Inferring structural variant cancer cell fraction

Marek Cmero1,2,3,4,5,7,8*, Ke Yuan6,7,8, Cheng Soon Ong8,9,10, Jan Schröder4, PCAWG Evolution and Heterogeneity Working Group, Niall M. Corcoran1,2, Tony Papenfuss4, Christopher M. Hovens1,2, Florian Markowetz7, Geoff Macintyre3,7* & PCAWG Consortium

We present SVclone, a computational method for inferring the cancer cell fraction of structural variant (SV) breakpoints from whole-genome sequencing data. SVclone accurately determines the variant allele frequencies of both SV breakends, then simultaneously estimates the cancer cell fraction and SV copy number. We assess performance using in silico mixtures of real samples, at known proportions, created from two clonal metastases from the same patient. We find that SVclone’s performance is comparable to single-nucleotide variant-based methods, despite having an order of magnitude fewer data points. As part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium, which aggregated whole-genome sequencing data from 2658 cancers across 38 tumour types, we use SVclone to reveal a subset of liver, ovarian and pancreatic cancers with subclonally enriched copy-number neutral rearrangements that show decreased overall survival. SVclone enables improved characterisation of SV intra-tumour heterogeneity.
T he clonal theory of cancer evolution\(^1\) posits that cancers arise from a single progenitor cell that has acquired mutations conferring selective advantage, resulting in the expansion of a genetically identical cell population or clone. As a cancer grows, a process akin to Darwinian species evolution emerges with subsequent genetically distinct populations arising from the founding clone via the continual acquisition of advantageous genomic aberrations. Consequently, tumours are likely to consist of a genetically heterogeneous combination of multiple cell populations, the extent of which has been revealed through the use of whole-genome sequencing\(^2\-4\). As clones can respond differently to therapy\(^5\), understanding this cellular diversity has important clinical implications\(^6\). The mutations belonging to each clone in a tumour can be interrogated using bulk whole-genome sequencing, with mutation detection subject to factors such as sequencing depth and quality, tumour cellularity and mutation copy number\(^6\). The expansion of each clone over the life of a tumour is encoded in the allele frequency of somatic mutations\(^7\). To characterise the clonal composition of a tumour, the variant allele frequency (VAF) must be converted to a cancer cell fraction (CCF), the fraction of cancer cells within which the variant is present. Events appearing in all cancer cells (CCF = 100%) are considered clonal and due to a pervasive expansion. Events appearing in a subset of cells (CCF < 100%) are considered subclonal and part of an ongoing expansion. Estimating the cell fraction of events is challenging, as the observed variant allele frequency depends on the amount of normal cell admixture (purity) and local copy number.

Given these challenges, previous computational approaches for estimating CCF have focused on individual facets of this complexity, commonly limiting their view to single-nucleotide variants (SNVs)\(^8\-13\) or somatic copy-number aberrations (SCNAs)\(^14\-16\). This has left the clonality of balanced rearrangements largely unexplored, despite their implication in oncogenic fusions\(^17\) and subclonal translocations conferring drug-resistant phenotypes\(^18\). While SNV-based approaches have provided solutions to the problem of downstream inference of mutation CCF, they cannot be used for structural variant (SV) breakpoint data as: (i) no complete and robust methodology exists yet to calculate VAFs from SVs (Fan et al.\(^19\) provides a limited framework that does not correct for DNA-gains or support all SV types), (ii) SVs themselves can cause copy-number changes (background copy numbers must therefore be inferred differently), (iii) SVs are composed of two ends, each with a potentially different VAF, and (iv) due to the relatively small number of data points (on average compared with SNVs), false-positive SVs greatly diminish clonality performance, hence a robust filtering methodology is required to consider only high-confidence SVs.

To address this gap, we present SVclone, an algorithmic approach that infers CCFs of SV breakpoints. It considers all types of large-scale structural variation (SV), including copy-number aberrant and copy-number neutral variation. The Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium has aggregated whole-genome sequencing data from 2658 cancers across 38 tumour types generated by the ICGC and TCGA projects. These sequencing data were re-analysed with standardised, high-accuracy pipelines to align to the human genome (reference build hs37d5) and identify germline variants and somatically acquired mutations, as described in\(^20\). Here we apply SVclone to these large-scale data to generate insight into patterns of clonality of structural variation across a large number of cancer types, and identify functionally important and clinically relevant observations.

**Results**

**Algorithm overview.** The SVclone algorithm consists of five steps: annotate, count, filter, cluster and post-assign. A graphical representation of the SVclone pipeline can be found in Fig. 1a. Here we briefly summarise each step with detailed explanations appearing in the Methods section.

**Annotate:** SV calls are required as input into the annotate step (single-nucleotide resolution paired SV loci), and the corresponding whole-genome sequencing file in BAM format. The annotate step determines the read directionality of SVs and classifies the SV type.

**Count:** The count step estimates the supporting and normal (non-supporting) read counts and computes SV VAFs.

**Filter:** The filter step removes low-quality SVs and those with missing information, and, given copy-number calls, infers the background copy number for each break-end.

**Cluster:** The cluster step simultaneously estimates the mutated copy number of SVs, the number of clusters and their respective CCF means. Allele frequencies from both break-ends of each SV are used to perform inference.

**Post assign:** The post-assign step (re)assigns variants a most-likely mutated copy number and CCF, given the previously obtained clustering configuration.

**Estimation of SV allele frequency.** SV variant allele frequencies can be estimated in the same way as SNVs: the number of variant reads divided by the total number of reads observed at the SV breakpoint. The challenge for SVs is that many reads are split across the breakpoint making extracting accurate estimates for these read counts difficult. To explore how best to deal with this challenge we simulated reads from SVs with known allele frequency, at varying tumour purity. We then implemented an optimised approach for computing a VAF from these read counts (see Methods). The simulations revealed that the VAF estimates were accurate, independent of purity, except for duplications (Fig. 1c). Duplications showed an increased normal read count due to DNA gains showing no loss of normal DNA (Fig. 1b). To account for this bias, we introduced a scaling factor that incorporates tumour purity to calibrate the supporting read counts. This corrected for the bias and showed accurate estimation of the underlying VAF (Fig. 1c).

**In silico subclonal mixing of tumours for validation.** Recently, a number of efforts have been made to simulate datasets with known subclonal structure to assess the performance of algorithms that infer the CCF of mutations\(^21\,\,22\). However, these have been limited to simulating SNVs and copy-number changes. To date, a gold standard dataset to test the performance SV cancer cell fraction inference does not exist. Therefore, we created a dataset of tumour samples with known SV subclonal structure. Rather than simulate SVs, we opted to mix two whole-genome sequenced samples from the same patient\(^23\), in silico, at known subclonal proportions (Fig. 2a). By mixing tumour sequence data, we maintained many of the noise characteristics of real sequence data. Our samples consisted of a set of three-cluster mixtures with SV and SNVs subsampled with known clonal frequencies at 10% increments, as well as four and five-cluster mixtures created by subsampling odd and even chromosomes at different frequencies (Fig. 2a). The prostate cancer samples used to create the mixtures had no evidence of subclonality (Supplementary Fig. 2d from Hong et al.\(^24\)), and had similar read coverage and tumour purity.

**Optimal cancer cell fraction versus ground truth.** Our in silico mixtures allowed us to explore some of the fundamental noise properties of CCF distributions. As the read counts supporting the SVs and SNVs in our mixed samples were subject to noise (approximately binomially distributed), we hypothesised that the resulting CCF estimates must also be noisy (approximately
Performance assessment. SVclone is chiefly designed to determine the CCF of SVs in a single, whole-genome sequenced tumour sample. Common downstream analyses of these data include analysing the number of subclonal populations in a sample and observing which SVs are clonal or subclonal. As such, we designed performance metrics to interrogate such variables including: cluster number error, mean cluster CCF error, mean variant CCF error, and sensitivity and specificity for calling a variant subclonal. As one of the key features of any CCF inference algorithm is to estimate the number of chromosome copies of a variant (known as multiplicity), we also observed the mean multiplicity error.

To our knowledge, no other method for estimating SV CCF exists for direct comparison. Instead we opted to compare to two representative, state of the art methods for estimating the CCF of SNVs, PyClone\(^\text{10}\), and copy number, Battenberg\(^\text{15}\), from single samples. In addition, we also ran SVclone in SNV clustering mode, which uses Ccube’s clustering model\(^\text{26}\). Performance is summarised in Fig. 3, and a breakdown of the performance under each measure can be found below.

Cluster number error: This metric indicates how effective the given clustering algorithms were at inferring the correct number of clusters. SVclone applied to the in silico mixtures was able to identify the correct number of clusters in 7 of 11 cases (Fig. 3). SVclone’s SNV clustering found the correct number of clusters in 5/11 cases, compared with PyClone’s 4/11, suggesting that SVs may have a slight advantage in identifying the correct number of underlying clusters.

Mean cluster CCF error: Mean cluster CCF error was generally higher in the SV data, with an average mean error of 0.0913, compared with 0.0412 and 0.0756 observed in the SNV data by SVclone and PyClone respectively. This is likely due to the variant number differences, as the comparatively larger number of SNVs is likely to lead to more accurate cluster CCF estimates.

Mean variant CCF error: Similarly, mean variant CCF error was slightly higher in the SV data than other methods. SV CCFs had an average mean error of 0.0873, compared with −0.034 for SVclone SNVs, −0.0213 for PyClone, and 0.0375 for Battenberg. Slightly higher error rates for SV CCFs are expected, given that the optimal (i.e. best obtainable given knowledge of cluster means) CCF mean errors averaged 0.0408 and 0.002 for SVs and SNVs respectively (Fig. 2c). Notably, while Battenberg performed on average better than SVclone in terms of mean variant CCF error for the three-cluster mixtures, SVclone performed better on the...
four- and five-cluster mixtures, demonstrating SVclone’s advantage in being able to consider >2 subclones. SVclone’s SV clustering and PyClone displayed similar mean error trends across the mixtures. Given the relatively smaller number of variants used in the clustering compared with SNVs, and the fewer data points used to infer fraction compared with SCNASs, SV CCF mean errors were in general comparable to other methodologies, with <0.05 absolute difference, on average, across the mixtures.

Sensitivity and specificity for calling a variant subclonal: SVclone’s SV estimates demonstrated similar sensitivity to SNVs when classifying a variant as subclonal, with an average sensitivity of 0.670 (compared with an SNV sensitivity of 0.6643). The SVs had a lower specificity (0.8852 vs. 0.952 with SNVs). PyClone displayed a lower specificity, but higher sensitivity than the other methods at 0.577 and 0.9687 respectively. Battenberg had the highest average sensitivity and specificity (sens = 0.747, spec = 0.9175), which is expected given the number of data points (germline SNVs) used by Battenberg to infer each copy-number fraction.

Multiplicity error: Multiplicity error represents the difference in the multiplicity inferred from clustering, compared with the inferred multiplicity given the ‘true’ CCF cluster mean (as multiplicity cannot directly be observed). As PyClone averages across all possible multiplicities, and does not directly estimate multiplicity, we did not consider PyClone for this metric. Average multiplicity errors were −0.0391 for SVs and 0.1029 for SNVs. The lower multiplicity error rate in SVs is likely due to the subclonal copy-number inference model (only SNVs with clonal copy numbers were considered), which allows for non-integer copy numbers. The mean multiplicity error for clonal SVs across the three-cluster mixtures was −0.1239, similar in absolute terms to the SNV multiplicity error (0.1029).

SVclone’s comparable performance to SNV-based clustering indicates that clonal structure can be effectively reconstructed with high concordance and accuracy, despite the relative deficit in variant number. This means that the clonal structure of a tumour can be inferred from SNVs and SVs independently and their results compared. However, if it is assumed that the clonal populations in a sample share the same SNVs and SVs, we have also provided an option to cluster both SVs and SNVs using the same clustering framework. This is particularly powerful when considering model-based post-assignment. SNV CCF posterior can be integrated with SV read counts’ likelihood to make assignment calls and vice versa (see Supplementary Fig. 1). By combining these data types overall performance can be increased.

Two of SVclone’s unique design features also warranted further performance assessment: (1) SVclone incorporates background SCNA states from both breakpoint ends into its clustering model; and (2) SVclone clusters variants in clonal and subclonal copy-number regions. Here, we sought to quantify the advantages of both approaches over ‘naive’ approaches which considered only one breakpoint for each SV, or used only variants in clonal copy-number regions.
Sensitivity and specificity are robust to SCNA noise. To investigate this point, we selected the single-end models respectively (see Methods). Considering the AUC indicates that the dual-end model showed a higher performance across the measured metrics compared with the dual-end model (where only one side is perturbed). The dual-end model was more robust to perturbation across all metrics for all perturbations except for cluster number (i.e. a variant is classified as clonal if present in both samples of the mixture, and subclonal otherwise).

![Fig. 3 Clustering performance metrics versus existing methods.](https://media.nature.com/naturecommunications/11/730/fig3.png)

Fig. 3 shows the effects of the perturbation experiments on the single-end model versus the dual-end model (where only one side is perturbed). The dual-end model was more robust to perturbation across all metrics for all perturbations except for cluster number with the CN − 1 experiment (where one extra cluster was called), subclonal classification sensitivity in the CN − 1 experiment and a slightly worse mean multiplicity error in the CN + 1 scenario. Interestingly, mean cluster CCF error was still lower in the over-clustered case. Importantly, the mean variant CCF error and mean multiplicity estimation effects are largely restricted to errors in variant-level estimation.

Finally, we compared SVclone’s performance using SVs in both clonal and subclonal copy-number regions, to clonal only.
Performance is summarised in Fig. 5. Utilising all available SVs improved the performance significantly across all metrics (apart from subclonal classification specificity) compared with clustering SVs with clonal background copy numbers states only.

Clonality analysis of 1705 whole-genome sequenced tumours. We applied SVclone to 1705 WGS samples from the pan-cancer analysis of whole genomes (PCAWG) project (dcc.icgc.org/pcawg)\(^{20,27}\), clustering both SVs and SNVs separately. An analysis of the clonality of putative driver SV events can be found in Dentro, et al.\(^{30}\) Here, we sought to observe any differences in the clonal structure of SVs compared with SNVs. Downstream analysis was performed on 23 tumour types showing subclonality (>10 SVs, and >10 SNVs, and sufficient power to detect subclonality (total n = 1169, see Methods).

A comparison of the fraction of subclonal SVs versus SNVs showed different patterns across tumour types (Fig. 6a). Tumour types showing a greater proportion of subclonal SVs versus SNVs included 100% of lung squamous cell carcinomas, and 92% of both colorectal adenomas and ovarian adenocarcinomas. In contrast, 23% of biliary adenocarcinomas had a greater proportion of subclonal SNVs versus SVs (Supplementary Table 1). Some cancers also contained subsets of samples with distinct patterns of clonality, for instance, liver cancers contained a cluster of 19 samples with high SV subclonality (≥50%) and low SNV subclonality (<30%).

One unique feature of SVclone is that it determines the clonality of copy-number neutral rearrangements (inversions and inter-chromosomal translocations). We applied a test for enrichment of subclonal copy-number neutral rearrangements across the PCAWG cohort. A total of 177 samples across 28 cancer types exhibited a subclonal copy-number neutral rearrangement (SCNR) pattern (e.g. Fig. 6c–f, see Supplementary Fig. 5 for the distribution of the pattern across histologies), with ovarian (n = 29, 25.7% of total ovarian), liver hepatocellular carcinoma (n = 26, 10.4% of liver samples) and pancreatic cancers (n = 18, 7.5% of total pancreatic) overrepresented in this set.

To test for potential clinical relevance of this SCNR pattern, we compared the overall survival of SCNR cases (n = 177), with high SV heterogeneity cases (n = 650), and all remaining cases (n = 447) for which overall survival was recorded, stratified on age, tumour histological subtype, and number of SVs. These groups showed significantly different survival probabilities (p = 0.006, likelihood-ratio test), with median survival times of 1236, 1470 and 2907 days, respectively (Fig. 6b). This resulted in a hazard ratio of 1.930 for SCNR cases, significantly higher compared with the baseline cohort (p = 0.0014, Z-test). In contrast, the high SV heterogeneity cases had a hazard ratio of 1.302 (p = 0.084, Z-test). Given the high number of ovarian samples within the SCNR cohort, we also considered whether fold-back inversions (FBI) were enriched, as they have been previously associated with poor prognosis\(^{29}\). We found no evidence for enrichment of FBIs (see Supplementary Fig. 6 and methods for further details), suggesting that the SCNR genotype might arise from an independent mechanism.

To test if these SCNR events were the result of a single complex rearrangement event (such as chromothripsis), or were simply a set of unrelated rearrangements, we looked for clustered events, and where possible, attempted to walk the derivative chromosome
suggesting that samples, along with KRAS and CTNNB1 (all FDR < 0.001), are enriched in this cohort (36.77% of high SV heterogeneity and complex rearrangements). However, an enrichment is consistent with the reported link between TP53 mutations and cancer drivers, which may cause the SCNR genotype. We considered a candidate bi-allelic hit as two separate mutation events affecting the same gene (copy-number loss, an SV within the gene body and/or an SNV/INDEL). We found that 62.15% of SCNR samples had at least one subclonal balanced rearrangement affecting a driver gene that was also affected by another mutation, almost double the rate found in the high SV heterogeneity cohort (32.15%) (Supplementary Table 2). This indicated that functionally-relevant consequences of the SCNR genotype are likely.

**Discussion**

Here we have presented an integrated method for inferring the cancer cell fraction of structural variation breakpoints, and have demonstrated the importance of considering the clonality of neutral rearrangements. In cancers where copy-number neutral rearrangements are common, a significant portion of the clonal landscape has remained, until now, unexplored.

Despite the successful applications of SVclone demonstrated here, it is important to consider some of its limitations. In this work, our clustering model considers all SVs as independent events despite the fact that in some cases these SVs may be part of the same complex rearrangement. Complex rearrangements are not identified by SVclone’s classification framework, however, users may specify their own types, if known. As more sophisticated methods for classifying complex SV events become available, this could be integrated into the algorithm framework. Another limitation to consider is that all CCF clustering-based methods are affected by the power to detect
variants and accurately estimate their VAFs. We present an extensive analysis investigating the effects of tumour purity, coverage and copy number (for SNVs) on the power to detect clones and subclonal mutations in Dentro et al.25, which is also applicable to SVs.

Inferring the evolutionary history of SVs from whole-genome sequence data is a challenging problem. One of the key goals in the field is to derive a clone tree that depicts the acquisition of SVs over time and their relationship to clonal expansions during tumour evolution. To achieve this, a number of key variables
must be inferred from the data: variant allele frequencies of SV breakpoints; number of DNA copies harbouring SV breakpoints (also known as multiplicity), the cancer cell fraction of SVs, cancer cell fraction of clones, and a clone phylogeny. No one method exists that can simultaneously infer all variables, but rather existing methods tackle subsets: Fan et al.: VAF31, WEAVER: VAF + clonal multiplicity32, TUSV: clonal multiplicity + clone CCF + phylogeny31 (additionally) clone copy number31, Meltos: VAF + phylogeny34, and SVclone: VAF + clonal multiplicity + approximate clone CCF + SV CCF. At present these methods need to be combined to achieve a more complete picture of the evolution of SVs (e.g. WEAVER + TUSV33 or SVclone + Meltos34). Thus, there remains an opportunity for further development of an algorithm that can simultaneously infer all variables.

Inferring the evolution of all variant classes, including SVs, SNVs, SCNAs, indels, and their respective clonality will ultimately be required to gain a more complete picture of the tumour heterogeneity landscape. We have presented an integrated software package for modelling the cancer cell fraction of structural variation breakpoints using single sample whole-genome sequencing data and have demonstrated its application by identifying patterns of subclonal variation. This software enables further exploration and quantification of tumour heterogeneity, and moves us closer to an integrated approach to modelling tumour heterogeneity.

Methods
Data input. The SVclone algorithm requires, at a minimum, a list of SV break-points and associated tumour BAM file. SV breakpoints can be provided as a VCF or as a tab-delimited file of paired single-nucleotide resolution break-ends. Using an SV caller with directionality of each break-end is recommended. The Socrates34 output format is natively supported and allows additional filtering by repeat type and average MAPQ. An associated paired-end, indexable whole-genome sequencing BAM file is required. In the filter step, copy-number information can be added in Battenberg15, ASCAT35 or PCAWG consensus format to aid in correcting VAFs. SNV input is also supported in multiple VCF formats (sanger, mutect, mutect call-stats and PCAWG consensus). Further details of input formats can be found in the repository README file.

SV annotation. To accurately calculate variant allele frequency (VAF) of structural variants, the following information is required: (i) the single-nucleotide location of loci comprising each breakpoint; (ii) the direction in which the break faces, i.e. whether the breakpoint is on the left (−) or the right side (+) of a locus that connects to the distant locus; and (iii) the classification of the SV. SV directionality affects read counting, as only reads on one side of each break-end will correspond to a specific breakpoint. SVclone incorporates basic methodology to infer the breakpoint direction (ii) and classification (iii) of the SV, however, we recommend using the information provided by the SV caller if it is available. SVclone will infer the directionality of each breakpoint by determining which side of the break-end has soft-clipped reads. If SVclone finds evidence of soft-clips lying on both sides of the break-end (i.e. at least 10% of soft-clipped reads support the opposite directionality), we consider the directionality for this break-end as mixed (i.e. multiple break-end pairs are involved for this event). If only one break-end of a pair has mixed directionality, the SV will be split into two events, one where the mixed-evidence locus is (−), and the other end is (+). If both ends have mixed directionality, we attempt to resolve this by searching the SV input for other SV events matching the SV break-ends, considering the following scenarios (see Supplementary Table 4 for a summary). We denote each SV as \(i \in \{1, 2\} \subseteq [u, l] \cap \{l, u\} \) where \(l = \) lower break-end locus and \(u =\) upper break-end locus, \(l < u\) if the chromosome is the same, or the lower of the chromosomes for inter-chromosomal translocations.

**SV directionality inference.** Directionality is determined for each SV as follows: (i) either \(l\) nor \(u\) matches any other event; the SV breakpoint is considered to be \((-,-)\); and a new SV breakpoint is created with directionality \((+,+); \) (ii) both \(l\) and \(u\) match (within a threshold): we consider one pair’s directionality as \((-,-)\) and the other pair’s as \((+,+);\) (iii) two matching breakpoints are found, each break-end matching one locus of each partner only: if the positional rankings of the three SV breakpoints (on a single chromosome) are \((1, 2, 3), (3, 1, 2)\), we consider this a translocation event, and assign the directions of \((+,+), (-,-), (-,-)) \); (iv) more than two matching breakpoints are found: the SV breakpoint is considered a complex event, and is discarded at the count step.

The directionality inference does not utilise local realignment of reads. The function is not intended to provide a mechanism for robust annotation. We recommend that directionality be inferred from the SV caller of choice.

**SV classification.** After resolving directionality, we employ a decision-tree based approach to classify SV events into categories if this information is unavailable from the SV caller (see Supplementary Fig. 7). We consider six simplified categories of rearrangements: inversions, deletions, tandem duplications, interspersed duplications and intra- and inter-chromosomal translocations. Inferences refer to a flipping of a segment of DNA, where the head of one segment joins the tail of another at both ends. Deletions are considered a loss of DNA at a locus where the flanking non-deleted segments join directly, without the intervening deleted sequence. Duplications are split into two categories: tandem and interspersed. The former category consists of a duplication joining tail to head immediately one after another. In the latter case, the duplication may be interspersed anywhere within the same chromosome. An intra-chromosomal translocation is similar to interspersed duplications, except that the original mobile element is deleted rather than retained. Inter-chromosomal translocations are defined as any joining event involving different chromosomes.

The classification heuristics are shown in Supplementary Fig. 8 and are summarised as: (i) inversion (INV): \((l, u)\) directionality matches, i.e. \((+,+)\) or \((-,-)\), and there are 1 or 2 breakpoints corresponding to the inversion event; (ii) deletion (DEL): \((l, u)\) directionality is \((+,+)\), where \(l < u\); (iii) tandem duplication (DUP) - breakpoint directionality is \((-,-)\), where \(l < u\); (iv) interspersed duplication (INTDUP): requires two breakpoints, \((l, u)\) and \((l, u)\) where \(l \neq u\) (within 100 bp) and \(u \neq u\) (one breakpoint has a tandem duplication signature and the other a deletion signature, i.e. \((l, u)\) and \((l, u)\); (v) intra-chromosomal translocation (TRX), the same as an interspersed duplications, except with the presence of a third breakpoint \((l, u)\), classified as a deletion that spans the mobile element: \(l = u\) and \((l, u)\); (vi) inter-chromosomal translocation (INTTRX) - the only criteria is that the chromosomes of \(l\) and \(u\) do not match, no directionality is considered.

**Read counting.** We consider three types of reads that cross the respective break-ends \((l, u)\): (within 6 bp): \(s_l = s_u + s_c\) supporting split reads at \(l\) and \(u\) respectively. These are variant reads (supporting the break) where one of the read-pairs lies across the break-end by a specified number of base-pairs, which must be greater than the soft-clip threshold (\(5\) by default for 100 bp reads). \(s_c\) supporting discordant (spanning) reads, i.e. reads that span across the \((l, u)\) breakpoint, where each read of the pair lies on one side of the break, effectively spanning the breakpoint (see Supplementary Fig. 9). The insert distance is calculated by both reads’ distance from their respective breakpoint at both ends. One of the reads may also be soft-clipped at the breakpoint, and still be counted as a supporting discordant read (these reads are counted under the spanning read category). In addition, the read orientation of both reads is also checked to ensure both reads are oriented towards the break (this is always the case for a true spanning read supporting the breakpoint). \(s_{ll}\): normal read count at \(l\) and \(u\) respectively. Either the read or the insert between the reads must lie across the breakpoint locus. These are reads derived from alleles not supporting the breakpoint. The outside ends of each read pair must overlap the breakpoint boundary by at least the specified base-pairs (10 by default for 100 bp reads) to be counted. Reads must not be soft-clipped above a small threshold (6 bp by default).
Supporting read calculation. Supporting reads are only counted if reads match the specified break-end directionality, this avoids double-counting of reads for events where reads are present at both sides of the breakpoint, such as inversions and translocations (these events consist of ≥2 breakpoints per event). All reads that are counted towards the supporting or normal read totals must have an insert size (fragment size) < \( \mu_{\text{ins}} + (3 \cdot \sigma_{\text{ins}}) \), where \( \mu_{\text{ins}} \) is the mean of the insert size and \( \sigma_{\text{ins}} \) is the standard deviation of the insert size. The (insert size for supporting spanning reads is considered the adjusted insert size for this criterion.) This is a quality-checking measure to ensure only high-confidence reads are counted. We consider both spanning and split reads together as the total supporting read count: \( b_2 = n + e \).

SV breakpoints where at any break-end the average depth exceeds \( 3 \cdot \max_{\text{depth}} \), are considered high depth regions and are ignored, where \( \lambda \) is the expected number of reads per locus and \( \max_{\text{depth}} \) is the maximum expected copy-number value (coverage and maximum expected copy number can be defined by the user). These breakpoints are likely caused by repetitive regions, and are not suitable for inference of clonal composition. Bed filtering has been incorporated to automatically ignore breaks falling within specific regions (to accommodate blacklists such as DAC—www.encodeproject.org/annotations/ENCSR636HFF/). In order to determine whether micro-homology was likely to play a large role in the read count processing, we analysed the distribution of breaks containing micro-homologies across the PCAWG samples used in the paper analysis (using PCAWG’s consensus SVs v6.6). We found that the mean and median micro-homology lengths were 1 and 2 bp respectively. Micro-homologies ≤ 6 bp in length are handled by the variable threshold used by the read counting step. We found that 6.17% of SVs had micro-homologies greater than 6 bp and ≤1% of SVs had micro-homologies greater than 20 bp. Given the minority of SVs affected, handling of longer micro-homologies is outside the scope of this work, and such SVs should be filtered out.

Non-supporting read calculation. For each SV, normal reads are counted at the break-ends resulting in two normal read count totals (\( n_{i,j} \)). In the case where the SV results in a gain of DNA (interspersed and tandem duplications), the normal read count must be adjusted. We consider the SV classification \( \kappa \) for an SV j, where \( \kappa \in \{\text{DEL, DUP, INT Dup, INV, TRX, INT RX}\} \) (respectively: deletions, duplications, interspersed duplications, inversions, translocations and inter-chromosomal translocations). We define two subsets \( \kappa_{\text{ins}} = \{\text{DEL, DUP, INT Dup}\} \) where normal reads at the variant population’s break-ends are unaffected at the variant allele, and \( \kappa_{\text{del}} = \{\text{INV, TRX}\} \) where there is a decrease in the number of reads at the variant population’s break-ends are replaced by supporting reads. We compute an adjustment factor, \( \Delta \) (AF) is the tumour content and \( n_{i,j} \) the tumour ploidy. The normal read counts of all DNA-gain events are then multiplied by this adjustment factor \( (\alpha_{i,j} = \alpha_{i,j} \cdot \Delta) \) if \( \kappa_{\text{ins}} \), while events that are not DNA-gains remain unadjusted.

Anomalous reads. Reads that cross the SV boundary but do not meet the requirements for split, spanning or normal reads are considered anomalous and do not contribute to read counts. Reads can be considered anomalous for numerous reasons: (i) the insert distance is greater than \( \mu_{\text{ins}} + (3 \cdot \sigma_{\text{ins}}) \), (ii) discordant reads do not face the break, (iii) the read is soft-clipped at both ends, (iv) the read is soft-clipped but is either not in the vicinity of the breakpoint, boundary or the soft-clip is below the threshold, or (v) the reads support the breakpoint in the opposite direction, but have not been called by the SV calling algorithm. To investigate points i-v, we investigated anomalous reads in the 100% purity deletion simulations, and flagged an average of 8.74 anomalous reads per breakpoint per 246.18 considered (3.57%) from the extracted regions around both break-ends of a breakpoint (these reads are proximal to the breakpoint and may not directly cross it). Upon manual inspection, we found that anomalous reads largely fell in the (iv) category, i.e. insufficient long soft-clips or the reads genuinely did not cross the breakpoint but were termed anomalous. Manual analysis uncovered no consistent under-counting of supporting reads.

Filtering variants. While tumours may contain several thousand unique mutations, typically SVs number in the dozens to low-hundreds (for instance, in breast cancers68). With typically 10-fold fewer variants, each variant utilised in clustering has a higher influence on the clustering results. A conservative approach to filtering is therefore required to minimise noise propagated through variants variant-specific read counts. The following filtering criteria have been implemented, with default values, to provide a baseline for minimising noise. These variants may be adjusted in cases by the user to tailor their noise thresholds to the samples under consideration. We filter on the following criteria:

Germline variants. The output from the count step for the corresponding patient’s germline sample can be filtered out any events where there is at least one supporting read in the germline for breakpoints that are considered the same event (both break-ends match directionality and are within 6 bp of each other).

SV size. If a breakpoint is on the same chromosome, SV size (\( u - l \)) must be larger than the fragment size (by default) as otherwise supporting and normal reads may be difficult to distinguish. This criterion is only considered for intra-chromosomal events.

Minimum support. The SV breakpoint must have at least one split and one spanning read supporting the break (\( u_1 > 1 \), \( c_1 > 1 \)). Custom minimum values can be specified.

Minimum depth. The minimum supporting + normal reads must be greater than the minimum depth for each break-end: \( (b_1 + s_1) > b_{\text{min}} \) and \( (b_2 + s_2) > b_{\text{min}} \) (Default \( b_{\text{min}} = 2 \)).

Copy-number state. If copy-number input is provided, either \( l \) or \( u \) must have a valid copy-number state for each variant. The major + minor copy numbers must be at least 1 for a site to be considered valid. Optionally, in some instances it may be appropriate to filter on several further criteria:

Copy-number neutral regions. Filters out variants with copy-number states that are not 1, 1 for major, minor alleles. Used if copy-number calls are unreliable and sufficient regions of neutral copy-number exist.

Subclonal copy-number regions. This filter may be invoked to remove any variants with subclonal copy-number states. This reduces the copy-number search space, which is useful for clustering high numbers of variants.

Assigning background copy-number states. Allele-specific copy-number variation can be supplied as input to SVclone in order to attach copy-number states to break-ends. We assign the estimated copy-number state that occurred before the SV occurred. For intra-chromosomal SVs, this involves obtaining the copy-number state upstream of the lower break-end and downstream of the upper break-end. For inter-chromosomal translocations, we obtain the copy number in the opposite direction of the break-end directionality. See Supplementary Fig. 10 for a conceptual schematic and Supplementary Table 5 for the mathematical representation.

Battenberg65 output format is preferential to carry subclonal CNAs, however, ASCAT35 is also supported. If no CNA information is supplied, the algorithm assumes that the total tumour copy number \( (\nu_{\text{tot}}) \) matches the normal copy number \( (\nu_{\text{norm}}) \), with no subclonality. For robustness of the algorithm results, it is recommended that copy-number information be supplied if available. If Battenberg input is defined, is the first solution set of segmentations in the input is considered. We define the total copy number as the sum of each clone’s copy number, weighted by the clonal fraction:

\[
\tau_{\text{tot},i,j} = \sum \rho_{r,i,j} \tau_{\text{tot},r,i,j}
\]

where \( \rho_{r,i,j} \) and \( \tau_{\text{tot},r,i,j} \) are the total copy number and copy-number fraction per copy number clone \( r \in 1, 2 \).

Clustering. The clustering step of SVclone simultaneously computes SV CCFs and clusters SVs of similar CCF, based on purity, ploidy and copy-number status of the normal, reference and tumour populations. SVclone uses a bespoke clustering algorithm that takes read counts and copy-number states at both break-ends of the same SV as input, and utilises a Bayesian mixture model, implemented using variational inference, to approximate posterior distributions for unknown parameters. The algorithm determines the number of clusters automatically and infers average CCF per cluster, as well as the multiplicity of each variant (the number of mutated chromosomal copy). The model extends our previous method, Ccube26, for estimating and cluster CCFs for SNVs by allowing it to deal with additional read and copy-number profiles from the two break-ends. This is achieved by assigning the two break-ends of an SV to the same CCF cluster. Below is a detailed description of our clustering method.

Read distribution. Let \( i \in 1, 2 \) and \( j \in 1, 2, \ldots, b \) be the indexes of break-ends and breakpoints respectively. We assume the supporting read counts from both breakpoints are independently distributed following two different Binomial distributions. The joint probability mass function of the supporting read counts is the following:

\[
p(\{b_i | d_i, f_j\}) = \prod_{i=1}^{n} \text{Binomial}(b_i | d_i, f_j),
\]

where \( b_i, d_i, \) and \( f_j \) denote the number of supporting reads, the number of normal reads, and the probability of observing one support read. The bold font variable are collections of these across both breakpoints, \( b_i = [b_{i,1}, b_{i,2}] \), \( d_i = [d_{i,1}, d_{i,2}] \), and \( f_j = [f_{j,1}, f_{j,2}] \).
We model the probability of sampling a variant read given variant locus $i$ at break-end $j$ as coming from a binomial distribution with trials $d$ (read depth $b_j + o$) and probability $f_{j,i,k}$:

$$f_{j,i,k} = \binom{d}{j} p \phi_k + e,$$

where $b_j = \sigma_j + \varepsilon_j$, $\epsilon_j$ is the number of split reads and $\sigma_j$ the number of spanning reads supporting the break. We assume the two breakpoints are conditionally independent of each other given the same CCF. In order to calculate $f_{j,i,k}$ we require the tumour purity estimate $t$ and copy-number information:

$$f_{j,i,k} = \frac{\pi(t-1)}{\pi + \pi(t)} + \frac{\pi(t)}{\pi + \pi(t)} = e,$$

where $\pi_{n_m,i}$ and $\pi_{n_m,1}$ are the total copy number of the normal and tumour population respectively and $\epsilon$ is the sequencing error constant. $\phi_k, k \in 1, ..., K$ represents the unknown CCF, and is indexed by $k$, representing the $k$th cluster. The other unknown parameter is $\pi_{n_m,i}$, the number of mutated chromosomal copies, also known as the multiplicity of the variant. See below for how these are inferred.

To test the appropriateness of the binomial distribution for SV allele numbers, the mapping is an approximation, in which the clonal copy number, the mapping is exact. In the presence of subclonal copy number, the mapping is an approximation, in which $\pi_{n_m,i}$ is replaced by the weighted average total tumour copy number. Here we provide a detailed description of the inference:

**Posterior inference.** We estimate the unknown $\phi_k$ and $\pi_{n_m,i}$ in Eq. (3) by variational inference [VI]. Specifically, the algorithm obtains a posterior distribution over $\phi_k$ and a point estimate of $\pi_{n_m,i}$ for $\phi_k$ we specify a Gaussian distribution as its prior. As a result, we obtain a maximum pseudo marginal likelihood estimator for $\pi_{n_m,i}$. The model employs a finite mixture model, hence, we introduce additional parameters such as the mixing coefficient $\pi_k$ and the cluster assignment variable $\zeta_{nm}$, which have the standard Dirichlet and Categorical prior respectively. We use this formulation for both clonal and subclonal copy-number settings. In regions of clonal copy number, the mapping is exact. In the presence of subclonal copy number, the mapping is an approximation, in which $\pi_{n_m,i}$ is replaced by the weighted average total tumour copy number. Here we provide a detailed description of the inference:

The variational inference method maximises the evidence lower bound (ELBO) of the marginal likelihood of the model:

$$\text{ELBO} = \log p(B, Z, \phi, M, H) \mid \pi \log p(Z) \log p(\pi) dZ d\phi d\pi$$

where $B = (b_j), Z = (Z_j), \phi = (\phi_k), \pi = (\pi_k), M = (m_{ij})$. We use $H$ to represent all fixed variables.

Assuming independence among the unknowns, $q(Z, \phi, \pi) = q(Z)q(\phi)q(\pi)$, the ELBO is maximised by the following solution:

$$q(Z) \propto \exp \left( \sum_{i \in V} \log p(Z_i, \phi, M_i, H_i) \right)$$

$$q(\phi) \propto \exp \left( \sum_{i \in V} \log p(B, \phi, \pi, M_i, H_i) \right)$$

$$q(\pi) \propto \exp \left( \sum_{i \in V} \log p(Z_i, \phi, \pi, M_i, H_i) \right)$$

The multiplicities of the post-assigned, $\tilde{m}_{i,post}$, are estimated as:

$$\tilde{m}_{i,post} = \arg\max_{m_{i}} \sum_{k \in K} \log p(b_j | \phi_k, f_{j,i,k})$$

We simulate SVs by first rearranging the reference genome to create an artificial genome containing SVs, and then simulating reads with SimSeq [26] from this rearranged reference. The reads were then mapped back to the original, unmodified reference genome using bowtie2 with the local alignment flag. SV size was randomly chosen among the size categories 300–500 bases. We generated 1000 paired-end reads with an average fragment size of 300 bp and an insert-size standard deviation of 20 bp. The SV events were assumed to always occur in the heterozygous fashion, hence the ‘true’ VAF was always considered to be half of the simulated purity value. To achieve the effect of differing purities, simulated normal reads were mixed with tumour samples with coverage equivalent to 0.8 and normal read coverage of $1 - \beta$ where $\beta$ represents the expected total read count at a locus. We ran simulations at 50x coverage, typical for WGS data by simulating $\frac{50X}{L}$ total reads per simulation where $L$ is the chromosome length (post rearrangement) and 300 is the fragment length. The number of reads generated for...
Prostate sample mixing: The metastatic sample bM (A) and gM (B) from Patient 001** were chosen due to their similar coverage (51.5x and 58.9x) and purity (49 and 46%). Previous work by Hong et al.** showed that these metastases shared a common ancestral clone, had no evidence of subclonality, and contained a number of private SVs and SNVs. Mixing two clonal metastases from the same patient has many advantages over spike-in approaches including: realistic sequencing noise, realistic subclonal mixing of SVs, SCNAs and SNVs, and a natural branching clonal architecture with both clonal and subclonal mutations present. We generated a total of nine samples with subclonal mixes of reads sampled at percentages 10–90, 20–80, 30–70, 40–60, 50–50, 60–40, 70–30, 80–20, and 90–10 for metastasis A and B, respectively. Three clusters are expected to be revealed upon mixing: shared variants present at 100% CCF, one cluster at bM’s mixture frequency and one cluster at gM’s mixture frequency. We also generate identities for both bM and gM mixed reads across all nine mixture proportions for each of the four clusters. The four cluster mixture was constructed by subsampling bM’s odd and even chromosomes separately at 20 and 60% respectively, and then mixing this with a 40% subsampled mixture from gM’s odd chromosomes only (effectively creating a mixture where odd and even chromosomes comprise 60% of one cluster and 40% of the other cluster at gM’s mixture frequency). We also generate identities for both bM and gM mixed reads across all nine mixture proportions for each of the four clusters. The four cluster mixture was constructed by subsampling bM’s odd and even chromosomes separately at 20 and 60% respectively, and then mixing this with a 40% subsampled mixture from gM’s odd chromosomes only (effectively creating a mixture where odd and even chromosomes comprise 60% of one cluster and 40% of the other cluster at gM’s mixture frequency). A merged variant list was created for SVs and SNVs, containing both the individual sample’s high-confidence calls. SV breakpoints were then run through SVclone’s complete pipeline, and SNVs were counted at each variant locus using the Genomon pipeline** for each mixture. Battenberg was run on each mixture to obtain SCNA data and purity estimates (which were used as the purity values for both SVclone and PyClone). A truth set was created for benchmarking purposes, constructed for SV, SNV and SCNA by determining whether the variant was unique to one sample, or shared in both.

In silico mixtures were created using the subsample and merge functions from SAMtools v1.2. copy numbers were obtained from Battenberg on each merged sample with default parameters. To construct the breakpoint list for input into SVclone’s annotate step, Socrates was run on the individual bM and gM samples, then run through SVclone’s annotate and count steps (using Socrates’ directions, filtered on simple and satellite repeats using the repeat-masker track [repeatmasker.org] and numpy average MAPQ of 20). The resulting bM and gM SVs were then merged and filtered against the germline. copy numbers were matched using corresponding Battenberg subclonal copy-number output. The merged SV list was used as the set of SV calls for the annotate step for each mix.

The reference and variant alleles were counted at each of the 9810 SNVs across the different SVclone files (Mixture variant calls from Hong et al.** were used with alleles recounted using the Samtools mpileup and pileup2base (https://github.com/riverlee/pileup2base) using a minimum quality and MAPQ cutoffs of 20 to count a base. Battenberg was run on each mixture and was used to provide copy-number information for each variant locus, as well as the purity estimate for both bM and gM. However, filtering out any variants in the different mixture proportion BAMs (gM only) resulted in a number of SVs with 0 reads in the pure bM and gM samples) (see Supplementary Table 6). We therefore randomly removed SVs with homogeneous background SCNAs until the rate of heterogeneity was 50%, resulting in 45 SVs with homogeneous and heterogeneous background SCNAs, and used these data for downstream experiments (see the SVclone_Rmarkdown notebook under code availability to replicate this analysis).

SCNAs were perturbed as follows: (i) CN − 1: major alleles were subtracted by one in the fraction A subclone in Battenberg. If the copy number was 1, subtracting one from the minor allele was attempted. If the copy-number state was 1-0 or 0-1, no modification was performed (only two SVs were unable to be changed); (ii) CN + 1: major alleles were incremented by one for the fraction A clone; (iii) Socrates was used to detect and correct copy-number changes. Performance metrics were calculated as usual with the two single-end metrics averaged out over the two runs (Supplementary Fig. 3).

Analysis of ICGC/TCGA pan-cancer samples. We utilised the pan-cancer analysis of whole genomes (PCAWG) October 2012 to 2016 consensus SNV call set, the v1.6 consensus SVs and the consensus subclonal copy numbers (19th of January 2017). For a detailed explanation on how these were generated, see**. Annotation and count were run using each sample’s associated mini-bam. Consensus purity and ploidy estimates (January 9th 2017) were used. Sample SVs and SNVs were run
separately through SVclone’s SV and SNV clustering model with default parameters.

We considered only white-listed PCAWG samples that had sufficient power to detect subclonality (number of reads per chromosome copy or NRPPC > 10; n = 1705, see Supplementary Note 1 for a list of samples). As a QC measure, we tested the association of SV number with sample purity (Supplementary Fig. 4), and found no evidence of any significant over-representation (using an FDR < 0.05 significance threshold). To determine whether SCNR samples were overrepresented for SNV clusters within each histology type and found no evidence of any significant over-representation (using an FDR < 0.05 significance threshold).

We tested each PCAWG sample for the enrichment of balanced rearrangements (inversions and inter-chromosomal translocations) below the CCF cutoff (0.7) using a hypergeometric test, with the alternative hypothesis of P(X ≥ x), where $x = \sum_{i=1}^{n} n_i$, $s$ refers to a given SV classification. Survival analysis was undertaken using the survival CRAN package (cran.r-project.org/package = survival). Hazard ratios were calculated using the Cox proportional hazards regression model, stratified by tier 4 tumour histology, age and the number of SVs in 1–100, 101–200 etc. bins. We used a hypergeometric test to determine whether any ICGC/TCGA contributors were overrepresented for SCNR samples within each histology type and found no evidence of any significant over-representation (using an FDR < 0.05 significance threshold).

In silico sample mixtures were generated from patient data derived from patient 001 coding driver genes (29th of September 2016). Patient-centric coding point mutation calling was performed from the consensus copy-number data and from the corresponding author upon reasonable request. A reporting summary for this article is available as a Supplementary Information file.

Code availability

SVclone software, user documentation, and example data can be downloaded from https://github.com/mcmero/SVclone. Cube clustering code can be found under https://github.com/keyesan/cube. Code for generating all figures in the manuscript and the in silico mixture samples can be found under https://github.com/mcmero/SVclone_Rmarkdown. Code for simulating SVs can be found under https://github.com/mcmero/se_sims_pipe.

Received: 1 December 2017; Accepted: 16 December 2019; Published online: 05 February 2020

References

1. Nowell, P. C. The clonal evolution of tumor cell populations. Science 194, 23–28 (1976).
2. Ding, L. et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 481, 506–510 (2012).
3. Gerlinger, M. et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat. Genet. 46, 225–233 (2014).
4. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer Cell 27, 15–26 (2015).
5. Aparicio, S. & Caldas, C. The implications of clonal genome evolution for cancer medicine. N. Engl. J. Med. 368, 842–851 (2013).
6. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).
7. Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of reconstructing the subclonal architecture of cancers. Cold Spring Harb. Perspect. Med. 7 (2017).
8. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. & Morris, Q. Inferring clonal evolution of tumours from single nucleotide somatic mutations. BMC Bioinform. 15, 35 (2014).
9. Hajirasouliha, I., Mahmoody, A. & Raphael, B. J. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. Bioinformatics 30(4), 178–186 (2014).
10. Roth, A. et al. PyClone statistical inference of clonal population structure in cancer. Nat. Methods 11, 396–398 (2014).
11. Fischer, A., Vázquez-García, I., Illingworth, C. J. R. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. Cell Rep. 7, 1740–1752 (2014).
12. Deshwar, A. G. et al. PhylOVS: Reconstructing subclonal composition and cancer evolution from whole genome sequencing of tumors. Genome Biol. 16, 35 (2015).
13. Popic, V. et al. Fast and scalable inference of multi-sample cancer lineages. Genome Biol. 16, 1–18 (2014).
14. Oesper, L., Mahmoody, A. & Raphael, B. J. ThetA: inferring clonal architecture from multi-sample cancer sequencing data. Genomes Data 213, 299 (2014).
15. Nik-Zainal, S. et al. The life history of 21 breast cancers. Cell 149, 994–1007 (2012).
16. Ha, G. et al. TITAN : inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Biol. 14, R80 (2013).
17. Fischer, A., Vázquez-García, I., Illingworth, C. J. R. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. Cell Rep. 7, 1740–11972 (2014).
18. Saito, M. et al. Development of lung adenocarcinomas with exclusive dependence on oncogene fusions. Cancer Res. 75, 2264–2272 (2015).
19. Patch, A.-M. et al. Whole-genome characterization of chemo-resistant ovarian cancer. Nature 521, 489–494 (2015).
20. Fan, X., Zhou, W., Chong, Z., Nakhleh, L. & Chen, K. Towards accurate characterization of clonal heterogeneity based on structural variation. BMC Bioinformatics 15, 299 (2014).
21. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature https://doi.org/10.1038/s41586-020-1909-6 (2020).
22. Saladoc, A. et al. A community effort to create standards for evaluating tumor subclonal reconstruction. Nat. Biotechnol. 38, 97–107 (2020).
23. Bhansali, V. et al. Quantifying the influence of mutation detection on tumour subclonal reconstruction. Preprint at https://doi.org/10.1101/418780 (2019).
24. Hong, M. K. H. & et al. Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. Nat. Commun. 6, 1–12 (2015).
25. Schroder, J. et al. Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. Bioinformatics 30, 1064–1072 (2014).

Data availability

In silico sample mixtures were generated from patient data derived from patient 001 from the Hong et al. study23. The data are available in the EGA Sequence Read Archive under accession EGA85008010008942.

Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium is described in ref 20 and available for download at https://dcc.icgc.org/releases/PCAWG. Additional information on accessing the data, including raw read files, can be found at https://dcc.icgc.org/pcaWG/data/.

In accordance with the data access policies of the ICGC and TCGA projects, most data that are in a public domain will be available to the public, and researchers will also need to obtain appropriate permissions. A list of consensus coding driver genes was obtained from the curated PCAWG Coding Catalog (29th of September 2016). Patient-centric coding point mutation calling was performed from the consensus copy-number data and from the corresponding author upon reasonable request. A reporting summary for this article is available as a Supplementary Information file.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Received: 1 December 2017; Accepted: 16 December 2019; Published online: 05 February 2020

Nature Communications | https://doi.org/10.1038/s41467-020-14351-8 | www.nature.com/naturecommunications
25. Dentro, S. C., Leshchiner, I., Haase, K. & Tarabichi, M. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. Preprint at https://doi.org/10.1101/312041 (2018).
26. Yuan, K., Macintyre, G. & Liu, W. Ccube: a fast and robust method for estimating cancer cell fractions. https://doi.org/10.1101/484402 (2018).
27. Gerstung, M. et al. The evolutionary history of 2,658 cancers. Nature https://doi.org/10.1038/s41586-019-1907-7 (2020).
28. Wang, Y. K. et al. Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. Nat. Genet. 49, 856–865 (2017).
29. Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. Cell 152, 1226–1236 (2013).
30. Rausch, T. et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell 148, 59–71 (2012).
31. Fan, X., Zhou, W., Chong, Z., Nakhleh, L. & Chen, K. Towards accurate characterization of clonal heterogeneity based on structural variation. BMC Bioinform. 15, 299 (2014).
32. Li, Y., Zhou, S., Schwartz, D. C. & Ma, J. Allele-specific quantification of structural variations in cancer genomes. Cell Syst. 3, 21–34 (2016).
33. Eaton, J., Wang, J. & Schwartz, R. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. Bioinformatics 34, i357–i365 (2018).
34. Ricketts, C. et al. Meltos: multi-sample tumor phylogeny reconstruction for structural variants. Bioinformatics https://doi.org/10.1093/bioinformatics/btz737 (2019).
35. Van Loo, P. et al. Allele-specific copy number analysis of tumors. Proc. Natl Acad. Sci. USA 107, 16910–16915 (2010).
36. Stephens, P. J. et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 462, 1005–1010 (2009).
37. Benidt, S. & Nettleton, D. SimSeq: a nonparametric approach to simulation of RNA-seq datasets. Bioinformatics 31, 2131–2140 (2015).
38. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
39. Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
40. Sabarimuthan, R. et al. The whole-genome panorama of cancer drivers. Preprint at https://doi.org/10.1101/190303 (2017).

Acknowledgements
We would like to thank Kangbo Mo for his assistance in developing the SV classification scheme during his Master studies, and Christoffer Flensburg for his helpful advice for the four- and five-cluster mixtures. F.M., G.M. and K.Y. would like to acknowledge the support of the University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited. G.M., K.Y. and F.M. were funded by CRUK grants CI4303/A17197 and A19274. G.M. was funded by CRUK grant A15973. This work was supported, in part, by NHMRC grants 1047581 and 1104010 to C.M.H. and 1024081 to N.M.C as well as a VCA early career seed grant 14010 to NMC. NMC was supported by a David Bickart Clinician Researcher Fellowship from the Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, and more recently by a Movember – Distinguished Gentleman’s Ride Clinician Scientist Award through the Prostate Cancer Foundation of Australia’s Research Program. M.C.: would like to acknowledge the support of the Cyberspace Foundation and the Endeavour Research Fellowship. We acknowledge the contributions of the many clinical networks across ICGC and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collation, realignment and harmonised variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects.

Author contributions
M.C.: methodology, analysis, software, visualisation, manuscript writing, editing and review; K.Y.: methodology, software, manuscript writing, editing and review; C.S.O.: methodology; J.S.: methodology; N.M.C.: supervision, manuscript editing and review; T.P.: supervision, methodology; C.M.H.: supervision, manuscript editing and review; F.M.: methodology, supervision, manuscript editing and review; G.M.: supervision, conceptualisation, methodology, analysis, manuscript writing, editing and review. The P.C.A.W.G. Evolution and Heterogeneity Working Group (led by P.S., Pvl and D.C.W.): analysis.

Competing interests
R.B. owns equity in Ampersa Therapeutics. G.G. receives research funds from IBM and Pharmacyclics and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSMSToT and POLY SOLVER. I.L. is a consultant for PACT Pharma. B.I.R. is a consultant at and has ownership interest (including stock, patents, etc.) in Medley Genomics. All the other authors have no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-14351-8.

Correspondence and requests for materials should be addressed to M.C. or G.M.

Peer review information Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PCAWG Evolution and Heterogeneity Working Group
David J. Adams11, Pavana Anur12, Rameen Beroukhim13,14,15, Paul C. Boutros16,17,18,19, David L. Bowtell20,21, Peter J. Campbell11,22, Shao long Cao23, Elizabeth L. Christie20, Yupeng Cun24, Kevin J. Dawson11, Jonas Demeulemeester25,26, Stefan C. Dentro11,26,27, Amit G. Deshwar28, Nilgun Donmez29,30, Ruben M. Drews31, Roland Eils32,33,34,5, Yu Fan23, Matthew W. Fittall26, Dale W. Garsed20,36, Moritz Gerstung37,38, Gad Getz13,15,39,40, Santiago Gonzalez37,38, Gavin Ha13, Kerstin Haase26, Marcin I mielinski41,42, Lara Jerman38,43, Yuan Ji44,45, Clemency Jolly26, Kortine Kleinheinz32,34, Juhee Lee46, Henry Lee-Six11, Ignaty Leshchiner13, Dimitri Livitz13, Salem Malikic29,30, Inigo Martincoren a11,
Thomas J. Mitchell,
Quaid D. Morris, Ville Mustonen, Layla Oesper, Martin Peifer,
Myron Peto, Benjamin J. Raphae, Daniel Rosebrock, Yulia Rubanova,
S. Cenk Sahinalp, Adriana Salcedo, Matthias Schlesner, Steven E. Schumacher,
Subhajit Sengupta, Ruian Shi, Seung Jun Shin, Paul T. Spellman,
Oliver Spiro, Lincoln D. Stein, Maxime Tarabichi, Peter Van Loo,
Shankar Vembu, Ignacio Vázquez-García, Wenyi Wang, David C. Wedge,
David A. Wheeler, Jeffrey A. Wintersinger, Tsun-Po Yang,
Xiaotong Yao, Kaixian Yu & Hongtu Zhu.