Intra-tumor heterogeneity and clonal exclusivity in renal cell carcinoma

Ariane L. Moore[1, 2], Jack Kuipers[1, 2], Jochen Singer[1, 2], Elodie Burcklen[1], Peter Schraml[3], Christian Beisel[1], Holger Moch[3] and Niko Beerenwinkel[1, 2]*

[1] Department of Biosystems Science and Engineering, ETH Zurich, Mattenstrasse 26, 4058 Basel, Switzerland
[2] SIB Swiss Institute of Bioinformatics, Mattenstrasse 26, 4058 Basel, Switzerland
[3] Department of Pathology and Molecular Pathology, University and University Hospital Zurich, Schmelzbergstrasse 12, 8091 Zurich, Switzerland
* Correspondence: niko.beerenwinkel@bsse.ethz.ch

Abstract

Tumorigenesis is an evolutionary process in which different clones evolve over time. Interactions between clones can affect tumor evolution and hence disease progression and treatment outcome. We analyzed 178 tumor samples in 89 clear cell renal cell carcinoma patients and found high intra-tumor heterogeneity with 62% of mutations detected in only one of two biopsies per patient. We developed a novel statistical test to identify gene pairs that are altered in co-occurring clones of the same tumor, including the pairs *TP53* and *MUC16*, as well as *BAP1* and *TP53*. The mutations in these gene pairs are clonally exclusive meaning that they occurred in different branches of the tumor phylogeny, suggesting a synergistic effect between the two clones carrying these mutations. Our analysis sheds new light on tumor development and implies that clonal interactions are common within tumors, which may eventually open up novel treatment strategies to improve cancer treatment.

Keywords

Intra-tumor heterogeneity, co-occurrence, mutual exclusivity, cooperation, clones, renal cell carcinoma, clonal exclusivity, mutualism, synergism, convergent evolution

Introduction

Tumors generally consist of genetically and phenotypically distinct cancer cell populations that evolve over time through a process that involves mutation and selection (Nowell, 1976). The different cancer cell populations, also called clones, accumulate mutations, and those with a selective advantage will expand and possibly outcompete the other cancer cell populations (Nowell, 1976). This process of mutation and selection can create a diverse set of clones within the tumor, a phenomenon called intra-tumor heterogeneity (Beerenwinkel et al., 2016, Dexter et al., 1978, Heppner, 1984). Intra-tumor heterogeneity poses a challenge to the treatment of cancer, because the number of different somatic mutations may be undervalued when taking just a single biopsy from a solid tumor or if there are different clones in the metastases (Gerlinger et al., 2014, Beerenwinkel et al., 2016). Identifying subclonal mutations is important because it has been shown that even low-frequency clones can carry markers of poor prognosis (Gerlinger et al., 2014), and they drive the process of metastasis. It has also been observed that the number of clones in a neoplasm can predict the probability of progression from precursor lesions to cancer (Maley et al., 2006). Moreover, a higher genetic diversity increases the chances for the tumor as a whole to adapt quickly to changes in selection pressures and to escape therapy and immune responses (Yates et al., 2015, Heppner, 1984). Hence, it is essential to quantify intra-tumor heterogeneity and to detect subclonal variants in order to improve the success of cancer therapies.

Intra-tumor heterogeneity is more than a bet-hedging strategy because tumors are evolving ecosystems in which the individual clones do not evolve independently, but may interact with each other in order to confer communal advantages (Tabassum and Polyak, 2015, Marusyk and Polyak, 2010, Heppner, 1984, Bonavia et al., 2011). Tabassum and Polyak summarize the different types of ecological interactions among cancer clones into negative and positive interactions (Tabassum and Polyak, 2015). For instance, a negative interaction between clones arises when they compete for nutrients or oxygen (Tabassum and Polyak, 2015). Conversely, positive interactions such as commensalism, synergism, and mutualism, drive tumor cell proliferation and ultimately favor greater intra-tumor heterogeneity (Tabassum and Polyak, 2015). Commensalism represents the case where one clone benefits from another one, without providing anything in return (Tabassum and Polyak, 2015, Marusyk and Polyak, 2010). This is the case, for instance, if one clone stimulates blood vessel growth, which supplies also surrounding cells of other clones with nutrients and oxygen (Axelrod et al., 2006). Kalas et al. demonstrated that cancer cells with a *Ras* mutation secrete factors that lead to down-regulation of a strong angiogenesis inhibitor in nearby normal cells (Kalas et al., 2005). Using a GFP reporter system, they showed that this paracrine signaling can overcome distances of more than 10 mm (Kalas et al., 2005), suggesting that interactions may occur not only between directly neighboring clones in solid tumors. Synergism and mutualism arise when two clones have mutually positive effects on each other and thereby increase the overall fitness of the whole tumor (Tabassum and Polyak, 2015, Marusyk and Polyak, 2010). Here, we refer to synergism and mutualism as cooperation. These processes are not to be confused with epistasis, where, for instance, two genes are mutated in the same cell or clone, and together lead to an unanticipated change in the phenotype (Wang et al., 2014).

Axelrod and colleagues hypothesized that two or more clones in a tumor could stably co-exist and cooperate based on sharable resources (Axelrod et al., 2006). Moreover, they suggest that this cooperative behavior can make the tumor more aggressive in a shorter time period (Axelrod et al., 2006). For example, two clones may each produce a different growth factor that the other clone is missing, and then cross-feed each other (Axelrod et al., 2006). In fact, such a division of labor was observed in breast cancer, where adjacent cancer cells expressed different growth factors and the respective growth factor receptor of their neighbor, which suggests that they share the resources and mutually benefit from each other (de Jong et al., 1998). Cooperation between two cancer clones was also observed in several mouse models showing that specific clones were together more proliferative and metastatic (Calbo et al., 2011, Marusyk et al., 2014). Cleary and colleagues found two different clones in a mouse model of breast cancer that stably co-exist and demonstrated strong selection for this co-existence indicating the advantage of the combined phenotype (Cleary et al., 2014). Many more studies reported clonal cooperations (Mateo et al., 2014, Zhang et al., 2015, Brumby and Richardson, 2003, Uhlirova et al., 2005, Chapman et al., 2014, Konen et al., 2017), as well as their involvement in resistance to treatment (Archetti, 2013, Miller et al., 1991). Thus, clonal cooperation is an important evolutionary force and it is crucial to elucidate its underlying mechanisms. Uncovering the principles of how cancer cell populations interact, will shed light on how to disrupt this process and potentially open up novel treatment strategies to improve personal cancer treatment. In order to investigate how clones co-exist and possibly interact, a systematic screening of subclonal mutation compositions is necessary. If a combination of two or more clones is beneficial, these clones will likely co-exist stably over time and not outcompete each other (Axelrod et al., 2006, Cleary et al., 2014). The analysis of subclonal mutations in several independent tumors allows for systematically assessing whether specific subclonal mutation combinations exist that occur at a higher rate than expected.

Here, we investigate the intra-tumor heterogeneity and subclonal mutation composition of clear cell renal cell carcinoma (ccRCC). Clear cell RCC is the most common type of kidney cancer and is known to be difficult to treat, as it is, for instance, mostly resistant to chemotherapy when metastasized (Choueiri and Motzer, 2017, Cancer Genome Atlas Research, 2013, Penticuff and Kyprianou, 2015). Several studies revealed that ccRCC is a genetically very heterogeneous disease (Gerlinger et al., 2012, Gerstung et al., 2012). The analysis of 25 single cells from one ccRCC patient revealed a large extent of genetic heterogeneity between the different cancer cells of the same tumor (Xu et al., 2012). A study of multiple tumor samples from eight ccRCC cases demonstrated that the diversity within tumors is in some cases as high as the diversity between patients (Martinez et al., 2013). Gerlinger and colleagues analyzed the intra-tumor diversity of 10 ccRCC patients, and also found that most driver mutations were subclonal (Gerlinger et al., 2014). This finding prompts a systematic analysis of the combinations of subclonal mutations in co-occurring clones in ccRCC. For this purpose, we investigated a cohort of 89 ccRCC patients (Fig. 1A). In the initial discovery phase, we analyzed two spatially separated biopsies and a matched normal sample from each of 16 ccRCC patients to provide an overview of the diversity and to inform the selection of genes for the second in-depth follow-up analysis. In the second phase, we used the constructed gene panel to sequence 826 genes at high coverage in 178 paired tumor samples and 89 matched normal samples from 89 ccRCC patients. The two paired biopsies per tumor and the high coverage allow a detailed investigation of the subclonal mutation composition. We developed a customized statistical framework, called GeneAccord, to examine the combinations of clones that co-exist in these tumors. We detected several pairs of mutated genes that are mutually exclusive within the clones of tumors and referred to as clonally exclusive (Fig. 1B). Such subclonal mutations are co-occurring in the same tumor, but occur in different clones.

Several methods have been proposed that detect co-occurrence of mutated genes or pathways in cancer patients (Jiang et al., 2016, Zhang et al., 2014, Dao et al., 2017). For instance, Dao et al. (Dao et al., 2017) proposed a framework that detects groups of genes that are united by a certain property such as co-occurrence, mutual exclusivity or functional relatedness, while exhibiting a different characteristic between the groups of genes. However, the aforementioned approaches do not refer to the individual clones within a tumor. They only distinguish between mutations being present or absent in a certain patient. Therefore, they do not allow for a distinction between epistatic effects from having two pathways mutated in the same cell, or having two different pathways mutated in two different clones that may complement each other's abilities and possibly cooperate. Jiang and colleagues performed clonal analysis only for one sample and found that the mutations in the two co-occurring pathways are actually in the same clone (Jiang et al., 2016). In this work, we specifically seek to find mutations that are assigned to different clones in order to uncover novel principles of the mechanisms of clonal co-existence and potential cooperation. We present an approach that systematically assesses subclonal mutation patterns in a large cohort of patients.

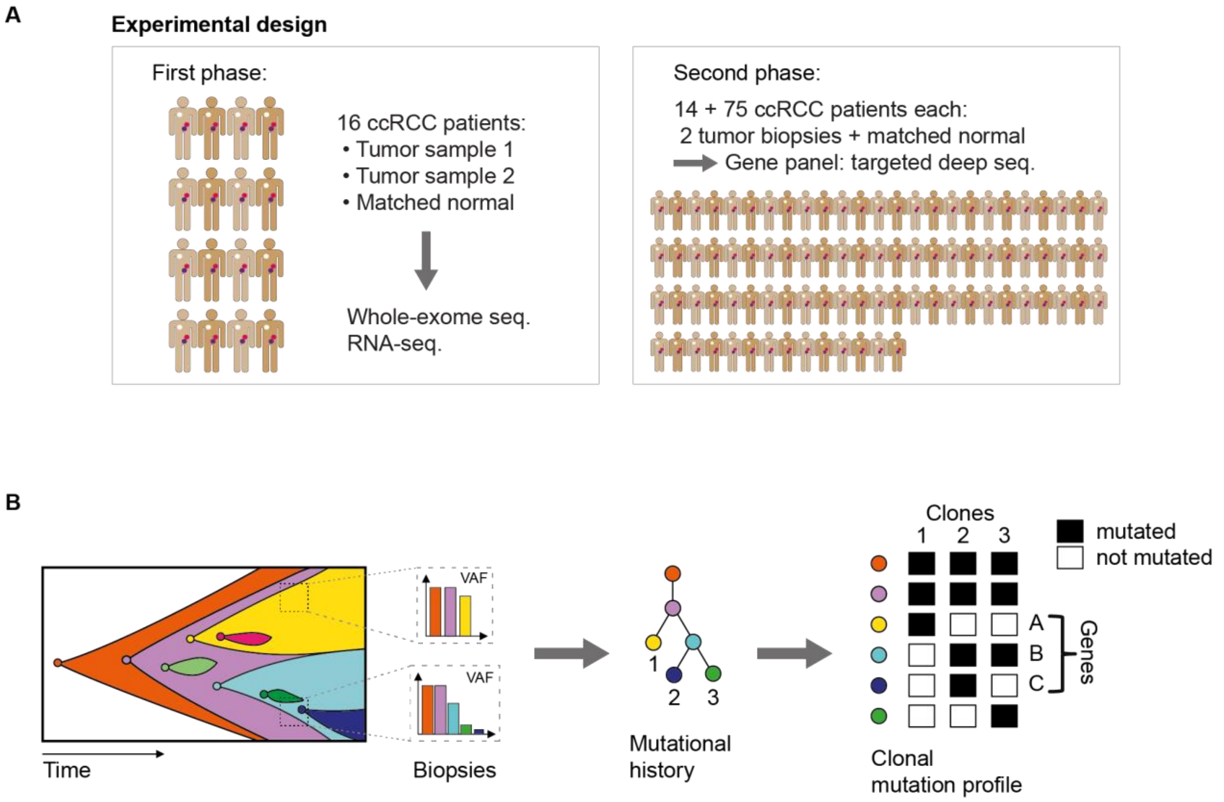


Figure 1. (A) Experimental design**.** The first phase includes 16 clear cell renal cell carcinoma (ccRCC) patients of which two spatially separated biopsies from the primary tumor and a matched normal sample were collected. Whole-exome sequencing and transcriptome sequencing was performed and the detected mutations informed the selection of genes for the panel of the second phase. The second phase includes an extended cohort of patients and the selected genes were targeted with higher coverage. From a total of 89 patients, we analyzed two spatially separated tumor biopsies and a matched normal sample per patient. Fourteen of the patients in this panel data set were also among the 16 from the first phase. **(B)** Intra-tumor heterogeneity and co-occurring tumor clones. Schematic representation of tumor development, co-existing clones, and their shared and private mutations. Left: The tumor initiates from a single cell that acquires a driver alteration (orange). A subsequent driver event (purple) leads to the spread of a new clone that eventually outcompetes the first one. Finally, the yellow and light blue driver alterations lead to clones that stably co-exist and possibly benefit from each other. The biopsies obtained consist of a heterogeneous collection of clones. The mutations show different variant allele frequencies (VAF). Right: The inference of clones from bulk sequencing data and their phylogenetic relationship uncovers the mutational history of the tumor. The mutation ordering can be represented in a mutation tree as depicted here. In particular, each mutation is assigned to one or more clones, depending on where in the tree it occurred. The yellow and light blue mutations have a mutually exclusive clone assignment, and hence are referred to as being clonally exclusive.

Results

We first provide an overview of the detected intra-tumor heterogeneity in the ccRCC cohort, and compare the findings to previous studies. Next, GeneAccord, a novel statistical test to detect co-occurring clones in tumors, is introduced and applied to the panel data set obtained from 89 ccRCC patients and also to the whole-exome data set in order to detect clonally exclusive pairs of pathways in these tumors.

Genetic and transcriptomic diversity of ccRCC

The analysis of intra-tumor heterogeneity and subclonal mutation patterns was performed on a cohort of a total of 89 ccRCC patients and it comprised of two phases (Fig. 1A). In an initial discovery phase, whole-exome and transcriptome sequencing data from paired tumor biopsies from 16 ccRCC patients plus one matched normal sample per patient were analyzed. The coverage of the whole-exome sequencing (WES) data was on average 85x. The mutations were called from the WES data set (Methods section), and between 29 to 130 single-nucleotide variants (SNVs) and insertions and deletions (indels) were detected per patient (Fig. 2A). The fraction of mutations that was only detected in one of the two biopsies from the same tumor was on average 40%, which indicates high levels of intra-tumor genetic diversity. These mutations are referred to as private, whereas mutations detected in both tumor samples of a patient are called shared.

From the RNA-sequencing (RNA-seq) data, the differentially expressed genes were called (Methods section). We found an average of 6,364 genes per patient to be upregulated and 6,598 genes downregulated (Fig. 2B). On the transcriptomic level, the intra-tumor heterogeneity was slightly reduced with an average of 31% of differentially expressed genes being detected only in one of the two biopsies. Pathway overrepresentation analysis was performed with the set of differentially expressed genes using the Reactome pathway database (Fabregat et al., 2018). Among the most overrepresented pathways are many pathways related to translation, signal transduction and growth factors (Fig. 2C). For instance, the signaling pathways involving the growth factors PDGF, VEGF, SCF, or the growth factor receptor EGFR are deregulated in all patients. In ccRCC, the vascular endothelial growth factor (VEGF) is often activated which is important for angiogenesis, cell growth, and survival (Choueiri and Motzer, 2017). In fact, *VEGFA*, is upregulated in all patients of this data set. The most overrepresented pathways related to translation are highly overrepresented in patients 4, 15, and 16. They are all enriched only privately in one tumor sample of patient 3 and patient 15, indicating that these deregulated processes are subclonal in these tumors.

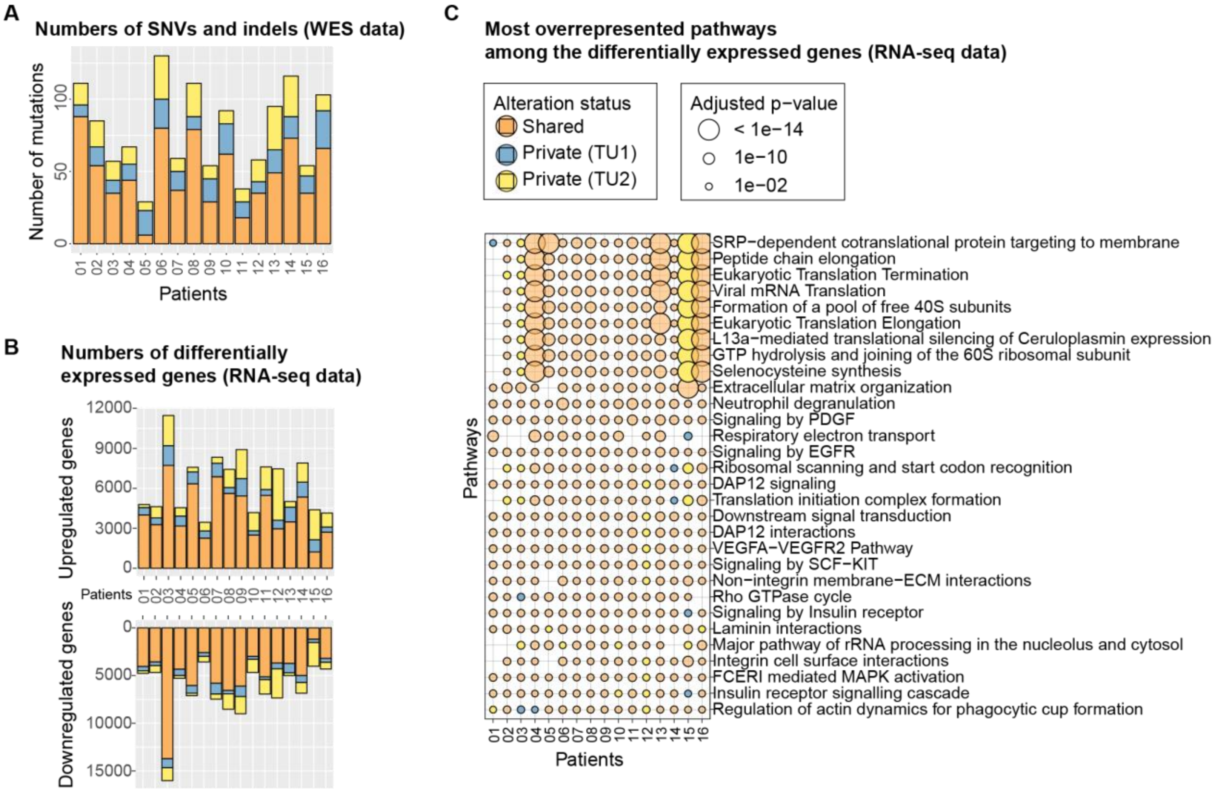


Figure 2. Genetic and transcriptomic diversity in 16 patients. **(A)** Number of shared (orange) and private (yellow, blue) mutations in the WES data set. A shared alteration was detected in both samples of a patient, whereas a private alteration was only found in one of the two samples. The two biopsies of the same tumor are labeled ‘TU1’ and ‘TU2’. **(B)** Number of differentially expressed genes in the RNA-seq data set. **(C)** The most overrepresented Reactome pathways among the differentially expressed genes. The color indicates the alteration status.

In the second phase, the cohort was extended to 89 ccRCC patients. From each, two spatially separated biopsies of the primary tumor and a matched normal sample were collected. This data set includes 14 of the patients from the WES data set, and 75 additional patients. A selection of 826 genes was compiled, which was informed by the mutated genes detected in the WES data set, as well as from the frequently mutated genes in a large ccRCC study from The Cancer Genome Atlas Research Network (TCGA) (Cancer Genome Atlas Research, 2013). Using this customized gene panel, we sequenced the 178 tumor and 89 normal samples at high depth. The deep coverage enables detection of low-frequency mutations, and the larger cohort provides increased statistical power such that rare subclonal mutation patterns can be detected. A large cohort is especially important for the analysis of co-occurring mutated genes, because most genes are mutated in only a small fraction of patients (Garraway and Lander, 2013), and therefore the chance of the same pair of mutated genes occurring in multiple patients is generally small. The cohort includes two spatially separated tumor biopsies from each patient in order to facilitate more robust inference of clones and their phylogenetic relationships. The sequenced reads contain unique molecular identifiers (UMI), which allows for the correction of potential sequencing errors (Methods section). The coverage of the panel sequencing (panel seq) data set was on average 933x, and after UMI consensus building and read filtering it was 93x.

The number of SNVs as well as indels in the panel seq data set was on average 22 per patient (Fig. 3, top panel). Pairwise comparison of the two biopsies from the same tumor revealed that on average 62% of the mutations in a patient were private to one of the two samples. This is in line with the observations made in another study, in which the intra-tumor heterogeneity in four renal cell carcinoma patients was analyzed, and where between 63 and 69% of mutations were not detected in all samples from the same tumor (Gerlinger et al., 2012). A previous study of TCGA including 499 ccRCC showed that the gene VHL is hit by a mutation in 51% of ccRCC cases (Cancer Genome Atlas Research, 2013). Furthermore, chromatin remodeling genes such as *PBRM1*, *SETD2* and *BAP1* are, with frequencies of 36%, 13%, and 10%, often mutated in ccRCC (Cancer Genome Atlas Research, 2013). These four most commonly mutated genes in ccRCC are also among the most frequently mutated genes in our data set (Fig. 3, bottom panel), and the frequencies of VHL, SETD2, and BAP1 are with 54%, 12%, and 13% comparable to the ones from the TCGA study. The gene *PBRM1* occurs in 19% of the patients in our cohort, which is less than the frequency reported from TCGA, which is 36% (Cancer Genome Atlas Research, 2013). Another ccRCC study by Sato et al. including 106 samples reported *PBRM1* to occur in 29% of cases (Sato et al., 2013), which is closer to the frequency in the present cohort. The mutations in *VHL* are known to occur early in tumor development (Pena-Llopis et al., 2013), which is in line with our observation that in 39 of 48 cases, the *VHL* mutations are shared between both tumor biopsies of a patient.

Among the most frequently mutated genes in this data set are also the mucins *MUC6*, *MUC16*, and *MUC3A*. In the TCGA study, the mucins *MUC4*, and *MUC16* were among the seven most frequently mutated genes (Cancer Genome Atlas Research, 2013). In a study of 67 ccRCCs, *MUC16* was among the five most recurrently mutated genes (Arai et al., 2014), and Sato et al. found *MUC16* to be the fourth most frequently mutated gene (Sato et al., 2013). Also, *MUC4* was among the significantly mutated genes in their study (Sato et al., 2013). *MUC4* is not in this gene panel, but the mutation frequencies of the mucins *MUC6*, *MUC16*, and *MUC3A* are with 42%, 38%, and 18% higher than those reported in TCGA (Cancer Genome Atlas Research, 2013). The majority of mutations in these genes in our cohort has a low variant allele frequency (VAF). Specifically, more than three quarters of the mutations in *MUC6*, *MUC16*, and *MUC3A* have a VAF below 10%, and almost half below 5%. We conjecture that the low frequency may be the reason why many mutations in these genes are missed in studies with lower read coverage, or only one biopsy per tumor. The analysis of 25 single cells from one ccRCC patient also showed that most mutations were only found in a small fraction of the tumor cells, suggesting that rare mutations may be missed with low read coverage in a bulk sample (Xu et al., 2012). Mucins may actually play an important role in ccRCC, as, for instance, several studies in the past years highlighted that the expression level of certain mucins is predictive of clinical outcome in ccRCC patients (Fu et al., 2016, NguyenHoang et al., 2016, Bai et al., 2015, Zhang et al., 2017). In the present cohort, the survival data and the expression status of the three mucins *MUC6*, *MUC16*, and *MUC3A* is available for eleven patients, and a log-rank test was performed to assess potential correlations between survival and expression status of these three genes. *MUC6* downregulation showed a significant correlation with decreased survival (p = 0.01; Log-rank test).

Previous studies of intra-tumor heterogeneity in ccRCC reported patterns of convergent phenotypic evolution in several genes including *BAP1*, *SETD2*, *PBRM1*, *PIK3CA*, *PTEN*, and *KDM5C* (Gerlinger et al., 2012, Gerlinger et al., 2014), where the genes were affected by multiple distinct mutations across the clones in the tumor. In this cohort, such a pattern of convergent phenotypic evolution can also be found. In patient 88, the gene *PBRM1* is hit by different mutations in the two tumor samples. One tumor sample, TU1, has a frameshift deletion, and a missense mutation, while the other tumor sample, TU2, has a missense mutation at a different loci in *PBRM1*. Both subclonal missense mutations of *PBRM1* are predicted to be deleterious according to the SIFT annotation (Sim et al., 2012).

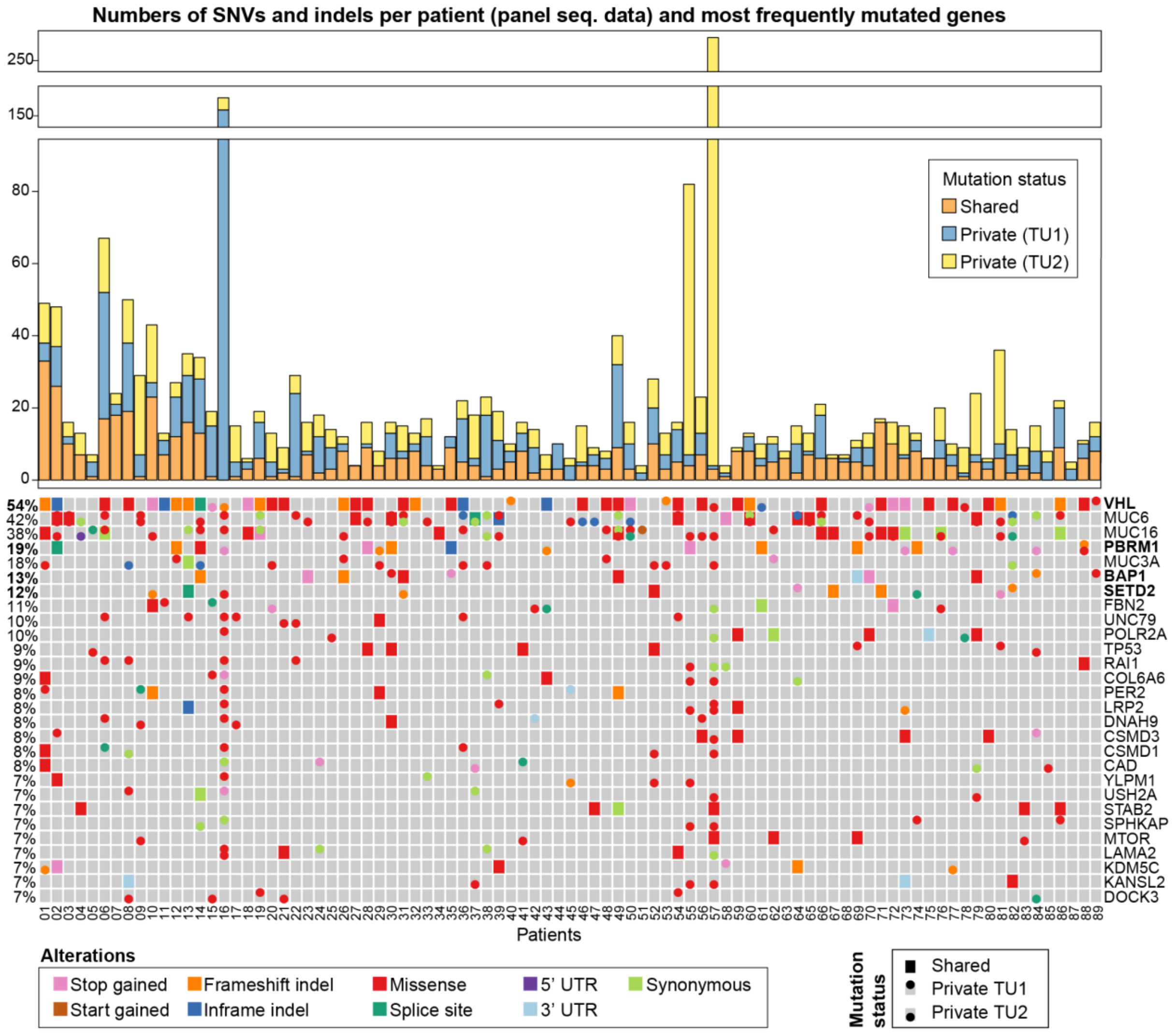


Figure 3. Genetic diversity in the panel seq data of 89 ccRCC patients. Top: Numbers of shared and private mutations in the data set. Bottom: The heatmap highlights the mutations that were detected in the most frequently mutated genes. The four most frequently altered genes in ccRCC are *VHL*, *PBRM1*, *BAP1*, and *SETD2* (Cancer Genome Atlas Research, 2013)

and are highlighted in bold. If a gene was hit by multiple mutations that all have the same status (Shared, Private TU1, or Private TU2), the following ordering is applied to prioritize which color is shown in the heatmap, starting with the highest priority: Stop gained, Start gained, Frameshift indel, Inframe indel, Missense, Splice site, Five prime UTR, Three prime UTR, Synonymous. That means, if a gene has, e.g., a missense and a synonymous mutation, the missense mutation will be displayed in the heatmap. The variants were annotated with SnpEff (Cingolani et al., 2012b) (Supplementary Table S1). Mutations in non-coding regions are omitted from the heatmap.

GeneAccord algorithm

In order to systematically analyze subclonal mutation combinations in co-occurring clones, we developed a statistical framework called GeneAccord to identify pairs of mutated genes or pathways that co-occur in the tumor but in different clones. That is, the pairs of mutated genes or pathways are mutated in the same patient, but their subclonal mutation profiles are mutually exclusive. The underlying rationale is that if two clones co-exist in a tumor and cooperate, for example, by sharing diffusible factors, they may have acquired a complimentary set of mutations to benefit each other, namely, one clone has mutations that the other one does not have, and vice versa. We refer to this situation as clonal exclusivity. More precisely, two mutated genes A and B, are called clonally exclusive, if one clone has a mutation in gene A, but no mutation in gene B, and another clone of the same tumor has a mutation in gene B, but no mutation in gene A (Fig. 1B).

Such a mutually exclusive pattern within the clone assignment of a tumor's set of mutated genes may occur at random, and therefore our statistical framework assesses whether a pattern of clonal exclusivity occurs more often than expected by chance alone across a cohort of patients. It makes use of a likelihood ratio test to compare the number of observed clonal exclusivities of a gene pair of interest across the cohort to the background distribution of the expected frequencies of clonal exclusivity (Supplementary Fig. S1, Methods section). The null hypothesis is that the gene pair of interest exhibits the pattern of clonal exclusivity as often as expected by chance when assuming independent occurrence of mutations. The alternative hypothesis is that the pattern occurs at a rate greater than random chance. The magnitude and type of this difference in the rates is quantified with a parameter Δ (Methods section). A positive parameter Δ indicates that the pair tends to be mutated in different clones more often than expected, whereas if it is negative, the gene pair tends to co-occur more often together in the same clone. All pairs with a positive parameter Δ are possible candidates for being significantly clonally exclusive. For those pairs the p-values of testing the alternative delta>0 versus the null delta=0 are computed. Multiple testing correction is done using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). If a pair is significantly clonally exclusive, it suggests that this specific clone configuration may confer a selective advantage, possibly through cooperation between the clones. GeneAccord is implemented as an R package available at https://github.com/cbg-ethz/GeneAccord and submitted to Bioconductor.

The input data of the algorithm are the mutated gene-to-clone assignments from a cohort of cancer patients, which were obtained by running phylogenetic tree inference methods. Reconstructing the evolutionary history of a tumor and detecting the clones is challenging (Beerenwinkel et al., 2015). For non-deterministic algorithms, repeated tree inference runs may lead to slightly different mutation-to-clone assignments. Therefore, our algorithm was designed to allow as input multiple gene-to-clone assignments per patient. They may have been generated by repeated tree reconstruction runs, or by sampling from the posterior distribution of trees. The tree inference methods designate the mutations to individual clones. The mutations can then be mapped to genes or pathways using existing pathway databases. Hence our statistical framework can be applied on the gene level, or on the pathway level to detect clonally exclusive pairs of pathways.

Clonal exclusivity in 89 ccRCC patients

The program Cloe (Marass et al., 2016) was applied to infer the phylogenetic history of each tumor and to obtain the mutated gene-to-clone assignments. The input to Cloe are mutations in copy number neutral regions. After filtering out mutations that are in potential copy number regions (Methods section), there were on average 17 mutations per patient. We performed 20 repeated runs of Cloe with different seeds in order to incorporate uncertainty in tree inference. The reconstruction of the phylogenetic history of each tumor maps each mutation to one or several clones. Mutations were mapped to genes and only those mutated genes were kept that were annotated to have a likely effect, that is, for instance, synonymous or intronic mutations were filtered out (Methods section). As a result, a total of 82 patients were left for which the mutated gene-to-clone assignments contained between two and 68 mutated genes per patient (Supplementary Fig. S2A). The average rates of clonal exclusivity, reflecting how often the clonal exclusivity pattern occurs in general, were computed for each patient (Methods section). The rates were between 0.0 and 0.675 (Supplementary Fig. S2B), and patients had between two to five clones. The distribution of the test statistic under the null hypothesis was computed (Supplementary Fig. S3, Methods section). A total of 362 mutated genes were among the gene-to-clone assignments. Across all patients, there were 6,270 pairs of genes, most of which were mutated in just one patient. A total of 325 pairs of genes were mutated together in two or more patients. The parameter Δ was computed for these 325 pairs, and 16 pairs had a positive parameter Δ, which were tested for significance of clonal exclusivity.

A total of five pairs were significant at the 5% level after correction for multiple testing (Supplementary Table S2), including the gene pairs {*MUC16*, *TP53*} (adjusted p-value = 0.01), and {*BAP1*, *TP53*} (adjusted p-value = 0.04). The most significant gene pair, *MUC16* and *TP53*, is clonally exclusive in both patients in which it is mutated, providing strong evidence for clonal exclusivity. The other top gene pairs are also mutated in two patients, but clonally exclusive only in one of them, while co-occurring in the same clone in the other patient. Therefore, they should be interpreted more carefully: The patients in which they are clonally co-occurring all have an average rate of clonal exclusivity of zero, indicating that the inferred phylogeny did not have a single branching. Hence, also the two mutated genes from a gene pair are in the same branch, and therefore clonally co-occurring in these patients. In such cases, it is also insightful to examine the variant allele frequencies (VAF) as explained below.

The most striking gene pair consists of *MUC16* and *TP53*, which is clonally exclusive in patients 5 and 81 (Fig. 4A). Both mutated genes were assigned to different branches of the phylogenetic tree. This is underlined by considering the VAF of the mutations in these two genes. The VAF of a mutation is computed as the fraction of reads that contain the mutation out of all reads covering the locus of the mutation. Assuming copy number neutrality, this quantity reflects the number of cells in the biopsy that contain the mutation. For instance, a VAF of around 50% indicates that the mutation may be heterozygous in all cells of the biopsy, or homozygous in half of the cells. Subclonal mutations generally have a lower VAF, since those mutations only occur in a subset of the cancer cells. The VAF of cancer mutations inform the clonal decomposition of the tumor. The mutations in *MUC16* and *TP53* were assigned to different clones, because their VAFs change into different directions between the two samples (Fig. 4B). For instance, one tumor sample of patient 5 with $\mathrm{VAF}_{\mathrm{MUC16}}$ = 3.5% and $\mathrm{VAF}_{\mathrm{TP53}}$ = 0% has a higher VAF in *MUC16*, while the other tumor sample with $\mathrm{VAF}_{\mathrm{MUC16}}$ = 0% and $\mathrm{VAF}_{\mathrm{TP53}}$ = 14.7% shows a higher VAF in *TP53*.

The second most significant pair is *BAP1* and *TP53*, which is clonally exclusive in one of two patients (Fig. 4C). More precisely, it is clonally exclusive in patient 84, and co-occurring in patient 30. The latter one has an average rate of clonal exclusivity of zero, which means that the inferred phylogenetic tree was always linear in all repeated runs of tree inference. In general, the tree inference favors parsimonious trees (Marass et al., 2016). If there is not enough evidence for a branching of the phylogenetic tree, the inferred tree will be linear. For patient 30, the change in the VAF between the two tumor samples is different for the two mutations in *BAP1* and *TP53* (Fig. 4D), supporting that they are actually in different clones. However, the coverage of the *BAP1* mutation locus is only 25x in the second tumor sample. The average coverage in the two samples of patient 30 is with 75x and 67x lower than the average coverage across all samples in general (93x). Additionally, patient 30 had only six mutations for the tree inference, and hence there may not be enough evidence to infer a tree with a branching that separates the two genes. A larger number of biopsies of the tumor would allow to infer the clonal composition with more certainty. *BAP1* and *TP53* were found to be mutated in co-existing clones of a ccRCC case before (Gerlinger et al., 2014).

The two more speculative clonally exclusive gene pairs {*MUC16*, *TP53*} and {*BAP1*, *TP53*} are examined in more detail in the discussion section, including possible interpretations on how the clones with these mutations could be synergistic.

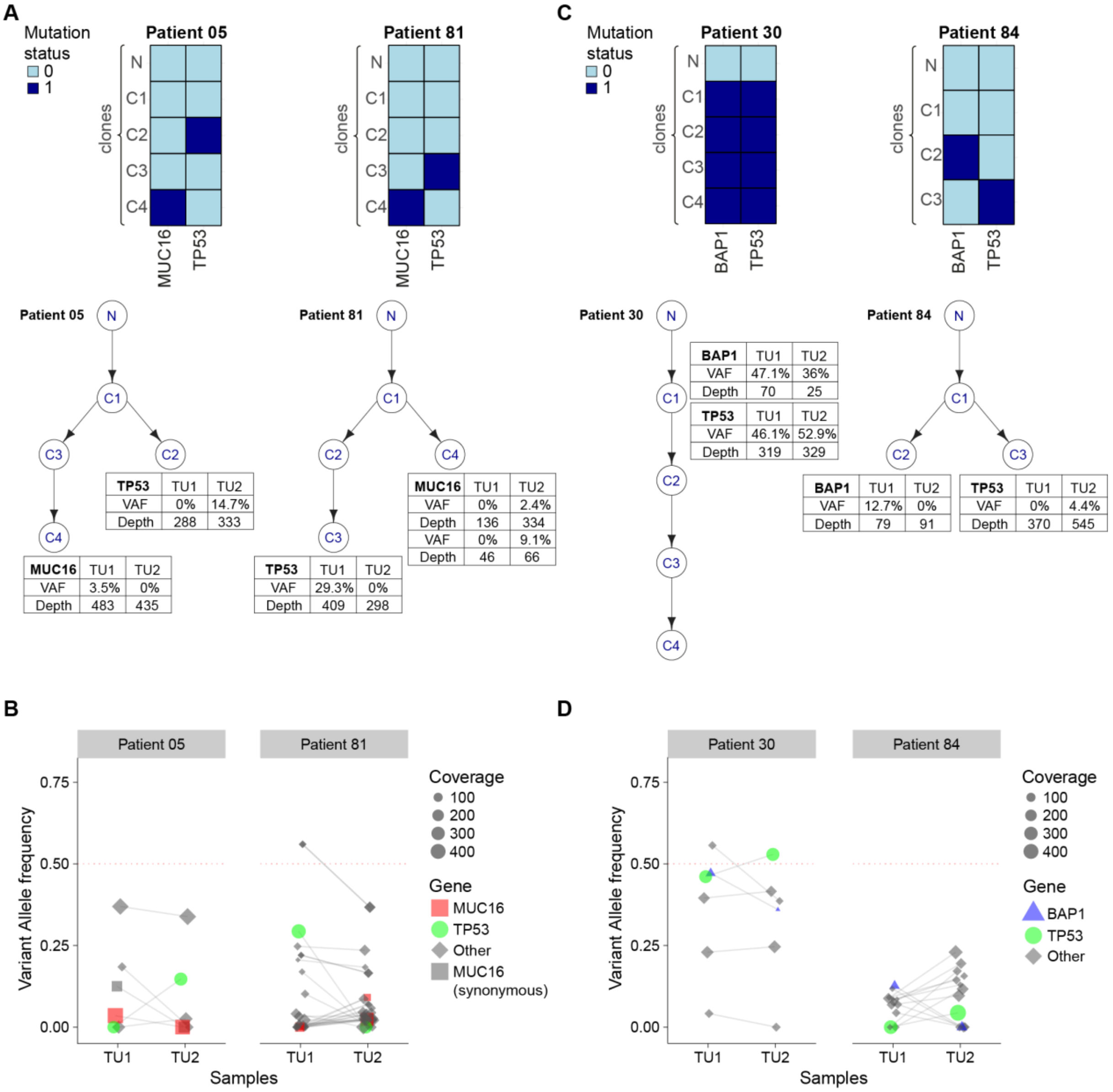
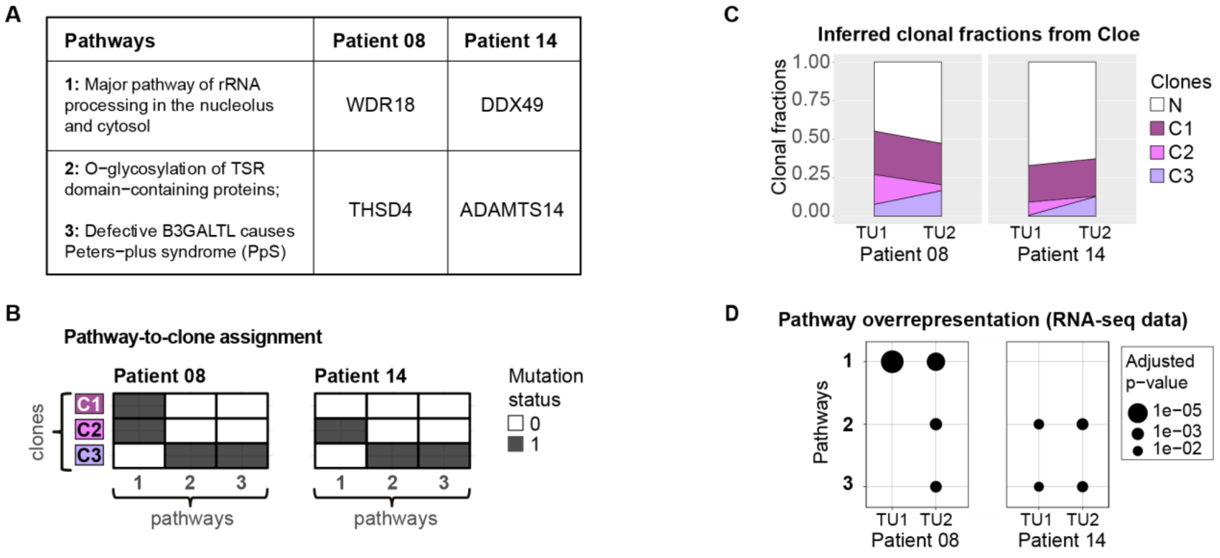


Figure 4. The two most striking gene pairs {*MUC16*, *TP53*} and {*BAP1*, *TP53*} are examined in more detail. **(A)** *MUC16* and *TP53* are mutated in patients 5 and 81. The heatmap and the phylogenetic tree from one of the tree inference runs indicate the clones in which *MUC16* and *TP53* are mutated. They are clonally exclusive in both patients. The depth and variant allele frequency (VAF) is listed for each of the mutations in *MUC16* and *TP53* in both samples, tumor sample 1 (TU1) and tumor sample 2 (TU2). In the heatmaps and phylogenetic trees, the clone N represents the normal or germline sample. **(B)** The VAF in the paired samples of patients 5 and 81 and in particular of the mutations in *MUC16* and *TP53*. The change in the VAF between the two samples is different for the two genes, which is an indication that they are in different clones. **(C)** *BAP1* and *TP53* are mutated in patients 30 and 84. They are clonally exclusive in patient 84 as indicated by the heatmap and phylogenetic tree. Patient 30 has an average rate of clonal exclusivity of 0, which reflects a linear tree, and hence no genes can be in different clones. **(D)** The VAF in the paired samples of patients 30 and 84. The change in the VAF between the two samples for the genes *BAP1* and *TP53* is different in both patients. The inferred tree is linear in patient 30, which places the two genes *BAP1* and *TP53* in the same clone. However, there is a difference in the VAF of *BAP1* and *TP53* also in the two samples of patient 30, which could reflect that they are in fact in different clones.

Pathway-level clonal exclusivity in 16 ccRCC patients

The WES data from paired tumor biopsies and matched normal samples of 16 ccRCC patients enables a clone pattern analysis on the pathway level. That is, the mutated genes can be mapped to pathways, and we can detect pathway pairs that are affected in several patients. The pathway-level analysis is feasible since, with the whole exome, all genes are covered and it is possible to determine whether a pathway is affected or not, unlike for the panel, where only a subset of genes is targeted. The mutations detected in the WES data set (Methods section) were assigned to clones with Cloe (Marass et al., 2016), mapped to genes, and subsequently, the genes were mapped to pathways using the Reactome pathway database (Fabregat et al., 2018). This procedure resulted in a total of 877 affected pathways. Across all patients, there were 71,422 pairs of pathways, most of which were affected in just one patient. A total of 5,433 pairs of pathways were simultaneously affected in two or more patients. Among these, when applying GeneAccord, 211 pairs had a parameter Δ that was greater zero and were tested for significance.

The two most striking clonally exclusive pathway pairs are {“Major pathway of rRNA processing in the nucleolus and cytosol” (referred to as pathway 1), “O-glycosylation of TSR domain-containing proteins” (pathway 2)}, as well as {pathway 1, “Defective B3GALTL causes Peters-plus syndrome (PpS)” (pathway 3)}, which are clonally exclusive in both patients in which they are affected (Fig. 5A, 5B, Supplementary Table S3). They are clonally exclusive in patient 8 and patient 14. Both patients have a very low average rate of clonal exclusivity of 0.012 and 0.008, respectively, on the pathway level, and hence this clone pattern is highly significant ($p < 10^{-5}$). Pathway 1 belongs to the category “Metabolism of RNA”, while pathway 2 falls into the class “Metabolism of proteins”, and pathway 3 is a disease pathway related to diseases of glycosylation (Fabregat et al., 2018). Pathway 1 was also significantly enriched among the differentially expressed genes in 13 of 16 patients of this cohort. The altered metabolism of RNA appears indeed to be important in ccRCC: Recently, gene expression analysis revealed that deregulation of genes in the RNA metabolic process was a significant and distinctive feature in ccRCC when compared to other subtypes of RCC (Ricketts et al., 2018). Also, mRNA processing was among the significantly affected pathways in a study of 240 ccRCC (Sato et al., 2013). The pathway pairs {1, 2} and {1, 3} are affected in the two patients in a different subset of genes (Fig. 5A). Hence this clonal exclusivity pattern is only detectable on the pathway level. Potential synergies between the two clones with these affected pathways are discussed below.



Pathways	Patient 08	Patient 14
1: Major pathway of rRNA processing in the nucleolus and cytosol	WDR18	DDX49
2: O−glycosylation of TSR domain−containing proteins; **3:** Defective B3GALTL causes Peters−plus syndrome (PpS)	THSD4	ADAMTS14

Figure 5. The two most striking clonally exclusive pathway pairs from the clonal exclusivity test when performed on the pathway level of the WES data set from 16 ccRCC patients. The pathway pairs are affected in patients 8 and 14 and in each patient, a different subset of genes is mutated. Hence this clonal exclusivity pattern is only detectable on the pathway level. **(A)** The table displays the genes that are mutated in these pathways. **(B)** The heatmaps illustrate in which clones the genes in these pathways are mutated. **(C)** The proportion of cells from the clones in each of the two samples from the two patients. The label ‘N’ represents the fraction of normal cells in the biopsy. **(D)** The data set also includes RNA-seq data from each sample. Among the differentially expressed genes in each sample, the pathways are significantly overrepresented in some of the samples.

Discussion

GeneAccord provides a statistical framework for detecting pairs of genes or pathways that are more often mutated in different clones of the same tumor than expected by chance. There are several possible biological explanations for such a mutational pattern. The two clones can have complementing phenotypes and cooperate or benefit from each other, for instance, by sharing of diffusible factors. Another possibility is that the two genes of a clonally exclusive pair are synthetically lethal and that both genes being mutated in the same cell would lead to a disadvantageous phenotype. Lastly, the two clones may also be the result of parallel convergent evolution where both clones exhibit the same phenotype by different mutations. In order to gain a better understanding of possible reasons for the clonally exclusive pattern, it is important to examine the biological functions of these genes and pathways in more detail.

Clonal exclusivity of MUC16 and TP53

TP53 is a well-known tumor suppressor and plays a role in growth arrest, apoptosis, and DNA repair (The UniProt, 2017). According to the COSMIC database, *TP53* was found to be mutated in ccRCC in 6% of 1,483 ccRCC cases (Forbes et al., 2017). *TP53* was among the significantly mutated genes in several ccRCC studies (Sato et al., 2013, Cancer Genome Atlas Research, 2013, Chen et al., 2016), and was associated with reduced survival (Ricketts et al., 2018, Girgin et al., 2001), which is confirmed in the cohort presented here (p = 0.01; Log-rank test). The p53 signaling pathway in general was also reported to be significantly affected in ccRCC (Sato et al., 2013). A previous study of the intra-tumor diversity in ten ccRCC cases revealed that *TP53* was one of the most extreme examples of genes being detected more often when sequencing multiple biopsies per tumor instead of one biopsy (Gerlinger et al., 2014). This observation suggests that *TP53* mutations may be a crucial subclonal event in ccRCC. Studies with RCC cell lines showed that p53 was repressed (Ku et al., 2013) and not functional, even when the gene *TP53* was not mutated (Gurova et al., 2004).

MUC16 is also called CA 125 and encodes a transmembrane glycoprotein (Hollingsworth and Swanson, 2004, Lucarelli et al., 2014). It is mutated in 8% and 14% of ccRCC cases in the TCGA study, and the 106 samples from Sato et al., respectively (Cancer Genome Atlas Research, 2013, Sato et al., 2013). *MUC16* is upregulated in lung cancer (Lakshmanan et al., 2017), pancreatic cancer (Haridas et al., 2011) and ovarian cancer, where it is used as a biomarker (Felder et al., 2014) and plays a crucial role in promoting tumorigenesis and metastases (Theriault et al., 2011). Moreover, several RCC studies highlighted that increased levels of CA 125 were associated with poor prognosis and more advanced stage, suggesting the potential use of this serum as a biomarker also in RCC (Grankvist et al., 1997, Lucarelli et al., 2014, Bamias et al., 2003).

Lakshmanan and colleagues recently demonstrated that the overexpression of *MUC16* is associated with increased tumor cell growth, cancer cell migration, and resistance to cytotoxic drugs in lung cancer (Lakshmanan et al., 2017). Furthermore, they discovered that MUC16 activates the signaling proteins JAK2, STAT3, and glucocorticoid receptor (GR), which in turn upregulate TSPYL5 (Lakshmanan et al., 2017). TSPYL5 was previously reported to be a suppressor of TP53 (Epping et al., 2011). Hence Lakshmanan et al. suggested that MUC16 exerts its chemoresistant, anti-apoptotic, and metastatic function by downregulating *TP53* and its target genes via the signaling cascade of JAK2, STAT3, GR, and TSPYL5 (Lakshmanan et al., 2017). In light of this fact, the tumors with *MUC16* and *TP53* mutated clones may have undergone convergent evolution where both clones developed their own mechanism of shutting down *TP53*, one with a mutation in *TP53* directly, and the other one via a mutation in *MUC16* that targets TP53 indirectly via other signaling genes. This explanation is in line with the fact that they were found to be mutually exclusive in breast cancer and endometrial carcinoma (Dao et al., 2017).

However, mutual exclusivity is not necessarily an indication that two genes are functionally redundant. This was exemplified by a study in which *BAP1* and *PBRM1* were found to be mutually exclusive in ccRCC (Pena-Llopis et al., 2013), and these genes seem to be involved in different processes: The tumors with either mutation were reported to have different gene expression signatures, and the patients showed different survival rates (Pena-Llopis et al., 2013). Therefore, alternative interpretations for the clonally exclusive pattern of *MUC16* and *TP53* are possible. MUC16 is a transmembrane protein, and might act as a cell-surface receptor that transfers external signals into the cell (Hollingsworth and Swanson, 2004). This could influence various processes such as cell-adhesion, which is important for metastasis and invasion (Hollingsworth and Swanson, 2004). Furthermore, MUC16 activates the signaling proteins JAK2 and STAT3 (Lakshmanan et al., 2017), which in turn are known to orchestrate angiogenesis by upregulating VEGF and basic fibroblast growth factor (bFGF) (Zhao et al., 2011). On the other hand, TP53 is involved in various signaling pathways such as Wnt signaling, MAPK signaling, and signaling by Interleukins (Kanehisa and Goto, 2000, Fabregat et al., 2018). Moreover, TP53 is known to interact with growth factors and in particular to be a regulator of several growth factors and growth factor receptors (Asschert et al., 1998). Hence, if the *MUC16* clone is engaged in metastasis, invasion, and angiogenesis, while the *TP53* clone promotes growth factors, the two clones could be complementary.

Clonal exclusivity of BAP1 and TP53

BAP1 is a tumor suppressor that is involved in DNA repair and among the most frequently mutated genes in ccRCC (Pena-Llopis et al., 2013, Cancer Genome Atlas Research, 2013). It is a histone deubiquitinase and associated with poor survival prognosis (Cancer Genome Atlas Research, 2013). In this cohort, *BAP1* mutations were also found to be associated with highest tumor grade ($p < 0.01$; Fisher's exact test) and decreased survival ($p = 0.06$; Log-rank test). *BAP1* and *TP53* were found to be mutually exclusive in ccRCC (Bi et al., 2016). According to the gene ontology resource, TP53 and BAP1 both perform chromatin and protein binding, and both are associated to biological processes such as “negative regulation of cell proliferation”, “regulation of cell cycle”, and “protein localization to mitochondrion” (The Gene Ontology, 2017, Ashburner et al., 2000). Both proteins are members of the Reactome pathways “Deubiquitination” and “DNA double strand break response” (Fabregat et al., 2018). Following the same reasoning as before, the clonally exclusive pattern may also have other reasons than similar functions. TP53 and BAP1 actually share interacting partners, such as BRCA1 (Jiang et al., 2011, Nishikawa et al., 2009) and BARD1 (Irminger-Finger et al., 2001, Nishikawa et al., 2009), which could imply an indirect signaling connection between the two genes and possibly between the clones. Gerlinger and colleagues also found *TP53* and *BAP1* to be mutated in co-existing clones in a ccRCC case and hypothesized that it represents parallel and divergent evolution within the tumor (Gerlinger et al., 2014).

Clonally exclusive pathway pairs

The two most striking pathway pairs include “Major pathway of rRNA processing in the nucleolus and cytosol” (pathway 1), which is clonally exclusive with “O-glycosylation of TSR domain-containing proteins” (pathway 2) and “Defective B3GALTL causes Peters-plus syndrome (PpS)” (pathway 3), respectively (Fig. 5A). The pathway pairs {1, 2} and {1, 3} are affected in the two patients in a different subset of genes (Fig. 5A), namely *WDR18* and *THSD4* in patient 8, and *DDX49* and *ADAMTS14* in patient 14. ADAMTS14 belongs to the ADAMTS protein family, which are secreted zinc metalloproteases that play a role in the extracellular matrix related to angiogenesis and cancer (Apte, 2009). THSD4 is also referred to as ADAMTS-like protein 6 and is also secreted to the extracellular matrix (The UniProt, 2017). Both, *ADAMTS14* and *THSD4* contain the thrombospondin type 1 repeat (TSR) domain (Matthews et al., 2009). The proteins with TSR domains can undergo O-fucosylation, a protein modification that plays a role in angiogenesis and Notch signaling (Adams and Tucker, 2000, Matthews et al., 2009, Moremen et al., 2012).

The mutated genes in the pathways 2 and 3, *ADAMTS14* and *THSD4*, are both secreted to the extracellular matrix, and therefore the clones with the affected pathways 2 and 3 may have an effect on other clones in the tumor. Whether there is a potential crosstalk between the clones with the pathway pairs {1, 2} and {1, 3} should be verified experimentally. To gain additional insight, it is useful to consider the RNA-seq data of these biopsies. Aside from inferring the mutation-to-clone assignment, the Cloe software also estimates the fractions of the clones in each sample (Fig. 5C). This is important in order to interpret possible changes on the transcriptomic level in the bulk samples. The RNA-seq data from the bulk samples allows to detect differential gene expression within all cells of the biopsy combined. Hence there is no clonal assignment of the transcriptomic changes. However, the clonal fractions estimated by Cloe hint towards the clones that are most abundant in the bulk sample. Among the differentially expressed genes in patients 8 and 14, the two most striking clonally exclusive pathways pairs are also enriched in some of the samples (Fig. 5D). The pathway 1 was mapped to clones 1 and 2 in patient 8 (Fig. 5B), which together have a clonal faction of 47.5% and 30.7% in samples TU1, and TU2, respectively (Fig. 5C). The pathway enrichment analysis shows that this pathway is also highly overrepresented in these two samples on the transcriptomic level (Fig. 5D). The pathways 2 and 3 were assigned to clone 3, which is, with 16.6%, most abundant in sample TU2 of patient 08 (Fig. 5C). Indeed, these pathways are also enriched among the differentially expressed genes in this sample (Fig. 5D). Pertaining to patient 14, no enrichment of pathway 1 can be found in either sample (Fig. 5D). Pathway 1 was assigned to clone 2, which was estimated to have a clonal fraction of only 8.6% and 0.2% in samples TU1 and TU2, respectively (Fig. 5C), which may explain why there is no signal detectable in the bulk transcriptome samples for this pathway. Pathways 2 and 3, however, were assigned to clone 3 (Fig. 5B), which has an estimated clonal fraction of 0.5% and 12.8% in samples TU1 and TU2, respectively. In both bulk RNA samples, pathways 2 and 3 are enriched. The fact that pathways 2 and 3 are also enriched in sample TU1 despite the small clonal fraction of 0.5% in this bulk sample shows that clone 3, which contains the affected pathways 2 and 3, may have an effect also on the other clones in this tumor.

In summary, we analyzed the intra-tumor heterogeneity in a cohort comprising a total of 178 tumor biopsies from 89 ccRCC patients. The investigation of the WES and RNA-seq data of 16 ccRCC patients revealed that the intra-tumor heterogeneity is very pronounced on the genetic level, while the transcriptomic level tends to show less heterogeneity. Yet still 31% of differentially expressed genes were detected in only one of the two biopsies from a tumor, and some pathways, especially related to translation, were only enriched subclonally in some patients. The extended cohort of 178 tumor biopsies from 89 ccRCC patients enabled a more detailed analysis. The high degree of intra-tumor heterogeneity commonly observed in ccRCC was also found in this cohort, where 62% of mutations were only detected in one of the two biopsies from a tumor. The four most frequently mutated genes in ccRCC, namely VHL, PBRM1, SETD2, and BAP1, were also among the most frequently mutated genes in this cohort. We found a pattern of convergent phenotypic evolution in the gene *PBRM1*, which was hit by two different deleterious missense mutations in each of the two biopsies of a patient. Furthermore, many mutations were identified in the mucins *MUC6*, *MUC16*, and *MUC3A*, which may indicate that these genes also play an important role in ccRCC development.

We have also introduced a novel approach to systematically analyze the subclonal mutation patterns in this large cohort of tumor patients. While many cancer-related mutually exclusive gene pairs have been previously identified across different tumors, finding such pairs within a single tumor has been elusive. Our statistical framework, GeneAccord, reveals pairs of mutated genes in co-existing clones. The co-existence in the same tumor suggests the possibility of a selective advantage of the combined clones possibly through cooperative interactions between them. A definite proof of such interactions will require future experimental studies *in vitro* or *in vivo*. Here, we have developed and applied a new computational approach to identify such gene (or pathway) pairs that are unlikely to be generated by random chance alone. As such, the GeneAccord method will also be useful in finding candidate cooperative tumor clones in other cancer patient cohorts.

Using our approach, the analysis of subclonal mutation patterns in the 89 ccRCC cases revealed several promising pairs of co-existing clones that harbor mutations in known cancer related genes. The most striking pairs of mutated genes in co-existing clones include {*MUC16*, *TP53*} and {*BAP1*, *TP53*}. Moreover, we applied the clonal exclusivity test to 16 ccRCC patients on the pathway level, and identified co-existing clones with affected pathways related to the metabolism of proteins and RNA. The identification of co-existing clones and their drivers may inform the design of combination treatments that target all clones within a tumor and will potentially improve treatment outcome.

Given the observation from a previous analysis of ten ccRCC that more biopsies lead to the detection of more subclonal driver alterations in almost all cases (Gerlinger et al., 2014), we emphasize that our analysis probably still underestimates the genetic diversity of ccRCC. We would obtain a clonal decomposition at an even higher resolution if more than two biopsies were included per patient. In this sense, the intra-tumor genetic heterogeneity and clonal exclusivity described in this study, most likely still represents only an improved lower bound of the true amount present in ccRCC.

A possible extension of GeneAccord would be to directly work with the posterior distribution of the inferred tree given the observed data, rather than taking as input a fixed collection of trees per patient. In that case, GeneAccord would need to be directly combined with the tree inference method. Another extension would be to allow also the detection of multiple genes rather than just pairs. That way, groups of genes could be identified that occur together in the same clone but are clonally exclusive with another group of genes. However, since most genes are only mutated in a small fraction of patients, this would require a very large cohort size.

Pertaining to the applicability of our framework, we note that the mutation-to-clone assignment, which is the input to our method, depends on the variant allele frequency, and therefore it is important to have sufficient depth of read coverage in the samples as well as ideally several samples per tumor. Also, a large cohort of patients is advantageous to identify rare clonal compositions. Such data sets are currently not abundant, but with decreasing sequencing costs (Muir et al., 2016), the cohorts of sequenced cancer patients will also increase in size, as well as breadth and depth of the sequenced regions, and number of biopsies per patient. This development will likely soon enable whole-exome or whole-genome sequencing also in large cohorts which will allow for simultaneous pathway and gene pair analyses on the same cohorts.

Additionally, the methods to infer the clonal genotypes of tumors improve constantly. For instance, with the advent of single-cell sequencing technologies (Wang and Navin, 2015), and methods to infer the phylogeny from single-cell data (Kuipers et al., 2017), the mutation-to-clone assignment will be even more reliable. In fact, combined single-cell sequencing of the whole exome and the transcriptome will even allow to assign differentially expressed genes to specific clones. This would enable a GeneAccord analysis on the transcriptomic level. For our algorithm, it does not matter from which omics layer the alterations come, it is only crucial to have a clonal assignment of all altered players. Therefore, with single-cell data, one could then also perform a combined analysis using various omics layers including differentially expressed genes, as well as copy number and epigenetic changes. This would allow obtaining a more holistic portrait of the subclonal alteration profiles and potentially reveal more synergies between clones that could inform the design of future treatments.

Finally, we reported a systematic analysis of subclonal mutation patterns in a large cohort of cancer patients including 178 ccRCC biopsies and revealed promising pairs of co-existing clones. In general, our statistical framework opens up a novel avenue for systematically analyzing the clone constellations in large cohort of patients and to pinpoint pairs of genes that are altered in co-existing clones. The analyses this approach enables will contribute towards a better understanding of the evolutionary forces beyond mutation and selection that drive tumor evolution and will help to improve treatment strategies to eradicate the tumor as a whole with all its clones and key players.

Materials and Methods

Clonal exclusivity test for mutated genes in different clones

We developed a statistical framework to find clonally exclusively altered gene or pathway pairs in cancer clones across patients. A gene or pathway pair is referred to as clonally exclusive in a tumor, if both members are mutated, but not in the same clone, i.e. the respective mutations occurred in different branches of the tumor phylogenetic tree. In the example of Figure 1B, the yellow and light blue mutations are assigned to a disjoint set of clones. Let us assume the yellow mutation occurs in gene A, the light blue mutation in gene B, and the dark blue mutation in gene C. The gene pair $\{A, B\}$ is clonally exclusive, and so is $\{A, C\}$, but not $\{B, C\}$. This pattern can of course also occur by chance. To decide whether two genes are clonally exclusive at a higher rate than expected, we developed the clonal exclusivity test, which is implemented as an R package called GeneAccord.

The required input data is a table with altered genes or pathways and their clone assignments in a binary fashion from a cohort of cancer patients (Supplementary Fig. S1A). For each patient, one may run the phylogenetic tree inference several times or take multiple samples from the posterior tree distribution, as this step involves uncertainty. Genes that were only assigned to clones in a subset of the collection of tree inferences of a patient may be excluded from further analysis to reduce the noise in the gene-to-clone assignments. We use the phylogenetic tree inference tool Cloe (Marass et al., 2016) and run it 20 different times with different seeds in order to account for uncertainty. Genes that are assigned to clones in less than 90% of the tree inference runs were removed for the GeneAccord analysis. After having computed one or several tree inferences for all patients, the following two steps are carried out for each patient in the cohort separate: First, the number of gene pairs that are mutated in different branches of the tree, i.e., that are clonally exclusive, is computed for each tree in the collection of tree inferences. Two genes being assigned to different branches of the phylogenetic tree is equivalent to them having a mutually exclusive clone mutation profile (Fig. 1B). We refer to

$$r = \frac{\text{\# of gene pairs on different branch}}{\text{\# of all gene pairs}}$$

as the rate of clonal exclusivity. This quantity depends on the tree structure and the number of clones in total, and hence may be different for each patient and for each tree. Let m_i be the mean of the clonal exclusivity rates across all trees from the collection for a patient i.

Second, for each gene pair $\{A, B\}$ in a patient i, the number of times the pair occurs across the collection of trees is computed and is referred to as $NR_{i,\{A,B\}}$. For instance, a gene pair could occur in 19 of 20 trees, if one of the genes was not assigned to a clone in one of the trees. Moreover, for each gene pair $\{A, B\}$, the number of times it was clonally exclusive across the collection of trees $NSE_{i,\{A,B\}}$ is computed. For instance, a gene pair can be clonally exclusive in 16 of 19 trees, and be assigned to the same clone in the other trees. Hence the quantity $\frac{NSE_{i,\{A,B\}}}{NR_{i,\{A,B\}}}$ is the fraction of trees the pair was found clonally exclusive in patient i. For each gene pair $\{A, B\}$ in patient i, the overall rate of clonal exclusivity is defined as

$$s_{i,\{A,B\}} = \frac{NSE_{i,\{A,B\}}}{NR_{i,\{A,B\}}} m_i + \left(1 - \frac{NSE_{i,\{A,B\}}}{NR_{i,\{A,B\}}}\right)(1 - m_i).$$

Hence $s_{i,\{A,B\}}$ contains the average rate of the patient, which is weighted by the fraction of trees, in which the specific pair was clonally exclusive. Therefore this overall rate takes into account the gene pair's clone constellation across the collection of trees. After the rates m and the values s for all pairs and all patients have been generated, the entire cohort is considered to compute the null distribution of the test statistic by decoupling gene pair labels between patients. The test statistic is the likelihood ratio of the null hypothesis and the alternative hypothesis. The null hypothesis H_0 assumes independence of mutations, and therefore states that the gene pair of interest occurs across the cohort in the clonal exclusivity pattern as often as expected according to the overall rates s. The likelihood under the null hypothesis for a pair $\{A, B\}$ is defined as

$$L_{0,\{A,B\}} = \prod_i s_{i,\{A,B\}}$$

Where the product runs over all patients in which the pair is mutated. The alternative hypothesis H_1 is that the gene pair is clonally exclusive at a rate which is different from the rates s. The alternative likelihood for the pair $\{A, B\}$ can be written as

$$L_{1,\{A,B\}} = \prod_i s^*_{i,\{A,B\}}$$

where $s^*_{i,\{A,B\}} = \frac{NSE_{i,\{A,B\}}}{NR_{i,\{A,B\}}} m^*_i + (1 - \frac{NSE_{i,\{A,B\}}}{NR_{i,\{A,B\}}})(1 - m^*_i)$. The parameter m^*_i is the alternative rate of clonal exclusivity and is linked to the rate m_i as

$$\operatorname{logit}(m^*_i) = \operatorname{logit}(m_i) + \Delta \qquad (1)$$

where the logit transformation is used to ensure that the rates are always between zero and one.

The parameter Δ quantifies the deviation from the expected rate for the pair of interest. It is estimated from the data using maximum likelihood. The test statistic is defined as

$$T = -2\log\left(\frac{L_{0,\{A,B\}}}{L_{1,\{A,B\}}}\right).$$

In order to obtain the expected distribution of the test statistic under the null hypothesis, its empirical cumulative distribution function (ECDF) is built using the following procedure. For each number of patients n that a pair can be mutated in, the test statistic under the null hypothesis is computed $100{,}000$ times. To compute the test statistic under the null hypothesis, n patients from the cohort are sampled. The sampling is random but proportional to the number of pairs in each patient. From each of these n patients, a gene pair is randomly chosen from which the quantity $\frac{NSE_{i,\{A,B\}}}{NR_{i,\{A,B\}}}$ as well as the average clonal exclusivity rate m_i is taken. The rate m_i is distorted by a beta distribution with mean m_i, and $M = \alpha + \beta = 1000$, in order to obtain a smooth distribution function. Next, the overall rate of clonal exclusivity $s_{i,\{A,B\}}$ is computed. From the n different rates s from the n patients, the test statistic T is computed. The distribution of the test statistic under the null hypothesis is computed separately for each number of patients n that a pair can be mutated in. This is owing to the fact that the null distribution of the test statistic may be different depending on in how many patients a gene pair is mutated (Supplementary Fig. S3).

Next, the observed test statistic of a gene pair of interest can be compared to the distribution of the test statistic under the null hypothesis to obtain a p-value. That is, the p-value of an observed test statistic t for a gene pair of interest is defined as

$$p = P_{H_0}(T > t) = 1 - F(t),$$

where F is the ECDF of the test statistic under the null hypothesis which was previously computed. We only test if a gene pair is mutated in more than one patient, since we do not have enough power otherwise. When comparing an observed test statistic to the null distribution, the number of patients in which the pair is mutated determines which of the previously generated null distributions is used.

We assessed the p-value distribution under the null hypothesis in order to verify that the p-values are uniformly distributed. To this end, we generated 5,000 simulated pairs by using the same strategy as described above for the generation of the ECDF. This was done for different numbers of patients n that the pairs are mutated in ($n = 2, 3, \ldots, 7$). The p-values are distributed uniformly (Supplementary Fig. S4).

The sign of the parameter Δ in equation (1) signifies the type of the clone pattern for a pair. That is, if the parameter Δ is greater zero, the pair tends to be clonally exclusive more often than the expected rate. Conversely, if the parameter Δ is less than zero, the pair tends to occur more often together in the same clone. Here, we are interested in pairs with a clonally exclusive mutation pattern, and hence we only test pairs where the parameter Δ is greater zero.

Data generation and analysis

We applied the clonal exclusivity test to analyze two data sets from a cohort of ccRCC patients. The tumors of these ccRCC patients have been classified according to the 2016 WHO classification (Moch et al., 2016).

The first data set encompasses two spatially separated primary tumor biopsies and one matched normal sample from each of sixteen clear cell renal cell carcinoma (ccRCC) patients. The whole exome was sequenced using the Illumina HiSeq 2000 system to obtain 101-base paired-end reads. The computational pipeline to analyze the data was a customized version of the NGS-Pipe framework (Singer et al., 2017) that included the following steps: Adapter clipping and trimming of low-quality bases with Trimmomatic (Bolger et al., 2014), alignment of the reads to the human reference genome version hg19 using bwa (Li and Durbin, 2009), as well as read processing with SAMtools (Li et al., 2009b), Picard tools (Picard), and bam-readcount. Reads were realigned locally around indels, and base qualities were recalibrated with the Genome Analysis Toolkit (GATK) (Van der Auwera et al., 2013). The single-nucleotide variants (SNVs) were called using the rank-combination (Hofmann et al., 2017) of deepSNV (Gerstung et al., 2012), JointSNVMix2 (Roth et al., 2012), MuTect (Cibulskis et al., 2013), SiNVICT (Kockan et al., 2017), Strelka (Saunders et al., 2012), and VarScan2 (Koboldt et al., 2012). The rank-combination is a method that combines the results of different variant callers by integrating the ranked lists of variants to generate a combined ranking (Hofmann et al., 2017). P-values of deepSNV were corrected for multiple testing with the R package IHW (Ignatiadis et al., 2016). Indels were called using SiNVICT (Kockan et al., 2017), Strelka (Saunders et al., 2012), VarDict (Lai et al., 2016), and VarScan2 (Koboldt et al., 2012), and combining them with the rank-combination (Hofmann et al., 2017). For copy number variant detection, the tool Sequenza (Favero et al., 2015) was employed. Mutations in copy number neutral regions were selected as input for Cloe (Marass et al., 2016), a tool to reconstruct the evolutionary history of a tumor and to assign the mutations to different clones. In order to account for the uncertainty in the phylogenetic tree inference, Cloe was run 20 times with different seeds.

Additionally, for the 48 samples from the initial 16 ccRCC, paired-end and single-end RNA-sequencing was performed on the Illumina HiSeq 2000 system to generate 101-base paired-end reads, and 51-base single-end reads for each sample. For the computational analysis, the NGS-Pipe framework was adapted (Singer et al., 2017). Reads were clipped and trimmed using Trimmomatic (Bolger et al., 2014), and alignment was performed with STAR (Dobin et al., 2013). Read counts for the genes were obtained with the program featureCounts (Liao et al., 2014). Differential gene expression analysis was done using DESeq2 (Love et al., 2014). The R package WebGestaltR (Wang et al., 2017) was applied to perform enrichment analysis.

The second data set is a panel sequencing data set. It comprises an extended cohort of patients from which a selected set of genes was sequenced at higher depth. The selection of 826 genes was informed by the mutated genes detected in the WES data set, as well as from the frequently mutated ccRCC genes in TCGA (Cancer Genome Atlas Research, 2013). We generated panel seq data from 89 ccRCC patients, including two spatially separated primary tumor biopsies, and a matched normal sample per patient. This data set comprises 14 of the patients from the WES data set, and 75 additional patients. The data was sequenced using the Illumina HiSeq 2500 system. The sequenced reads contain unique molecular identifiers (UMI), and this allows for the correction of potential sequencing errors. Reads with identical UMIs, which are mapped to the same genomic position, come from the same DNA molecule, and therefore, the consensus sequence can be built, and the variants can be called with higher confidence.

The pipeline for analyzing the sequencing data was again a customized version of the NGS-Pipe framework (Singer et al., 2017) including the following steps: Raw reads were clipped and trimmed using the tool SeqPurge (Sturm et al., 2016). Reads were aligned to the human reference genome version hg19 with the aligner bwa (Li and Durbin, 2009). Reads were further processed using SAMtools (Li et al., 2009a), Picard tools (Picard) and bam-readcount. Local realignment around indels was done with GATK (Van der Auwera et al., 2013). We used the software UMI-tools (Smith et al., 2017) to group reads with identical UMI and identical mapping position together, and an in-house tool to build the consensus sequence and thereby correct sequencing errors. Our in-house tool takes the grouped reads with identical UMI and identical mapping position and attempts to generate the consensus sequence from these grouped reads. If the reads contain contradicting bases at a nucleotide position, it is masked with the base 'N'. The SNV and indel calling was similar as for the WES data set. Some of our samples are from formalin fixed paraffin embedded (FFPE) material. FFPE samples are known to harbor artificial C>T and G>A alterations (Wong et al., 2014, Oh et al., 2015). They occur mostly at lower frequencies in the range between 1-10% variant allele frequency (VAF), since the DNA damage occurs at different genomic positions in different cells (Wong et al., 2014, Yost et al., 2012). To remove potential FFPE artifacts, we filtered out C>T and G>A mutations that had a VAF < 10%. The tool Cloe (Marass et al., 2016) for the tree inference requires as input mutations in copy number neutral regions. In order to filter out mutations that are in potential copy number variant regions, mutations that are within 4000 bps of an imbalanced heterozygous germline SNP were filtered out. An imbalanced heterozygous germline SNP is a SNP that has VAF between 40-60% in the normal sample, but in the tumor sample the VAF is out of these bounds, indicating a potential copy number change.

In the WES and panel seq data sets, in order to perform quality control we used the programs Qualimap (Okonechnikov et al., 2016) and FastQC (Andrews, 2010). For annotation of the variants, SnpSift (Cingolani et al., 2012a) and SnpEff (Cingolani et al., 2012b) as well as the data bases COSMIC (Forbes et al., 2016) version 80, and dbSNP (Sherry et al., 2001) version 138 were used. For the subsequent GeneAccord analysis, only genes with a potential impact were kept. While mutations such as synonymous or intronic variants are informative for the tree inference, they were not of interest for the clonal exclusivity test. More precisely, the annotation program SnpEff classifies variants into four categories based on the potential impact of the mutation (Cingolani et al., 2012b). These are, in descending order of importance, the categories 'HIGH', 'MODERATE', 'LOW', and 'MODIFIER'. Examples of the category 'HIGH' would be frameshift indel. Missense mutations and inframe indels are classified as 'MODERATE'. The category 'LOW' includes synonymous and splice region mutations. Variants that are annotated as 'upstream', 'intronic', or 'UTR region' fall into the class 'MODIFIER'. For the GeneAccord analysis, we keep mutations that are in the category 'HIGH', 'MODERATE', and from the class 'LOW' we keep all variants with the exception of: synonymous variants, or mutations that are annotated as the case where a start codon mutates into another start codon, or analogous for stop codon. To sum up, we keep variants such as missense, frameshift or inframe indel or variants in splice regions, but filter out variants that are synonymous, intronic or in the UTR regions.

For the data analysis in R (Team, 2018) as well as for visualizing results, several R packages were used including biomaRt (Durinck et al., 2005, Durinck et al., 2009), caTools (Tuszynski, 2014), dplyr (Hadley Wickham, 2017), ggplot2 (Wickham, 2009), ggpubr (Kassambara, 2017), gtools (Gregory R. Warnes, 2015), maxLik (Toomet, 2011), tibble (Kirill Müller, 2018), magrittr (Stefan Milton Bache, 2014), reshape2 (Wickham, 2007), RColorBrewer (Neuwirth, 2014), ComplexHeatmap (Gu et al., 2016), and survival (Therneau, 2015).

Abbreviations

ccRCC: Clear cell renal cell carcinoma; SNV: Single- nucleotide variant; FFPE: Formalin-Fixed Paraffin-Embedded; GATK: Genome Analysis Toolkit; VAF: Variant allele frequency; WES: Whole-exome sequencing; ECDF: Empirical cumulative distribution function; TCGA: The Cancer Genome Atlas Research Network; UMI: Unique molecular identifiers; RNA-seq: RNA-sequencing; panel seq: panel sequencing; TU1: tumor sample 1; TU2: tumor sample 2

Funding

Part of this work was supported by SystemsX.ch IPhD Grant SXPHI0_142005, ERC Synergy Grant 609883, Horizon 2020 SOUND Grant Agreement no 633974, and SNF grant 310030_166391 (to HM).

Author contributions

Conceptualization: A.L.M., J.K., N.B., C.B., P.S., H.M.; Methodology: A.L.M., J.K., N.B.; Software: A.L.M., J.S.; Formal analysis: A.L.M.; Investigation: A.L.M., E.B., N.B., C.B., P.S., H.M.; Resources: N.B., C.B., P.S., H.M.; Writing – original draft: A.L.M.; Writing – review & editing: all authors; Visualization: A.L.M.; Supervision: J.K., N.B., C.B., P.S., H.M.; Project administration: N.B., C.B., P.S., H.M.; Funding Acquisition: N.B., C.B., H.M.

Availability of data and material

The generated read data from the clear cell renal cell carcinoma samples and the matched normal samples have been deposited in the European Nucleotide Archive under accession numbers ERP108328 and ERP108326. GeneAccord is implemented as an R package available at https://github.com/cbg-ethz/GeneAccord and is submitted to Bioconductor.

Ethics approval

This study was approved by the cantonal commission of ethics of Zurich (KEK-ZH-nos. 2013-0629 and 2014-0604). The retrospective use of sequencing data obtained from normal and tumor tissues of 91 ccRCC patients is in accordance with the Swiss Law ("Humanforschungsgesetz"), which, according to Article 34, allows the use of biomaterial and patient data for research purposes without informed consent under certain conditions that include the present cases. Law abidance of this study was reviewed and approved by the ethics commission of the Canton Zurich.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to thank Francesco Marass for support with visualizing the variant allele frequencies.

References

ADAMS, J. C. & TUCKER, R. P. 2000. The thrombospondin type 1 repeat (TSR) superfamily: diverse proteins with related roles in neuronal development. *Dev Dyn,* 218**,** 280-99.

ANDREWS, S. 2010. FastQC: a quality control tool for high throughput sequence data.

APTE, S. S. 2009. A disintegrin-like and metalloprotease (reprolysin-type) with thrombospondin type 1 motif (ADAMTS) superfamily: functions and mechanisms. *J Biol Chem,* 284**,** 31493-7.

ARAI, E., SAKAMOTO, H., ICHIKAWA, H., TOTSUKA, H., CHIKU, S., GOTOH, M., MORI, T., NAKATANI, T., OHNAMI, S., NAKAGAWA, T., FUJIMOTO, H., WANG, L., ABURATANI, H., YOSHIDA, T. & KANAI, Y. 2014. Multilayer-omics analysis of renal cell carcinoma, including the whole exome, methylome and transcriptome. *Int J Cancer,* 135**,** 1330-42.

ARCHETTI, M. 2013. Evolutionary game theory of growth factor production: implications for tumour heterogeneity and resistance to therapies. *Br J Cancer,* 109**,** 1056-62.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

ASSCHERT, J. G., VELLENGA, E., DE JONG, S. & DE VRIES, E. G. 1998. Mutual interactions between p53 and growth factors in cancer. *Anticancer Res,* 18**,** 1713-25.

AXELROD, R., AXELROD, D. E. & PIENTA, K. J. 2006. Evolution of cooperation among tumor cells. *Proceedings of the National Academy of Sciences,* 103**,** 13474-13479.

BAI, Q., LIU, L., LONG, Q., XIA, Y., WANG, J., XU, J. & GUO, J. 2015. Decreased expression of mucin 18 is associated with unfavorable postoperative prognosis in patients with clear cell renal cell carcinoma. *Int J Clin Exp Pathol,* 8**,** 11005-14.

BAMIAS, A., CHORTI, M., DELIVELIOTIS, C., TRAKAS, N., SKOLARIKOS, A., PROTOGEROU, B., LEGAKI, S., TSAKALOU, G., TAMVAKIS, N. & DIMOPOULOS, M. A. 2003. Prognostic significance of CA 125, CD44, and epithelial membrane antigen in renal cell carcinoma. *Urology,* 62**,** 368-73.

BEERENWINKEL, N., GREENMAN, C. D. & LAGERGREN, J. 2016. Computational Cancer Biology: An Evolutionary Perspective. *PLoS Comput Biol,* 12**,** e1004717.

BEERENWINKEL, N., SCHWARZ, R. F., GERSTUNG, M. & MARKOWETZ, F. 2015. Cancer evolution: mathematical models and computational inference. *Syst Biol,* 64**,** e1-25.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300 %@ 0035-9246.

BI, M., ZHAO, S., SAID, J. W., MERINO, M. J., ADENIRAN, A. J., XIE, Z., NAWAF, C. B., CHOI, J., BELLDEGRUN, A. S., PANTUCK, A. J., KLUGER, H. M., BILGUVAR, K., LIFTON, R. P. & SHUCH, B. 2016. Genomic characterization of sarcomatoid transformation in clear cell renal cell carcinoma. *Proc Natl Acad Sci U S A,* 113**,** 2170-5.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics***,** btu170.

BONAVIA, R., INDA, M. M., CAVENEE, W. K. & FURNARI, F. B. 2011. Heterogeneity maintenance in glioblastoma: a social network. *Cancer Res,* 71**,** 4055-60.

BRUMBY, A. M. & RICHARDSON, H. E. 2003. scribble mutants cooperate with oncogenic Ras or Notch to cause neoplastic overgrowth in Drosophila. *EMBO J,* 22**,** 5769-79.

CALBO, J., VAN MONTFORT, E., PROOST, N., VAN DRUNEN, E., BEVERLOO, H. B., MEUWISSEN, R. & BERNS, A. 2011. A functional role for tumor cell heterogeneity in a mouse model of small cell lung cancer. *Cancer Cell,* 19**,** 244-56.

CANCER GENOME ATLAS RESEARCH, N. 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature,* 499**,** 43-9.

CHAPMAN, A., FERNANDEZ DEL AMA, L., FERGUSON, J., KAMARASHEV, J., WELLBROCK, C. & HURLSTONE, A. 2014. Heterogeneous tumor subpopulations cooperate to drive invasion. *Cell Rep,* 8**,** 688-95.

CHEN, F., ZHANG, Y., SENBABAOGLU, Y., CIRIELLO, G., YANG, L., REZNIK, E., SHUCH, B., MICEVIC, G., DE VELASCO, G., SHINBROT, E., NOBLE, M. S., LU, Y., COVINGTON, K. R., XI, L., DRUMMOND, J. A., MUZNY, D., KANG, H., LEE, J., TAMBOLI, P., REUTER, V., SHELLEY, C. S., KAIPPARETTU, B. A., BOTTARO, D. P., GODWIN, A. K., GIBBS, R. A., GETZ, G., KUCHERLAPATI, R., PARK, P. J., SANDER, C., HENSKE, E. P., ZHOU, J. H., KWIATKOWSKI, D. J., HO, T. H., CHOUEIRI, T. K., HSIEH, J. J., AKBANI, R., MILLS, G. B., HAKIMI, A. A., WHEELER, D. A. & CREIGHTON, C. J. 2016. Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma. *Cell Rep,* 14**,** 2476-89.

CHOUEIRI, T. K. & MOTZER, R. J. 2017. Systemic Therapy for Metastatic Renal-Cell Carcinoma. *N Engl J Med,* 376**,** 354-366.

CIBULSKIS, K., LAWRENCE, M. S., CARTER, S. L., SIVACHENKO, A., JAFFE, D., SOUGNEZ, C., GABRIEL, S., MEYERSON, M., LANDER, E. S. & GETZ, G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology,* 31**,** 213-219.

CINGOLANI, P., PATEL, V. M., COON, M., NGUYEN, T., LAND, S. J., RUDEN, D. M. & LU, X. 2012a. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in genetics,* 3.

CINGOLANI, P., PLATTS, A., WANG LE, L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. & RUDEN, D. M. 2012b. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin),* 6**,** 80-92.

CLEARY, A. S., LEONARD, T. L., GESTL, S. A. & GUNTHER, E. J. 2014. Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. *Nature,* 508**,** 113-7.

DAO, P., KIM, Y. A., WOJTOWICZ, D., MADAN, S., SHARAN, R. & PRZYTYCKA, T. M. 2017. BeWith: A Between-Within method to discover relationships between cancer modules via integrated analysis of mutual exclusivity, co-occurrence and functional interactions. *PLoS Comput Biol,* 13**,** e1005695.

DE JONG, J. S., VAN DIEST, P. J., VAN DER VALK, P. & BAAK, J. P. 1998. Expression of growth factors, growth-inhibiting factors, and their receptors in invasive breast cancer. II: Correlations with proliferation and angiogenesis. *J Pathol,* 184**,** 53-7.

DEXTER, D. L., KOWALSKI, H. M., BLAZAR, B. A., FLIGIEL, Z., VOGEL, R. & HEPPNER, G. H. 1978. Heterogeneity of tumor cells from a single mouse mammary tumor. *Cancer Res,* 38**,** 3174-81.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.

DURINCK, S., MOREAU, Y., KASPRZYK, A., DAVIS, S., DE MOOR, B., BRAZMA, A. & HUBER, W. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics,* 21**,** 3439-40.

DURINCK, S., SPELLMAN, P. T., BIRNEY, E. & HUBER, W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc,* 4**,** 1184-91.

EPPING, M. T., MEIJER, L. A., KRIJGSMAN, O., BOS, J. L., PANDOLFI, P. P. & BERNARDS, R. 2011. TSPYL5 suppresses p53 levels and function by physical interaction with USP7. *Nat Cell Biol,* 13**,** 102-8.

FABREGAT, A., JUPE, S., MATTHEWS, L., SIDIROPOULOS, K., GILLESPIE, M., GARAPATI, P., HAW, R., JASSAL, B., KORNINGER, F., MAY, B., MILACIC, M., ROCA, C. D., ROTHFELS, K., SEVILLA, C., SHAMOVSKY, V., SHORSER, S., VARUSAI, T., VITERI, G., WEISER, J., WU, G., STEIN, L., HERMJAKOB, H. & D'EUSTACHIO, P. 2018. The Reactome Pathway Knowledgebase. *Nucleic Acids Res,* 46**,** D649-D655.

FAVERO, F., JOSHI, T., MARQUARD, A. M., BIRKBAK, N. J., KRZYSTANEK, M., LI, Q., SZALLASI, Z. & EKLUND, A. C. 2015. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol,* 26**,** 64-70.

FELDER, M., KAPUR, A., GONZALEZ-BOSQUET, J., HORIBATA, S., HEINTZ, J., ALBRECHT, R., FASS, L., KAUR, J., HU, K., SHOJAEI, H., WHELAN, R. J. & PATANKAR, M. S. 2014. MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress. *Mol Cancer,* 13**,** 129.

FORBES, S. A., BEARE, D., BOUTSELAKIS, H., BAMFORD, S., BINDAL, N., TATE, J., COLE, C. G., WARD, S., DAWSON, E. & PONTING, L. 2016. COSMIC: somatic cancer genetics at high-resolution. *Nucleic acids research,* 45**,** D777-D783 %@ 0305-1048.

FORBES, S. A., BEARE, D., BOUTSELAKIS, H., BAMFORD, S., BINDAL, N., TATE, J., COLE, C. G., WARD, S., DAWSON, E., PONTING, L., STEFANCSIK, R., HARSHA, B., KOK, C. Y., JIA, M., JUBB, H., SONDKA, Z., THOMPSON, S., DE, T. & CAMPBELL, P. J. 2017. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res,* 45**,** D777-D783.

FU, H., LIU, Y., XU, L., CHANG, Y., ZHOU, L., ZHANG, W., YANG, Y. & XU, J. 2016. Low Expression of Mucin-4 Predicts Poor Prognosis in Patients With Clear-Cell Renal Cell Carcinoma. *Medicine (Baltimore),* 95**,** e3225.

GARRAWAY, L. A. & LANDER, E. S. 2013. Lessons from the cancer genome. *Cell,* 153**,** 17-37.

GERLINGER, M., HORSWELL, S., LARKIN, J., ROWAN, A. J., SALM, M. P., VARELA, I., FISHER, R., MCGRANAHAN, N., MATTHEWS, N., SANTOS, C. R., MARTINEZ, P., PHILLIMORE, B., BEGUM, S., RABINOWITZ, A., SPENCER-DENE, B., GULATI, S., BATES, P. A., STAMP, G., PICKERING, L., GORE, M., NICOL, D. L., HAZELL, S., FUTREAL, P. A., STEWART, A. & SWANTON, C. 2014. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet,* 46**,** 225-33.

GERLINGER, M., ROWAN, A. J., HORSWELL, S., LARKIN, J., ENDESFELDER, D., GRONROOS, E., MARTINEZ, P., MATTHEWS, N., STEWART, A., TARPEY, P., VARELA, I., PHILLIMORE, B., BEGUM, S., MCDONALD, N. Q., BUTLER, A., JONES, D., RAINE, K., LATIMER, C., SANTOS, C. R., NOHADANI, M., EKLUND, A. C., SPENCER-DENE, B., CLARK, G., PICKERING, L., STAMP, G., GORE, M., SZALLASI, Z., DOWNWARD, J., FUTREAL, P. A. & SWANTON, C. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med,* 366**,** 883-92.

GERSTUNG, M., BEISEL, C., RECHSTEINER, M., WILD, P., SCHRAML, P., MOCH, H. & BEERENWINKEL, N. 2012. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun,* 3**,** 811.

GIRGIN, C., TARHAN, H., HEKIMGIL, M., SEZER, A. & GUREL, G. 2001. P53 mutations and other prognostic factors of renal cell carcinoma. *Urol Int,* 66**,** 78-83.

GRANKVIST, K., LJUNGBERG, B. & RASMUSON, T. 1997. Evaluation of five glycoprotein tumour markers (CEA, CA-50, CA-19-9, CA-125, CA-15-3) for the prognosis of renal-cell carcinoma. *Int J Cancer,* 74**,** 233-6.

GREGORY R. WARNES, B. B., AND THOMAS LUMLEY 2015. gtools: Various R Programming Tools. *R package version 3.5.0. https://CRAN.R-project.org/package=gtools*.

GU, Z., EILS, R. & SCHLESNER, M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics,* 32**,** 2847-9.

GUROVA, K. V., HILL, J. E., RAZORENOVA, O. V., CHUMAKOV, P. M. & GUDKOV, A. V. 2004. p53 pathway in renal cell carcinoma is repressed by a dominant mechanism. *Cancer Res,* 64**,** 1951-8.

HADLEY WICKHAM, R. F., LIONEL HENRY AND KIRILL MÜLLER 2017. dplyr: A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr.

HARIDAS, D., CHAKRABORTY, S., PONNUSAMY, M. P., LAKSHMANAN, I., RACHAGANI, S., CRUZ, E., KUMAR, S., DAS, S., LELE, S. M., ANDERSON, J. M., WITTEL, U. A., HOLLINGSWORTH, M. A. & BATRA, S. K. 2011. Pathobiological implications of MUC16 expression in pancreatic cancer. *PLoS One,* 6**,** e26839.

HEPPNER, G. H. 1984. Tumor heterogeneity. *Cancer research,* 44**,** 2259-2265.

HOFMANN, A. L., BEHR, J., SINGER, J., KUIPERS, J., BEISEL, C., SCHRAML, P., MOCH, H. & BEERENWINKEL, N. 2017. Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. *BMC Bioinformatics,* 18**,** 8.

HOLLINGSWORTH, M. A. & SWANSON, B. J. 2004. Mucins in cancer: protection and control of the cell surface. *Nat Rev Cancer,* 4**,** 45-60.

IGNATIADIS, N., KLAUS, B., ZAUGG, J. B. & HUBER, W. 2016. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat Methods,* 13**,** 577-80.

IRMINGER-FINGER, I., LEUNG, W. C., LI, J., DUBOIS-DAUPHIN, M., HARB, J., FEKI, A., JEFFORD, C. E., SORIANO, J. V., JACONI, M., MONTESANO, R. & KRAUSE, K. H. 2001. Identification of BARD1 as mediator between proapoptotic stress and p53-dependent apoptosis. *Mol Cell,* 8**,** 1255-66.

JIANG, J., YANG, E. S., JIANG, G., NOWSHEEN, S., WANG, H., WANG, T., WANG, Y., BILLHEIMER, D., CHAKRAVARTHY, A. B., BROWN, M., HAFFTY, B. & XIA, F. 2011. p53-dependent BRCA1 nuclear export controls cellular susceptibility to DNA damage. *Cancer Res,* 71**,** 5546-57.

JIANG, T., SHAHAM, U., PARISI, F., HALABAN, R., SAFONOV, A., KLUGER, H., WEISSMAN, S., CHANG, J. & KLUGER, Y. 2016. Methods for detecting co-mutated pathways in cancer samples to inform treatment selection. *bioRxiv*, 082552.

KALAS, W., YU, J. L., MILSOM, C., ROSENFELD, J., BENEZRA, R., BORNSTEIN, P. & RAK, J. 2005. Oncogenes and Angiogenesis: down-regulation of thrombospondin-1 in normal fibroblasts exposed to factors from cancer cells harboring mutant ras. *Cancer Res,* 65, 8878-86.

KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res,* 28, 27-30.

KASSAMBARA, A. 2017. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.1.6. https://CRAN.R-project.org/package=ggpubr.

KIRILL MÜLLER, H. W. 2018. tibble: Simple Data Frames. R package version 1.4.2. https://CRAN.R-project.org/package=tibble.

KOBOLDT, D. C., ZHANG, Q., LARSON, D. E., SHEN, D., MCLELLAN, M. D., LIN, L., MILLER, C. A., MARDIS, E. R., DING, L. & WILSON, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research,* 22, 568-576.

KOCKAN, C., HACH, F., SARRAFI, I., BELL, R. H., MCCONEGHY, B., BEJA, K., HAEGERT, A., WYATT, A. W., VOLIK, S. V., CHI, K. N., COLLINS, C. C. & SAHINALP, S. C. 2017. SiNVICT: ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA. *Bioinformatics,* 33, 26-34.

KONEN, J., SUMMERBELL, E., DWIVEDI, B., GALIOR, K., HOU, Y., RUSNAK, L., CHEN, A., SALTZ, J., ZHOU, W., BOISE, L. H., VERTINO, P., COOPER, L., SALAITA, K., KOWALSKI, J. & MARCUS, A. I. 2017. Image-guided genomics of phenotypically heterogeneous populations reveals vascular signalling during symbiotic collective cancer invasion. *Nat Commun,* 8, 15078.

KU, B. M., KIM, D. S., KIM, K. H., YOO, B. C., KIM, S. H., GONG, Y. D. & KIM, S. Y. 2013. Transglutaminase 2 inhibition found to induce p53 mediated apoptosis in renal cell carcinoma. *FASEB J,* 27, 3487-95.

KUIPERS, J., JAHN, K. & BEERENWINKEL, N. 2017. Advances in understanding tumour evolution through single-cell sequencing. *Biochim Biophys Acta,* 1867, 127-138.

LAI, Z., MARKOVETS, A., AHDESMAKI, M., CHAPMAN, B., HOFMANN, O., MCEWEN, R., JOHNSON, J., DOUGHERTY, B., BARRETT, J. C. & DRY, J. R. 2016. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res,* 44, e108.

LAKSHMANAN, I., SALFITY, S., SESHACHARYULU, P., RACHAGANI, S., THOMAS, A., DAS, S., MAJHI, P. D., NIMMAKAYALA, R. K., VENGOJI, R., LELE, S. M., PONNUSAMY, M. P., BATRA, S. K. & GANTI, A. K. 2017. MUC16 Regulates TSPYL5 for Lung

Cancer Cell Growth and Chemoresistance by Suppressing p53. *Clin Cancer Res,* 23, 3906-3917.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25, 1754-60.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009a. The sequence alignment/map format and SAMtools. *Bioinformatics,* 25, 2078-2079.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009b. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25, 2078-9.

LIAO, Y., SMYTH, G. K. & SHI, W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics,* 30, 923-30.

LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol,* 15, 550.

LUCARELLI, G., DITONNO, P., BETTOCCHI, C., VAVALLO, A., RUTIGLIANO, M., GALLEGGIANTE, V., LAROCCA, A. M., CASTELLANO, G., GESUALDO, L., GRANDALIANO, G., SELVAGGI, F. P. & BATTAGLIA, M. 2014. Diagnostic and prognostic role of preoperative circulating CA 15-3, CA 125, and beta-2 microglobulin in renal cell carcinoma. *Dis Markers,* 2014, 689795.

MALEY, C. C., GALIPEAU, P. C., FINLEY, J. C., WONGSURAWAT, V. J., LI, X., SANCHEZ, C. A., PAULSON, T. G., BLOUNT, P. L., RISQUES, R. A., RABINOVITCH, P. S. & REID, B. J. 2006. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet,* 38, 468-73.

MARASS, F., MOULIERE, F., YUAN, K., ROSENFELD, N. & MARKOWETZ, F. 2016. A phylogenetic latent feature model for clonal deconvolution. *The Annals of Applied Statistics,* 10, 2377-2404 %@ 1932-6157.

MARTINEZ, P., BIRKBAK, N. J., GERLINGER, M., MCGRANAHAN, N., BURRELL, R. A., ROWAN, A. J., JOSHI, T., FISHER, R., LARKIN, J., SZALLASI, Z. & SWANTON, C. 2013. Parallel evolution of tumour subclones mimics diversity between tumours. *J Pathol,* 230, 356-64.

MARUSYK, A. & POLYAK, K. 2010. Tumor heterogeneity: causes and consequences. *Biochim Biophys Acta,* 1805, 105-17.

MARUSYK, A., TABASSUM, D. P., ALTROCK, P. M., ALMENDRO, V., MICHOR, F. & POLYAK, K. 2014. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature,* 514, 54-8.

MATEO, F., MECA-CORTES, O., CELIA-TERRASSA, T., FERNANDEZ, Y., ABASOLO, I., SANCHEZ-CID, L., BERMUDO, R., SAGASTA, A., RODRIGUEZ-CARUNCHIO, L., PONS, M., CANOVAS, V., MARIN-AGUILERA, M., MENGUAL, L., ALCARAZ, A., SCHWARTZ, S., JR., MELLADO, B., AGUILERA, K. Y., BREKKEN, R., FERNANDEZ, P. L., PACIUCCI, R. & THOMSON, T. M. 2014. SPARC mediates metastatic cooperation between CSC and non-CSC prostate cancer cell subpopulations. *Mol Cancer,* 13**,** 237.

MATTHEWS, L., GOPINATH, G., GILLESPIE, M., CAUDY, M., CROFT, D., DE BONO, B., GARAPATI, P., HEMISH, J., HERMJAKOB, H., JASSAL, B., KANAPIN, A., LEWIS, S., MAHAJAN, S., MAY, B., SCHMIDT, E., VASTRIK, I., WU, G., BIRNEY, E., STEIN, L. & D'EUSTACHIO, P. 2009. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res,* 37**,** D619-22.

MILLER, B. E., MACHEMER, T., LEHOTAN, M. & HEPPNER, G. H. 1991. Tumor subpopulation interactions affecting melphalan sensitivity in palpable mouse mammary tumors. *Cancer Res,* 51**,** 4378-87.

MOCH, H., CUBILLA, A. L., HUMPHREY, P. A., REUTER, V. E. & ULBRIGHT, T. M. 2016. The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs-Part A: Renal, Penile, and Testicular Tumours. *Eur Urol,* 70**,** 93-105.

MOREMEN, K. W., TIEMEYER, M. & NAIRN, A. V. 2012. Vertebrate protein glycosylation: diversity, synthesis and function. *Nat Rev Mol Cell Biol,* 13**,** 448-62.

MUIR, P., LI, S., LOU, S., WANG, D., SPAKOWICZ, D. J., SALICHOS, L., ZHANG, J., WEINSTOCK, G. M., ISAACS, F., ROZOWSKY, J. & GERSTEIN, M. 2016. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol,* 17**,** 53.

NEUWIRTH, E. 2014. RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer.

NGUYENHOANG, S., LIU, Y., XU, L., CHANG, Y., ZHOU, L., LIU, Z., LIN, Z. & XU, J. 2016. High mucin-7 expression is an independent predictor of adverse clinical outcomes in patients with clear-cell renal cell carcinoma. *Tumour Biol,* 37**,** 15193-15201.

NISHIKAWA, H., WU, W., KOIKE, A., KOJIMA, R., GOMI, H., FUKUDA, M. & OHTA, T. 2009. BRCA1-associated protein 1 interferes with BRCA1/BARD1 RING heterodimer activity. *Cancer Res,* 69**,** 111-9.

NOWELL, P. C. 1976. The clonal evolution of tumor cell populations. *Science,* 194**,** 23-8.

OH, E., CHOI, Y. L., KWON, M. J., KIM, R. N., KIM, Y. J., SONG, J. Y., JUNG, K. S. & SHIN, Y. K. 2015. Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples. *PLoS One,* 10**,** e0144162.

OKONECHNIKOV, K., CONESA, A. & GARCIA-ALCALDE, F. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics,* 32**,** 292-4.

PENA-LLOPIS, S., CHRISTIE, A., XIE, X. J. & BRUGAROLAS, J. 2013. Cooperation and antagonism among cancer genes: the renal cancer paradigm. *Cancer Res,* 73**,** 4173-9.

PENTICUFF, J. C. & KYPRIANOU, N. 2015. Therapeutic challenges in renal cell carcinoma. *Am J Clin Exp Urol,* 3**,** 77-90.

PICARD. *Picard. Accessed 31 August 2017.* [Online]. Available: http://broadinstitute.github.io/picard/ [Accessed].

RICKETTS, C. J., DE CUBAS, A. A., FAN, H., SMITH, C. C., LANG, M., REZNIK, E., BOWLBY, R., GIBB, E. A., AKBANI, R., BEROUKHIM, R., BOTTARO, D. P., CHOUEIRI, T. K., GIBBS, R. A., GODWIN, A. K., HAAKE, S., HAKIMI, A. A., HENSKE, E. P., HSIEH, J. J., HO, T. H., KANCHI, R. S., KRISHNAN, B., KWAITKOWSKI, D. J., LUI, W., MERINO, M. J., MILLS, G. B., MYERS, J., NICKERSON, M. L., REUTER, V. E., SCHMIDT, L. S., SHELLEY, C. S., SHEN, H., SHUCH, B., SIGNORETTI, S., SRINIVASAN, R., TAMBOLI, P., THOMAS, G., VINCENT, B. G., VOCKE, C. D., WHEELER, D. A., YANG, L., KIM, W. T., ROBERTSON, A. G., CANCER GENOME ATLAS RESEARCH, N., SPELLMAN, P. T., RATHMELL, W. K. & LINEHAN, W. M. 2018. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Rep,* 23**,** 313-326 e5.

ROTH, A., DING, J., MORIN, R., CRISAN, A., HA, G., GIULIANY, R., BASHASHATI, A., HIRST, M., TURASHVILI, G., OLOUMI, A., MARRA, M. A., APARICIO, S. & SHAH, S. P. 2012. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics,* 28**,** 907-13.

SATO, Y., YOSHIZATO, T., SHIRAISHI, Y., MAEKAWA, S., OKUNO, Y., KAMURA, T., SHIMAMURA, T., SATO-OTSUBO, A., NAGAE, G., SUZUKI, H., NAGATA, Y., YOSHIDA, K., KON, A., SUZUKI, Y., CHIBA, K., TANAKA, H., NIIDA, A., FUJIMOTO, A., TSUNODA, T., MORIKAWA, T., MAEDA, D., KUME, H., SUGANO, S., FUKAYAMA, M., ABURATANI, H., SANADA, M., MIYANO, S., HOMMA, Y. & OGAWA, S. 2013. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat Genet,* 45**,** 860-7.

SAUNDERS, C. T., WONG, W. S., SWAMY, S., BECQ, J., MURRAY, L. J. & CHEETHAM, R. K. 2012. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics,* 28**,** 1811-7.

SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res,* 29**,** 308-11.

SIM, N. L., KUMAR, P., HU, J., HENIKOFF, S., SCHNEIDER, G. & NG, P. C. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res,* 40, W452-7.

SINGER, J., RUSCHEWEYH, H. J., HOFMANN, A. L., THURNHERR, T., SINGER, F., TOUSSAINT, N. C., NG, C. K. Y., PISCUOGLIO, S., BEISEL, C., CHRISTOFORI, G., DUMMER, R., HALL, M. N., KREK, W., LEVESQUE, M., MANZ, M. G., MOCH, H., PAPASSOTIROPOULOS, A., STEKHOVEN, D. J., WILD, P., WUST, T., RINN, B. & BEERENWINKEL, N. 2017. NGS-pipe: a flexible, easily extendable, and highly configurable framework for NGS analysis. *Bioinformatics*.

SMITH, T., HEGER, A. & SUDBERY, I. 2017. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research,* 27, 491-499 %@ 1088-9051.

STEFAN MILTON BACHE, H. W. 2014. magrittr: A Forward-Pipe Operator for R. R package version 1.5. https://CRAN.R-project.org/package=magrittr.

STURM, M., SCHROEDER, C. & BAUER, P. 2016. SeqPurge: highly-sensitive adapter trimming for paired-end NGS data. *BMC Bioinformatics,* 17, 208.

TABASSUM, D. P. & POLYAK, K. 2015. Tumorigenesis: it takes a village. *Nat Rev Cancer,* 15, 473-83.

TEAM, R. C. 2018. R: A Language and Environment for Statistical Computing. https://www.R-project.org/. *R Foundation for Statistical Computing, Vienna, Austria*.

THE GENE ONTOLOGY, C. 2017. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res,* 45, D331-D338.

THE UNIPROT, C. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res,* 45, D158-D169.

THERIAULT, C., PINARD, M., COMAMALA, M., MIGNEAULT, M., BEAUDIN, J., MATTE, I., BOIVIN, M., PICHE, A. & RANCOURT, C. 2011. MUC16 (CA125) regulates epithelial ovarian cancer cell growth, tumorigenesis and metastasis. *Gynecol Oncol,* 121, 434-43.

THERNEAU, T. M. 2015. A Package for Survival Analysis in S. version 2.38, URL: https://CRAN.R-project.org/package=survival.

TOOMET, A. H. A. O. 2011. maxLik: A package for maximum likelihood estimation in R. *Computational Statistics,* 26, 443-458.

TUSZYNSKI, J. 2014. caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc. R package version 1.17.1. https://CRAN.R-project.org/package=caTools.

UHLIROVA, M., JASPER, H. & BOHMANN, D. 2005. Non-cell-autonomous induction of tissue overgrowth by JNK/Ras cooperation in a Drosophila tumor model. *Proc Natl Acad Sci U S A,* 102, 13123-8.

VAN DER AUWERA, G. A., CARNEIRO, M. O., HARTL, C., POPLIN, R., DEL ANGEL, G., LEVY-MOONSHINE, A., JORDAN, T., SHAKIR, K., ROAZEN, D., THIBAULT, J., BANKS, E., GARIMELLA, K. V., ALTSHULER, D., GABRIEL, S. & DEPRISTO, M. A. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics,* 43**,** 11 10 1-33.

WANG, J., VASAIKAR, S., SHI, Z., GREER, M. & ZHANG, B. 2017. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res,* 45**,** W130-W137.

WANG, X., FU, A. Q., MCNERNEY, M. E. & WHITE, K. P. 2014. Widespread genetic epistasis among cancer genes. *Nat Commun,* 5**,** 4828.

WANG, Y. & NAVIN, N. E. 2015. Advances and applications of single-cell sequencing technologies. *Mol Cell,* 58**,** 598-609.

WICKHAM, H. 2007. Reshaping Data with the reshape Package. *Journal of Statistical Software,* 21**,** 1-20.

WICKHAM, H. 2009. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*.

WONG, S. Q., LI, J., TAN, A. Y., VEDURURU, R., PANG, J. M., DO, H., ELLUL, J., DOIG, K., BELL, A., MACARTHUR, G. A., FOX, S. B., THOMAS, D. M., FELLOWES, A., PARISOT, J. P., DOBROVIC, A. & COHORT, C. 2014. Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. *BMC Med Genomics,* 7**,** 23.

XU, X., HOU, Y., YIN, X., BAO, L., TANG, A., SONG, L., LI, F., TSANG, S., WU, K., WU, H., HE, W., ZENG, L., XING, M., WU, R., JIANG, H., LIU, X., CAO, D., GUO, G., HU, X., GUI, Y., LI, Z., XIE, W., SUN, X., SHI, M., CAI, Z., WANG, B., ZHONG, M., LI, J., LU, Z., GU, N., ZHANG, X., GOODMAN, L., BOLUND, L., WANG, J., YANG, H., KRISTIANSEN, K., DEAN, M., LI, Y. & WANG, J. 2012. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell,* 148**,** 886-95.

YATES, L. R., GERSTUNG, M., KNAPPSKOG, S., DESMEDT, C., GUNDEM, G., VAN LOO, P., AAS, T., ALEXANDROV, L. B., LARSIMONT, D., DAVIES, H., LI, Y., JU, Y. S., RAMAKRISHNA, M., HAUGLAND, H. K., LILLENG, P. K., NIK-ZAINAL, S., MCLAREN, S., BUTLER, A., MARTIN, S., GLODZIK, D., MENZIES, A., RAINE, K., HINTON, J., JONES, D., MUDIE, L. J., JIANG, B., VINCENT, D., GREENE-COLOZZI, A., ADNET, P. Y., FATIMA, A., MAETENS, M., IGNATIADIS, M., STRATTON, M. R., SOTIRIOU, C., RICHARDSON, A. L., LONNING, P. E., WEDGE, D. C. & CAMPBELL, P. J. 2015. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med,* 21**,** 751-9.

YOST, S. E., SMITH, E. N., SCHWAB, R. B., BAO, L., JUNG, H., WANG, X., VOEST, E., PIERCE, J. P., MESSER, K., PARKER, B. A., HARISMENDY, O. & FRAZER, K. A. 2012.

Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res,* 40**,** e107.

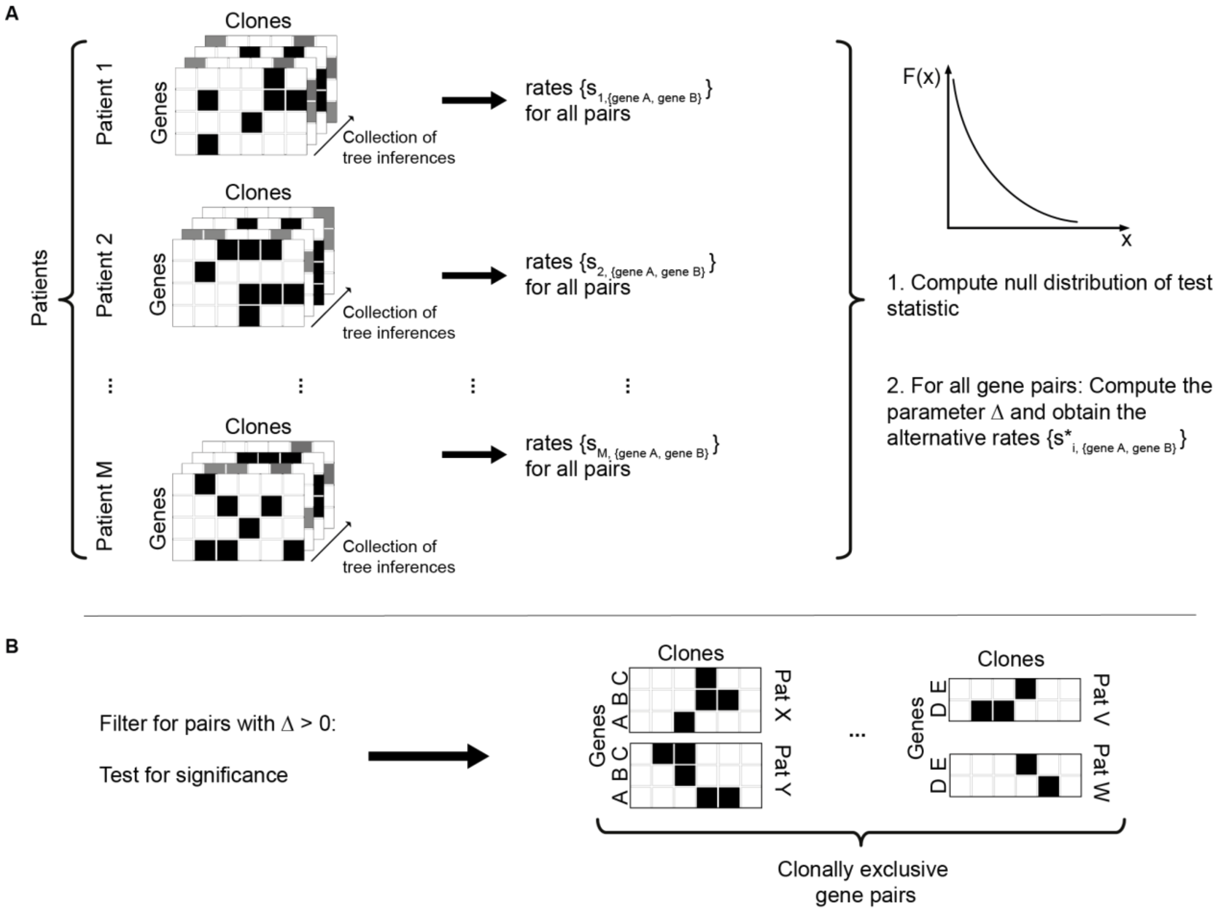
ZHANG, H., LIU, Y., XIE, H., LIU, W., FU, Q., YAO, D., XU, J. & GU, J. 2017. High mucin 5AC expression predicts adverse postoperative recurrence and survival of patients with clear-cell renal cell carcinoma. *Oncotarget,* 8**,** 59777-59790.

ZHANG, J., WU, L. Y., ZHANG, X. S. & ZHANG, S. 2014. Discovery of co-occurring driver pathways in cancer. *BMC Bioinformatics,* 15**,** 271.

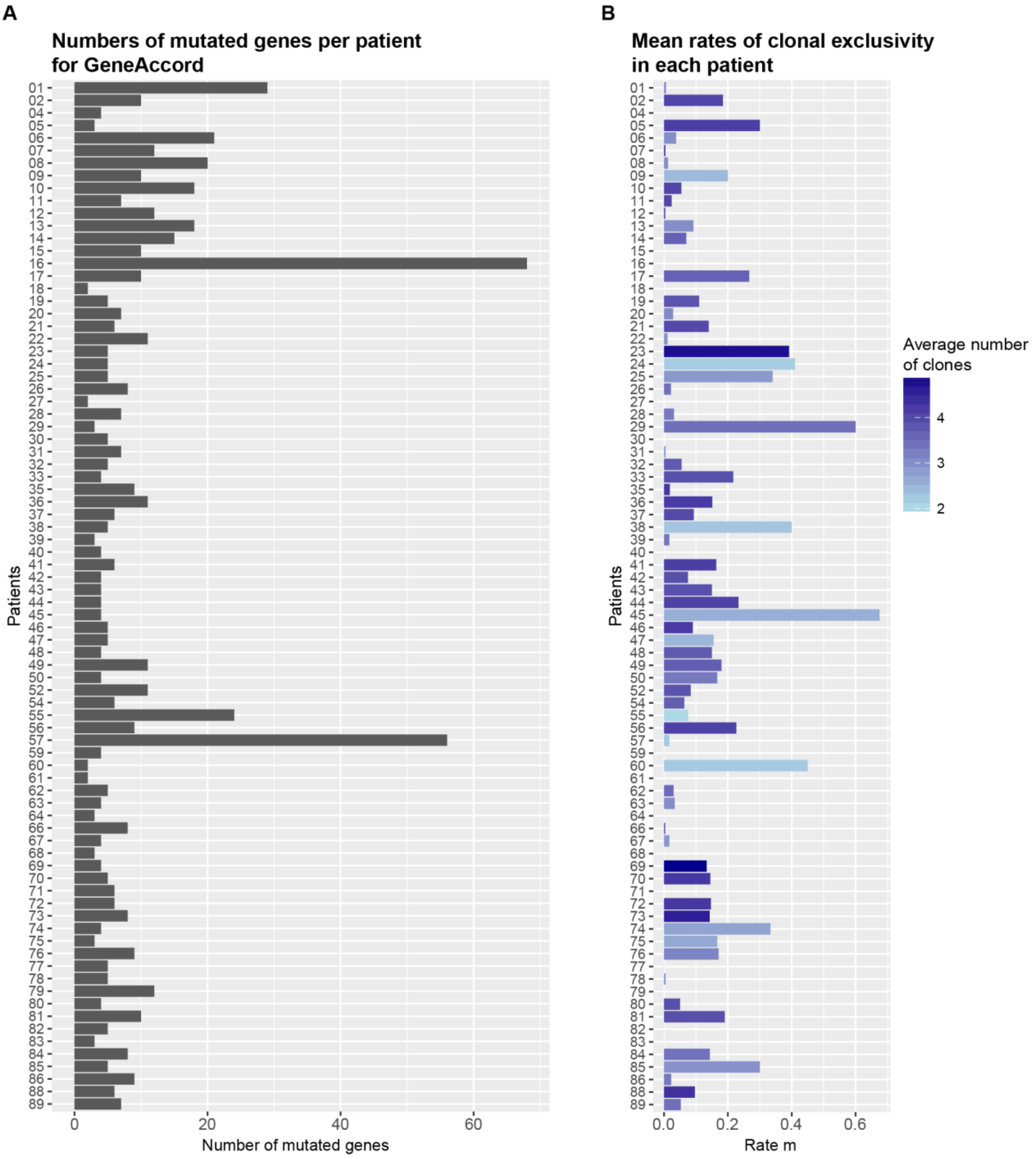
ZHANG, M., TSIMELZON, A., CHANG, C.-H., FAN, C., WOLFF, A., PEROU, C. M., HILSENBECK, S. G. & ROSEN, J. M. 2015. Intratumoral heterogeneity in a Trp53-null mouse model of human breast cancer. *Cancer discovery,* 5**,** 520-533 %@ 2159-8274.

ZHAO, M., GAO, F. H., WANG, J. Y., LIU, F., YUAN, H. H., ZHANG, W. Y. & JIANG, B. 2011. JAK2/STAT3 signaling pathway activation mediates tumor angiogenesis by upregulation of VEGF and bFGF in non-small-cell lung cancer. *Lung Cancer,* 73**,** 366-74.

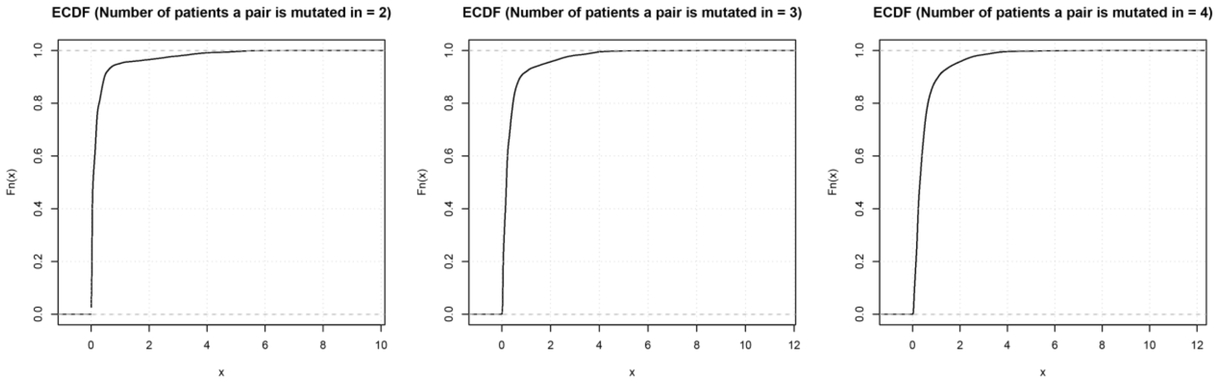
Supplementary Material


A
Patients
Patient 1
Clones
Genes
Collection of tree inferences
rates {$s_{1,\{\text{gene A, gene B}\}}$}
for all pairs
Patient 2
Clones
Genes
Collection of tree inferences
rates {$s_{2,\{\text{gene A, gene B}\}}$}
for all pairs
Patient M
Clones
Genes
Collection of tree inferences
rates {$s_{M,\{\text{gene A, gene B}\}}$}
for all pairs
F(x)
x
1. Compute null distribution of test statistic
2. For all gene pairs: Compute the parameter Δ and obtain the alternative rates {$s^*_{i,\{\text{gene A, gene B}\}}$}
B
Filter for pairs with Δ > 0:
Test for significance
Clones
Genes
A B C
Pat X
A B C
Pat Y
...
Clones
Genes
D E
Pat V
D E
Pat W
Clonally exclusive gene pairs

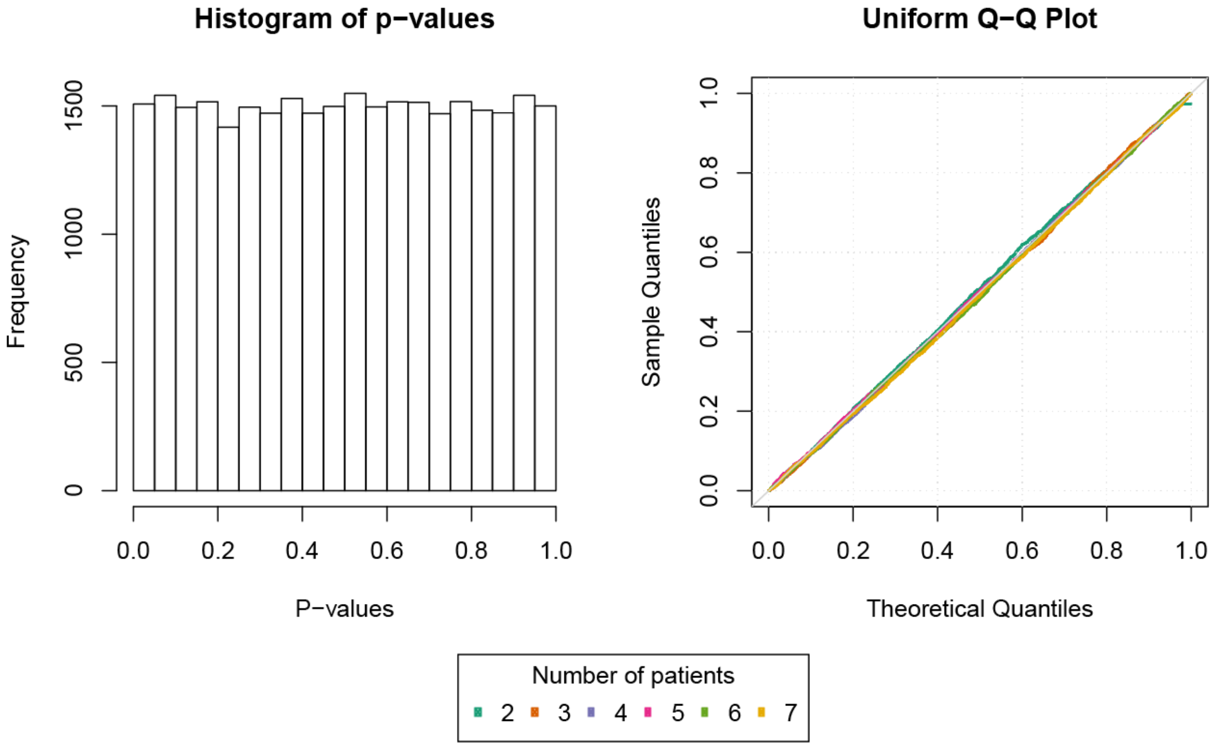
Supplementary Figure S1. GeneAccord algorithm. Schematic overview of the procedure of GeneAccord to detect clonally exclusive pairs of mutated genes. **(A)** The input data are the mutated gene-clone assignments from a cohort of cancer patients and from a collection of trees to take into account uncertainty in the phylogenetic tree inference. The first step is to compute the overall rates of clonal exclusivity for all gene pairs for each patient separately. These rates reflect the expected prevalence of the clonal exclusivity pattern in each patient and for each gene or pathway pair. From these, the distribution of the test statistic under the null hypothesis of the likelihood ratio test is computed. For each gene pair, the parameter Δ is then estimated using maximum likelihood. **(B)** A positive parameter Δ indicates that the pair tends to be mutated in different clones more often than expected. We filter for such pairs and test them for significance.



Supplementary Figure S2. (A) Numbers of mutated genes per patient that were used as input for GeneAccord. The mutations were assigned to clones through Cloe, and were mapped to genes afterwards. For GeneAccord, only mutations with a potential impact were considered, that is, for instance, synonymous or intronic mutations were filtered out (Methods section). Seven patients had only zero or one gene-to-clone assignment (patients 03, 34, 51, 53, 58, 65, 87), and therefore the GeneAccord analysis was done with the remaining 82 patients. **(B)** Average rates of clonal exclusivity per patient. The average rate and number of clones with at least one nonsynonymous mutation are computed from all 20 trees. The average number of clones ranges from 2.0 to 4.95. The rate of clonal exclusivity of a patient is the fraction of gene pairs that are in different branches of the phylogenetic tree. For instance, if the rate is zero, it means that the tree was linear without any branches, and that all genes assigned to the clones are in the same lineage, as is the case for patient 30.



Supplementary Figure S3: The empirical cumulative distribution function (ECDF) of the test statistic of the likelihood ratio test under the null hypothesis. The test statistic was computed 100,000 times for each of the cases of a pair being mutated in two to four patients. The pairs were sampled randomly from the patients, and the average rate of clonal exclusivity of each patient was distorted with a beta distribution in order to obtain a smooth ECDF. The distribution of the test statistic under the null hypothesis is computed separately for each number of patients n that a pair can be mutated in. When comparing an observed test statistic to the null distribution, the number of patients in which the pair is mutated determines which of the previously generated null distributions is used.



Supplementary Figure S4: The p-value distribution under the null hypothesis. Displayed is the histogram in the left panel and the Q-Q plot in the right panel. To simulate a pair under the null hypothesis, patients were randomly sampled. From each of these patients, the quantity $\frac{NSE_{i,\{A,B\}}}{NR_{i,\{A,B\}}}$ and the mean clonal exclusivity rate m_i were used to compute the overall rate of clonal exclusivity $s_{i,\{A,B\}}$, where the rate m_i is distorted by a beta distribution. Subsequently, the test statistic and p-value are computed. We distinguish between the cases, where pairs are mutated in 2, 3, …, 7 patients (different colors in the right panel). For each of these, a total of 5,000 pairs was simulated.

Supplementary Table S1. The variant categories from Fig. 3 and the corresponding annotations from SnpEff

Variant category in Fig. 3	SnpEff annotation
Start lost	start_lost&conservative_inframe_deletion, start_lost
Stop lost	stop_lost, frameshift_variant&stop_lost
Stop gained	stop_gained, stop_gained&disruptive_inframe_deletion, stop_gained&splice_region_variant
Start gained	5_prime_UTR_premature_start_codon_gain_variant
Frameshift indel	frameshift_variant, frameshift_variant&splice_region_variant, frameshift_variant&splice_donor_variant&splice_region_variant&intron_variant
Inframe indel	disruptive_inframe_deletion, disruptive_inframe_insertion, inframe_insertion, inframe_deletion, conservative_inframe_insertion, conservative_inframe_deletion
Missense	missense_variant&splice_region_variant, missense_variant
Splice site	splice_donor_variant&splice_region_variant&3_prime_UTR_variant&intron_variant, splice_region_variant, splice_region_variant&synonymous_variant splice_region_variant&non_coding_exon_variant, splice_region_variant&intron_variant, splice_region_variant&non_coding_transcript_exon_variant, splice_acceptor_variant&intron_variant, splice_donor_variant&intron_variant
Protein interacion loci	structural_interaction_variant, protein_protein_contact
5 prime UTR	5_prime_UTR_variant, 5_prime_UTR_truncation&exon_loss_variant
3 prime UTR	3_prime_UTR_variant
Synonymous	synonymous_variant
Ignored	intron_variant
Sequence feature	sequence_feature
Ignored	non_coding_exon_variant, non_coding_transcript_variant, non_coding_transcript_exon_variant
Ignored	intergenic_region, upstream_gene_variant, downstream_gene_variant, TF_binding_site_variant(no gene associated to this annotation; it is a "motif annotation"), intragenic_variant

The variants were annotated to genes with the program SnpEff (Cingolani et al., 2012b). It concatenates annotations with '&' in case several effects are true for the same mutation and transcript. The same gene can be hit by several mutations, or the same mutation can have several annotations for different transcripts. Therefore, in order to visualize the mutations in a heatmap like in Fig. 3, an order of importance needed to be defined, which is the following starting with highest priority: 'Start lost', 'Stop lost', 'Stop gained', 'Start gained', 'Frameshift indel', 'Inframe indel', 'Missense', 'Splice site', 'Protein interaction loci', '5 prime UTR', '3 prime UTR', 'Synonymous', 'Sequence feature'. This means that, if a gene was hit by two mutations, where one is synonymous and the other one is a missense mutation, only the missense mutation will be shown in the heatmap.

Supplementary Table S2. The 16 gene pairs with a positive parameter delta of the clonal exclusivity test applied to the 89 ccRCC

Gene A	Gene B	P-value	Adjusted p-value	Mutated in (rate)	Clonally exclusive in
MUC16	*TP53*	< 0.01	0.01	Patient 05 (0.3); Patient 81 (0.19)	Patient 05; Patient 81
BAP1	*TP53*	0.01	0.04	Patient 30 (0); Patient 84 (0.14)	Patient 84
CACNA1S	*MAP3K3*	0.01	0.04	Patient 16 (0); Patient 36 (0.15)	Patient 36
CSMD1	*MAP3K3*	0.01	0.04	Patient 16 (0); Patient 36 (0.15)	Patient 36
USP36	*CNOT1*	0.02	< 0.05	Patient 16 (0); Patient 49 (0.18)	Patient 49

PBRM1	*BAP1*	0.02	0.05	Patient 30 (0);
				Patient 35 (0.02);
				Patient 84 (0.14)
					Patient 84

MUC16	*TRO*	0.03	0.05	Patient 26 (0.02); Patient 56 (0.23)	Patient 56
DNAH9	*MUC16*	0.04	0.05	Patient 06 (0.04); Patient 56 (0.23)	Patient 56
DNAH9	*WNT5B*	0.04	0.05	Patient 06 (0.04); Patient 17 (0.27)	Patient 17
DNAH9	*UNC79*	0.04	0.05	Patient 06 (0.04); Patient 17 (0.27)	Patient 17
WNT5B	*RARA*	0.04	0.05	Patient 06 (0.04); Patient 17 (0.27)	Patient 17
RARA	*UNC79*	0.04	0.05	Patient 06 (0.04); Patient 17 (0.27)	Patient 17
PBRM1	*CSMD3*	0.04	0.05	Patient 73 (0.14); Patient 84 (0.14)	Patient 84
CSMD3	*ARHGAP35*	0.05	0.06	Patient 02 (0.18); Patient 84 (0.14)	Patient 02
USP36	*YLPM1*	0.06	0.06	Patient 16 (0); Patient 45 (0.68)	Patient 45

SETD2	*PIK3CA*	0.07	0.07	Patient 10 (0.05);
				Patient 74 (0.33);
				Patient 82 (0)
					Patient 74

The 16 gene pairs from the panel sequencing data set whose estimate of the parameter delta is greater zero. The column ‘Mutated in (rate)’ lists the patients in which the pair was mutated in. The average rate of clonal exclusivity of the respective patient is shown in parenthesis. The column ‘Clonally exclusive in’ lists the patients in which the pair was also clonally exclusive in the majority of the trees.

Supplementary Table S3. The ten most striking pathway pairs with a positive parameter delta of the clonal exclusivity test applied to the 16 ccRCC on the pathway level

Pathway A	Pathway B	P-value	Adjusted p-value	Affected in (rate)	Clonally exclusive in
Defective B3GALTL causes Peters-plus syndrome (PpS)	Major pathway of rRNA processing in the nucleolus and cytosol	0	0	Patient 08 (0.012); Patient 14 (0.008)	Patient 08; Patient 14
O-glycosylation of TSR domain-containing proteins	Major pathway of rRNA processing in the nucleolus and cytosol	0	0	Patient 08 (0.012); Patient 14 (0.008)	Patient 08; Patient 14
Cobalamin (Cbl, vitamin B12) transport and metabolism	PRDM12	0	0.02	Patient 04 (0.003); Patient 15 (0.008)	Patient 15
COPI-dependent Golgi-to-ER retrograde traffic	Major pathway of rRNA processing in the nucleolus and cytosol	0.001	0.026	Patient 03 (0.029); Patient 14 (0.008)	Patient 14
Kinesins	Major pathway of rRNA processing in the nucleolus and cytosol	0.001	0.026	Patient 03 (0.029); Patient 14 (0.008)	Patient 14
Major pathway of rRNA processing in the nucleolus and cytosol	MHC class II antigen presentation	0.001	0.026	Patient 03 (0.029); Patient 14 (0.008)	Patient 14

Activation of Matrix Metalloproteinases	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	0.001	0.026	Patient 01 (0.015);
				Patient 10 (0.017);
				Patient 15 (0.008)
					Patient 15

Ion homeostasis	NCAM1 interactions	0.002	0.026	Patient 02 (0.003);
				Patient 06 (0.047);
				Patient 15 (0.008)
					Patient 02

E3 ubiquitin ligases ubiquitinate target proteins	MUC2	0.002	0.026	Patient 06 (0.047); Patient 13 (0.065)	Patient 13
Laminin interactions	MUC2	0.005	0.026	Patient 06 (0.047); Patient 13 (0.065)	Patient 13

The ten most striking pathway pairs from the WES data set whose estimate of the parameter delta is greater zero. The column ‘Affected in (rate)’ lists the patients in which the pair was affected in. The average rate of clonal exclusivity of the respective patient is shown in parenthesis. The column ‘Clonally exclusive in’ lists the patients in which the pair was also clonally exclusive in the majority of the tree inference runs. In case a gene could not be mapped to a pathway, as is the case for PRDM12 and MUC2, the gene identifiers were retained.