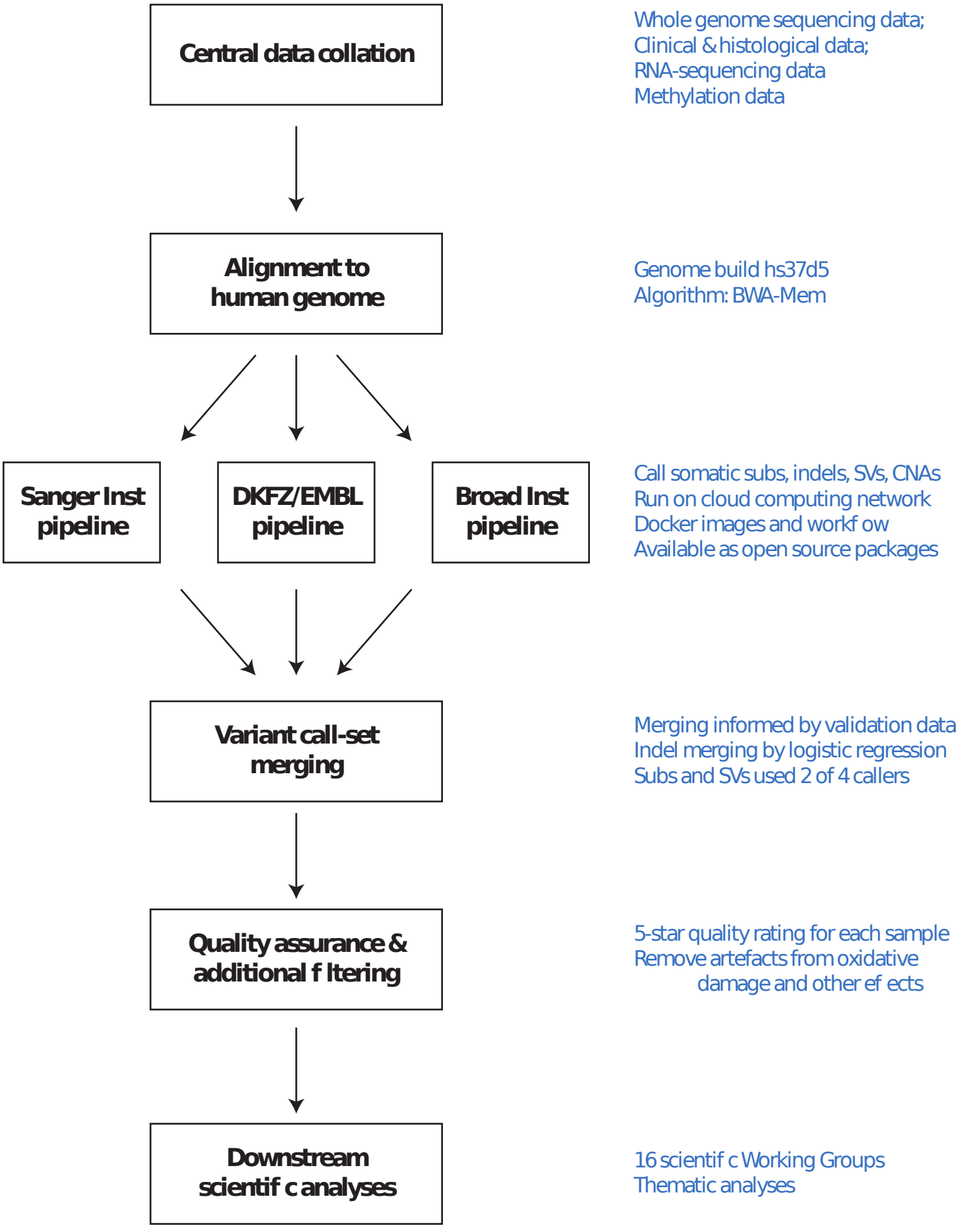


Figure 2

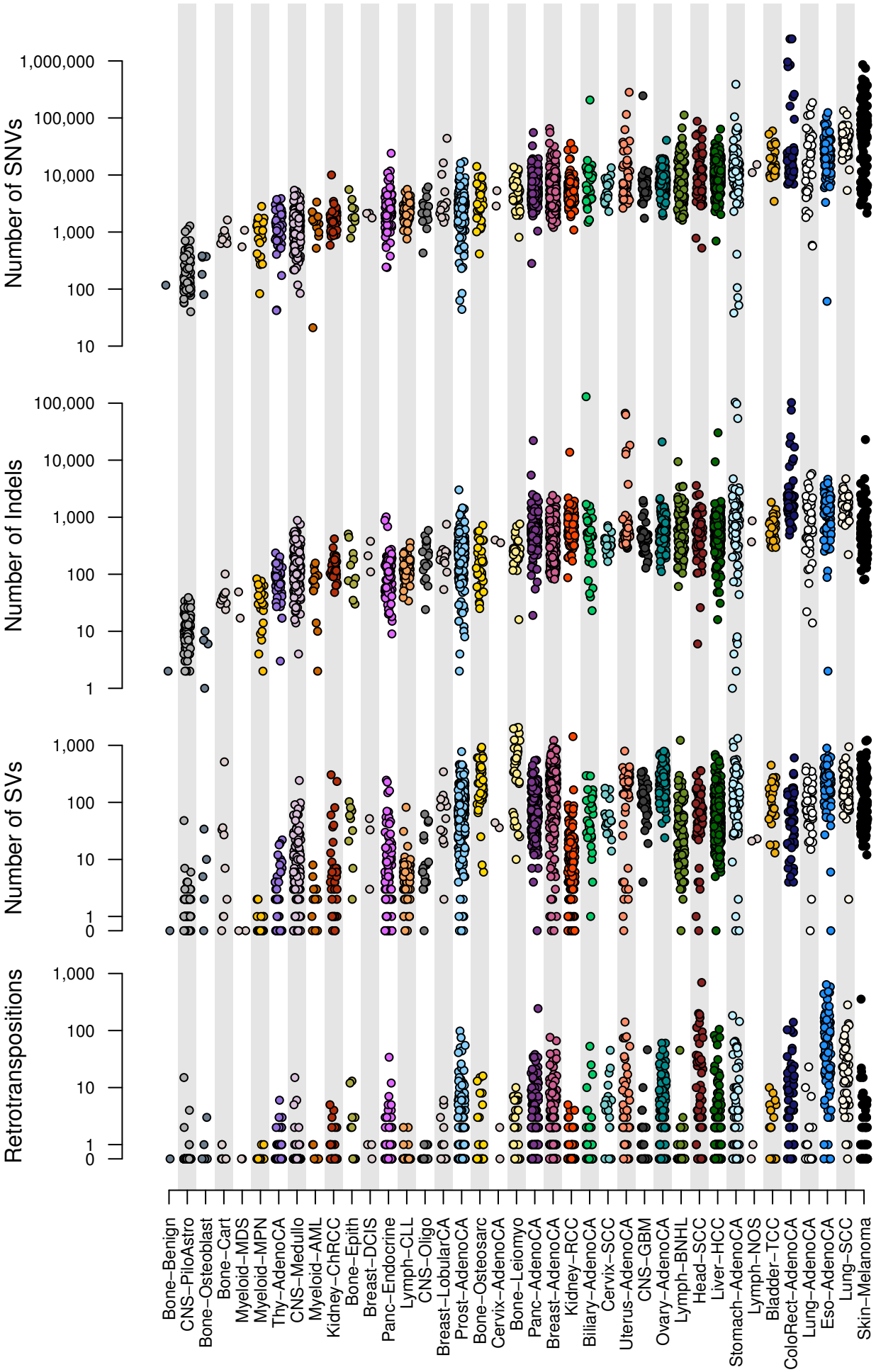
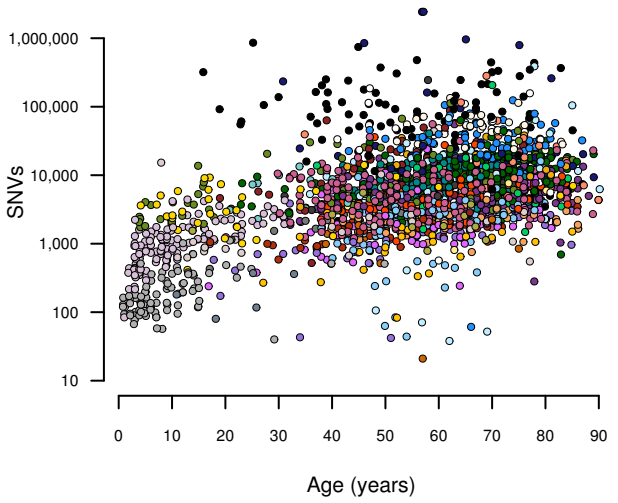
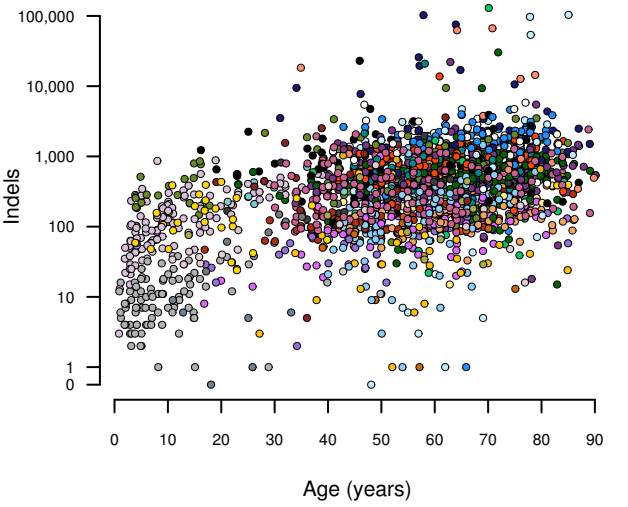
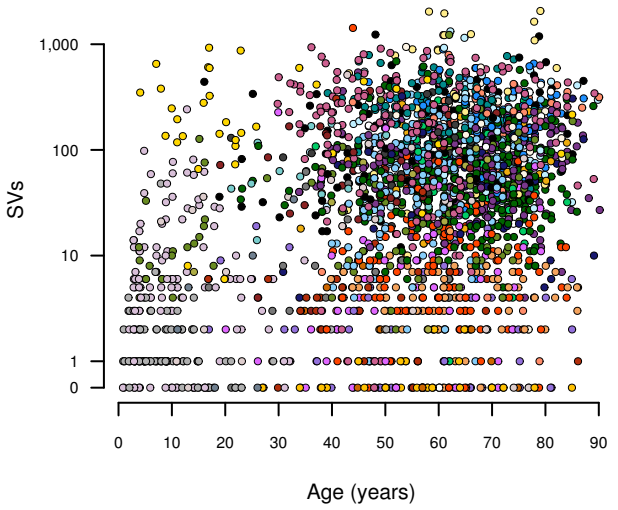
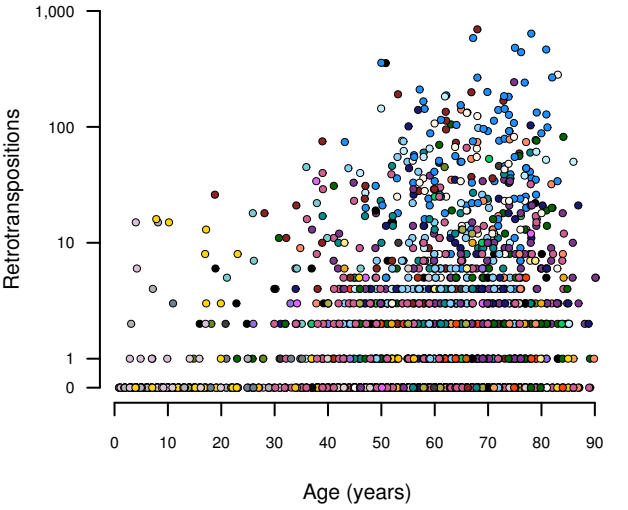


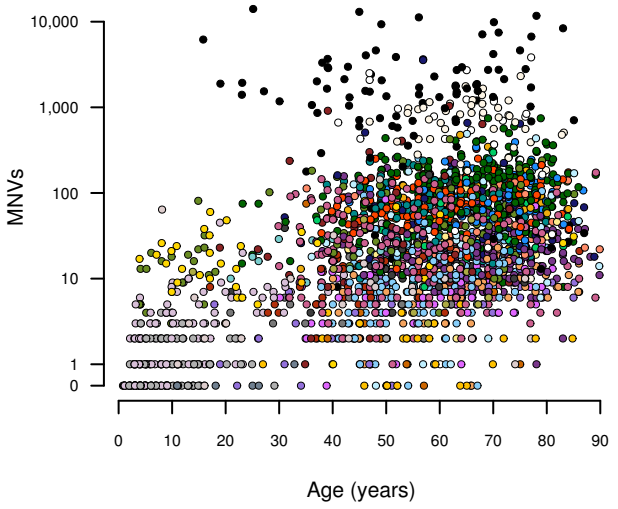
Figure 3

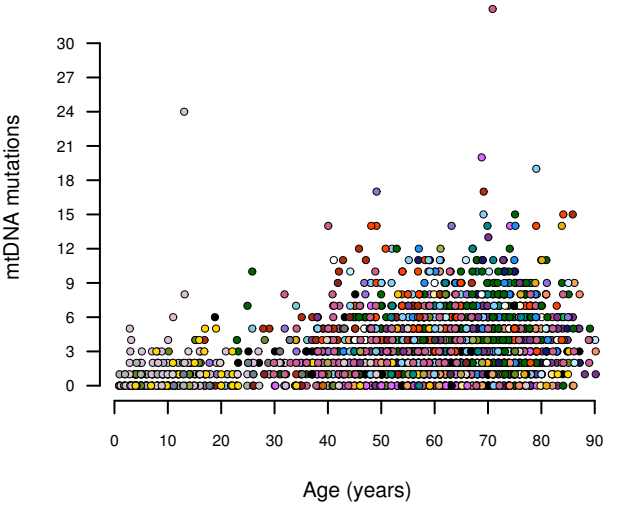

SNVs
1,000,000
100,000
10,000
1,000
100
10
0 10 20 30 40 50 60 70 80 90
Age (years)


Indels
100,000
10,000
1,000
100
10
1
0
0 10 20 30 40 50 60 70 80 90
Age (years)


SVs
1,000
100
10
1
0
0 10 20 30 40 50 60 70 80 90
Age (years)


Retrotranspositions
1,000
100
10
1
0
0 10 20 30 40 50 60 70 80 90
Age (years)


MNVs
10,000
1,000
100
10
1
0
0 10 20 30 40 50 60 70 80 90
Age (years)


mtDNA mutations
30
27
24
21
18
15
12
9
6
3
0
0 10 20 30 40 50 60 70 80 90
Age (years)

Supplementary Figure 1