Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types

Stefan C. Dentro[1,2,3,#], Ignaty Leshchiner[4,#], Kerstin Haase[1,#], Maxime Tarabichi[1,2,#], Jeff Wintersinger[5,#], Amit G. Deshwar[5,#], Kaixian Yu[6,#], Yulia Rubanova[5,#], Geoff Macintyre[8,#], Ignacio Vázquez-García[2,7], Kortine Kleinheinz[9,10], Dimitri G. Livitz[4], Salem Malikic[11], Nilgun Donmez[11], Subhajit Sengupta[12], Jonas Demeulemeester[1,13], Pavana Anur[14], Clemency Jolly[1], Marek Cmero[15], Daniel Rosebrock[4], Steven Schumacher[4], Yu Fan[6], Matthew Fittall[1], Ruben M. Drews[8], Xiaotong Yao[16,17], Juhee Lee[18], Matthias Schlesner[9], Hongtu Zhu[6], David J. Adams[2], Gad Getz[4], Paul C. Boutros[5,19], Marcin Imielinski[16,17], Rameen Beroukhim[4], S. Cenk Sahinalp[20], Yuan Ji[12,21], Martin Peifer[22], Inigo Martincorena[2], Florian Markowetz[8], Ville Mustonen[23], Ke Yuan[8,24], Moritz Gerstung[25], Paul T. Spellman[14], Wenyi Wang[6,#], Quaid D. Morris[5,#], David C. Wedge[3,26,#,*], Peter Van Loo[1,13,#,*], on behalf of the PCAWG Evolution and Heterogeneity Working Group[27] and the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network.

[1]The Francis Crick Institute, London, United Kingdom; [2]Wellcome Trust Sanger Institute, Cambridge, United Kingdom; [3]Big Data Institute, University of Oxford, Oxford, United Kingdom; [4]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [5]University of Toronto, Toronto, Canada; [6]The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [7]University of Cambridge, Cambridge, United Kingdom; [8]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom; [9]German Cancer Research Center (DKFZ), Heidelberg, Germany; [10]Heidelberg University, Heidelberg, Germany; [11]Simon Fraser University, Vancouver, Canada; [12]NorthShore University HealthSystem, Evanston, IL, USA; [13]Department of Human Genetics, University of Leuven, Leuven, Belgium; [14]Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA; [15]University of Melbourne, Melbourne, Australia; [16]Weill Cornell Medicine, New York, NY, USA; [17]New York Genome Center, New York, NY, USA;

[18]University of California Santa Cruz, Santa Cruz, CA, USA; [19]Ontario Institute for Cancer Research, Toronto, Canada; [20]Indiana University, Bloomington, IN, USA; [21]The University of Chicago, Chicago, IL, USA; [22]University of Cologne, Cologne, Germany; [23]University of Helsinki, Helsinki, Finland; [24]University of Glasgow, Glasgow G12 8RZ, United Kingdom; [25]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom; [26]Oxford NIHR Biomedical Research Centre, Oxford, United Kingdom.

[#]These authors contributed equally.

[*]To whom correspondence may be addressed:

David C. Wedge, Big Data Institute, University of Oxford, Li Ka Shing Centre for Health Information and Discovery, Old Road Campus, Oxford OX3 7LF, United Kingdom. Tel: +44 (0) 1865 289 610, e-mail: David.Wedge@bdi.ox.ac.uk.

Peter Van Loo, The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, United Kingdom. Tel: +44 (0) 20 3796 1719, e-mail: Peter.VanLoo@crick.ac.uk.

[27]A list of members of the PCAWG Evolution and Heterogeneity Working Group can be found at the end of the manuscript.

Summary

Continued evolution in cancers gives rise to intra-tumour heterogeneity (ITH), which is a major mechanism of therapeutic resistance and therefore an important clinical challenge. However, the extent, origin and drivers of ITH across cancer types are poorly understood. Here, we extensively characterise ITH across 2,778 cancer whole genome sequences from 36 cancer types. We demonstrate that nearly all tumours (95.1%) with sufficient sequencing depth contain evidence of recent subclonal expansions and most cancer types show clear signs of positive selection in both clonal and subclonal protein coding variants. We find distinctive subclonal patterns of driver gene mutations, fusions, structural variation and copy-number alterations across cancer types. Dynamic, tumour-type specific changes of mutational processes between subclonal expansions shape differences between clonal and subclonal events. Our results underline the importance of ITH and its drivers in tumour evolution and provide an unprecedented pan-cancer resource of extensively annotated subclonal events, laying a foundation for future cancer genomic studies.

Introduction

Cancers accumulate somatic mutations as they evolve[1,2]. Some of these are driver mutations that convey fitness advantages to their host cells and can lead to clonal expansions[3-6]. Late clonal expansions or incomplete selective sweeps result in distinct cellular populations and manifest as intra-tumour heterogeneity (ITH)[1]. Clonal mutations are shared by all cancer cells whereas subclonal mutations are present only in some.

ITH represents an important clinical challenge, as it provides genetic variation fuelling cancer progression and can lead to the emergence of therapeutic resistance[7-9]. Subclonal drug resistance and associated driver mutations are common[10-15]. ITH can impact precision medicine trial design[16], predict progression[17], and can be directly prognostic. For example, ITH of copy number aberrations (CNAs) is associated with increased risk of relapse in non-small cell lung cancer[18], head and neck cancer[19,20] and glioblastoma multiforme[21].

ITH can be characterised from massively parallel sequencing data[10,11,22-24], as the cells comprising a clonal expansion share a unique set of driver and passenger mutations that occurred in the expansion-initiating cell. Each mutation within this shared set is present in the same proportion of tumour cells (known as cancer cell fraction, CCF), which may be estimated by adjusting the mutation allele frequencies for local copy number changes and sample purity. Subsequent clustering of mutations based on their CCF (see Dentro *et al.*[25] for a recent review) yields a sample's 'subclonal architecture', i.e. estimates of the number of tumour cell populations in the sequenced sample, the CCF of each population, and assignments of mutations to each population (subclone).

Previous pan-cancer efforts used these principles to characterise subclonal events, but have been limited to exomes, which restricts the number and resolution of somatic mutation calls and ignores structural variation[26]. Two recent studies using pan-cancer data from The Cancer Genome Atlas found that actionable driver mutations are often subclonal[11], and that ITH has broad prognostic value[26]. Williams and colleagues[27] proposed that neutral evolutionary dynamics are responsible for the observed ITH in a large proportion of cancers, although the test statistics developed there have been shown to poorly discriminate neutral evolution from selection[28,29]. To date, ITH

remains poorly characterised across cancer types, and there is substantial uncertainty concerning the selective pressures operating on subclonal populations.

Recent studies have used multi-region exome or targeted sequencing to characterise ITH in detail in specific cancer types[18,30]. Due to the 'illusion of clonality'[31] variants found clonal in one sample may be subclonal in other samples, and therefore, single-sample analyses may underestimate the amount of ITH. Importantly, however, any mutations detected as subclonal in any single sample, will remain subclonal when additional samples are assayed. Therefore, through assaying single cancer samples, a robust lower limit of ITH can be established.

Here, we assess ITH, its origin and drivers, and its role in tumour development, across 2,778 tumours from 36 histologically distinct cancer types. Our study is built on the International Cancer Genome Consortium's Pan-Cancer Analysis of Whole Genomes (PCAWG) initiative, which represents the largest dataset of cancer whole-genome sequences to date[32]. Whole-genome sequencing data provides 1-2 orders of magnitude more point mutations, greater resolution to detect CNAs and the ability to call structural variants (SVs). Combined, these greatly increase the breadth and depth of our ITH analyses. Building on the high-quality consensus calls generated by the PCAWG consortium, we find pervasive ITH across cancer types. In addition, we observe clear signs of positive selection in detected subclones, we identify subclonal driver mutations in known cancer genes and find changes in mutational signature activity across cancer types, which combined provide detailed insight into tumour evolutionary dynamics.

Results

Consensus-based characterisation of intra-tumour heterogeneity in 2,778 cancers

We set out to paint detailed portraits of ITH across cancer types, including SNVs, indels, SVs and CNAs, as well as subclonal drivers, subclonal selection, and mutational signatures. We leveraged PCAWG initiative dataset, encompassing 2,778 whole-genome sequences across 36 distinct histological cancer types[32].

We applied an ensemble of six state-of-the-art copy number callers and 11 subclonal reconstruction methods and developed approaches to integrate their calls into a high-confidence consensus (**Fig. 1a**, **Supplementary Methods**). As previous studies report high sensitivity of subclonal reconstruction methods to the quality of copy number calls[26], we devised a robust consensus approach to copy number calling. In addition to breakpoints called by the six CNA callers, we incorporated SVs into our consensus call set, improving sensitivity and obtaining breakpoints with base-pair resolution (**Supplementary Methods**). Consensus purity and ploidy were determined, and correlate strongly with a recent cross-omics analysis of tumour purity[33] (**Supplementary Fig. 1**). We identify samples that have undergone whole-genome duplication, as they separate from other samples when comparing tumour ploidy and the extent of loss of heterozygosity (**Fig. 1b**, **Supplementary Methods**). These samples exhibit synchronous chromosomal gains (see our companion paper[34]), further validating the purity and ploidy estimates. Consensus copy number calls were assigned 'tiers', based on the level of agreement between different callers. On average, we reached consensus on 93% of the genome (**Fig. 1c**, **Supplementary Methods**).

Consensus copy number profiles, SNVs and purity estimates served as input to 11 subclonal SNV-clustering methods, the results of which were combined into a single reconstruction for each tumour. We validated three consensus approaches on two independent simulated datasets and assessed their robustness on the real data. Consensus performance was comparable to the best individual methods on both simulated datasets and the top-performing individual methods also displayed high similarity scores (**Fig. 1d**, **Supplementary Methods**). In contrast, on the real data, the highest similarities were observed only between consensus methods (**Fig. 1d**). Using one simulated dataset with 965 samples, we evaluated the performances of consensus

approaches over all 2,035 possible combinations of 11 individual methods, and observed that the most robust performance, when the best callers are not known *a priori*, is achieved in having all 11 callers combined (**Supplementary Methods**). Hence, we used the output of one of our consensus methods as the basis for our global assignment strategy (**Supplementary Methods**), obtaining the number of detectable subclonal expansions, the fraction of subclonal SNVs, indels, SVs and CNAs, as well as the assignment of SNVs, indels and SVs to subclones.

Portraits of intra-tumour heterogeneity across cancer types

We find pervasive ITH across all 36 cancer types. Subclonal expansions are evident in 95.1% of the 1,868 tumour samples for which our analysis is powered to detect subclones with CCF > 30% (**Fig. 2**, **Supplementary Fig. 2**, **Supplementary Methods**). Importantly, these estimates, based on single sample reconstruction, provide a lower bound of the number of subclonal mutations and the true proportion of cancers with ITH is likely to be even higher. In contrast to nearly all primary tumour samples, only half of melanoma metastases had detectable subclones (96.7% of 1,801 vs 51% of 67 samples). Surprisingly, metastases of other cancer types all contained detectable subclones (100%, n = 42). Similar to primary tumours, melanoma recurrence samples show a high degree of ITH (**Fig. 2**, **Supplementary Fig. 3a**). An approach orthogonal to clustering of SNVs confirmed that melanoma metastases contain significantly less subclonal signal (p-value = 8.4×10^{-5}, **Supplementary Fig. 3b**, **Supplementary Methods**).

The patterns of ITH across SNVs, indels, SVs and CNAs paint a characteristic portrait for each histological cancer type (**Fig. 2**). While some cancer types have limited ITH across these different types of somatic variants (e.g. lung cancers, squamous cell carcinomas and liposarcomas), others show abundance of ITH in some somatic variant types, but nearly none in others (e.g. kidney cancers and pancreatic neuroendocrine tumours show high subclonal burden across somatic variant types, except CNAs) (**Fig. 2**). We noticed an anti-correlation between the number of SNVs and the average fraction of subclonal SNVs across cancer types (**Fig. 2**), yet this relation does not hold on the level of individual tumours (**Supplementary Methods**). The proportions of subclonal indels and SNVs are strongly correlated (R^2 = 0.89).

SVs follow a similar trend (R^2 = 0.64 with SNVs), except for lung squamous cell carcinoma and kidney papillary carcinoma, which show higher fractions of subclonal SVs than SNVs (**Fig. 2**, **Supplementary Fig. 4**). In contrast, the average proportions of subclonal large-scale CNAs and SNVs are only weakly correlated (R^2 = 0.33).

These findings highlight the high prevalence of ITH across cancer types. Nearly all primary tumours, irrespective of cancer type, have undergone recent subclonal expansions giving rise to detectable subclonal populations. In addition, we find that the average proportions of subclonal SNVs, indels, SVs and CNAs are highly variable across cancer types. These observations accentuate different ITH portraits, suggesting distinct evolutionary narratives of each histological cancer type. Further, among the primary tumours of each cancer type, we find substantial diversity in the fraction of subclonal mutations.

The landscape of subclonal driver mutations

We leveraged the comprehensive whole-genome view of driver events in these cancer genomes[35] to gain insight into clonal *vs.* subclonal drivers. Out of 4,211 high-confidence driver mutations in 360 genes, we find 699 subclonal ones (SNVs and indels) across 196 genes (**Fig. 3a**). However, 74% of samples with at least one subclone (1,499 / 2,038), and 79% of all detected subclones (2,148 / 2,724), contain no identified subclonal driver SNVs or indels. In contrast, only 29% of samples (770 / 2,658) lack identified clonal driver SNVs or indels.

Overall, the landscape of subclonal driver mutations indicates that specific genes are recurrently hit in subclones across cancer types (**Fig. 3a**). For example, the *PTEN* tumour suppressor is commonly found subclonally mutated in both pancreatic and stomach adenocarcinomas. Interestingly, mutations in some driver genes that are exclusively clonal in most cancer types, are predominantly subclonal in others. For example, we find subclonal driver mutations in *TP53* in CLL and thyroid cancers; *PIK3CA* in pilocytic astrocytomas, melanomas and prostate cancers; and *KRAS* in pilocytic astrocytomas and cervical cancers.

Several tumour types have higher average numbers of subclonal known drivers per sample, suggesting greater subclonal diversity (**Fig. 3a**). Gene set analysis

(**Supplementary Methods**) revealed enrichment of subclonal mutations in genes responsible for chromatin regulation and transcriptional activity, suggesting an important role in later cancer progression. Indeed, we found that *ARID1A*, *PBRM1*, *KMT2C/D* and *SETD2* were highly enriched for subclonal driver mutations. Splicing factor *SF3B1* was also often subclonally mutated, and tumour suppressor *SMAD4* was subclonally aberrated in breast and pancreatic neuroendocrine tumours.

To assess the potential impact of ITH on clinical decisions, we identified actionable subclonal driver mutations. We reasoned that targeting mutations that are not present in all tumour cells will likely result in ineffective treatment. Restricting our analysis to genes and mutations for which inhibitors are available[35], we find that 11.7% of tumours with sufficient coverage harbour an identified subclonal driver that is clinically actionable (**Fig. 3b**). Among them, 5.1% of tumours show targetable driver mutations only in subclones, while the remaining 6.6% show both subclonal and clonal targetable drivers. When considering only tumours with at least one actionable event, we find that 20.7% of tumours contain at least one subclonal actionable driver, of which about half (9.1% of tumours) show only subclonal actionable events. As our results represent lower bound estimates of the subclonality at the level of the whole tumour, this suggests that targeted therapy would yield an incomplete response in at least 20% of cases. These results highlight the importance of assessing clonality of targeted mutations.

Subclones contain driver mutations that are under positive selection

Selective pressures acting on the coding regions of cancer genomes can be quantified using the dN/dS ratio, which compares the rates of non-synonymous and synonymous mutations[36]. A dN/dS ratio larger than 1 indicates positive selection, while smaller ratios characterise negative selection, and dN/dS ≈ 1 points towards neutral evolutionary dynamics (or, theoretically, approximately equal amounts of positive and negative selection).

Previously, dN/dS > 1, i.e., evidence of positive selection, has been shown for cancer driver genes[37]. When analysing clonal mutations in our dataset, we confirm this signature of selection within a set of 566 well-established driver genes

(**Supplementary Methods**). When specifically assaying our consensus subclonal mutations for the same set of drivers, we observed a dN/dS > 1 for nonsense, missense and splice-site SNVs (**Fig. 3c**). This indicates that driver mutations, rather than neutral evolutionary dynamics[27], frequently shape subclonal expansions. This is further supported by the identification of dN/dS ratios > 1 in subclonal mutations of tumours reportedly shaped by neutral evolutionary dynamics[27] (**Fig. 3d**). The 95% confidence intervals of dN/dS for subclonal mutations lay above 1 only in a subset of cancer types (**Fig. 3e**), in large part due to power limitations: cancer types with no mutation types showing dN/dS > 1 in subclonal mutations also had significantly lower numbers of samples available (*p*-value = 1.2×10^{-3}, Mann-Whitney U test).

SV clonality reveals how rearrangements influence tumour development and progression

Having established the presence of many subclonal driver SNVs, and a broad correlation between the proportions of subclonal SNVs and SVs, we then sought to examine patterns of subclonality among candidate SV driver mutations.

We defined an SV to be a candidate driver if it was associated with significantly recurrent breakpoints (see companion paper[38]) at non-fragile sites. All other SVs were deemed passengers (**Supplementary Methods**).

We found substantial variation in the clonality of driver SVs across cancer types, implying cancer type-specific roles for SVs in tumour development and/or progression (**Fig. 4a-c**). Nearly half of the samples (45%; 575 / 1,273) and all of the 28 cancer types analysed contained subclonal driver SVs. However, in nine of these cancer types, including B-cell non-Hodgkin lymphoma and melanoma, more than 75% of the candidate driver SVs were clonal, suggesting a role for SV drivers in tumour initiation but not progression. In four of these nine, the driver SVs were significantly more clonal than the passenger SVs (**Fig. 4c**, $p < 0.05$, difference of weighted medians, permutation testing), suggesting that the acquisition of early driver SVs was not caused by general genomic instability, nor did genomic instability cause the acquisition of further SV drivers. Similarly, driver SVs were also significantly more clonal than passengers in four of the remaining 19 cancer types. In contrast,

pancreatic neuroendocrine cancers and leiomyosarcomas had just over 50% of their driver SVs appearing subclonally, suggesting initiation in these cancer types was potentially driven by non-SV mutations, with subsequent ITH driven by SVs. In line with this, these two types have a relatively low number of subclonal SNV drivers (**Fig. 3a**) and no significant difference between the clonality of driver and passenger SVs. The remaining tumour types showed substantial evidence for both clonal and subclonal SV drivers, suggesting SVs can drive tumour initiation and progression in these cancer types.

Despite differences in clonality of driver SVs among cancer types, all loci containing recurrent breakpoints showed both clonal and subclonal breaks (**Supplementary Fig. 5a**), suggesting the same SVs can drive either tumour initiation or progression in different cancer types. Nonetheless, certain loci showed a preference for clonal or subclonal SVs (**Fig. 4d**, q-value < 0.05, rank-based permutation test). Candidate drivers targeted by predominately clonal SVs included *PTPRB*, *KIAA0125* (mainly in lymphomas, **Supplementary Fig. 5b**), *CDKN2A/B*, *TERT*, *MAP3K11*, *CCND1*, and *KCNU1*. Predominately subclonal targets included a gene-poor region on chromosome 4, and another region on chromosome 13 containing *RB1*, in agreement with previous studies linking *RB1* loss to tumour progression in liver[39], liposarcoma[40,41], and breast cancer[42].

To further understand how clonality impacts gain-of-function driver SVs across cancer types, we specifically focused on previously known and curated oncogenic driver fusion SVs (as described in COSMIC curated fusions [http://cancer.sanger.ac.uk/cosmic/fusion[43]]). We compared the clonality of fusions in this curated list of drivers with other unknown or out-of-frame fusion events, as well as with the overall pattern of SV clonality in studied samples. Known driver fusions were more likely to be clonal (p-value = 0.0284, Fisher's exact test, **Fig. 4e**) with some recurrent fusions appearing exclusively clonal or highly enriched for clonal events (*CCDC6-RET*, *BRAF-KIAA1549*, *ERG-TMPRSS2*), pointing to a model where gain-of-function SVs tend to appear early rather than late during tumour development.

Complex phylogenies among subclones revealed by whole genome sequencing

Whole-genome sequencing provides us with an opportunity to explore and reconstruct additional patterns of subclonal structure by performing mutation-to-mutation read phasing to assess evolutionary relationships of subclonal lineages (**Fig. 5a,b)**. Two subclones can be either linearly related to each other (parent-child relationship), or have a common ancestor, but develop on branching lineages (sibling subclones). Establishing evolutionary relationships between subclones is challenging on single-sample sequencing data due to the limited resolution to separate subclones and the uncertainties on their CCF estimates. We can however examine pairs of SNVs in WGS data that are covered by the same read pairs to reconstruct this relationship. Specifically, in haploid regions, if two SNVs are found in multiple non-overlapping read pairs, then they cannot belong to the same cell, suggesting a branching sibling lineage. In our series, we find that, of 84 tumours with sufficient mutation pairs and power, 42 (50%) show such in-*trans* SNV pairs in haploid regions (**Supplementary Methods**), suggesting that in at least 50% of tumours, branching subclonal lineages can be detected (**Fig. 5a**).

Similarly, in-*cis* SNV pairs (on same allele) support collinear subclones: when an SNV occurs only on a subset of read pairs that support another SNV, it means they belong in a parent-child relationship and thus indicate two successive subclonal expansions (**Supplementary Methods**). Using pairs of mutations confidently assigned to the same cluster, we find evidence that 44% (86 of 196) of tumours carry such in-*cis* SNV pairs, suggesting that we can further subdivide CCF clusters into multiple collinear lineages (**Fig. 5b**). These analyses illustrate frequent complex patterns of multiple subclonal expansions, exposed by whole-genome sequencing.

We further corrected the number of mutations in subclones detected by mutation clustering, by accounting for a detection bias introduced by somatic variant calling[23]. Specifically, in subclones with lower CCFs, some proportion of SNVs will be missed, causing an underestimation of the number of associated mutations and an overestimation of their subclones' CCFs (somewhat akin to the "winner's curse"). The larger number of SNVs revealed by WGS permits us to characterize and correct for these biases. We developed two methods to do this, validated them on simulated data (**Supplementary Methods**, **Supplementary Fig. 6**) and combined them to

correct the number of SNVs and the CCF of each subclone. We estimate that, on average, 14% of SNVs in detectable subclones are below the somatic caller detection limits (**Fig. 5c,d**), while in subclones with CCF < 30%, on average 21% of SNVs are missed. We therefore extrapolate that approximately 14% of subclonal drivers in detected subclones would be missed due to the limitations of mutation calling in this series.

Patterns of subclonal mutation signature activity changes across cancers

Mutational processes can differ in their activity between clonal and subclonal lineages[11]. To explore the subclonal dynamics of mutational signatures in detail, we examined subclonal mutations for changes in signature activity. We reasoned that when, for example, a mutational process is activated during tumour growth or specific subclonal expansion, only the post-expansion mutations will carry the corresponding mutational signature. Such signature activity change points can therefore be identified in SNVs that are rank-ordered by their CCFs estimates[44] (**Supplementary Methods**). Of the 2,488 samples with sufficient SNVs to perform this analysis, 1,897 (76.1%) had an activity change of at least 6% in at least one signature (a conservative threshold established *via* permutation and bootstrapping analyses, **Supplementary Methods**). We detect an average of 1.76 mutational signature activity transitions per sample.

Overall, mutational signature activity is remarkably stable. The most often changing signature (Signature 7, UV-light exposure) is variable in approximately 60% of the cases in which it is active (**Fig. 6a**). Across the dataset, we find that lifestyle-associated mutational signatures (Signatures 4, tobacco smoking, and 7, UV light exposure), and Signatures 9 (Pol η activity on AID lesions) and 12 (aetiology unknown) decrease in activity from clonal to subclonal in over half the tumours in which these signatures are active. When only considering pairs of signatures that change in the same tumour, we see that 6 out of the top 10 pairs involve Signature 5 (aetiology unknown but hypothesised to reflect lower-fidelity DNA repair pathways[45]). Such changes are often anti-correlated, suggesting that one of the mutational processes is changing at the proportional expense of others.

Evaluating signature trajectories per cancer type (**Fig. 6a**), we observe a gradually changing picture. In melanoma metastases, Signature 7 always decreases and Signature 5 increases. In contrast, in head-and-neck cancers, most signature activity changes go both up and down in similar, relatively low proportions of tumours. On average, signature activity changes are modest in size, with the maximum average exposure change recorded in CLL (29%, Signature 9). Some changes are observed across many cancer types - e.g., Signatures 5 and 40, of unknown aetiology - while others are found in only one or a few cancer types. For example, in hepatocellular carcinomas, we observe an increase in Signature 35 and a decrease in 12 (both aetiology unknown), and in oesophageal adenocarcinomas, we see an increase in Signature 3 (double-strand break-repair) and a decrease in 17 (aetiology unknown).

Average signature activity change across cancers of the same type is often monotonous along CCF (**Fig. 6b**, **Supplementary Fig. 7**). CLLs and lung adenocarcinomas initially see a sharp change in signature activity when transitioning from clonal to subclonal mutations, but activity of the signatures appears to remain stable within subclonal mutations (**Fig. 6b**). In contrast, oesophageal adenocarcinomas show a steady decrease in Signature 17 activity, whilst thyroid adenocarcinomas often contain a continuing increase in Signature 2 and 13 (APOBEC) activity. These patterns are consistent at a single sample level, for example in individual CLL tumours (**Fig. 6c**).

Mutation signature activity changes mark subclonal boundaries

We next compared the mutational signature change points (shifts in activity) with detected subclones and reasoned that these would correspond well if the emergence of subclones is associated with changes in mutational process activity. In such a scenario, we expect that the signature change points coincide with the CCF boundaries between subclones, assuming that clustering partitioned the SNVs accurately. In accordance with previous studies that highlight changes in signature activity between clonal and subclonal mutations[11,18], we find that between 36% and 53% of clone–subclone boundaries and between 43% and 59% of subclone-subclone boundaries coincide with a region of activity change (**Fig. 6d, Supplementary Methods**). This not only validates our clustering approach, but also demonstrates that

subclonal expansions are often associated with changes in signature activity. It further suggests that increased ITH would correspond to greater activity change. Indeed, the samples with the largest changes in activity tend to be the most heterogeneous (**Fig. 6e**). Conversely, 49% of changes per sample are not within our window of subclonal boundaries (**Fig. 6f**), suggesting that some detected CCF clusters represent multiple subclonal lineages, which could not be separated by single-sample clustering (**Supplementary Methods**).

Discussion

We have painted detailed portraits of ITH and subclonal selection for 36 cancer types, using SNVs, indels, SVs, CNAs, driver mutations and mutational signatures, leveraging the largest set of whole-genome sequenced tumour samples compiled to date. Remarkably, although these single-region-based results provide only a lower bound estimate of ITH, we detected subclonal tumour cell populations in 96.7% of 1,801 primary tumours. Individual subclones in the same tumour frequently exhibited differential activity of mutational signatures, implying that successive waves of subclonal expansion can act as witnesses of temporally and spatially changing mutational processes. We extensively characterised the clonality of SNVs, indels, SVs, and CNAs. For SNVs, we identified patterns of subclonal driver mutations in known cancer genes across 36 tumour types and average rates of subclonal driver events per tumour[10,11,14,18]. Analysis of dN/dS ratios revealed clear signs of positive selection across the detected subclones and across cancer types. Indels showed clonality patterns highly correlated with SNVs. For SVs, we analysed both candidate driver and passenger events, revealing different models of how SVs influence tumour initiation and progression. Clonality estimates from CNAs suggest a complementary role of chromosomal instability and mutagenic processes in driving subclonal expansions.

Evaluation of dN/dS ratios revealed that tumours classified as evolving neutrally according to the approach described by Williams *et al.*[27], contain subclones under positive selection, as previously reported[29]. Although our analyses do not exclude the possibility that a small fraction of tumours evolve under very weak or no selection, they show that selection is widespread across cancer types, with few exceptions. Recent methodological advances to test the neutral model based on explicit tumour growth models have emerged and could shed further light on the evolutionary dynamics of individual tumours through single[46] and multiple[47] tumour biopsies.

Our findings thus support and extend Nowell's model of clonal evolution[1]: as neoplastic cells proliferate under chromosomal and genetic instability, some of their daughter cells acquire mutations that convey further selective advantages, allowing them to become precursors for new subclonal lineages. Here, we have demonstrated

that this process is ongoing up to and beyond diagnosis, in virtually all tumours and cancer types.

Our observations highlight a considerable gap in knowledge about the drivers of subclonal expansions. Specifically, only 21% of the 2,724 detected subclones have a currently known SNV or indel driver mutation. Thus, late tumour development is either driven largely by different mechanisms – copy number alterations, genomic rearrangements[18,48] or epigenetic alterations – or most late driver mutations remain to be discovered. In support of the latter, our companion study[34] finds that late driver mutations occur in a more diverse set of genes than early drivers. For now, the landscape of subclonal drivers remains largely unexplored due to limited resolution and statistical power to detect recurrence of subclonal drivers. Nonetheless, each tumour type has its own characteristic patterns of subclonal SNVs, indel, SVs and CNAs, revealing distinct evolutionary narratives. Tumour evolution does not end with the last complete clonal expansion, and it is therefore important to account for ITH and its drivers in clinical studies.

We show that regions of recurrent rearrangements, harbouring likely driver SVs, also exhibit subclonal rearrangements. This suggests that improved annotations must be sought for both SVs and SNVs, in order to comprehensively catalogue the drivers of subclonal expansion. By combining analysis of SV clonality with improved annotations of candidate SV drivers[38], we highlight tumour types that would benefit from further characterisation of subclonal SV drivers, such as pancreatic neuroendocrine cancers and leiomyosarcomas.

These observations have a number of promising clinical implications. For example, there was subclonal enrichment of SVs causing *RB1* loss across multiple cancer types, expanding on the known behaviour of *RB1* mutations in breast cancer[42]. These SVs could be linked to known resistance mechanisms to emerging treatments (e.g. CDK4/6 inhibitors in breast[42] and bladder[49] cancer). If profiled in a resistance setting, they may provide a pathway to second-line administration of cytotoxic therapies such as cisplatin or ionizing radiation, which show improved efficacy in tumours harbouring *RB1* loss[50].

Our results show rich subclonal architectures, with both linear and branching evolution in many cancers. This suggests that driver mutations either reinforce or

compete with each other depending on the background in which they arise, in an evolutionary regime called clonal interference[5]. Given the pivotal role that positive selection plays in the evolution of cancer, further work is needed to characterise the full spectrum of cancer subclones and understand their fitness distribution. Meanwhile, results in controlled laboratory evolution can shed light on adaptive dynamics in the presence of genetic heterogeneity[51,52]. As the fitness distribution ultimately defines the rules for the evolutionary dynamics that ensue, future work should incorporate integrative analyses of clonal genotype and fitness to build a unified view of the selective constraints on cancer genomes.

Our study builds upon a wealth of data of cancer whole-genome sequences generated under the auspices of the International Cancer Genome Consortium and The Cancer Genome Atlas, allowing detailed characterisation of ITH from single tumour samples across 36 cancer types. It builds on consensus reconstructions of CNAs and subclones from 6 and 11 individual methods, respectively. In establishing these reconstructions, we found that each method makes errors that are corrected by the consensus. Our consensus-building tools and techniques thus provide a set of best practices for future analyses of tumour whole genome sequencing data. As multi-region sequencing strategies are better powered to infer detailed ITH compared to single-sample studies[18,47,53], future detailed pan-cancer analyses of ITH would greatly benefit from multi-region whole-genome sequencing approaches.

Methods summary

Consensus copy number analysis

As the basis for our subclonal architecture reconstruction, we needed a confident copy number profile for each sample. To this end, we applied six copy number analysis methods (ABSOLUTE, ACEseq, Battenberg, CloneHD, JaBbA and Sclust) and combined their results into a robust consensus (see **Supplementary Methods** for details). In brief, each individual method segments the genome into regions with constant copy number, then calculates the copy number of both alleles for the genomic location. Some of the methods further distinguish between clonal and subclonal copy number states, i.e. a mixture of two or more copy number states within a region. Disagreement between methods mostly stems from either difference in the segmentation step, or uncertainty on whole genome duplication (WGD) status. Both issues were resolved using our consensus strategy.

To identify a set of consensus breakpoints, we combined the breakpoints reported by the six methods with the consensus structural variants (SVs). If a hotspot of copy number breakpoints could be explained by an SV, we removed the copy number breakpoints in favour of the base-pair resolution SV. The remaining hotspots were merged into consensus calls to complement the SV-based breakpoints. This combined breakpoint set was then used as input to all methods in a second pass, where methods were required to strictly adhere to the provided breakpoints.

Allele-specific copy number states were resolved by assessing agreement between outputs of the individual callers. A consensus purity for each sample was obtained by combining the estimates of the copy number methods with the results of the subclonal architecture reconstruction methods that infer purity using only SNVs.

Each copy number segment of the consensus output was rated with a star-ranking representing confidence.

To create a subclonal copy number consensus, we used three of the copy number methods that predicted subclonal states for segments and reported the subclonal state if at least two methods were in agreement.

Consensus subclonal architecture clustering

We applied 11 subclonal reconstruction methods (BayClone-C, Ccube, CliP, cloneHD, CTPsingle, DPClust, PhylogicNDT, PhyloWGS, PyClone, Sclust, SVclone). Most were developed or further optimised during this study. Their outputs were combined into a robust consensus subclonal architecture (see **Supplementary Methods** for details). During this procedure, we used the PCAWG consensus SNVs and indels [Synapse ID syn7118450] and SVs [syn7596712].

The procedure to create consensus architectures consisted of three phases: a run of the 11 callers on a subset of SNVs that reside on copy number calls of high-confidence, merging of the output of the callers into a consensus and finally assignment of all SNVs, indels and SVs.

Each of the 11 subclonal reconstruction callers outputs the number of mutation clusters per tumour, the number of mutations in each cluster, and the clusters' proportion of (tumour) cells (cancer cell fraction, CCF). These data were used as input to three orthogonal approaches to create a consensus: WeMe, CSR and CICC. The results reported in this paper are from the WeMe consensus method, but all three developed methods lead to similar results, and were used to validate each other (**Supplementary Methods**).

The consensus subclonal architecture was compared to the individual methods on two independent simulation sets, one 500-sample for training and one 965-sample for validation, and on the real PCAWG samples to evaluate robustness. The metrics by which methods were scored take into account the fraction of clonal mutations, number of mutational clusters and the root mean square error (RMSE) of mutation assignments. To calculate the overall performance of a method, ranks of the three metrics were averaged per sample.

Across the two simulated datasets, the scores of the individual methods were variable, whereas the consensus methods were consistently among the best across the range of simulated number of subclones, tumour purity, tumour ploidy and sequencing depth. The highest similarities were observed among the consensus and the best individual methods in the simulation sets, and among the consensus methods in real data, suggesting stability of the consensus in the real set. Increasing the number of

individual methods input to the consensus consistently improved performance and the highest performance was obtained for the consensus run on the full 11 individual methods, suggesting that each individual method has its own strengths that are successfully integrated by the consensus approaches (**Supplementary Methods**).

All SNVs, indels and SVs were assigned to the clusters that were determined by the consensus subclonal architecture using MutationTimer[34]. Each mutation cluster is modelled by a beta-binomial and probabilities for each mutation belonging to each cluster are calculated.

Not only did this process result in the final consensus subclonal architecture, it also timed mutations relative to copy number gains (**Supplementary Methods**).

SV clonality analysis

Due to the difficulty in determining SV VAFs from short read sequence data, and subsequent CCF point estimation[54], we elected to explore patterns of putative driver SV clonality using subclonal *probabilities*, allowing us to account for uncertainty in our observations of SV clonality (**Supplementary Methods**). After excluding unpowered samples, highly mutated samples, and cancer types with less than ten powered samples (**Supplementary Methods**), we analysed 125,920 consensus SVs from 1,517 samples, across 28 cancer types. SVs were divided into candidate driver SVs and candidate passenger SVs using annotations from a companion paper[38]. SVs were considered candidate drivers if they were annotated as having significantly recurrent breakpoints (SRBs) at non-fragile sites, and candidate passenger SVs otherwise (**Supplementary Methods**).

Subclonal probabilities of driver and passenger SVs across tumour types were observed using weighted median and interquartile ranges (**Supplementary Methods**). Any tumour types with interquartile ranges exceeding subclonal probabilities of 0.5 were considered as having evidence of subclonal SVs. Permutation testing was used to determine significant differences in the weighted medians between driver and passenger SVs (**Supplementary Methods**). To test if any genomic loci were enriched for clonal or subclonal SVs across cancer types, we employed a GSEA-like[55] rank-based permutation test (**Supplementary Methods**).

"Winner's curse" correction

Because somatic mutation callers require a minimum coverage of supporting reads, in samples with low purity and/or small subclones, the reported CCF values and cluster sizes will be biased. As variants observed in a lower number of reads have a higher probability to be missed by somatic mutation callers, rare subclones will show lower apparent mutation numbers and higher apparent CCF values. We refer to this effect as the "Winner's curse". To adjust mutational clusters both in size and in CCF, we developed two methods, PhylogicCorrectBias and SpoilSport. Results from both methods were integrated to produce a consensus correction, and our correction approach was validated on simulated data (**Supplementary Methods**).

Mutation signatures trajectory analysis

Given the mutational signatures obtained from PCAWG [syn8366024], we used TrackSig[44] to fit the evolutionary trajectories of signature activities. Mutations were ordered by their approximate relative temporal order in the tumour, by calculating a pseudo-time ordering using CCF and copy number. Time-ordered mutations were subsequently binned to create time points on a pseudo-timeline to which signature trajectories can be mapped.

At each time point, mutations were classified into 96 classes based on their trinucleotide context and a mixture of multinomial distributions was fitted, each component describing the distribution of one active signature. Derived mixture component coefficients correspond to mutation signature activity values, reflecting the proportion of mutations in a sample that were generated by a mutational process. By applying this approach to every time point along a sample's evolutionary timeline, a trajectory showing the activity of signatures over time was obtained.

We applied likelihood maximisation and the Bayesian Information Criterion to simulations to establish the optimal threshold at which signature activity changes can be detected. This threshold was determined to be 6%. Subsequently, a pair of adjacent mutation bins was marked as constituting a change in activity if the absolute difference in activity between the bins of a at least one signature was greater than the threshold.

Signature trajectories were mapped to our subclonal reconstruction architectures by dividing the CCF space according to the proportion of mutations per time point belonging to a mutation cluster determined by the consensus reconstruction. By comparing distances in pseudo-time between trajectory change points and cluster boundaries, change points were classified as "supporting" a boundary if they are no more than three bins apart.

References

1 Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).

2 Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature* **300**, 143-149 (1982).

3 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724, doi:10.1038/nature07943 (2009).

4 Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17-37, doi:10.1016/j.cell.2013.03.002 (2013).

5 Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306-313, doi:10.1038/nature10762 (2012).

6 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

7 Maley, C. C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet* **38**, 468-473, doi:10.1038/ng1768 (2006).

8 Mroz, E. A. *et al.* High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma. *Cancer* **119**, 3034-3042, doi:10.1002/cncr.28150 (2013).

9 McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**, 613-628, doi:10.1016/j.cell.2017.01.018 (2017).

10 Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714-726, doi:10.1016/j.cell.2013.01.019 (2013).

11 McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci Transl Med* **7**, 283ra254, doi:10.1126/scitranslmed.aaa1408 (2015).

12 Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883-892, doi:10.1056/NEJMoa1113205 (2012).

13 Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353-357, doi:10.1038/nature14347 (2015).

14 Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med* **21**, 751-759, doi:10.1038/nm.3886 (2015).

15 Shaw, A. T. *et al.* Resensitization to Crizotinib by the Lorlatinib ALK Resistance Mutation L1198F. *N Engl J Med* **374**, 54-61, doi:10.1056/NEJMoa1508887 (2016).

16 Hiley, C., de Bruin, E. C., McGranahan, N. & Swanton, C. Deciphering intratumor heterogeneity and temporal acquisition of driver events to refine precision medicine. *Genome Biol* **15**, 453, doi:10.1186/s13059-014-0453-8 (2014).

17 Maley, C. C. *et al.* The combination of genetic instability and clonal expansion predicts progression to esophageal adenocarcinoma. *Cancer Res* **64**, 7629-7633, doi:10.1158/0008-5472.CAN-04-1738 (2004).

18 Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med* **376**, 2109-2121, doi:10.1056/NEJMoa1616288 (2017).

19 Rocco, J. W. Mutant allele tumor heterogeneity (MATH) and head and neck squamous cell carcinoma. *Head Neck Pathol* **9**, 1-5, doi:10.1007/s12105-015-0617-1 (2015).
20 Mroz, E. A. & Rocco, J. W. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncol* **49**, 211-215, doi:10.1016/j.oraloncology.2012.09.007 (2013).
21 Brastianos, P. K. *et al.* Resolving the phylogenetic origin of glioblastoma via multifocal genomic analysis of pre-treatment and treatment-resistant autopsy specimens. *npj Precision Oncology* **1**, 33, doi:10.1038/s41698-017-0035-9 (2017).
22 Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci U S A* **105**, 13081-13086, doi:10.1073/pnas.0801523105 (2008).
23 Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007, doi:10.1016/j.cell.2012.04.023 (2012).
24 Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A* **110**, 4009-4014, doi:10.1073/pnas.1219747110 (2013).
25 Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harb Perspect Med* **7**, doi:10.1101/cshperspect.a026625 (2017).
26 Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* **22**, 105-113, doi:10.1038/nm.3984 (2016).
27 Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat Genet* **48**, 238-244, doi:10.1038/ng.3489 (2016).
28 Noorbakhsh, J. & Chuang, J. H. Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures. *Nat Genet* **49**, 1288-1289, doi:10.1038/ng.3876 (2017).
29 Tarabichi, M. *et al.* Neutral tumor evolution? *bioRxiv*, doi:10.1101/158006 (2017).
30 Turajlic, S. *et al.* Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell* **173**, 595-610 e511, doi:10.1016/j.cell.2018.03.043 (2018).
31 de Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251-256, doi:10.1126/science.1253462 (2014).
32 Campbell, P. J., Getz, G., Stuart, J. M., Korbel, J. O. & Stein, L. D. Pan-cancer analysis of whole genomes. *bioRxiv*, doi:10.1101/162784 (2017).
33 Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat Commun* **6**, 8971, doi:10.1038/ncomms9971 (2015).
34 Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *bioRxiv*, doi:10.1101/161562 (2017).
35 Sabarinathan, R. *et al.* The whole-genome panorama of cancer drivers. *bioRxiv*, doi:10.1101/190330 (2017).
36 Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886, doi:10.1126/science.aaa6806 (2015).

37 Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e1021, doi:10.1016/j.cell.2017.09.042 (2017).
38 Wala, J. A. *et al.* Selective and mechanistic sources of recurrent rearrangements across the cancer genome. *bioRxiv*, doi:10.1101/187609 (2017).
39 Bollard, J. *et al.* Palbociclib (PD-0332991), a selective CDK4/6 inhibitor, restricts tumour growth in preclinical models of hepatocellular carcinoma. *Gut* **66**, 1286-1296, doi:10.1136/gutjnl-2016-312268 (2017).
40 Takahira, T. *et al.* Alterations of the RB1 gene in dedifferentiated liposarcoma. *Mod Pathol* **18**, 1461-1470, doi:10.1038/modpathol.3800447 (2005).
41 Schneider-Stock, R. *et al.* Significance of loss of heterozygosity of the RB1 gene during tumour progression in well-differentiated liposarcomas. *The Journal of Pathology* **197**, 654-660, doi:doi:10.1002/path.1145 (2002).
42 Condorelli, R. *et al.* Polyclonal RB1 mutations and acquired resistance to CDK 4/6 inhibitors in patients with metastatic breast cancer. *Ann Oncol*, doi:10.1093/annonc/mdx784 (2017).
43 Yoshihara, K. *et al.* The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* **34**, 4845, doi:10.1038/onc.2014.406 (2014).
44 Rubanova, Y. *et al.* TrackSig: reconstructing evolutionary trajectories of mutation signature exposure. *bioRxiv*, doi:10.1101/260471 (2018).
45 Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet* **48**, 600-606, doi:10.1038/ng.3557 (2016).
46 Williams, M. J. *et al.* Quantification of subclonal selection in cancer from bulk sequencing data. *bioRxiv*, doi:10.1101/096305 (2016).
47 Sun, R. *et al.* Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat Genet* **49**, 1015-1024, doi:10.1038/ng.3891 (2017).
48 Mamlouk, S. *et al.* DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer. *Nat Commun* **8**, 14093, doi:10.1038/ncomms14093 (2017).
49 Pan, Q. *et al.* CDK4/6 Inhibitors in Cancer Therapy: A Novel Treatement Strategy for Bladder Cancer. *Bladder Cancer* **3**, 79-88, doi:10.3233/BLC-170105 (2017).
50 Knudsen, E. S. & Knudsen, K. E. Tailoring to RB: tumour suppressor status and therapeutic response. *Nat Rev Cancer* **8**, 714-724, doi:10.1038/nrc2401 (2008).
51 Levy, S. F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181-186, doi:10.1038/nature14279 (2015).
52 Vazquez-Garcia, I. *et al.* Clonal Heterogeneity Influences the Fate of New Adaptive Mutations. *Cell Rep* **21**, 732-744, doi:10.1016/j.celrep.2017.09.046 (2017).
53 Alizadeh, A. A. *et al.* Toward understanding and exploiting tumor heterogeneity. *Nat Med* **21**, 846-853, doi:10.1038/nm.3915 (2015).
54 Cmero, M. *et al.* SVclone: inferring structural variant cancer cell fraction. *bioRxiv*, doi:10.1101/172486 (2017).
55 Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the*

National Academy of Sciences **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

Figure legends

Figure 1

Consensus-based characterisation of intra-tumour heterogeneity. (**a**) Schematic representation of intra-tumour heterogeneity (ITH) reconstruction from sequencing data. (**b**) Samples with and without whole-genome duplications separate in two clusters according to their consensus ploidy and the fraction of the genome showing loss of heterozygosity. (**c**) Agreement between the six copy number callers using a multi-tier consensus copy number calling approach. The three lines denote the fraction of the genome at which agreement is reached at different levels of confidence: (near-)complete agreement on both alleles of clonal copy number, a strict majority agreement on both alleles of clonal copy number and (near-)complete or strict majority agreement on both alleles of rounded subclonal copy number (see **Supplementary Methods**). At the third level, agreement is reached on an average 93% of the genome. (**d**) Heatmap of the average pairwise similarities of subclonal architectures identified by 11 individual, 3 consensus, and 3 control reconstruction methods. Each method is represented by one coloured square on the diagonal. On rows and columns, each method is compared to all other methods. The upper triangle shows the average pairwise similarities on the 2,778 PCAWG samples, the lower triangle shows the same on a validation set of 965 simulated samples. In the leftmost column similarities are computed against the truth of the simulated set. Colour intensities scale with the similarities and were normalised separately for PCAWG, simulations and truth.

Figure 2

Overview and characterisation of ITH across cancer types. Evidence of ITH is shown for 1,868 samples with sufficient power to detect subclones at CCF > 30% (see **Supplementary Methods**). Only primary tumours and representative samples[32] from multi-sample cases are shown, except for melanoma, which holds only metastatic samples. Top to bottom: barplot showing the fraction of samples with given number of subclones; scatterplots showing the fractions of subclonal SNVs, indels, SVs and subclonal arm-level CNAs (the latter two mutation types are only plotted for samples

that have at least 5 events, sample order is determined by increasing fraction of subclonal SNVs and conserved in the other three panels); violin plots showing the total mutation burden and overall fraction of the genome that does not have a copy number state of 1+1, or 2+2 in WGD samples; heatmaps showing the fraction of tumour samples with whole genome duplications and the mean power to identify subclones per cancer types (number of reads per clonal copy – nrpcc, see **Supplementary Methods).**

Figure 3

SNV and indel drivers and subclonal selection. (**a**) Heat map of the fraction of subclonal driver mutations in different cancer types. Square size scales with the number of samples containing that specific driver mutation. Marginal bar plots represent the aggregated subclonal driver proportion by gene (right) or cancer type (top). A track on the left displays gene set and pathway annotations for driver genes, highlighting an enrichment of subclonally mutated drivers in specific gene classes, such as chromatin remodelling. (**b**) Survey of clinically actionable driver mutations across cancer types, stratified by clonal status. (**c**) dN/dS values for clonal and subclonal SNVs in 566 established cancer genes across all primary tumours. Values for missense, nonsense, splice site, and all mutations are shown, along with the 95% confidence intervals. (**d**) dN/dS values further stratified for "neutral" and "non-neutral" tumours according to Williams *et al.*[27] based on an R^2 cutoff of the linear fit between cumulative number of SNVs at a given allelic frequency f (M(f)) and $\frac{1}{f}$. (**e**) Cancer and mutation types for which dN/dS is significantly greater than 1 (95% confidence intervals>1) for clonal and subclonal mutations. Cancer types are ordered by the total number of samples (N).

Figure 4

Clonality of significantly recurrent breakpoints and gene fusions. (a) Number of structural variants per sample grouped by cancer type. Boxplots show the interquartile range and dots represent values for individual samples. The vertical dashed line shows the pan-cancer average per patient. **(b)** Proportion of SVs falling in regions of

recurrent breakpoints (SRBs, significantly recurrent breakpoints). Boxplots show the interquartile range and dots represent values for individual samples. The vertical dashed line shows the pan-cancer average per patient. **(c)** Subclonal probabilities of SVs grouped by cancer type and divided into two categories: significantly recurrent breakpoints (candidate driver SVs) and non-recurrent breakpoints (candidate passenger SVs). The triangles and circles represent median probabilities of being subclonal, weighted by the number of reads per chromosome copy as a measure of the subclonal detection power (**Supplementary Methods**). The lines represent the interquartile range. * marks significant differences between candidate driver and passenger medians ($q < 0.05$, permutation test, effect size > 0.05). Cancer types with clonal driver SVs (to the left) suggest that SVs play a role in cancer initiation and early progression, whereas cancers with subclonal driver SVs (to the right) suggest a stronger role in driving cancer heterogeneity. Cancer types spanning the whole probability range may indicate a role for SVs throughout cancer development. **(d)** Clonal and subclonal enrichment of loci containing recurrent breakpoints. SVs were ranked by their weighted subclonal probability and those falling within a recurrently hit locus are shaded in the middle panel. Those appearing above the line had probabilities < 50% and below ≥ 50. Coloured lines (according to tumour type) represent breaks that contributed to the leading edge of the enrichment test, other SVs are shown in grey. Genes on the left were previously reported as the likely candidate driver at each locus[38]. The q-values represent the multiple testing-adjusted probability of achieving an enrichment score greater than the observed score, under a permutation test. **(e)** Clonality of driver gene fusions versus non-driver fusions.

Figure 5

Further characterisation of ITH using mutation phasing and "winner's curse" correction. (**a**, **b**) Proportion of tumours with evidence of branching and linear phylogenies, through analysis of phased reads of variants in-*trans* (**a**) or in-*cis* (**b**) among tumours with sufficient phased reads. Error bars are +/- the binomial standard deviation at the associated ratio and the total number of tumours. (**c**, **d**) Correction results for the "winner's curse"-like effect in all mutational clusters identified in the

study. Subclonal clusters show a shift to larger CCF values after correction (**c**) and the majority of clusters are estimated to contain additional missed SNVs (**d**).

Figure 6

Subclonal boundaries are associated with changes in mutation signature activity. (**a**) Mutational signature changes across cancer types. Bar graphs show the proportion of tumours in which signature (pairs) change and radial plots provide a view per cancer type. Each radial contains the signatures that are active in at least 5 tumours and change (≥ 6%) in at least 3 tumours. The left and right side of the radial represent signatures that become less and more active, respectively. The height of a wedge represents the average activity change (log scale), the colour denotes the signature and the transparency shows the fraction of tumours in which the signature changes (as a proportion of the tumours in which the signature is active). Signatures are sorted around the radial (top-to-bottom) by maximum average activity change. (**b**) Average signature trajectories for selected cancer types. Each line is coloured by signature and corresponds to the average activity across tumours of this cancer type in which the signature is active. The width of the line represents the number of tumours that are represented. Mutations are split into clonal and subclonal, visually divided by a red vertical line. (**c**) Signature trajectories for selected individual CLL tumours. Each line corresponds to an activity trajectory derived from a bootstrap sample of SNVs. The grey vertical grid represents the mutation bins. These are coloured grey when a significant change in signature activity is detected. Red vertical lines represent consensus subclonal mutation clusters. (**d**) The fraction of signature change points that coincide with boundaries between mutation clusters, as compared to what is expected when randomly placing change points. (**e**) The number of subclones detected in tumours grouped by the maximum detected signature activity change. (**f**) The number of tumours in which evidence of additional subclones is detected beyond those identified through clustering of SNVs.

Supplementary Figure Legends

Supplementary Figure 1

Validation of consensus purity values. The lower triangle shows pairwise scatterplots of the purities obtained through expression profiles of a panel of immune and stromal genes (ESTIMATE), somatic copy number data (ABSOLUTE), leukocyte unmethylation (LUMP), image analysis by haematoxylin and eosin staining (H&E staining), and consensus purity as derived by Aran *et al.*[33] (CPE). The top triangle shows the respective Pearson correlation coefficients and the number of samples that have both purity estimates available.

Supplementary Figure 2

Power analysis of the consensus subclonal architecture approach. (**a**) Our ability to detect subclones depends, not on the number of detected SNVs, but on the number of reads per clonal copy (nrpcc) available. This metric takes tumour purity, ploidy and sequencing coverage into account (see **Supplementary Methods**). We control for this effect by including only tumours with nrpcc $\geq$ 10. In these tumours, we should be sufficiently powered to detect a subclone at a CCF as low as 30% (see **Supplementary Methods**). This becomes clear from (**b**) which shows the minimum CCF of the detected clusters in each tumour against the number of reads per chromosome copy.

Supplementary Figure 3

Overview and characterisation of ITH across metastases and recurrences. (**a**) Overview of high-powered metastatic and recurrent samples that were excluded from **Fig. 2** (except for melanomas). Top-to-bottom: bar plot showing the fraction of samples with the indicated number of identified subclones; violin plots of the fractions of subclonal SNVs and arm level CNAs (for samples that have at least 5 events). The fraction of subclonal CNAs represents the number of subclonal arm-level events, out of all arm-level events across the genome. (**b**) An orthogonal approach not relying on mutation clustering (see **Supplementary Methods**) was applied to

conservatively quantify the proportion of subclonal SNVs in melanoma metastases. We observe that tumours identified as clonal contain a significantly lower proportion of subclonal SNVs (*p*-value = 8.4×10^{-5}, Kolmogorov-Smirnov test), in line with findings obtained through clustering of SNVs.

Supplementary Figure 4

Correlation in ITH between SNVs, indels, CNAs and SVs by cancer type. Evidence of ITH is shown for 1,868 samples with sufficient power to detect subclones above 30% CCF (see **Supplementary Methods**), as in **Fig. 2**. Pairwise scatterplots in the upper triangle show the fractions of subclonalSNVs, indels, CNAs and SVs per tumour sample. Pearson's correlation coefficient, *R*, is separately computed for each panel across all samples. Panels on the diagonal show the kernel density estimate of the distribution of subclonal fractions. In the lower triangle, each point shows the median subclonal fraction per cancer type and intervals indicate the interquartile range. Panels only include samples with at least 5 arm-level CNAs (1,217 / 1,868) and at least 5 SVs (1,405 / 1,868).

Supplementary Figure 5

Clonality analysis of significantly recurrent breakpoints. (**a**) Number and clonality of SVs observed in 52 loci with significantly recurrent breakpoints (SRBs)[38]. SVs with a subclonal probability larger than 50% were considered subclonal and clonal otherwise. (**b**) Proportion of cancer types contributing to the enrichment of clonal or subclonal SVs in a locus (see **Fig. 4d**). The genes on the y-axis represent the most likely driver gene for each locus[38].

Supplementary Figure 6

Evaluation of "winner's curse" correction (WCC) on simulated data. Corrected cellular prevalence (consensus from two correction methods) shows good concordance with the true cellular prevalence from simulated samples.

Supplementary Figure 7

Summary signature trajectories per cancer type. The average trajectories for mutational signatures were calculated across tumours of the same cancer type. The colour of the line denotes the signature and its width reflects the number of contributing tumours. The trajectories have been centred around the activity at the boundary between clonal and subclonal mutations in order to highlight relative changes in signature activity.

Acknowledgements

This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001202), the UK Medical Research Council (FC001202), and the Wellcome Trust (FC001202). This project was enabled through access to the MRC eMedLab Medical Bioinformatics infrastructure, supported by the Medical Research Council (grant number MR/L016311/1). MT and JD are postdoctoral fellows supported by the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie Grant Agreement No. 747852-SIOMICS and 703594-DECODE). JD is a postdoctoral fellow of the Research Foundation – Flanders (FWO). IVG was supported by a Wellcome Trust PhD fellowship (grant number WT097678). SM is funded by a Vanier Canada Graduate Scholarship. SCS is supported by the NSERC Discovery Frontiers Project, "The Cancer Genome Collaboratory" and by NIH GM108308. DJA is supported by Cancer Research UK. FM, GM and KeY would like to acknowledge the support of the University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited. GM, KeY and FM are funded by CRUK core grants C14303/A17197 and A19274. SSe and YJ are supported by NIH R01 CA132897. HZ is supported by grant NIMH086633 and an endowed Bao-Shan Jing Professorship in Diagnostic Imaging. PTS is supported by U24CA210957 and 1U24CA143799. WW is supported by the U.S. National Cancer Institute (1R01 CA183793 and P30 CA016672). DCW is funded by the Li Ka Shing foundation. PVL is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute. We gratefully acknowledge Nicholas McGranahan and Charles Swanton for valuable comments on our manuscript.

Members of the PCAWG Evolution and Heterogeneity Working Group

Stefan C. Dentro[1,2,3,*], Ignaty Leshchiner[4,*], Moritz Gerstung[5,*], Clemency Jolly[1,*], Kerstin Haase[1,*], Maxime Tarabichi[1,2,*], Jeff Wintersinger[6,*], Amit Deshwar[6,*], Kaixian Yu[7,*], Santiago Gonzalez[5,*], Yulia Rubanova[6,*], Geoff Macintyre[8,*], David J. Adams[2], Pavana Anur[9], Rameen Beroukhim[4], Paul C. Boutros[6,10], David D. Bowtell[11,12], Peter J. Campbell[2], Shaolong Cao[7], Elizabeth L. Christie[11], Marek Cmero[13], Yupeng Cun[14], Kevin Dawson[2], Jonas Demeulemeester[1,15], Nilgun Donmez[16], Ruben M. Drews[8], Roland Eils[17,18], Yu Fan[7], Matthew Fittall[1], Dale W. Garsed[11,13], Gad Getz[4], Gavin Ha[4], Marcin Imielinski[19,20], Lara Jerman[5,21], Yuan Ji[22,23], Kortine Kleinheinz[17,18], Juhee Lee[24], Henry Lee-Six[2], Jialu Li[7], Dimitri G. Livitz[4], Salem Malikic[16], Florian Markowetz[8], Inigo Martincorena[2], Thomas J. Mitchell[2,25], Ville Mustonen[26], Layla Oesper[27], Martin Peifer[14], Myron Peto[9], Benjamin J. Raphael[28], Daniel Rosebrock[4], S. Cenk Sahinalp[29], Adriana Salcedo[10], Matthias Schlesner[17], Steven Schumacher[4], Subhajit Sengupta[22], Ruian Shi[6], Seung Jun Shin[7,30], Lincoln D. Stein[10], Ignacio Vázquez-García[2,25], Shankar Vembu[6], David A. Wheeler[31], Tsun-Po Yang[14], Xiaotong Yao[19,20], Ke Yuan[8,32], Hongtu Zhu[7], Wenyi Wang[7,#], Quaid D. Morris[6,#], Paul T. Spellman[9,#], David C. Wedge[3,33,#], Peter Van Loo[1,15,#]

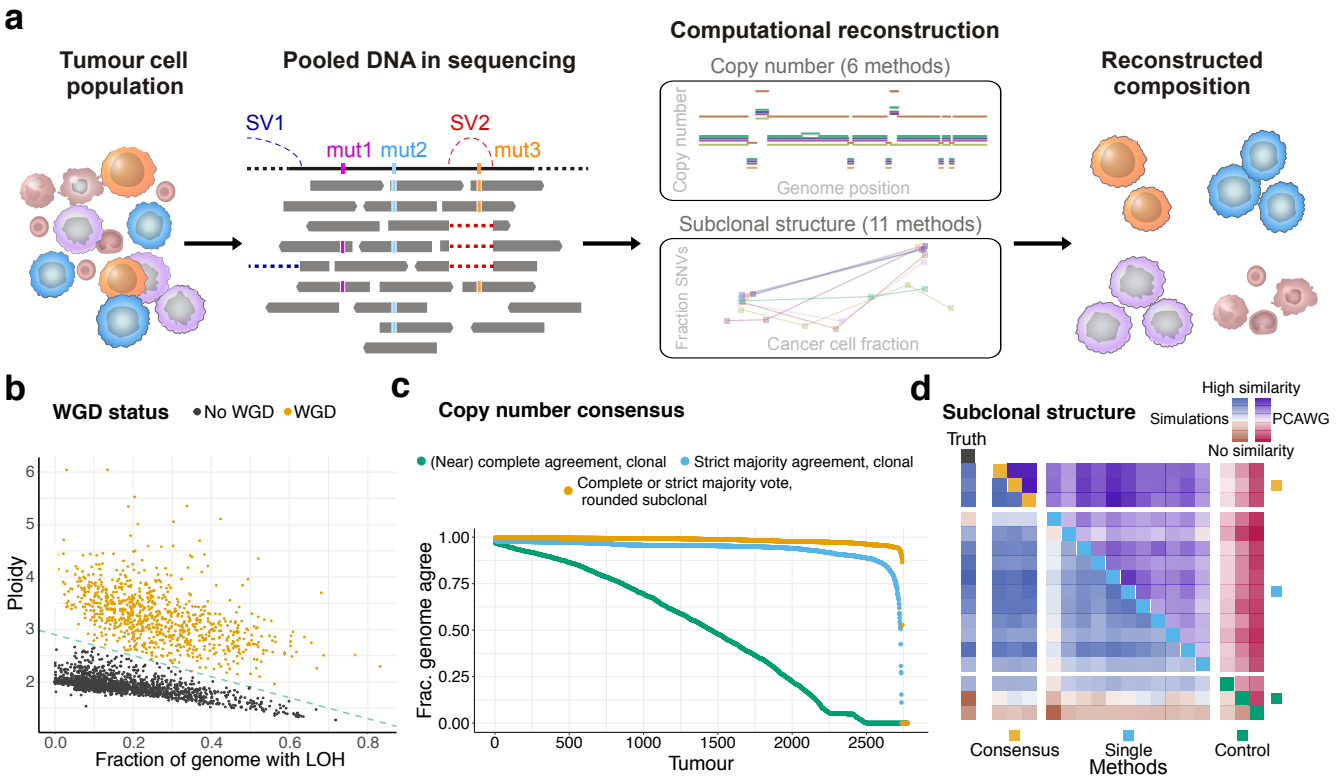
[1]The Francis Crick Institute, London NW1 1AT, United Kingdom; [2]Wellcome Trust Sanger Institute, Cambridge CB10 1SA, United Kingdom; [3]Big Data Institute, University of Oxford, Oxford OX3 7LF, United Kingdom; [4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [5]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Cambridge CB10 1SD, United Kingdom; [6]University of Toronto, Toronto, ON M5S 3E1, Canada; [7]The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; [8]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, United Kingdom; [9]Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR 97231, USA; [10]Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada; [11]Peter MacCallum Cancer Centre, Melbourne, VIC 3000, Australia; [12]Garvan Institute of Medical Research, Sydney, NSW 2010, Australia; [13]University of Melbourne, Melbourne, VIC 3010, Australia; [14]University of Cologne, 50931 Cologne, Germany; [15]Department of Human Genetics, University of Leuven, B-3000 Leuven, Belgium; [16]Simon Fraser University, Burnaby, BC V5A1S6, Canada; [17]German Cancer

Research Center (DKFZ), 69120 Heidelberg, Germany; [18]Heidelberg University, 69120 Heidelberg, Germany; [19]Weill Cornell Medicine, New York, NY 10065, USA; [20]New York Genome Center, New York, NY 10013, USA; [21]University of Ljubljana, 1000 Ljubljana, Slovenia; [22]NorthShore University HealthSystem, Evanston, IL 60201, USA; [23]The University of Chicago, Chicago, IL 60637, USA; [24]University of California Santa Cruz, Santa Cruz, CA 95064, USA; [25]University of Cambridge, Cambridge CB2 0QQ, United Kingdom; [26]University of Helsinki, 00014 Helsinki, Finland; [27]Carleton College, Northfield, MN 55057, USA; [28]Princeton University, Princeton, NJ 08540, USA; [29]Indiana University, Bloomington, IN 47405, USA; [30]Korea University, Seoul, 02481, Republic of Korea; [31]Baylor College of Medicine, Houston, TX 77030, USA; [32]University of Glasgow, Glasgow G12 8RZ, United Kingdom; [33]Oxford NIHR Biomedical Research Centre, Oxford OX4 2PG, United Kingdom.

*: These authors contributed equally

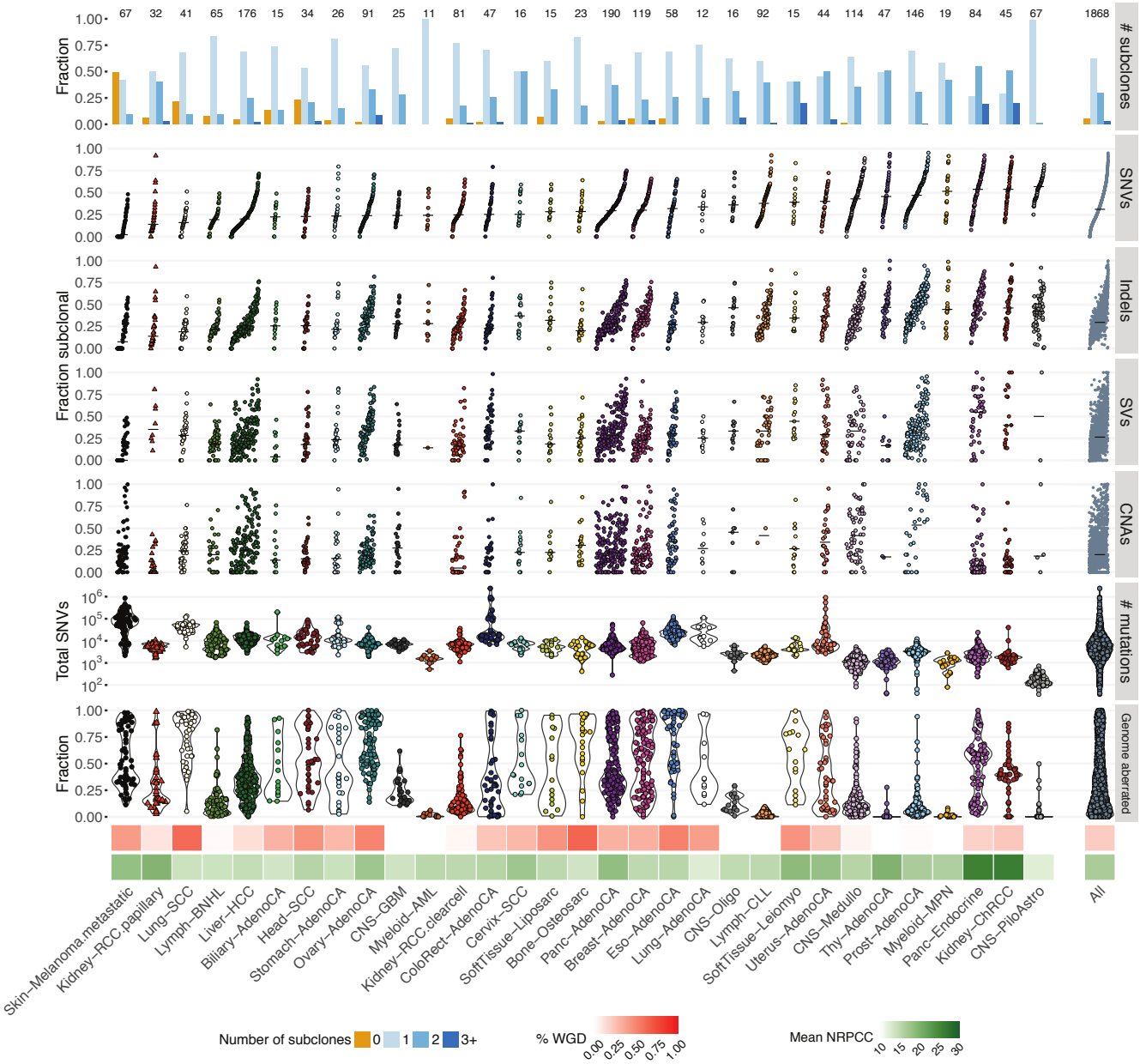
[#]: These authors jointly directed the work

Figure 1



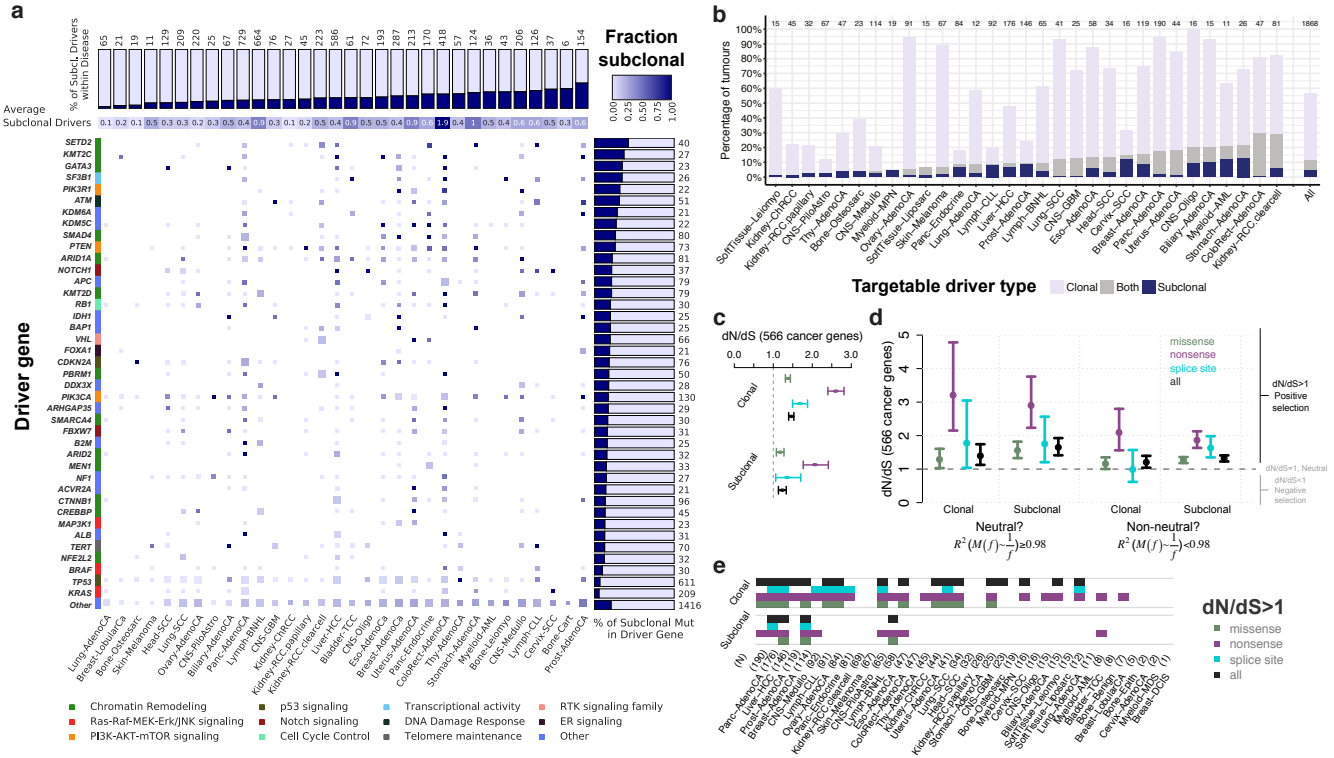
Consensus-based characterisation of intra-tumour heterogeneity. (a) Schematic representation of intra-tumour heterogeneity (ITH) reconstruction from sequencing data. (b) Samples with and without whole-genome duplications separate in two clusters according to their consensus ploidy and the fraction of the genome showing loss of heterozygosity. (c) Agreement between the six copy number callers using a multi-tier consensus copy number calling approach. The three lines denote the fraction of the genome at which agreement is reached at different levels of confidence: (near-)complete agreement on both alleles of clonal copy number, a strict majority agreement on both alleles of clonal copy number and (near-)complete or strict majority agreement on both alleles of rounded subclonal copy number (see Supplementary Methods). At the third level, agreement is reached on an average 93% of the genome. (d) Heatmap of the average pairwise similarities of subclonal architectures identified by 11 individual, 3 consensus, and 3 control reconstruction methods. Each method is represented by one coloured square on the diagonal. On rows and columns, each method is compared to all other methods. The upper triangle shows the average pairwise similarities on the 2,778 PCAWG samples, the lower triangle shows the same on a validation set of 965 simulated samples. In the leftmost column similarities are computed against the truth of the simulated set. Colour intensities scale with the similarities and were normalised separately for PCAWG, simulations and truth.

Figure 2



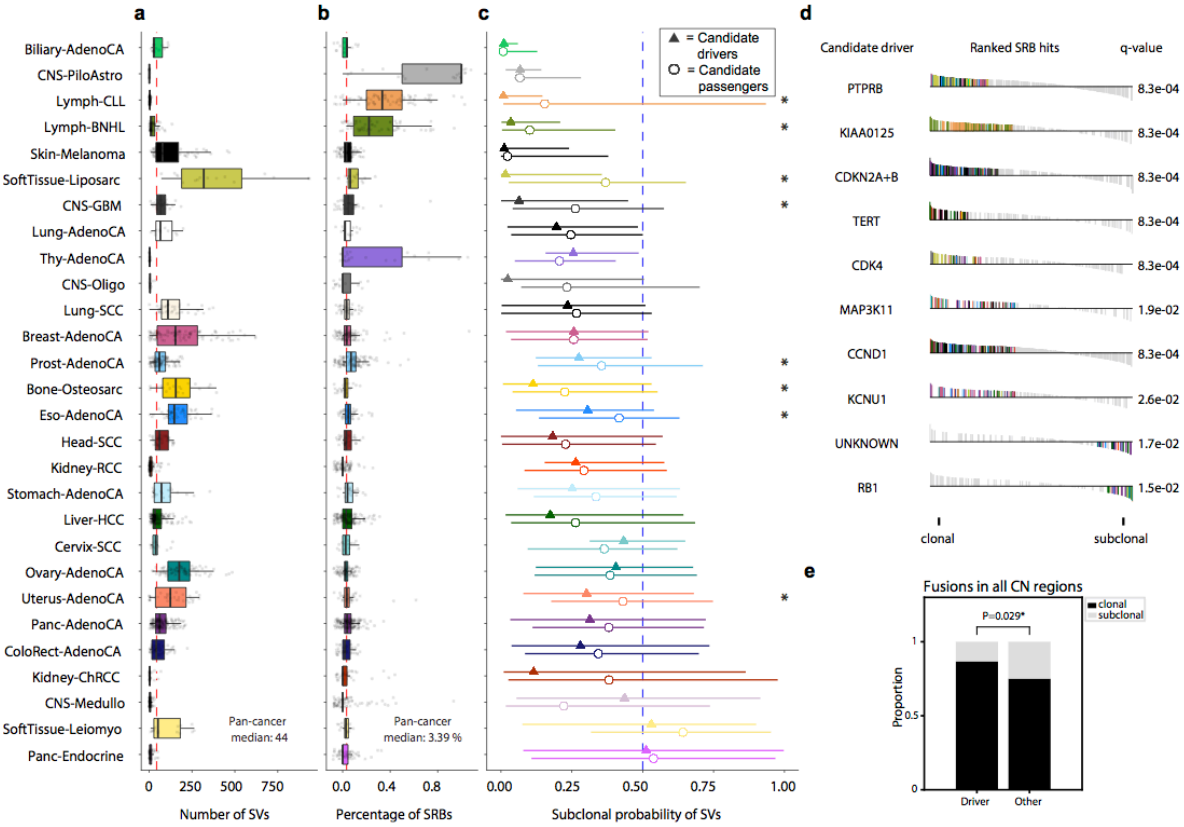
Overview and characterisation of ITH across cancer types. Evidence of ITH is shown for 1,868 samples with sufficient power to detect subclones at CCF > 30% (see Supplementary Methods). Only primary tumours and representative samples30 from multi-sample cases are shown, except for melanoma, which holds only metastatic samples. Top to bottom: barplot showing the fraction of samples with given number of subclones; scatterplots showing the fractions of subclonal SNVs, indels, SVs and subclonal arm-level CNAs (the latter two mutation types are only plotted for samples that have at least 5 events, sample order is determined by increasing fraction of subclonal SNVs and conserved in the other three panels); violin plots showing the total mutation burden and overall fraction of the genome that does not have a copy number state of 1+1, or 2+2 in WGD samples; heatmaps showing the fraction of tumour samples with whole genome duplications and the mean power to identify subclones per cancer types (number of reads per clonal copy – nrpcc, see Supplementary Methods).

Figure 3



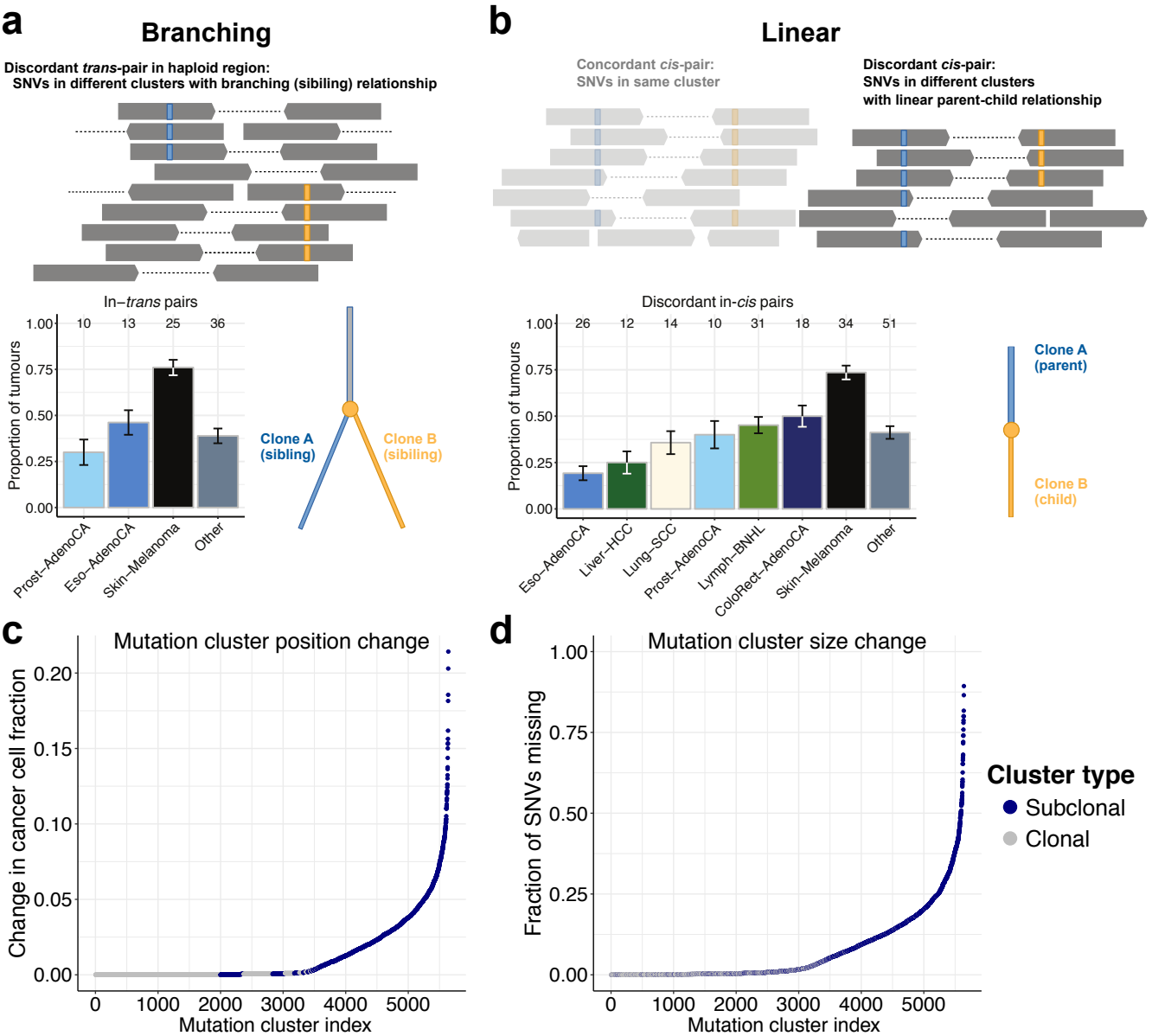
SNV and indel drivers and subclonal selection. (a) Heat map of the fraction of subclonal driver mutations in different cancer types. Square size scales with the number of samples containing that specific driver mutation. Marginal bar plots represent the aggregated subclonal driver proportion by gene (right) or cancer type (top). A track on the left displays gene set and pathway annotations for driver genes, highlighting an enrichment of subclonally mutated drivers in specific gene classes, such as chromatin remodelling. (b) Survey of clinically actionable driver mutations across cancer types, stratified by clonal status. (c) dN/dS values for clonal and subclonal SNVs in 566 established cancer genes across all primary tumours. Values for missense, nonsense, splice site, and all mutations are shown, along with the 95% confidence intervals. (d) dN/dS values further stratified for “neutral” and “non-neutral” tumours according to Williams et al. based on an R2 cutoff of the linear fit between cumulative number of SNVs at a given allelic frequency f (M(f)) and 1/f. (e) Cancer and mutation types for which dN/dS is significantly greater than 1 (95% confidence intervals>1) for clonal and subclonal mutations. Cancer types are ordered by the total number of samples (N).

Figure 4



Clonality of significantly recurrent breakpoints and gene fusions. (a) Number of structural variants per sample grouped by cancer type. Boxplots show the interquartile range and dots represent values for individual samples. The vertical dashed line shows the pan-cancer average per patient. (b) Proportion of SVs falling in regions of recurrent breakpoints (SRBs, significantly recurrent breakpoints). Boxplots show the interquartile range and dots represent values for individual samples. The vertical dashed line shows the pan-cancer average per patient. (c) Subclonal probabilities of SVs grouped by cancer type and divided into two categories: significantly recurrent breakpoints (candidate driver SVs) and non-recurrent breakpoints (candidate passenger SVs). The triangles and circles represent median probabilities of being subclonal, weighted by the number of reads per chromosome copy as a measure of the subclonal detection power (Supplementary Methods). The lines represent the interquartile range. * marks significant differences between candidate driver and passenger medians ($q < 0.05$, permutation test, effect size > 0.05). Cancer types with clonal driver SVs (to the left) suggest that SVs play a role in cancer initiation and early progression, whereas cancers with subclonal driver SVs (to the right) suggest a stronger role in driving cancer heterogeneity. Cancer types spanning the whole probability range may indicate a role for SVs throughout cancer development. (d) Clonal and subclonal enrichment of loci containing recurrent breakpoints. SVs were ranked by their weighted subclonal probability and those falling within a recurrently hit locus are shaded in the middle panel. Those appearing above the line had probabilities < 50% and below ≥ 50. Coloured lines (according to tumour type) represent breaks that contributed to the leading edge of the enrichment test, other SVs are shown in grey. Genes on the left were previously reported as the likely candidate driver at each locus. The *q*-values represent the multiple testing-adjusted probability of achieving an enrichment score greater than the observed score, under a permutation test. (e) Clonality of driver gene fusions versus non-driver fusions.

Figure 5



Further characterisation of ITH using mutation phasing and "winner's curse" correction. (a, b) Proportion of tumours with evidence of branching and linear phylogenies, through analysis of phased reads of variants in-trans (a) or in-cis (b) among tumours with sufficient phased reads. Error bars are +/- the binomial standard deviation at the associated ratio and the total number of tumours. (c, d) Correction results for the "winner's curse"-like effect in all mutational clusters identified in the study. Subclonal clusters show a shift to larger CCF values after correction (c) and the majority of clusters are estimated to contain additional missed SNVs (d).

Figure 6

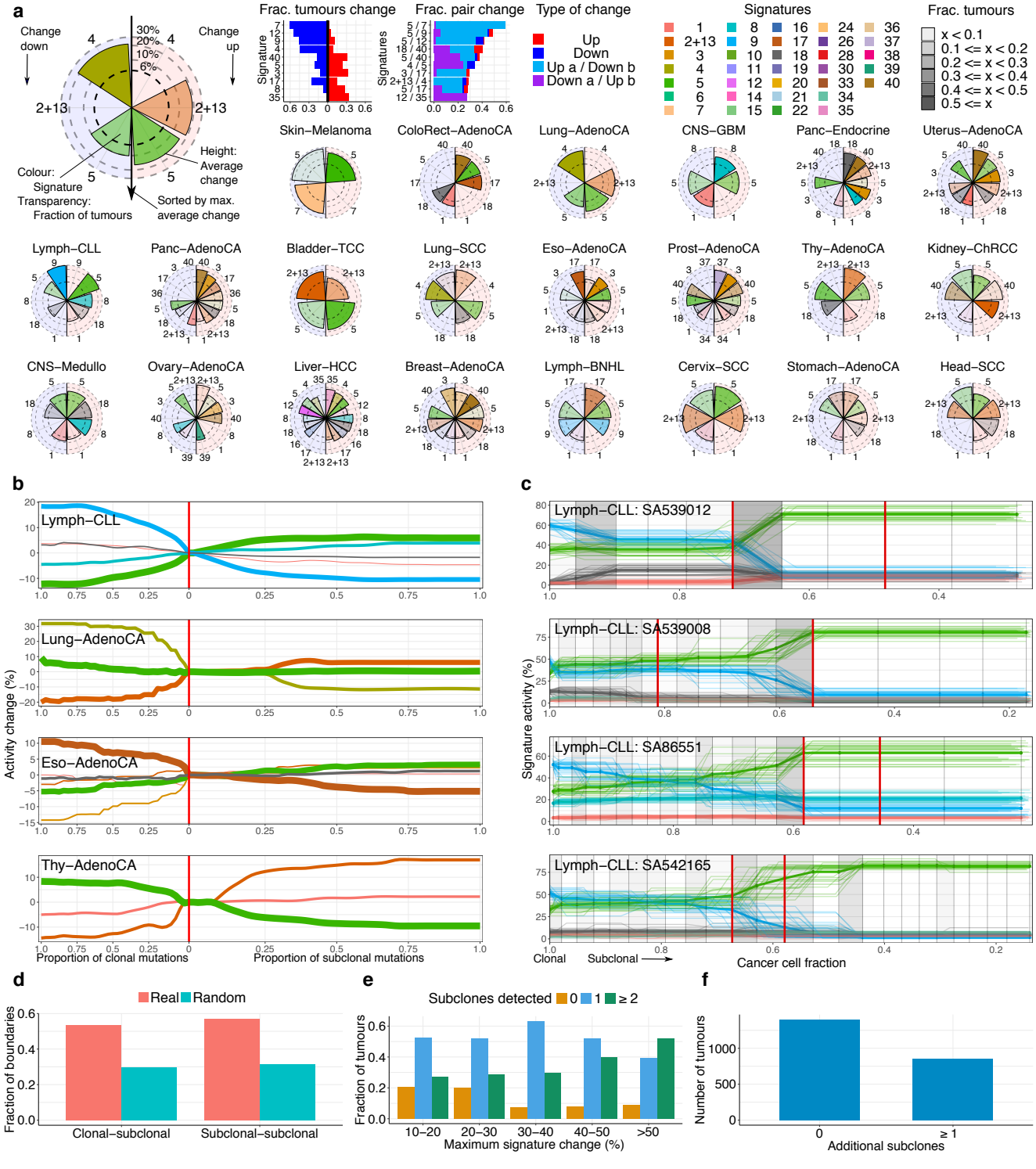
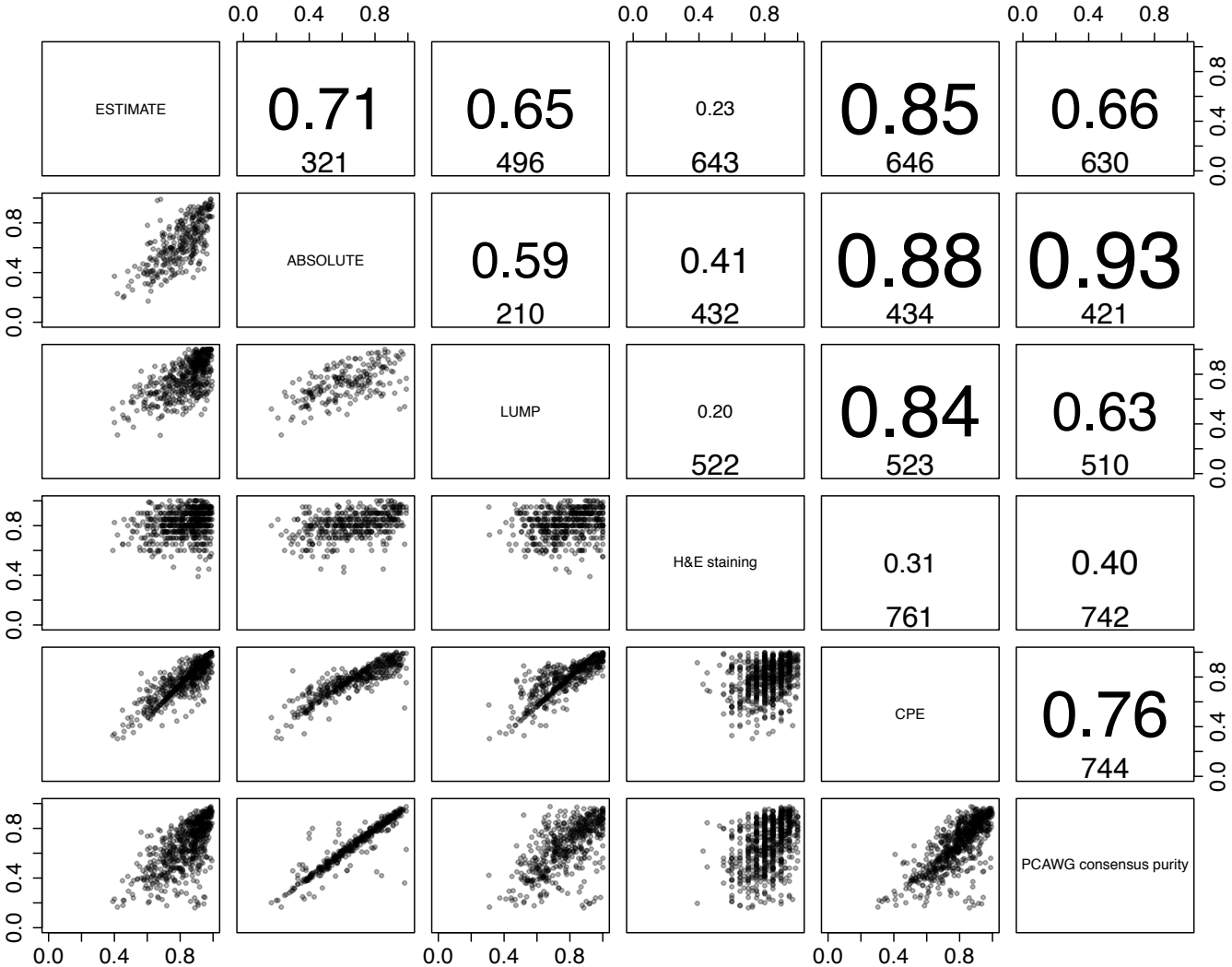

a
Change down
Change up
30%
20%
10%
6%
2+13
Height: Average change
Colour: Signature
Transparency: Fraction of tumours
Sorted by max. average change
Frac. tumours change
Frac. pair change
Signature
Signatures
Type of change
Up
Down
Up a / Down b
Down a / Up b
Signatures
Frac. tumours
x < 0.1
0.1 <= x < 0.2
0.2 <= x < 0.3
0.3 <= x < 0.4
0.4 <= x < 0.5
0.5 <= x
Skin–Melanoma
ColoRect–AdenoCA
Lung–AdenoCA
CNS–GBM
Panc–Endocrine
Uterus–AdenoCA
Lymph–CLL
Panc–AdenoCA
Bladder–TCC
Lung–SCC
Eso–AdenoCA
Prost–AdenoCA
Thy–AdenoCA
Kidney–ChRCC
CNS–Medullo
Ovary–AdenoCA
Liver–HCC
Breast–AdenoCA
Lymph–BNHL
Cervix–SCC
Stomach–AdenoCA
Head–SCC
b
Lymph–CLL
Lung–AdenoCA
Eso–AdenoCA
Thy–AdenoCA
Activity change (%)
Proportion of clonal mutations
Proportion of subclonal mutations
c
Lymph–CLL: SA539012
Lymph–CLL: SA539008
Lymph–CLL: SA86551
Lymph–CLL: SA542165
Signature activity (%)
Clonal
Subclonal
Cancer cell fraction
d
Real
Random
Fraction of boundaries
Clonal–subclonal
Subclonal–subclonal
e
Subclones detected
0
1
≥ 2
Fraction of tumours
10–20
20–30
30–40
40–50
>50
Maximum signature change (%)
f
Number of tumours
0
≥ 1
Additional subclones

Figure 6

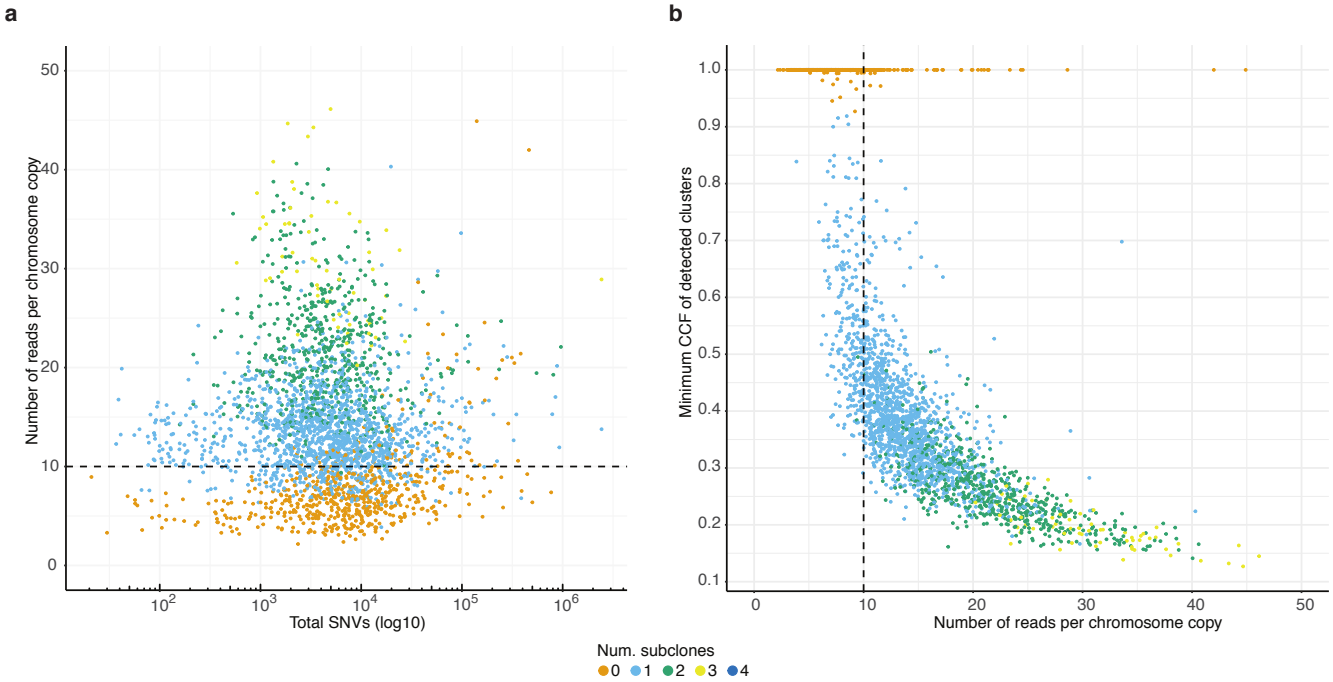
Subclonal boundaries are associated with changes in mutation signature activity. (a) Mutational signature changes across cancer types. Bar graphs show the proportion of tumours in which signature (pairs) change and radial plots provide a view per cancer type. Each radial contains the signatures that are active in at least 5 tumours and change (≥ 6%) in at least 3 tumours. The left and right side of the radial represent signatures that become less and more active, respectively. The height of a wedge represents the average activity change (log scale), the colour denotes the signature and the transparency shows the fraction of tumours in which the signature changes (as a proportion of the tumours in which the signature is active). Signatures are sorted around the radial (top-to-bottom) by maximum average activity change. (b) Average signature trajectories for selected cancer types. Each line is coloured by signature and corresponds to the average activity across tumours of this cancer type in which the signature is active. The width of the line represents the number of tumours that are represented. Mutations are split into clonal and subclonal, visually divided by a red vertical line. (c) Signature trajectories for selected individual CLL tumours. Each line corresponds to an activity trajectory derived from a bootstrap sample of SNVs. The grey vertical grid represents the mutation bins. These are coloured grey when a significant change in signature activity is detected. Red vertical lines represent consensus subclonal mutation clusters. (d) The fraction of signature change points that coincide with boundaries between mutation clusters, as compared to what is expected when randomly placing change points. (e) The number of subclones detected in tumours grouped by the maximum detected signature activity change. (f) The number of tumours in which evidence of additional subclones is detected beyond those identified through clustering of SNVs.

Supplementary Figure 1



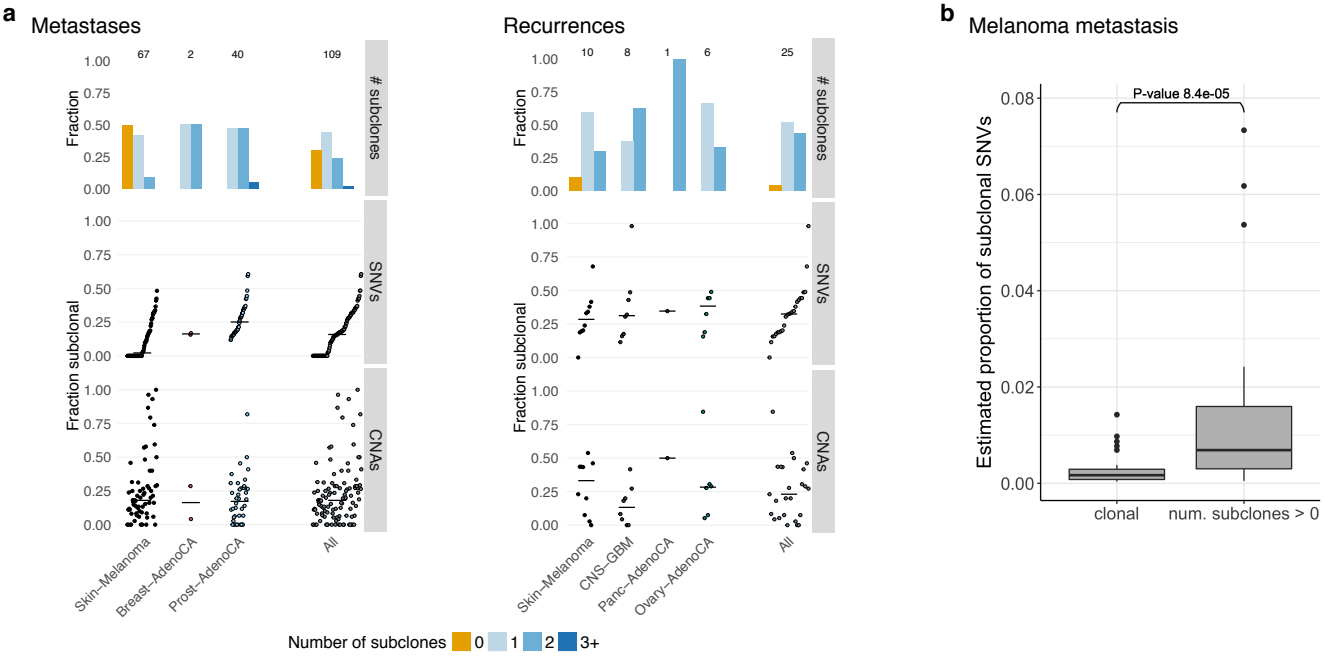
Validation of consensus purity values. The lower triangle shows pairwise scatterplots of the purities obtained through expression profiles of a panel of immune and stromal genes (ESTIMATE), somatic copy number data (ABSOLUTE), leukocyte unmethylation (LUMP), image analysis by haematoxylin and eosin staining (H&E staining), and consensus purity as derived by Aran et al.31 (CPE). The top triangle shows the respective Pearson correlation coefficients and the number of samples that have both purity estimates available.

Supplementary Figure 2



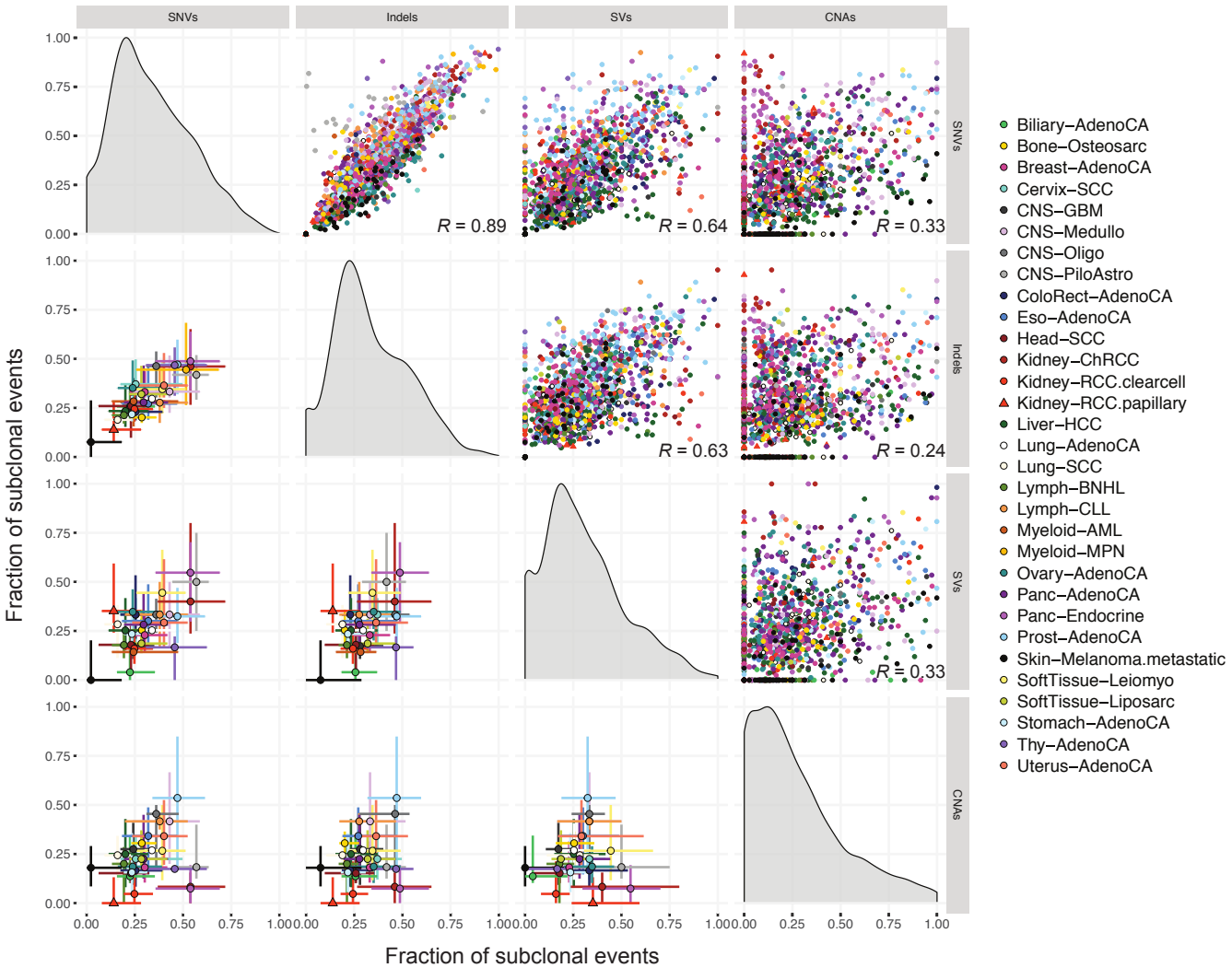
Power analysis of the consensus subclonal architecture approach. (a) Our ability to detect subclones depends, not on the number of detected SNVs, but on the number of reads per clonal copy (nrpcc) available. This metric takes tumour purity, ploidy and sequencing coverage into account (see Supplementary Methods). We control for this effect by including only tumours with nrpcc ≥ 10. In these tumours, we should be sufficiently powered to detect a subclone at a CCF as low as 30% (see Supplementary Methods). This becomes clear from (b) which shows the minimum CCF of the detected clusters in each tumour against the number of reads per chromosome copy.

Supplementary Figure 3



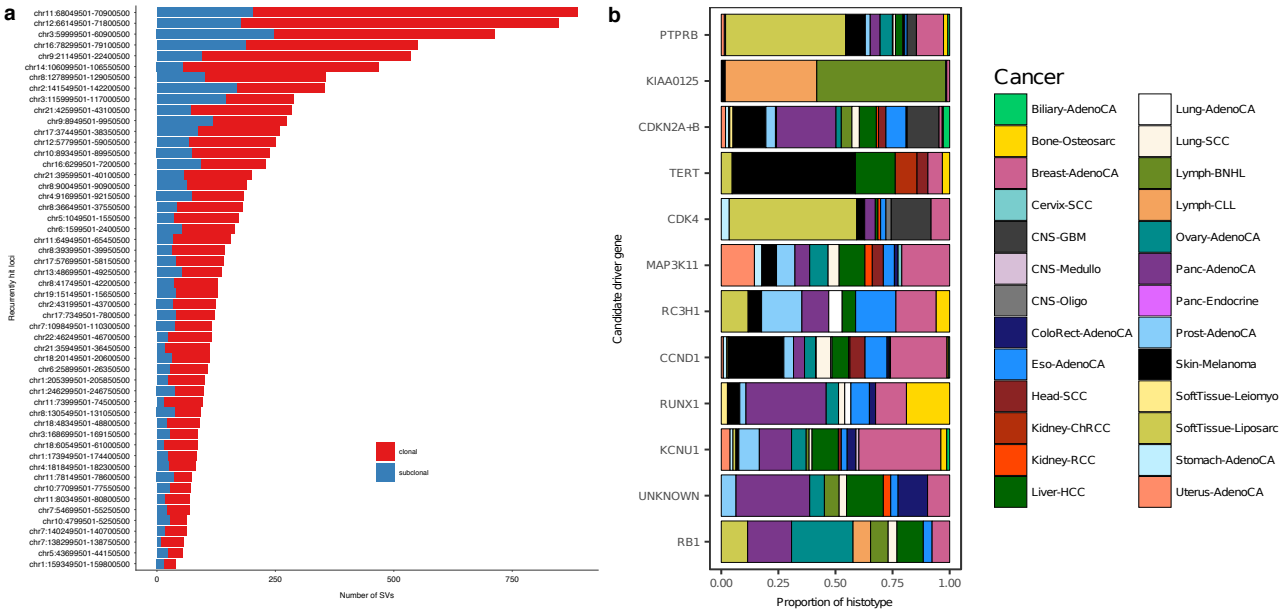
Overview and characterisation of ITH across metastases and recurrences. (a) Overview of high-powered metastatic and recurrent samples that were excluded from Fig. 2 (except for melanomas). Top-to-bottom: bar plot showing the fraction of samples with the indicated number of identified subclones; violin plots of the fractions of subclonal SNVs and arm level CNAs (for samples that have at least 5 events). The fraction of subclonal CNAs represents the number of subclonal arm-level events, out of all arm-level events across the genome. (b) An orthogonal approach not relying on mutation clustering (see Supplementary Methods) was applied to conservatively quantify the proportion of subclonal SNVs in melanoma metastases. We observe that tumours identified as clonal contain a significantly lower proportion of subclonal SNVs (p-value = 8.4x10^-5, Kolmogorov-Smirnov test), in line with findings obtained through clustering of SNVs.

Supplementary Figure 4



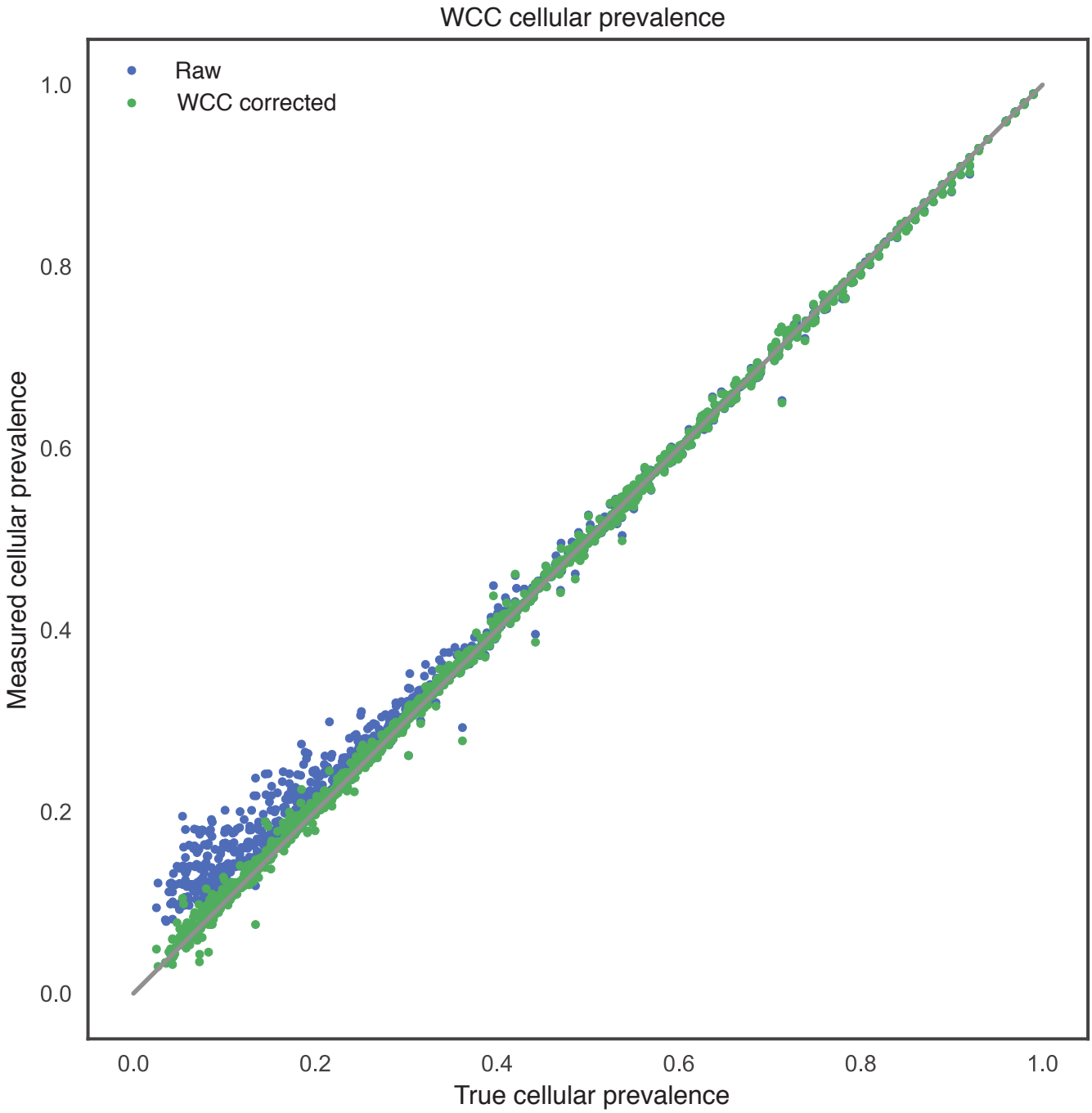
Correlation in ITH between SNVs, indels, CNAs and SVs by cancer type. Evidence of ITH is shown for 1,868 samples with sufficient power to detect subclones above 30% CCF (see Supplementary Methods), as in Fig. 2. Pairwise scatterplots in the upper triangle show the fractions of subclonalSNVs, indels, CNAs and SVs per tumour sample. Pearson's correlation coefficient, R, is separately computed for each panel across all samples. Panels on the diagonal show the kernel density estimate of the distribution of subclonal fractions. In the lower triangle, each point shows the median subclonal fraction per cancer type and intervals indicate the interquartile range. Panels only include samples with at least 5 arm-level CNAs (1,217 / 1,868) and at least 5 SVs (1,405 / 1,868).

Supplementary Figure 5



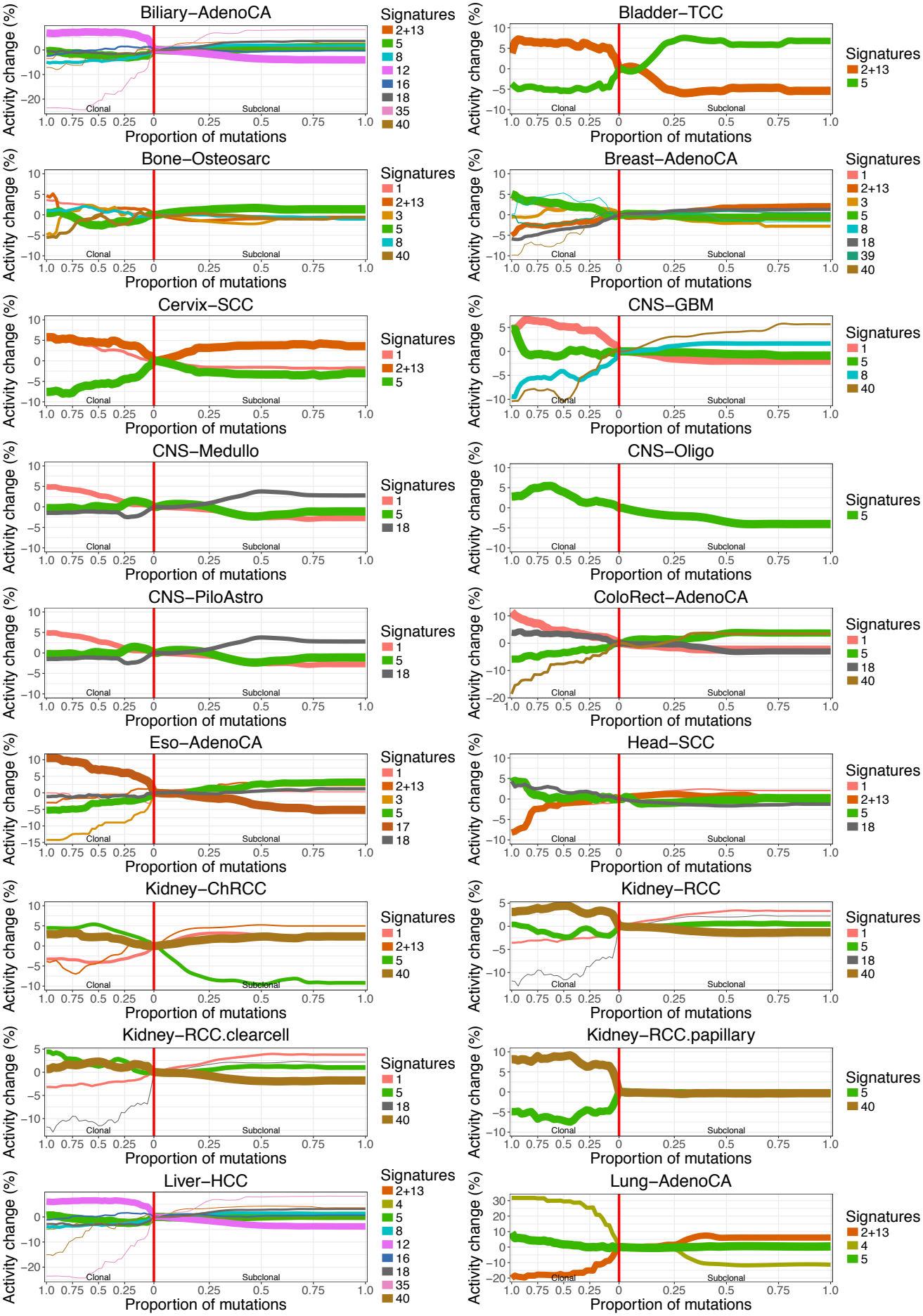
Clonality analysis of significantly recurrent breakpoints. (a) Number and clonality of SVs observed in 52 loci with significantly recurrent breakpoints (SRBs). SVs with a subclonal probability larger than 50% were considered subclonal and clonal otherwise. (b) Proportion of cancer types contributing to the enrichment of clonal or subclonal SVs in a locus (see Fig. 4d). The genes on the y-axis represent the most likely driver gene for each locus.

Supplementary Figure 6

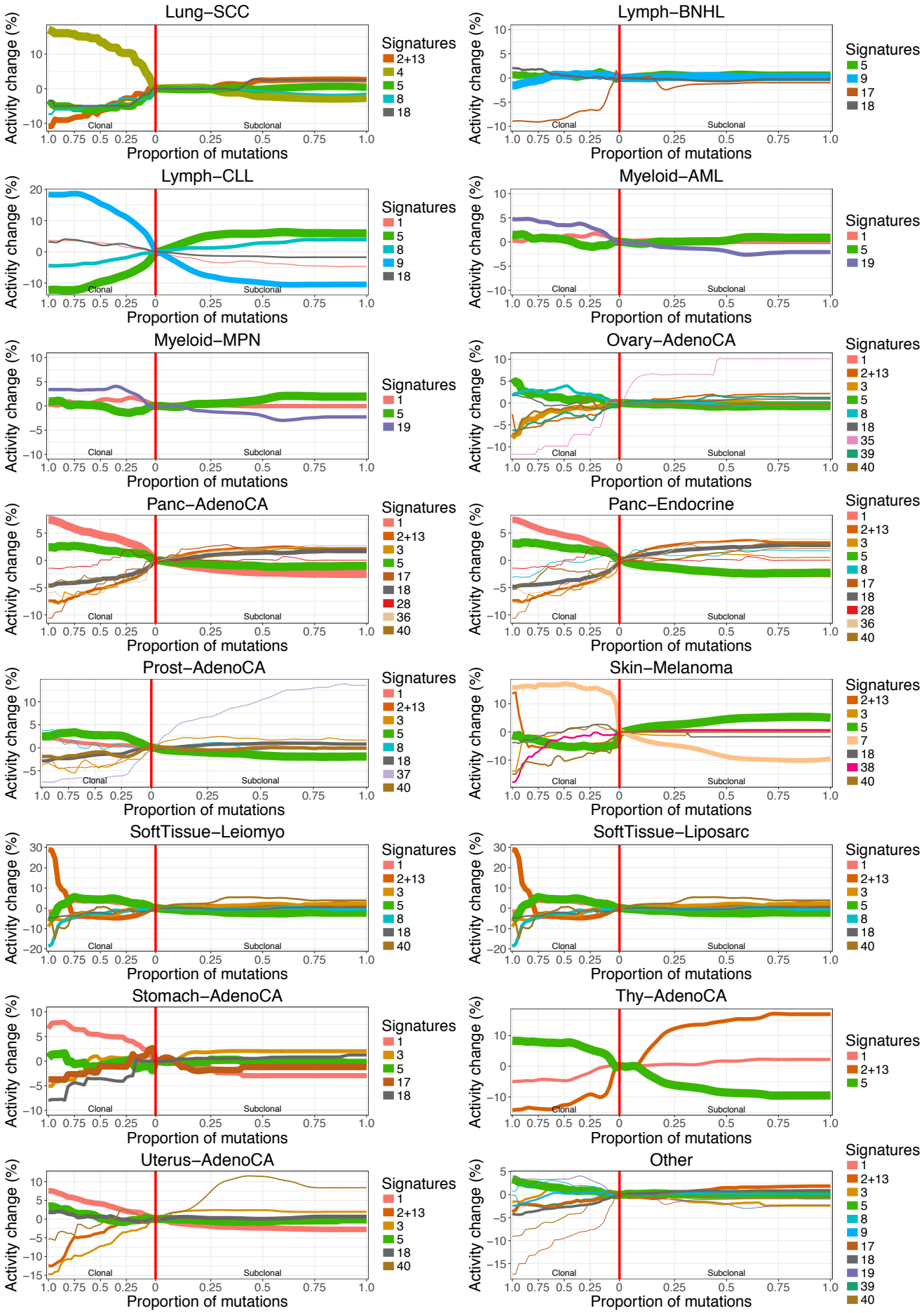


Evaluation of "winner's curse" correction (WCC) on simulated data. Corrected cellular prevalence (consensus from two correction methods) shows good concordance with the true cellular prevalence from simulated samples.

Supplementary Figure 7



Supplementary Figure 7



Supplementary Figure 7

Summary signature trajectories per cancer type. The average trajectories for mutational signatures were calculated across tumours of the same cancer type. The colour of the line denotes the signature and its width reflects the number of contributing tumours. The trajectories have been centred around the activity at the boundary between clonal and subclonal mutations in order to highlight relative changes in signature activity.