Chapter 7

A pan-cancer overview of tumour heterogeneity

7.1 Introduction

In the previous chapters I have introduced methods to analyse copy number and the subclonal architecture of cancers. Individual methods were introduced as well as approaches to construct a consensus copy number profile and a consensus subclonal architecture. These approaches were applied to the 2,778 cancer genomes contained within the ICGC PCAWG project (a description of how the consensus subclonal architecture was obtained can be found in the next section). This chapter describes the pan-cancer landscape of intra-tumour heterogeneity and evolution, as it emerges from the consensus results. The chapter, for the most part, covers the results that will be reported in Dentro et al. (manuscript in preparation) and also contains a high-level overview of results described in Gerstung et al. (2017), which is attached to this thesis as Appendix A. The results reported in this chapter are the culmination of my Ph.D. and are the result of a process in which I have been deeply involved over the last 3.5 years. These results also represent the outcome of a long standing collaboration between members of the PCAWG Evolution and Heterogeneity working group, without whom this project could not have succeeded. The figures in this chapter will appear in Dentro et al. Figs. 7.1 and 7.2 have been created by Kerstin Haase and Figs. 7.7 and 7.8 are by Maxime Tarabichi. All figures are used with permission. Fig. 7.4 is inspired by the driver figure made by Ignaty Leshchiner for the Dentro et al. manuscript.

7.2 Methods

We set out to obtain a robust consensus subclonal architecture for every tumour based on the consensus approaches introduced previously. We first applied the consensus copy number procedure to combine profiles from the six different copy number callers (ABSOLUTE, Carter et al. (2012); ACEseq, Kleinheinz et al. (2017); Battenberg, Nik-Zainal et al. (2012a); cloneHD, Fischer et al. (2014), Sclust Cun et al. (2018) and JaBbA (manuscript in preparation)) into a robust, high confidence consensus. Every copy number profile consists of a series of segments with each an assigned confidence level. Not every segment in every genome is of high confidence and an incorrect copy number call for a single segment could cause a subclonal architecture method to call a spurious mutation cluster, as the CCF values of mutations on that copy number segment would be incorrectly calculated from the VAF. I therefore created a subset of high confidence segments by ordering the segments of each tumour by their confidence level and select segments from the top until at least 75% of the genome was covered. The 11 subclonal architecture callers were restricted to use copy number and SNVs in the selected regions only.

The 11 subclonal architecture callers (BayClone-C, Sengupta et al. (2015); cloneHD, Fischer et al. (2014); CTPSingle, Donmez et al. (2016); DPClust, Bolli et al. (2014), Phylogic, Landau et al. (2013); PhyloWGS, Deshwar et al. (2015); PyClone, Roth et al. (2014); SVclone, Cmero et al. (2017), Ccube (manuscript in preparation), CliP (manuscript in preparation) and Sclust, Cun et al. (2018)) produced three key features to describe every tumour in the data set: The number of mutation clusters identified, properties of those clusters (the estimated number of mutations and the proportion of tumour cells that each cluster represents) and mutation assignments (either probabilistic or hard assignments). The three consensus subclonal architecture procedures were applied to produce a consensus set of mutation clusters described by a location (proportion of tumour cells estimate) and size (number of SNVs that the cluster contains).

I then applied the MutationTimer pipeline (Gerstung et al., 2017) to assign all available consensus SNVs (including the SNVs that were previously excluded when selecting highly confident copy number segments) and all indels and SVs for which allele frequencies were available. MutationTimer assumes each mutation cluster can be modelled by a beta-binomial and calculates probabilities for each mutation belonging to each cluster whilst also taking into account the size of the mutation clusters. It produced the final consensus subclonal architecture with the aforementioned key features, while also performing timing of mutations relative to gains to classify mutations in *clonal early*, *clonal not specified*, *clonal late* and *subclonal*. MutationTimer does this by evaluating the multiplicity state of a mutation and the copy number of the segment on which the mutation resides. If the mutation is on a gained

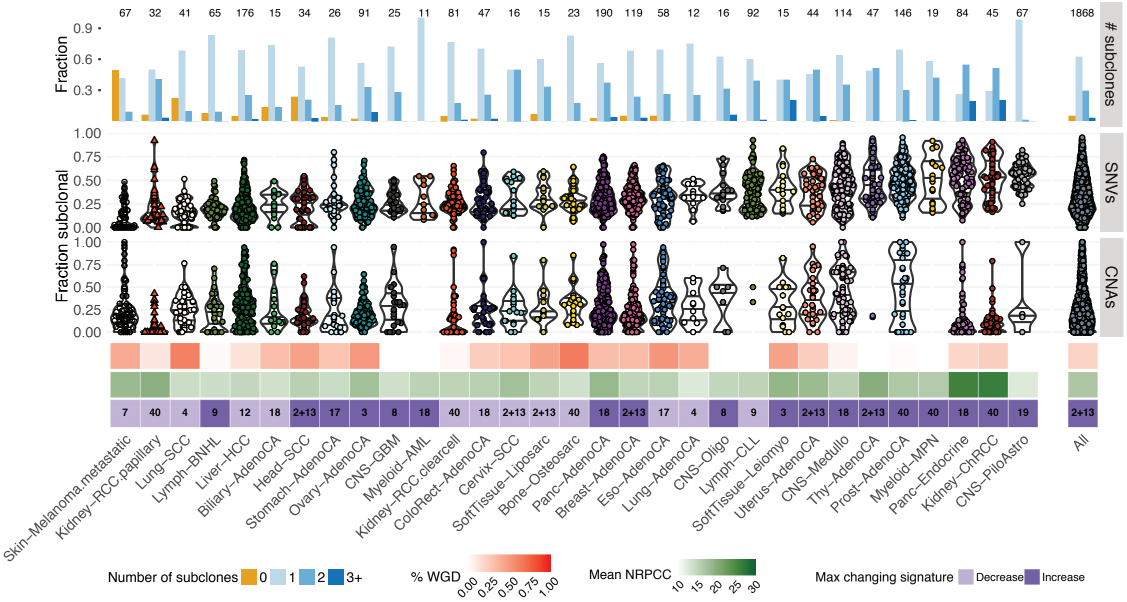


Fig. 7.1 A pan-cancer overview of intra-tumour heterogeneity. The top bar shows the total number of tumours per cancer type and the proportion of tumours where we identify zero (orange), one, two or three+ subclones (shades of blue). Below are the proportion of subclonal SNVs and CNAs, where the CNAs represent tumours with at least 5 whole chromosome arm CNA events. Whole genome duplication rate is show in boxes coloured red (high duplication rate) to white (low) and average reads per chromosome copy (introduced in section 6.4) is shown in shades of green. The most changing mutational signature is shown in purple, where the number refers to the COSMIC signature.

chromosome and has a multiplicity greater than one it is *clonal early*, if the mutation is on a gained chromosome and has a multiplicity equal to one it is *clonal late*, if the mutation falls in a region that has normal diploid copy number or is a loss it will classify clonal mutations as *clonal not specified* and if the cluster with the highest assignment probabality is subclonal, then the mutation is assigned *subclonal*.

The results reported in this chapter are from the WeMe consensus method, but due to the high similarity between the consensus subclonal reconstruction methods we could have chosen CSR or CICC and have shown the same basic results (Yu et al. 2017, manuscript in preparation).

7.3 Nearly all primary tumours contain detectable subclones

Figure 7.1 shows the pan-cancer overview of intra-tumour heterogeneity (ITH) that our analysis reveals. The figure contains all tumours with a number of reads per chromosome copy of 10 or more (which allows us to find a subclone at 30% of tumour cells or higher) to exclude tumours where not enough subclonal signal is obtained due to a combination of purity, ploidy and sequencing coverage. We selected all primary tumours and reduced multi-sample cases to their preferred tumour (a label that is provided by the PCAWG consortium). As there are only 2 primary melanoma tumours available in PCAWG, we instead included melanoma metastasis. The remaining metastasis and relapse tumours are discussed in the next section. The figure shows fractions of subclonal SNVs and CNAs. The fraction of subclonal CNAs indicates the number of arm-level subclonal CNAs over the total number of arm-level CNAs per tumour. A tumour is only included in the CNA plot if it's profile contains at least 5 arm-level events in total. Cancer types are sorted by the median proportion of subclonal SNVs.

The overview across 36 histologically distinct cancer types reveals that 96.7% of the 1,801 primary tumours contain at least one subclone. Patterns of ITH differ markedly between types of cancer: Prostate, uterus and esophageal adenocarcinomas show high proportions of both subclonal SNVs and CNAs. Kidney chromophobe and pancreatic endocrine tumours also show high proportions of subclonal SNVs, but differ from the previous group by containing few subclonal CNAs. On the other hand, hepatocellular carcinomas and squamous cell carcinomas of head-and-neck and lung contain low proportions of subclonal SNVs, but high proportions of subclonal CNAs. Finally, in osteosarcomas we find a high proportion of subclonal CNAs and varying degrees of subclonal SNVs. These findings suggest that tumour types exhibit their own, distinct, evolutionary narratives.

7.4 Metastatic melanomas are often clonal

In stark contrast to the high proportion of primary tumours with at least one subclone, we observe that over half of metastatic melanomas are clonal (Fig. 7.1). A comparison to metastasis of other cancer types available in this data set suggests that this might be a unique property (Fig. 7.2, left), although, the other cancer types are represented by a low number of cases and the observation would need to be verified in a larger cohort. There are only two breast and ten prostate metastasis cases available (the prostate tumours are from the multi-sample study Gundem et al. (2015)).

In contrast, melanoma relapse tumours are as heterogeneous as relapse cases from other types of cancer (Fig. 7.2, right). These findings may highlight properties of metastatic melanomas, for example that these metastasis belong to a group of rapidly developing melanoma tumours (Liu et al., 2006).

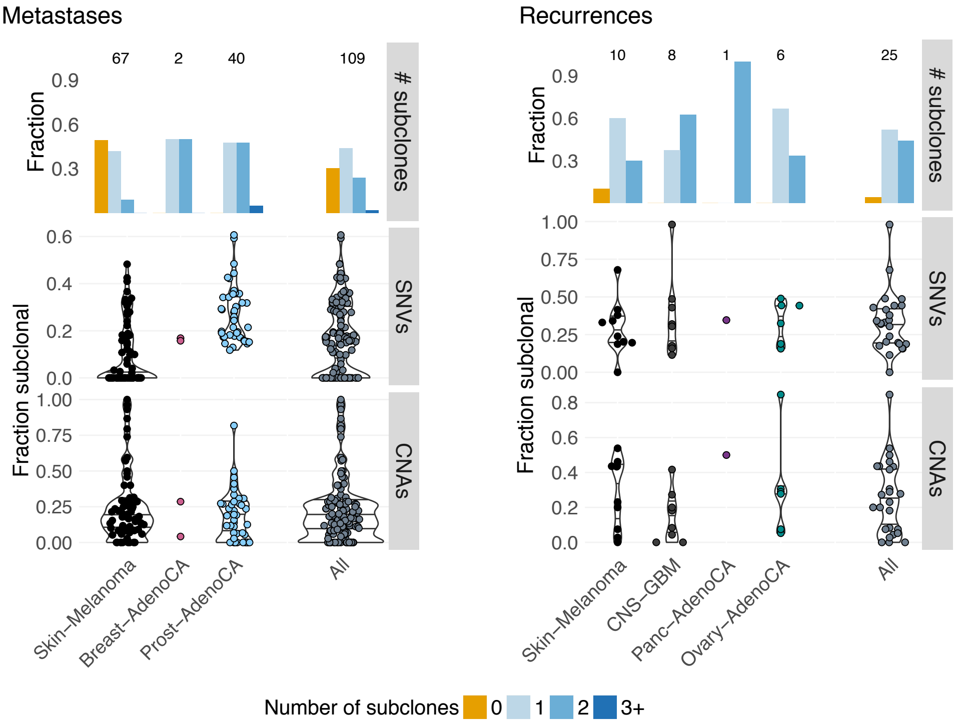


Fig. 7.2 The proportions of subclones, fraction of subclonal SNVs and CNAs found in metastasis and relapse tumours.

7.5 Subclonal driver mutations in known cancer genes

I used the catalogue of driver mutations identified in the PCAWG cohort by Sabarinathan et al. (2017) to obtain an overview of clonal and subclonal drivers in the PCAWG cohort. A number of filters have been applied to obtain the pan-cancer picture of subclonal drivers (Fig. 7.3). I excluded CNA drivers, as they are not assigned in our consensus subclonal architectures, and also excluded SV assignments, as their CCF values tend to be of quite variable quality. Furthermore, tumours with low reads per chromosome copy have been removed, multi-sample cases have been reduced to their PCAWG preferred sample and relapse and metastasis cases have been filtered out (apart from melanomas). After filtering there are 4,152 identified drivers remaining, spanning 362 different genes. 1,423 of 1,865 (76%) tumours contain an identified driver, but only 24% of tumours and 20% of subclones contain at least one subclonal driver.

Figure 7.4 provides a pan-cancer overview of the top 30 genes with subclonal drivers, identified by summing their probability of being subclonal. Each square is sized depending on the proportion of tumours of that cancer type containing a driver in that gene and the

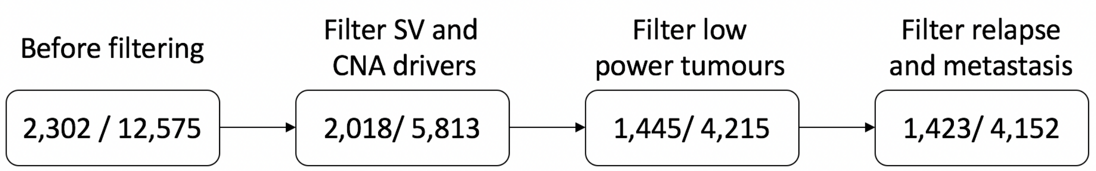


Fig. 7.3 Flow diagram showing the filters applied to obtain the driver overview. Each square shows two values, separated by a divider: First the number of unique tumours that are left in the data set after a filter has been applied and second the number of driver contained within those samples.

square is coloured depending on the proportion of tumours in which the driver is subclonal (again identified by summing probabilities of being clonal and subclonal), where darkblue means high proportion subclonal and lightblue means a high proportion clonal. Figure 7.5 shows the same data that is provided in each square in Fig. 7.4, but instead of showing the proportion of subclonal drivers, they are provided as counts to show the total number of tumours that support each square. Bars have been greyed out when they represent fewer than six tumours.

TP53, *TERT* and *VHL* show large, light squares for a number of cancer types, which means that these drivers are often activated early and are thus frequently observed as clonal. The figure does not contain any large, dark squares, which means that no gene is primarily identified as a subclonal driver in large numbers of tumours. The presence of small dark squares shows that driver mutations in known cancer genes can appear late during tumour evolution, but only in small proportions of tumours. This suggests that early drivers may be constrained to a select number of genes, while the spectrum of drivers becomes more diverse as tumours evolve further.

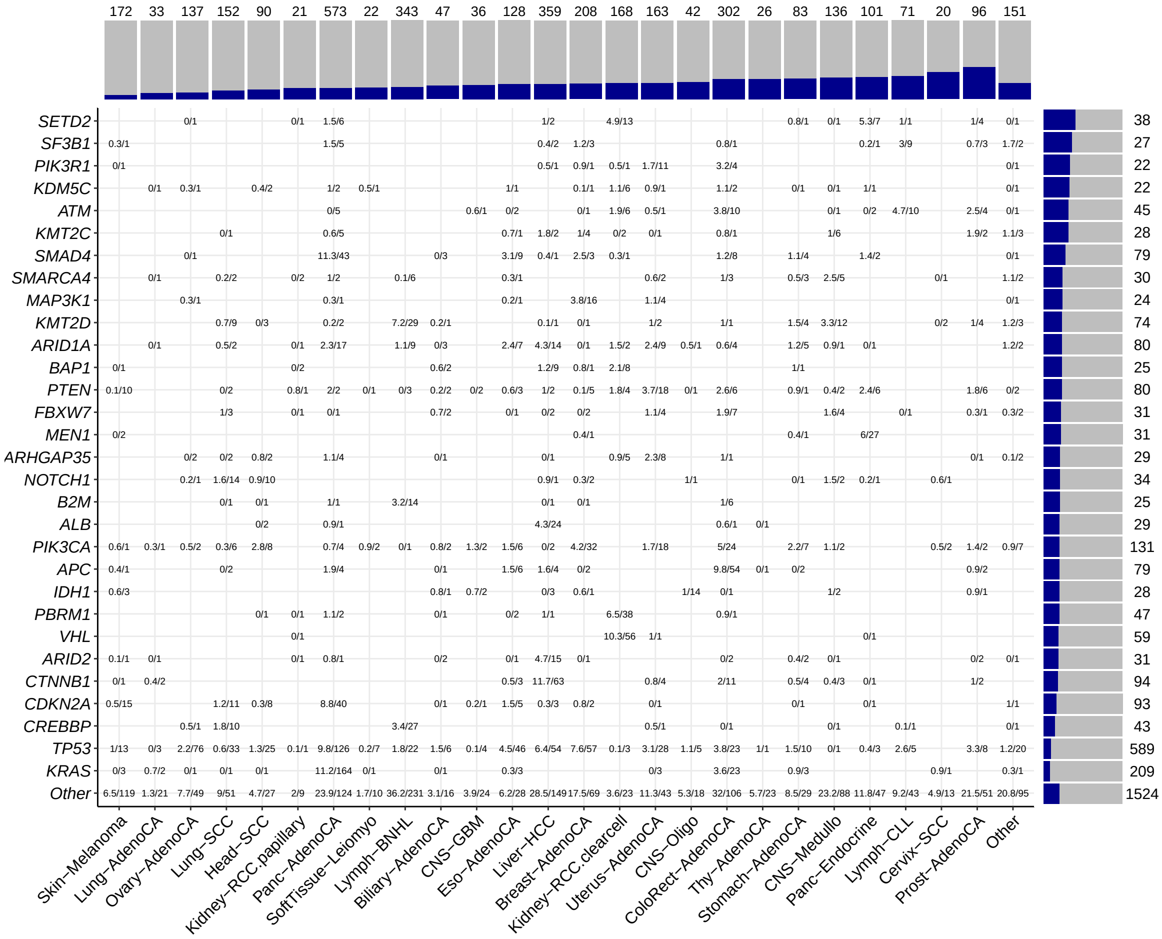


Fig. 7.4 A pan-cancer overview of subclonal drivers. Each cell contains two values, separated by a divider: The sum of the subclonal probabilities and the total number of drivers identified in the gene and the cancer type. Bars at the edges show the proportions of tumours with clonal (grey) and subclonal (dark blue) drivers for cancer types (top) and genes (side). The top 30 genes are shown, obtained by summing the subclonal probability of all drivers per gene.

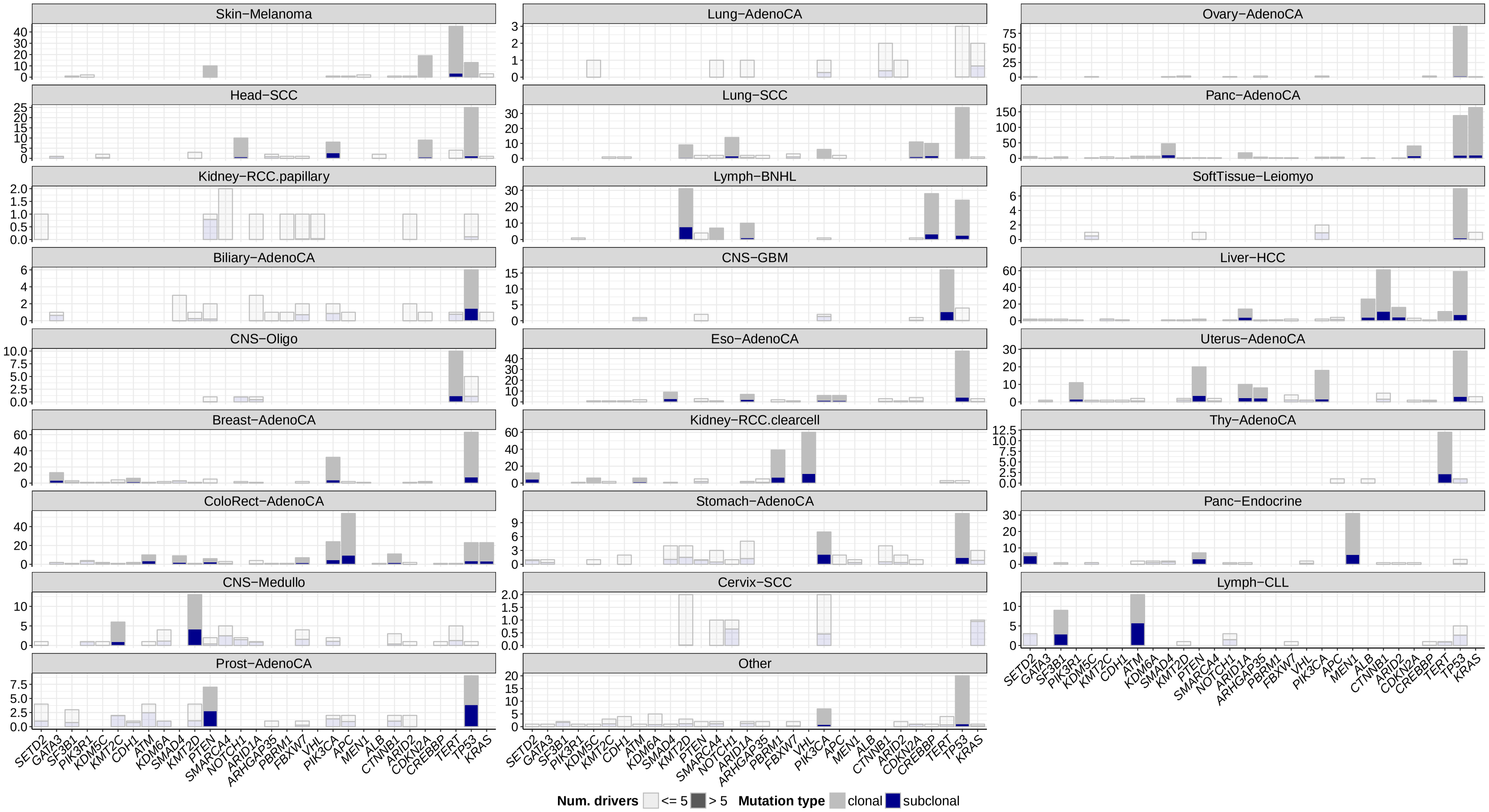


Fig. 7.5 Counts of samples per cancer type, per top 30 drivers shown in Fig. 7.4. Bars are greyed out when fewer than 5 tumours are found for a cancer type and gene combination.

Figure 7.5 shows that several genes are subclonal in high proportions of relatively few tumours: *SETD2* in pancreatic endocrine tumours, *ATM* in CLL cases and *PTEN* and *TP53* in prostate cancers. Other genes are at the top of the list of subclonal driver genes pan-cancer wide, but are supported by low numbers across cancer types: *GATA3*, *KMT5C* and *CDH1*.

These findings suggest that a large proportion of late drivers have yet to be identified. They show that known driver genes can be active late during tumour evolution, which may suggest that a particular environment is required for a driver to be effective.

If is likely however that the picture painted in this section is just the tip of the iceberg. Mutations identified as clonal in the sequencing sample may in fact not be carried by all tumour cells. In the scenario of small biopsies one may be painting a picture of the small biopsy only and therefore overestimate the number of mutations that are clonal, and what is determined to be subclonal to consist of relatively minor subclones. When a large portion of the tumour is obtained for sequencing one can be more confident about whether it represents the whole tumour, however subclones will represent major cellular populations at this scale. How the sequencing samples were obtained exactly has not been recorded as part of the PCAWG effort and it is therefore unclear how this affects the painted picture.

It is important to note however that what is subclonal at any point within the tumour, is in fact subclonal in the whole tumour. The painted picture therefore represents a conservative estimate of the amount of subclonality, regardless of the sampling strategy. Finally, our findings have been partially reported elsewhere, with *SETD2* often appearing as subclonal in a detailed study of kidney cancers (Turajlic et al., 2018) and driver mutations in chromatin remodellers shown to correlate with later transitions between tumour progression states in melanomas (Shain et al., 2018).

7.6 14% of mutations are undetected

We applied our methods to correct for the *winner's curse* (which were introduced in section 6.5) and estimate the number of mutations that are unaccounted for given a tumour's subclonal architecture. Figure 7.6 shows the correction applied for cluster position (left) and cluster size (right). Clonal clusters (shown in grey) often are corrected very little, highlighting that typically (nearly) all clonal mutations are detectable above 30X sequencing coverage. Subclonal clusters however can be corrected extensively with many mutations falling below the detection limit, in line with the observations on simulated data in the previous chapter. The average correction across all tumours (i.e. the difference between total detected mutations and total estimated mutations) is 14%, suggesting that 14% of mutations have been missed because they fell below the detection limit.

7.7 Clear signs of positive selection in subclonal mutations

We observe clear signs of positive selection within clonal and subclonal mutations and for missense, nonsense and splice-site SNVs (Fig. 7.7). Inspection of driver mutations reveals that the detected subclones contain driver mutations in known cancer genes (Fig. 7.4).

We next looked for signs of positive selection in both the clonal and subclonal mutations by analysing the ratio of non-synonymous and synonymous mutations, an approach often referred to as dN/dS. A ratio larger than 1 is considered a sign of positive selection, a

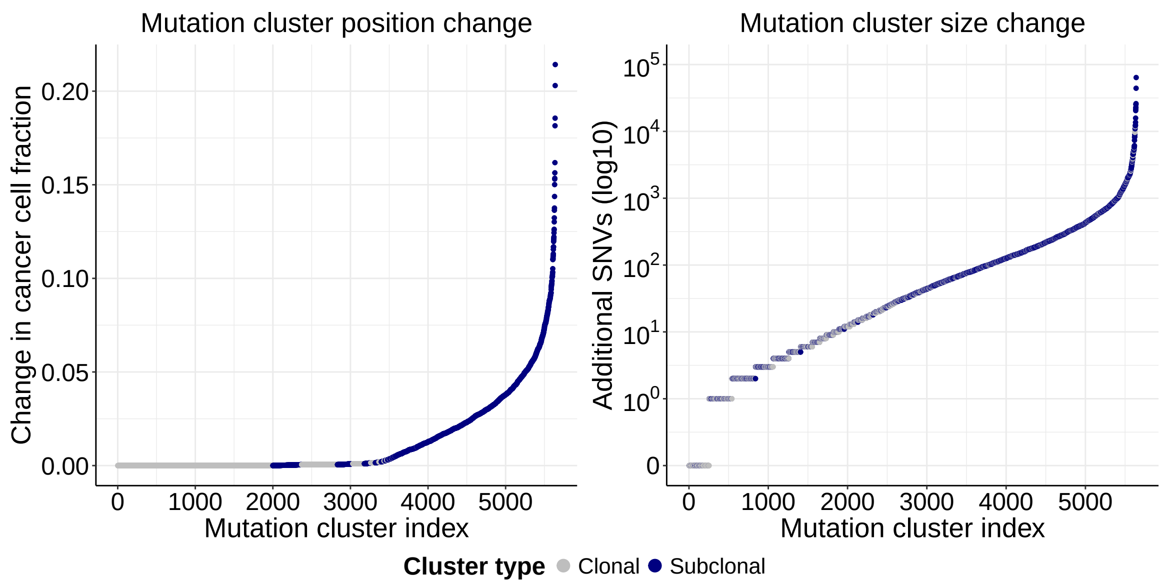


Fig. 7.6 The correction applied to cluster location (left) and cluster size (right). The average correction across all tumours is 14%. Clonal clusters are often adjusted very little, while subclonal clusters can be adjusted considerably, in line with observations on simulations in the previous chapter.

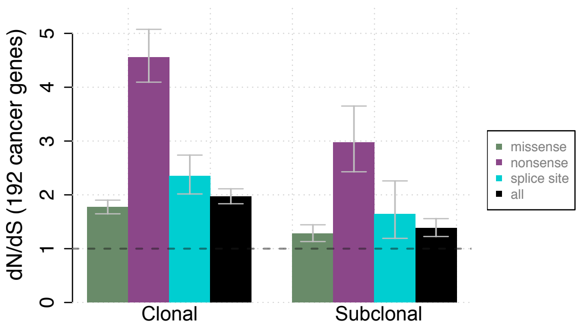


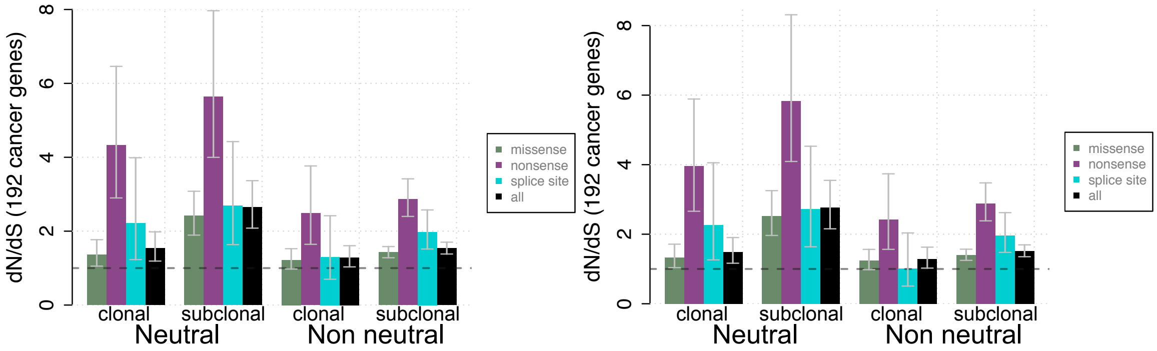
Fig. 7.7 dN/dS values for clonal and subclonal SNVs across all primary tumours as described by Martincorena et al. Values for missense, nonsense and all mutations are shown, along with the 95% percentage intervals. Positive selection is observed in all mutation classes.

ratio below 1 represents negative selection, while a ratio of 1 can mean either no selection (neutral) or an equal amount of positive and negative selection. dN/dS ratios have been used extensively in the field of evolutionary biology and have recently been adapted to study selection in cancer genomes (Greenman et al., 2006; Martincorena et al., 2016). We used the approach published in Martincorena et al. (2017) that models tri-nucleotide contexts and considers additional non-synonymous mutations beyond missense mutations, such as nonsense and splice-site mutations and indels, and has been shown to accurately recapitulate existing knowledge about cancer drivers (Martincorena et al., 2017). The analysis was performed on the 192 cancer genes in the COSMIC Cancer Gene Census v80 (Futreal et al., 2004) to provide a conservative estimate of positive selection.

Recently there has been discussion in the field of tumour evolution about whether positive selection may no longer be present and that further evolution of these tumours occurs due to genetic drift (Sottoriva et al., 2015; Williams et al., 2016). Williams et al. (2016) recently proposed a test that can be used on bulk whole genome sequencing data to identify tumours for which this is the case. The test is based on the principle that the further one zooms into a neutrally evolving tumour, the number of subclones and mutations increases at an exponential rate. If these neutral subclones are captured in a sequencing sample, one therefore expects the number of mutations to increase at an exponential rate as the VAF distribution goes to zero. Williams et al. (2016) propose to test the mutation VAF space of a sequencing sample against an exponential curve (which I'll denote as a "1/f" tail) and a high correlation between the VAF tail and the "1/f" tail would indicate the tumour is evolving neutrally. Williams et al. (2016) recommend a correlation of over 0.98 indicates neutral evolution.

The test and results are however not without controversy (Noorbakhsh and Chuang, 2017; Tarabichi et al., 2017). We therefore applied this principle to the PCAWG data set to identify neutrally evolving tumours and did so separately for all mutations and for mutations in unaltered copy number regions (one copy of the maternal and paternal alleles). A tumour was called neutral when the cumulative VAF space yielded a correlation of over 0.98 with a "1/f" tail. This test identified 557 tumours as neutrally evolving (531 on all mutations and 499 tumours when considering only mutations in normal copy number). The tumours that are identified as neutrally evolving have significantly higher reads per chromosome copies (p-value $8.74*10^{-90}$, Mann-Whitney-U test), which may suggest the number identified is an underestimate, as many tumours in the PCAWG dataset it is not possible to identify sufficient subclonal mutations.

We next applied the dN/dS pipeline to the mutations in these tumours, which identified positive selection in both clonal and subclonal mutations in neutral and non-neutral tumours (Fig. 7.8). Tumours identified as neutrally evolving also contain subclonal driver mutations



(a) Mutations in CNA regions

(b) Mutations in normal copy number

Fig. 7.8 Tumours were classified into neutral and non-neutral according to the rationale described by Williams et al. (2016) dN/dS values for clonal and subclonal SNVs were derived separately across all primary tumours in those two groups, as described by Martincorena et al. (2017). Values for missense, nonsense and all mutations are shown, along with the 95% percentage intervals. The same figure is shown after grouping neutral and non-neutral tumours based on SNVs in the diploid genome only.

in known cancer genes. 345 of the 557 identified tumours contain at least one identified driver. The tumours contain a total of 893 driver mutations, of which 114 have a probability > 0.95 of being subclonal. We find subclonal driver mutations in *TP53* and *PTEN* (6 each), *SETD2* (4), *ATM*, *FBXW7*, *KIT*, *NF1*, *SF3B1* and *TGFBR2* (3), 16 genes with 2 subclonal drivers and 48 with 1.

These findings show that tumours identified by the "1/f"-tail test contain subclones under positive selection and that the identified clonal expansions contain driver mutations in known cancer genes.

7.8 Subclonal clinically actionable events

I considered driver mutations (SNVs and indels) in genes and pathways for which drugs are either developed or in development to look specifically for tumours with subclonal targetable driver mutations (as predicted by Cancer Genome Interpreter (Tamborero et al., 2018)). A patient with a targetable driver mutation could in the near future be prescribed a targeted therapy, but the therapy is inherently flawed if the targetable mutation is not shared by all tumour cells. In this analysis we excluded all metastasis and relapse tumours, except melanomas. For multi-sample cases we only considered the PCAWG provided preferred sample for each donor.

Our consensus subclonal architecture approach produces probabilistic cluster assignments for each mutation and identifies a mutation cluster as clonal (the clonal cluster has CCF

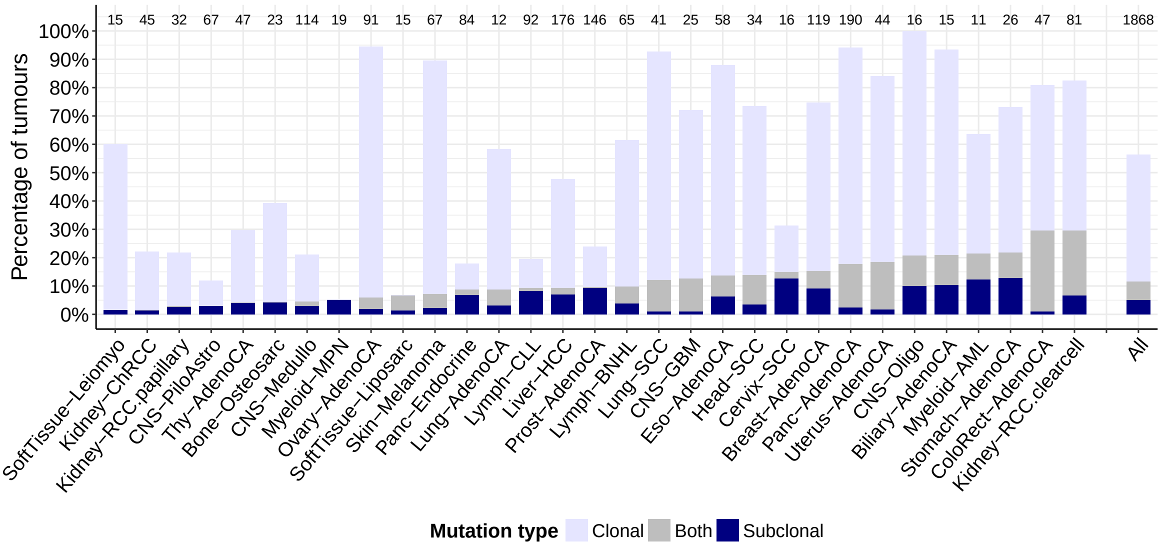


Fig. 7.9 Clinically actionable driver mutations were surveyed across the cohort (Sabarinathan et al., 2017) and assigned a probability of being clonal or subclonal. Per cancer type probabilities are combined to provide a fraction of tumours that contain only clonal actionable drivers, only subclonal actionable drivers or both (see supplementary methods). On average 11.7% of tumours contain at least one subclonal actionable driver, while in 5.1% of tumours we found that all actionable drivers are subclonal. Cancer types show markedly different proportions, ranging from 4.3% of thyroid cancers with at least one subclonal actionable driver to 29.7% of kidney clear cell carcinoma cases.

of 1, while a subclonal cluster has a CCF < 1). Through the consensus I can establish the probability whether a mutation is clonal or subclonal. The procedure is as follows: For each sample, I establish the probability that all actionable mutations are clonal, all actionable mutations are subclonal and the probability of observing at least one pair of clonal and subclonal targetable events.

The probability (p) of observing all n actionable mutations as clonal is:

$$\prod_{i=1}^{n} p_{i,clonal} \tag{7.1}$$

The probability of (p) of observing all n actionable mutations as subclonal is:

$$\prod_{i=1}^{n} p_{i,subclonal} \tag{7.2}$$

Then the probability of observing at least one pair of actionable mutations where one is clonal and one is subclonal is:

$$1-\left(\prod_{i=1}^{n} p_{i,clonal}+\prod_{i=1}^{n} p_{i,subclonal}\right) \tag{7.3}$$

The three probabilities were summed to create the three classes per type of cancer: *Clonal*, *Subclonal* and *Both*.

Through this analysis I find that 11.7% of tumours have an identified subclonal driver mutation that is clinically actionable (Fig. 7.9). In 5.1% of tumours, I find targetable driver mutations only in subclones and 6.6% of tumours contain both a clonal and a subclonal target. These estimates are likely a lower bound as tumours are only represented by a single sample, which likely shows (depending on how the samples were obtained) either local heterogeneity or large subclones, and mutations that appear to be clonal in one area of the tumour may in fact be subclonal overall when they are not present in another region.

These findings suggest it is important to consider the clonal status of a targetable mutation before treatment is started. Prescribing a drug that targets a mutation not carried by all tumour cells is certain to be ineffective. Meanwhile, in 6.6% both a clonal and a subclonal targetable mutation is found. In this scenario, clonality assessment would highlight the clonal mutation as the best candidate.

However, for clonality analysis to be truly informative in clinical application, one must be highly confident that a mutation that appears clonal is indeed carried by all tumour cells. Ultimately, it may prove impossible to truly establish a mutation is carried by every tumour cell as it would require assessing every tumour cell. Clonality assessment strategies may therefore be limited to identifying subclonal targetable mutations (a mutation that is subclonal in one region of the tumour is subclonal overall) that would provide ineffective treatment options.

7.9 Evidence of additional heterogeneity

Several studies have shown (Jamal-Hanjani et al., 2017; Sun et al., 2017) that multi-region sequencing is better powered to detect subclones, compared to single-region sequencing approaches. We reasoned that some of the subclones that cannot be reliably disentangled on single-region sequencing may leave a trace that can be detected in a single sample. Mutation clusters may be merged during a single-region based subclonal reconstruction when multiple subclones appear at a similar CCF. We therefore explored two aspects that could be informative about the number of additional subclones within a sequencing sample: Subclonal

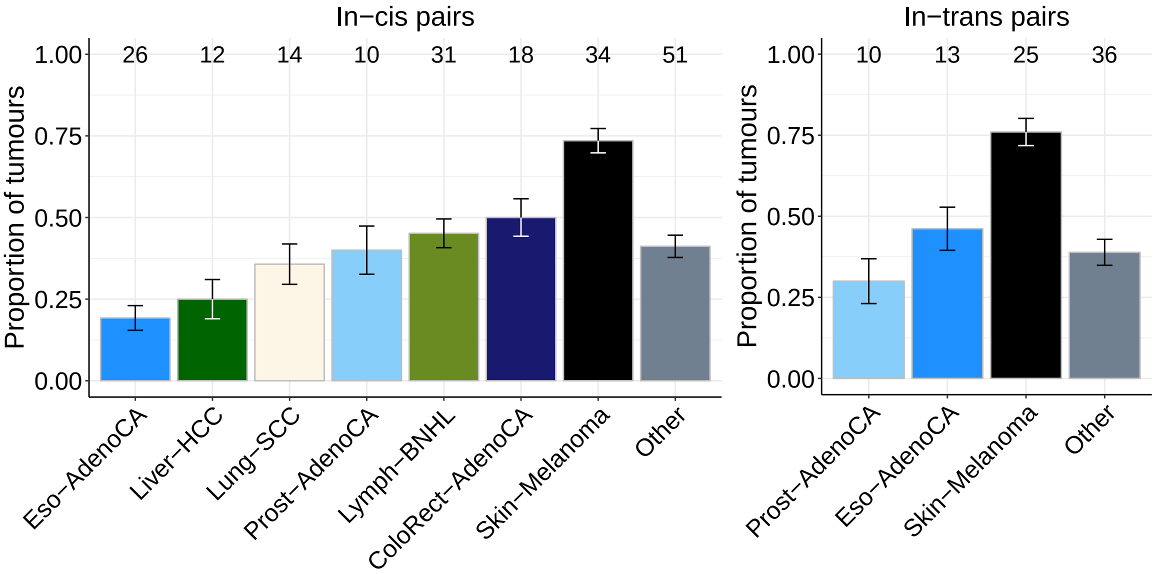


Fig. 7.10 Fraction of tumours where phased mutations provide evidence of additional heterogeneity for tumours where the mutations are *in-cis* (left, co-linear) or *in-trans* (right, branching) Error bars represent the binomial standard deviation of the total number of tumours for each type of cancer and the associated ratios.

mutational signature changes that are not close to a boundary between mutation clusters and evidence of mutations assigned to the same subclone that cannot have occurred in the same cell.

Within our working group, Yulia Rubanova has developed an approach (termed Tracksig) to estimate changes in mutational signature activity in approximate-time ordered mutations. Tracksig bins mutations and orders the bins by pseudo-time (mutations on two chromosome copies have occurred before mutations on the same segment carried by one chromosome copy, subclonal mutations occur after clonal mutations, etc) and subsequently detects in which pseudo-time bin mutational signature exposures change (Rubanova et al. 2017, manuscript in preparation). Yulia has applied her algorithm to the PCAWG data, which reveals that 37.4% of tumours had a signature exposure change of at least 10%, while 30.1% of signature changes correspond to a boundary between the mutations from a clone / subclone and 39.7% represent boundaries between subclones. We further find that an average of about 0.5 changes per sample are not within a subclone boundary, which suggests that additional subclones are measured by the sequencing data but have not been detected.

Within our working group Amit Deshwar has looked into pairs of mutations that cannot have occurred in the same cell, building on data that I generated. I generated counts for mutation pairs that fall within 700bp that could be spanned by a single read pair. For these mutations, it is possible to determine whether both mutations fall on the same chromosome copy (i.e. are phased) by examining the read pairs that cover both. A mutually exclusive pair of mutations (mutations are *in-trans*) that cannot have occurred in a single cell is measured as a pair of mutations (a,b) without a read-pair that report both variant alleles, while some reads report the variant allele of a and the reference allele of b and vice-versa, and the pair fall in a genomic region where only a single chromosome copy is available. In contrast, mutations where some read-pairs report both variant alleles and some pairs report only one (mutations are *in-cis*) the mutations represent clusters in an ancestral relationship. When considering phased mutation pairs Amit finds that in 44% (86 of 196) of tumours there is evidence of mutually exclusive mutations (Fig. 7.10).

These findings highlight that the found amounts of ITH, as is reported in Fig. 7.1, are an lower bound for the amount of ITH available in the sequencing samples.

7.10 Cancer types follow individual evolutionary narratives

We next characterised the evolutionary histories of the 2,658 tumour cases in the PCAWG dataset, described in full in Gerstung et al. (2017) and attached to this thesis as Appendix A. This section describes results that are the culmination of work by Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner and Santiago Gonzalez. In brief: three different mutation timing analysis were performed. (1) A classification of mutations into clonal early (gained mutations, on more than 1 chromosome copy), clonal late (mutations on a gained chromosome, but on 1 copy only), clonal unspecified (mutations in non-gained copy number regions) or subclonal.

(2) Timing of copy number gains was performed by accounting for multiplicity states (for example, a high ratio of multiplicity two mutations on a gained chromosome suggests the gain was late (Fig. A1.3)) and (3) CNAs were timed relatively against each other by league model analysis (these models pitch every pair of detected CNAs against each other, and like a sports league, build a league table out of all the matchups). We also overlayed mutational signature activity, and through the use of clock-like signatures we convert timing analysis into real time estimates.

We find that driver mutations predominantly occur early and are therefore observed as clonal (Fig. 7.11a and b). Drivers in *TP53* and *KRAS*, for example, are 5-9x more likely to occur early than late clonal. For *TP53* this effect is independent of tumour type (Fig. 7.11c). In general, the diversity in driver genes increases as tumour evolution progresses: 50% of all early driver mutations are found in 12 genes, while late clonal and subclonal drivers occur across 39 and 36 genes respectively (Fig. 7.11d). These findings suggest that early drivers occur within a specific set of genes, while late drivers are more diverse, giving rise to the "long tail" of driver genes active in low proportions of tumours.

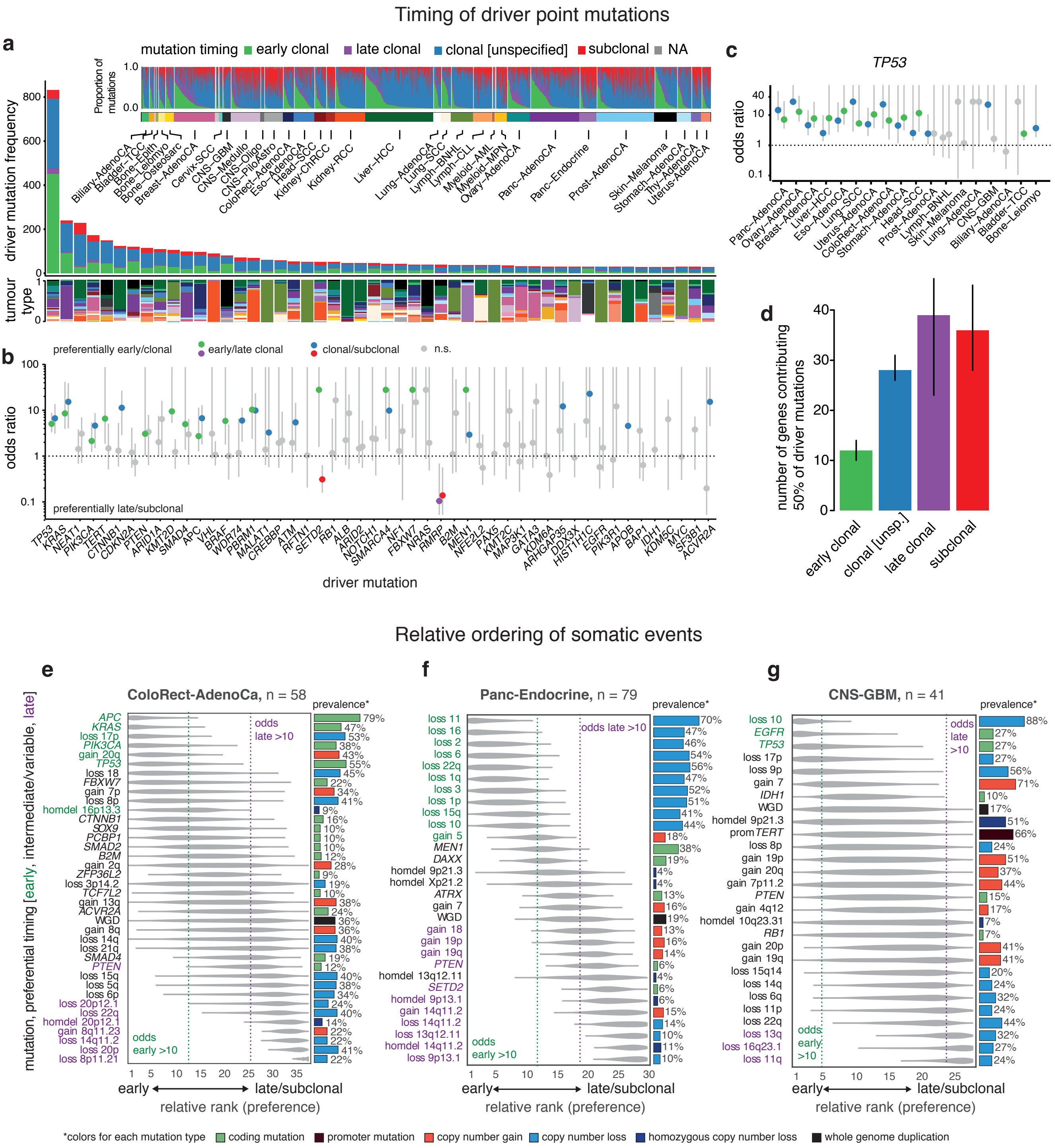
Our relative timing of events (Fig. 7.11e) reveals that in colorectal adenocarcinoma *APC* has the highest odds of occurring early, followed by *KRAS*, loss of 17p, *TP53*, loss of 8p, gain of 8q, in concordance with the progression model proposed by Fearon and Vogelstein (1990). For pancreatic endocrine cancers the relative timing suggests that in these tumours losses are frequently early, followed by driver mutations in *MEN1* and *DAXX* and a whole genome duplication (Fig. 7.11f). In glioblastoma cases we find that loss of chromosome 10 and driver mutations in *TP53* and *EGFR* are typically early, preceding early gains of chromosomes 7, 19 and 20 (Fig. 7.11g).

The timing of gains reveals that, pan-cancer wide, copy number gains typically occur during the second half of tumour evolution. But cancer types show marked differences (Fig. A2): Glioblastoma tumours show consistent early gains of chromosomes 7, 19 and 20, while medulloblastomas contain early gains of 17q. Gains are typically early in glioblastoma, medulloblastoma and pancreatic neuroendocrine cancers, late in squamous cell lung cancers and melanomas, while they occur during broad periods in other cancers.

Analysis of mutational process activity early, clonal and late reveals that signature activity typically changes by less than 30%, which indicates that signature activity is relatively constant during tumour evolution (Fig. 7.12a). Life style associated signatures, such as 4 (smoking associated) in lung adenocarcinoma, 7 (UV-light) in melanoma and 12 (aetiology unknown) in liver cancers, typically decrease in activity late, while signatures 2 and 13 typically increase in activity (Fig. A4b). The clock-like mutational signature 1 was used to infer real time estimates: whole genome duplications typically appear 2-11 years before diagnosis (Fig. 7.12b), while the most recent common ancestor appears six months to six years before we observe the tumour (Fig. 7.12c).

These analysis combined confirm previous knowledge about the distribution of driver genes and the classic progression model of colorectal adenocarcinoma, and reveal that cancer types follow distinct evolutionary patterns.

Figure 3: Timing of driver mutations and relative ordering of somatic events



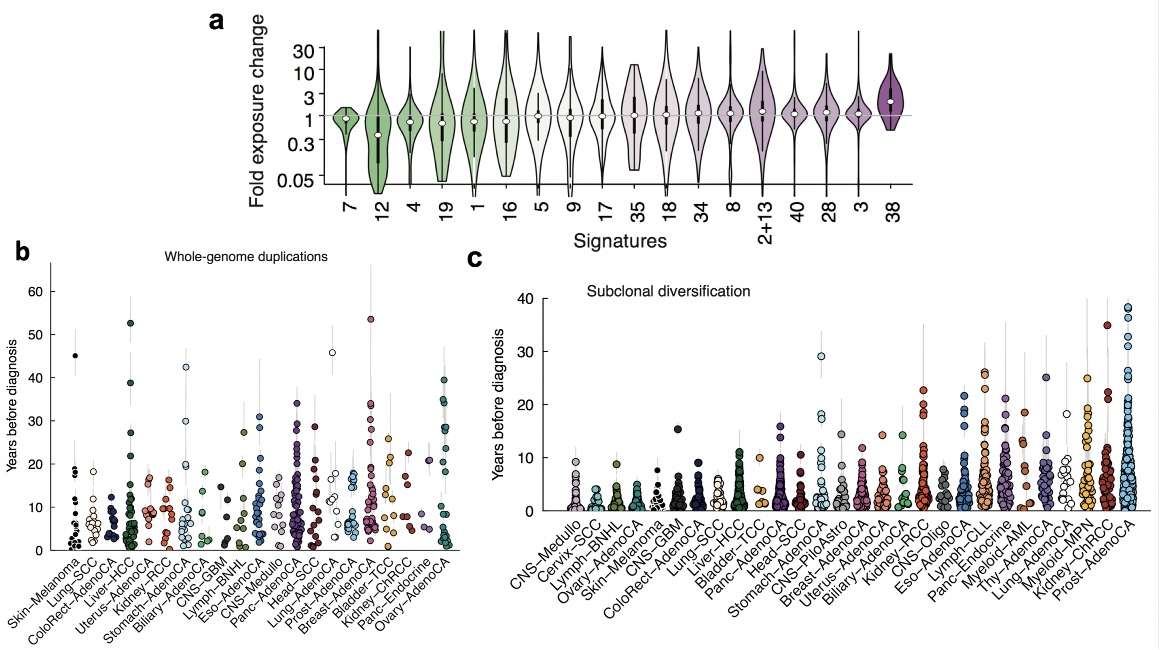


Fig. 7.12 This figure is a combination of various figures from Gerstung et al. (2017). (a) Fold changes in signature exposures between early and late clonal stages for all tumours. Each violin shows the distribution of exposure changes across tumour types in one signature. Signatures are sorted by the ratio of tumours with a positive signature change. (b) Time of occurrence of whole genome duplications in individual patients, split by tumour type, based on CpG>TpG mutations and patient age. Results are shown for a 5x acceleration of the mutation rate. (c) Timing of subclonal diversification using CpG>TpG mutations in individual patients.