Inferring clonal composition from multiple tumor biopsies

Manica, Matteo ; Kim, Hyunjae Ryan ; Mathis, Roland ; Chouvarine, Philippe ; Rutishauser, Dorothea ; De Vargas Roditi, Laura ; Szalai, Bence ; Wagner, Ulrich ; Oehl, Kathrin ; Saba, Karim ; Pati, Arati ; Saez-Rodriguez, Julio ; Roy, Angshumoy ; Parsons, Donald W ; Wild, Peter J ; Martínez, María Rodríguez ; Sumazin, Pavel

Abstract: Knowledge about the clonal evolution of a tumor can help to interpret the function of its genetic alterations by identifying initiating events and events that contribute to the selective advantage of proliferative, metastatic, and drug-resistant subclones. Clonal evolution can be reconstructed from estimates of the relative abundance (frequency) of subclone-specific alterations in tumor biopsies, which, in turn, inform on its composition. However, estimating these frequencies is complicated by the high genetic instability that characterizes many cancers. Models for genetic instability suggest that copy number alterations (CNAs) can influence mutation-frequency estimates and thus impede efforts to reconstruct tumor phylogenies. Our analysis suggested that accurate mutation frequency estimates require accounting for CNAs-a challenging endeavour using the genetic profile of a single tumor biopsy. Instead, we propose an optimization algorithm, Chimæra, to account for the effects of CNAs using profiles of multiple biopsies per tumor. Analyses of simulated data and tumor profiles suggested that Chimæra estimates are consistently more accurate than those of previously proposed methods and resulted in improved phylogeny reconstructions and subclone characterizations. Our analyses inferred recurrent initiating mutations in hepatocellular carcinomas, resolved the clonal composition of Wilms' tumors, and characterized the acquisition of mutations in drug-resistant prostate cancers.

ARTICLE OPEN
Check for updates

Inferring clonal composition from multiple tumor biopsies

Matteo Manica [1,2], Hyunjae Ryan Kim[3], Roland Mathis[1], Philippe Chouvarine[3], Dorothea Rutishauser[4], Laura De Vargas Roditi[4], Bence Szalai[5], Ulrich Wagner[4], Kathrin Oehl[4], Karim Saba [4], Arati Pati[3], Julio Saez-Rodriguez [5,6], Angshumoy Roy[3], Donald W. Parsons[3], Peter J. Wild[7 ✉], María Rodríguez Martínez [1 ✉] and Pavel Sumazin[3 ✉]

Knowledge about the clonal evolution of a tumor can help to interpret the function of its genetic alterations by identifying initiating events and events that contribute to the selective advantage of proliferative, metastatic, and drug-resistant subclones. Clonal evolution can be reconstructed from estimates of the relative abundance (*frequency*) of subclone-specific alterations in tumor biopsies, which, in turn, inform on its composition. However, estimating these frequencies is complicated by the high genetic instability that characterizes many cancers. Models for genetic instability suggest that copy number alterations (*CNAs*) can influence mutation-frequency estimates and thus impede efforts to reconstruct tumor phylogenies. Our analysis suggested that accurate mutation frequency estimates require accounting for CNAs—a challenging endeavour using the genetic profile of a single tumor biopsy. Instead, we propose an optimization algorithm, *Chimæra*, to account for the effects of CNAs using profiles of multiple biopsies per tumor. Analyses of simulated data and tumor profiles suggested that Chimæra estimates are consistently more accurate than those of previously proposed methods and resulted in improved phylogeny reconstructions and subclone characterizations. Our analyses inferred recurrent initiating mutations in hepatocellular carcinomas, resolved the clonal composition of Wilms' tumors, and characterized the acquisition of mutations in drug-resistant prostate cancers.

INTRODUCTION

Pan-cancer tumor profiling has identified recurrent alterations that are associated with tumor etiology at the loci of thousands of genes but the interpretation of genetic alterations remains a major challenge[1–3]. Knowledge about the clonal evolution of tumors can point to genetic alterations that both contribute to tumorigenesis, indicate prognostically relevant intratumoral variability, and point to refractory tumor subclones[4,5]. Specifically, clonal evolution—depicted as a phylogenetic tree in Fig. 1a—can help to identify alterations that play a role in tumor initiation as well as those that confer a selective advantage to altered tumor cells. Moreover, information about its subclone composition is important for predicting cancer's potential for drug resistance and metastasis, which vary across tumor subclones[6] and are the key determinants of patient outcome. Consequently, tumor-subclone characterization is essential for designing personalized therapies that target all tumor subclones and may hold the key to predicting tumor progression, metastases, drug sensitivity, and patient outcome.

Current methods that rely on DNA-profiling to reconstruct clonal evolution of tumors can be classified into two categories: methods that primarily rely on single-cell profiles[7–10] and those that computationally resolve mixtures of subclones from molecular profiles of bulk tumor cells, i.e., profiles of pools of cells that originate from a common malignant lesion[11–14]. Single-cell DNA sequencing can produce more definitive estimates of the proportion (*frequencies*) of tumor cells that contain each genetic alteration and more complete profiles of tumor subclones, including information about the co-occurrence of alterations within each subclone. Its primary disadvantage is operational: the availability of high-quality tumor samples that permit single-cell

isolation and profiling as well as the accuracy and cost associated with parallel sequencing DNA from a multitude of cells per tumor. Moreover, improving the accuracy of single-cell mutation profiling remains challenging due to limited material availability in single cells[15]; this is not likely to improve as future sequencing technologies focus on profiling formalin-fixed paraffin-embedded (*FFPE*) tumor samples[16,17]. Alternatively, single-cell RNA or protein profiling can help indicate tumor subclones, but these assays may not directly point to key driving genetic alterations.

Focusing on single-nucleotide somatic variants (*SNVs*; or simply mutations), we sought to reconstruct clonal evolution from DNA profiles of genetically unstable cancers. This entails deconvolving mutation frequencies, mutation-subclone associations, and CNAs from DNA profiles—including both whole-exome sequencing (*WES*) and panel-based (targeted) sequencing assays—that produce average estimates across cellular ensembles (Fig. 1b–d). One approach to improve the accuracy of these deconvolutions is to profile multiple biopsies from the same tumor across time points[18] or across regions[6,19]. This approach relies on two key assertions: (1) that genetic alterations that are specific to the same tumor subclone will co-occur with the same frequency across biopsies, and (2) that the clonal composition of heterogeneous regions varies, i.e., multiple sampling will allow for the aggregation and deconvolution of the frequencies of most mutations with improved power. It is important to note that mutations that underwent convergent evolution[20] do not violate these assertions and will not be aggregated with other mutations from the same tumor subclone because of differing frequency estimates across biopsies. It is also important to note that accurate deconvolution must account for tumor purity, and our efforts—including production of simulated data to compare leading methods and

[1]IBM Research—Zurich, 8803 Rüschlikon, Switzerland. [2]Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [3]Texas Children's Cancer Center, Baylor College of Medicine, Houston, TX, USA. [4]Pathology and Molecular Pathology, University Hospital Zurich, Zurich, Switzerland. [5]RWTH Aachen University, Faculty of Medicine, Joint Research Centre for Computational Biomedicine, Aachen, Germany. [6]Institute for Computational Biomedicine, Heidelberg University Hospital, Heidelberg, Germany. [7]Senckenberg Institute of Pathology, University Hospital Frankfurt, Frankfurt am Main, Germany. ✉email: Peter.Wild@kgu.de; mrm@zurich.ibm.com; sumazin@bcm.edu
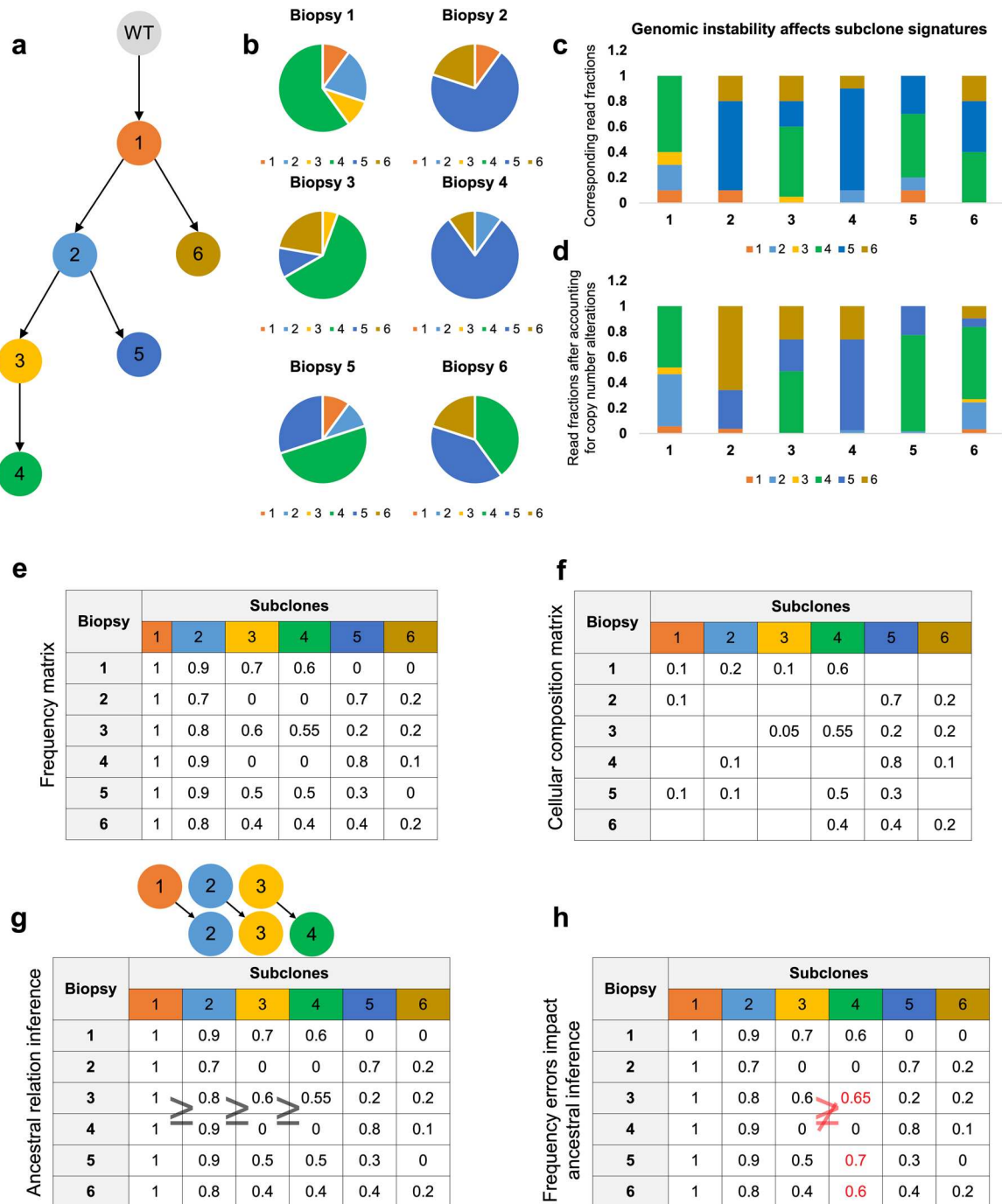
Fig. 1 A simulated footprint of the clonal evolution of a tumor, as observed in genetic profiles of multiple biopsies. a Tumor phylogeny composed of six dominant tumor subclones that make up the **b** cellular composition of six tumor biopsies. **c** Read fractions in a DNA profiling assay that are associated with these subclones and **d** corrected fractions after accounting for CNAs. **e** Frequencies of the variant allele, for each mutation defining a clone in each biopsy, are higher for ancestral clones. **f** Variant allele frequencies are linked to the cellular composition of the tumor and depend on its associated phylogeny. **g** Ancestral relations can be inferred by comparing subclone frequency vectors; e.g., Subclone 3 frequencies are greater or equal to those of Subclone 4 across all biopsies, suggesting that Subclone 3 may be ancestral to Subclone 4. **h** However, errors in frequency estimates (red) can complicate efforts to infer ancestry and tumor-phylogeny reconstruction.

analyses of tumor profiles—account for differences tumor purity across samples.

A central challenge for estimating mutation frequencies in tumors with unstable genomes is accounting for the effects of CNAs that can alter *mutated-read fractions*. These are observed in profiles of tumor biopsies that are composed of tumor cells with wild-type and mutated alleles as well as tumor-adjacent cells (Fig. 1e, f). In turn, inaccurate mutation-frequency estimates can

contribute to erroneous associations between mutations and tumor subclones as well as errors in phylogeny reconstructions (Fig. 1g, h). We describe the mutation-frequency inference problem as that of inferring tumor subclone frequencies and associating mutations with subclones. Consequently, we describe the tumor-phylogeny reconstruction problem as that of inferring ancestral relations between tumor subclones. The main challenges for addressing the mutation-frequency inference problem are to

aggregate co-occurring mutations across biopsies, estimate the frequency of each aggregate in every biopsy, and identify partial orders across aggregates that are consistent across biopsies. When viewed this way, each tumor subclone could be associated with a frequency vector that describes the proportion of cells containing its mutations in each biopsy. Ancestral order between two subclones could then be established based on (probabilistic) comparisons between their corresponding mutation frequencies. Ancestral order inference requires confident frequency assignment to the majority of mutations based on observed mutated-read fractions, and inference methods can be compared based on the number of mutations with frequency estimates, the accuracy of these estimates, and their accuracy at aggregating sister mutations that initiate the same clones.

We studied the mutation-frequency inference problem as a function of genetic instability and proposed the inference method Chimæra to improve these estimates and subsequent phylogeny reconstruction. Chimæra uses an optimization process to resolve the parameters of a natural model for the effects of CNAs on mutated-read fractions and is unique in its emphasis on the simultaneous inference of mutation frequencies and CNAs. We report on a comparison of Chimæra's accuracy to that of other mutation-frequency and heterogeneity inference methods on simulated DNA-profiling data of genetically unstable tumors and on biopsy subsets of thirteen tumors, including liver, kidney, and prostate cancers[21–23]. Each of these tumors was profiled in 4–10 regions, and three of the prostate cancers profiled were surveyed across multiple time points. We showed that Chimæra's inferences can be used to identify the key mutations that are associated with increased subclone proliferation, drug response, and tumor grade, as well as to infer the ancestral relations between tumor subclones that harbor these mutations.

RESULTS

We describe the results of our efforts to evaluate inference method accuracy on simulated data and to reconstruct phylogenies based on WES and targeted sequencing assays of tumor biopsies. Our analyses highlighted the challenges in inferring cancer mutation frequencies from these assays and the benefits of methods that rely on profiles of multiple biopsies per tumor to characterize tumor subclones and tumor evolution by tracking mutation aggregates.

Simulation of DNA profiling data

We used phylogeny models—with sizes ranging from three to twelve tumor subclones, twenty to fifty somatic mutations per subclone, and varying degrees of genetic instability—to generate simulated DNA profiles. ABSOLUTE[24], AncesTree[12], EXPANDS[25], PhyloWGS[14,26], SCHISM[13], and Chimæra were then used to reconstruct phylogenies based on simulated data. Each method inferred ancestral relations between mutation pairs, and errors were estimated as the combined frequencies of false-positive and false-negative predictions. ABSOLUTE infers tumor purity and malignant cell ploidy directly from the analysis of somatic DNA alterations by fitting estimates of copy-ratio of both homologous chromosomes with a Gaussian mixture model, where components were centered at the discrete concentration-ratios implied by an initial frequentist estimation. AncesTree characterizes the clonal evolution of tumors based on the probabilistic model for errors in observed read fractions and infers phylogeny matrices using integer linear programming. EXPANDS clusters mutations based on their cell-frequency probability distributions; clusters are next extended by members with similar distributions and pruned based on statistical confidence by comparing the cluster maxima and peaks observed outside the core region. PhyloWGS reconstructs phylogenies based on a model for simple somatic

mutations in addition to a correction for CNAs, all based on a single biopsy per tumor. SCHISM takes as input mutation cellularity estimations and mutation clustering inferred by other methods and uses a generalized likelihood ratio to infer lineage precedence and lineage divergence. A genetic algorithm is then used to build phylogenetic trees.

Accuracy of mutation-frequency estimation based on simulated data

Our initial efforts to compare accuracy between methods based on analyses of phylogenies of size three revealed variable success rates, with some methods showing consistently poor accuracy. EXPANDS and PhyloWGS, which were designed to reconstruct phylogenies using profiles of one biopsy per tumor, and ABSOLUTE, which is best known and most effective for estimating tumor purity, had consistently poor accuracy in our simulations—with results statistically indistinguishable from random inferences. SCHISM and AncesTree had better or comparable performance than these three methods in every simulated instance. For example, the magnitude of frequency inference errors by ABSOLUTE, which processed profiles of multiple biopsies per tumor, were more than double those of SCHISM and analyses required manual parameter optimization. However, ABSOLUTE had good accuracy for inferring tumor purity in our synthetic data. SCHISM and AncesTree do not explicitly account for the full range of observed CNAs in tumors, but they were accurate in 100% of our tested instances with three tumor subclones. Consequently, we focused on accuracy comparisons between inferences by SCHISM, AncesTree, and Chimæra on phylogenies composed of 6–12 tumor subclones. Moreover, our analysis suggested that more than one biopsy per tumor was required to accurately approximate mutation frequencies and CNAs at these mutation loci.

We compared the accuracy of SCHISM, AncesTree, and Chimæra on phylogenies that were adapted from a precompiled library that was generated both manually and using CITUP[27]; see Fig. 2b and Table S1 for representative phylogenies. Each somatic mutation was associated with a trio of copy numbers—δ^s, δ^s_w, and δ^s_m (Fig. 2a)—that were taken from truncated normal distributions with means $\mu \in \{1, 2, 3\}$, where $\mu = 1$ corresponds to no copy number changes and standard deviation $\sigma \in \{0, 1, 2, 3\}$; $\sigma = 0$. was used only when $\mu = 1$. The resulting copy numbers modeled a range of genetic instability conditions that were in line with observed CNAs in TCGA-profiled prostate, hepatocellular, breast carcinomas (HCC and BRCA in Fig. 2d, e); we assumed no linkage between simulated CNAs of any mutations, and we omitted prostate adenocarcinoma (PRAD) curves from Fig. 2 for readability. In addition, we added up to 10% of wildtype reads for all simulated mutations to account for the potential inclusion of nontumor cells in biopsied samples (WT subclone in Fig. 1a). Total coverage for each allele—i.e., the number of reads covering both the wildtype and mutated variants of a specific nucleotide—was taken by sampling mutation coverage values from prostate and Wilms' tumor biopsies profiled here. Finally, once idealized counts were available for both mutated and wildtype alleles, we simulated duplication or loss of up to 5% of the observations according to a uniform distribution. To simulate multiple regions per tumor, we repeated each biopsy simulation using the same simulation parameters but with distinct cellular composition vectors to produce simulated profiles of 6–12 biopsies per tumor, as depicted in Fig. 2c. The availability of six biopsies per tumor increases the likelihood that mutations can be aggregated and subclone mutation frequencies can be compared to infer ancestral relations and are in line with pre-existing datasets, including those reported here. The selection of six biopsies is a compromise between clinical feasibility and the power needed to infer mutation frequencies and phylogenies.
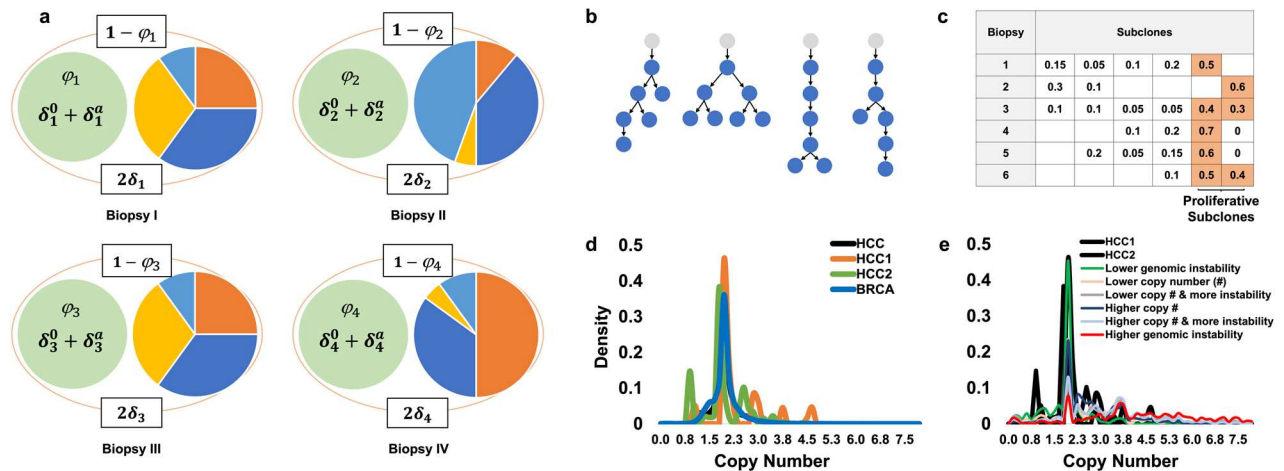
Fig. 2 Our model for the effects of copy number alterations on mutated-read fractions in DNA profiles. a For each mutation in each biopsy s, the mutated-read fraction is a function of the true proportion of profiled cells with the mutation (mutation frequency) φ_s, the copy number of the reference allele in cells without the mutation δ_s, and the copy numbers of the reference and the mutated allele in tumor cells with the mutation, δ_s^0 and δ_s^a, respectively. The frequency of tumor and WT cells without the mutation is $(1 - \varphi_s)$ and the two reference alleles can be assumed to have a combined copy number of $2\delta_s$. **c–e** To compare inference methods we synthetically generated profiling data based on parametrized simulated copy number distributions. **b** Shown are representative phylogenies and **c** a representative cellular composition matrix, as well as **d** density plots of average copy numbers across profiles of TCGA-profiled hepatocellular (HCC) and breast (BRCA) carcinomas. These include the distribution of copy numbers in two individual HCCs (HCC1 and HCC2) and across all HCCs (black) and all BRCAs (blue); copy numbers ranged from 0 to >260×. **e** Simulated copy numbers, including simulations of more-stable and less-stable cancers, ranged from 0× to 15× copies.

AncesTree accepts no external input when estimating mutation frequencies, but SCHISM can be guided by externally inferred mutation frequencies and clusters. SCHISM's implementation includes its own selected clustering methods, and these were also used to compare accuracy. Chimæra can also be guided by externally inferred mutation frequencies and clusters, but by default, it uses a clustering approach modeled after hdbscan[28]. When comparing SCHISM and Chimæra performance on synthetic data, we clustered mutations using their native clustering approaches and with tclust[29] optimization subroutines including ElbowSSE, Entropy, GMD, Mclust, and SDIndex[30,31]. We compared the accuracy of methods and pipelines on 2000 simulated assays, including both with and without modeled genetic instability by varying mutation copy numbers. The accuracy of SCHISM estimates was better, on average, than that of AncesTree, but it was relatively sensitive to clustering optimization methods, with SDIndex outperforming other methods, including SCHISM's native implementation. Comparatively, Chimæra estimates were less dependent on clustering methods and significantly outperformed estimates by SCHISM with SDIndex ($p < 1E-16$ by U-test).

When using its native clustering approach Chimæra exhibited lower accuracy than implementations using tclust (Fig. 3a), but it estimated frequencies for a significantly larger number of mutations (Fig. 3b). The number of mutations with no frequency estimates by Chimæra was 2-fold less than that of the next best method, which allowed for dramatically improved phylogeny reconstruction in both simulated data and profiled cancers. Inference accuracy, for both SCHISM and Chimæra, was anticorrelated with copy number variability across biopsies. We used CNA variability as a surrogate for genetic instability and quantified it using the coefficient of variation of mutation copy numbers across biopsies, which followed truncated normal distributions (Fig. 3c). However, while Chimæra inferences were affected by copy number variability, they were independent of the actual magnitude of CNAs (Fig. 3d). This, in turn, suggested that instability across biopsies is a key challenge for estimating mutation frequencies. We note that both the SCHISM and tclust-based Chimæra pipelines failed to cluster 40% of mutations in our synthetic data. Moreover, while Chimæra assigned frequencies to

all clustered mutations and made predictions for each simulate cancer, SCHISM did not successfully estimate mutation frequencies for some simulated genomes and cancer profiles. To account for this, accuracy comparisons in Fig. 3 relied on only those mutations that had assigned frequencies by all methods. In total, our analysis suggests that mutation frequency estimation is more challenging for genomes with high CNA variability (Fig. 3c, d). All data—including supplementary tables and analyses—are available at Chimæra's GitHub repository.

The number of regions profiled per tumor dictates algorithm convergence

Chimæra requires at least two profiled regions per tumor for making tumor subclone predictions, and while SCHISM can predict tumor subclones based on a single region, its accuracy improves with the number of profiled regions per tumor. To test the benefit of profile-multiplicity per tumor, we compared Chimæra and SCHISM analyses of multiregion WES profiles of 13 tumors using only subsets of the available tumor profiles for each analysis. Tumor profiles included profiles of nine hepatocellular carcinomas (HCCs)[32], three high-risk Wilms' tumors, and a castrate-resistant prostate cancer (CRPC). Each tumor was profiled in 5–10 distinct regions, and we compared the number of subclones detected by each method in each multiregion subset as a function of the number of regions, with a minimum of two WES profiles per tumor.

All HCCs were profiled in five regions, which permitted testing predictions in 2-size, 3-size, and 4-size subsets of the assays for each tumor; Wilms' tumors were profiled in six and eight regions; and the CRPC in ten regions. Our results (Fig. 4) suggested that, for most HCCs and Wilms' tumors, Chimæra-analysis of four tumor regions resulted in a similar number of subclones as profiling five tumor areas, however, profiling two tumor areas was insufficient for predicting tumor subclones accurately with Chimæra. Predictions for our CRPC, which had the highest genomic instability and mutation burden of the 13 tumors, largely converged with seven profiled regions. Interestingly, because mutations that were clustered together by Chimæra were associated with sufficiently different frequencies by SCHISM, the number of subclones
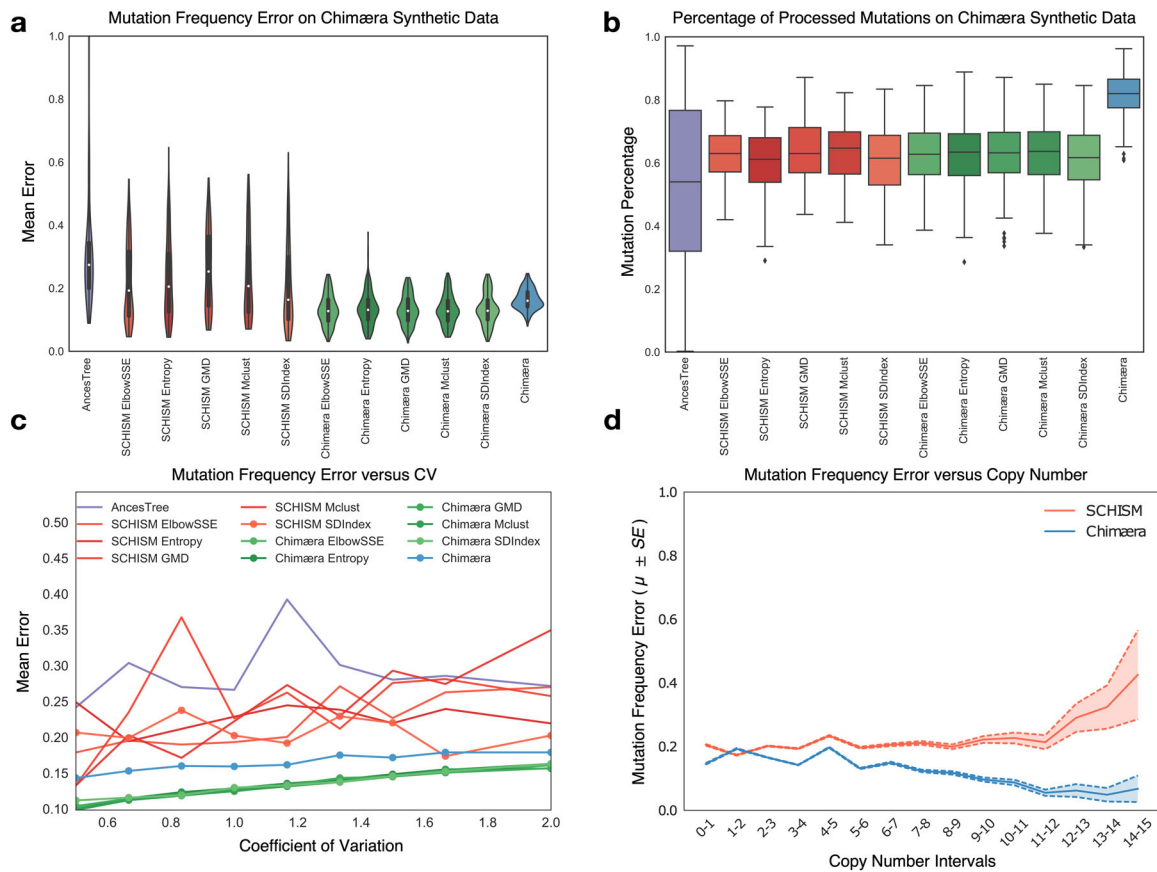
a

Mutation Frequency Error on Chimæra Synthetic Data

b

Percentage of Processed Mutations on Chimæra Synthetic Data

c

Mutation Frequency Error versus CV

d

Mutation Frequency Error versus Copy Number

Fig. 3 Performance of mutation frequency estimates on simulated data. a Performance—measured by mean error across simulated WES datasets from genomes with varying mutation copy numbers—of mutation-frequency estimates by AncesTree (purple), SCHISM (red), and Chimæra (green and blue); SCHISM and Chimæra were evaluated using multiple clustering methods in an effort to improve their accuracy, with SDIndex (SCHISM) and ElbowSSE (Chimæra) producing top accuracy, respectively. In blue, are reported estimates for the published Chimæra, which uses *hdbscan*. **b** No method is able to estimate mutation frequencies for every mutation; however, Chimæra assigns frequencies for over 80% of simulated mutations, compared to an average of 60% or fewer for other methods. **c** Errors in frequency estimates were correlated with genetic instability, which was measured here as the coefficient of variation within copy number distributions used in simulated WES profiles. Inferences by some methods were consistently better than others; e.g., SCHISM with SDIndex clustering outperformed AncesTree inferences. Chimæra clearly outperformed all the other methods regardless of the clustering strategy. **d** While copy-number variability in the same sample was correlated with inference errors, the absolute magnitude of copy numbers had no significant effect on Chimæra's performance. We report results for Chimera (hdbscan) and SCHISM with SDIndex (a representative that resembles results with other clustering methods). Standard errors are reported. Mean error is the mean of L1 distances between true and estimated mutation frequencies after normalizing for the number of biopsies tested.

predicted by SCHISM were often dramatically greater, and SCHISM analyses did not produce subclone predictions for some tumors and often required more regions to converge (Fig. 4b). A detailed description of tumor subclone predictions in each tumor context follows. All data and analyses are given in Table S2.

Phylogeny inference in HCC

HCCs are high-risk liver tumors that are known to have high genetic instability[23]. We used Chimæra to infer mutation frequencies and ancestral relations between HCC subclones based on WES profiles of nine HBV-positive HCCs[32]. In total, we obtained mutated-read fractions and CNA estimates for 1424 mutation candidates in 9 tumors and 43 tumor samples. Table S3 lists the data input to Chimæra, including mutation-frequency and CNA estimates for each mutation; it also details the outcome of the analyses described below.

Chimæra inferred frequencies estimates for 60% (858/1424) of all mutations, reconstructing phylogenetic trees for each tumor and predicting initiating clones and proliferative subclones; see representative trees in Fig. 5a–c. In contrast, SCHISM inferred mutations frequency for 18% of the identified mutations.

Interestingly, 100% (9/9) of the HBV-positive HCCs had predicted initiating mutations in WNT-signaling pathway genes (Fig. 5d). An examination of 102 TCGA-profiled HBV-positive HCCs[23] suggested that 74% (75/102) of samples carried mutations in WNT-signaling pathway genes, and that the majority of these samples (76%) had WNT-signaling pathway mutations with mutated-read fractions above 25%—corresponding to mutations that are potentially present in the majority of cells.

To test whether WNT-signaling pathway genes were enriched for mutations—and particularly high-frequency mutations with mutated-read fractions above 25%—we calculated the proportion of tumors with such mutations in each of 186 KEGG pathways in MSigDB[32]. The most enriched pathways by p-value and mutated-sample fraction are shown in Fig. 5e. *p*-values were estimated using permutation testing, where for each pathway, random same-size gene sets were generated using KEGG pathway genes and the mutated-sample fraction taken to generate a null distribution. WNT-signaling was the most enriched pathway, and most of the remaining enriched pathways significantly overlapped it ($p < 0.01$, FET). To correct for this overlap[33]—where pathways that overlap another pathway that is mutated in many samples are identified as significant—we recalculated enrichment significance

a Chimæra Inferences

Patient	Tumor Type	2	3	4	5	6	7	8	9	All Regions
HCC5647	HCC	23.5	18.5	11						8
HCC6046	HCC	6	7	8						7
HCC6952	HCC	7	5	5						6
HCC7608	HCC	6.5	2	2						2
HCC8010	HCC	4	3	3						2
HCC8031	HCC	5.5	3.5	3						4
HCC8213	HCC	7	6	6						6
HCC8257	HCC	5	4.5	5						4
HCC8392	HCC	2.5	2	2						2
CG118	Wilms	4	3	3	3					3
CG163	Wilms	5	4	4	4					2
CG565	Wilms	5	5	4	3	3	3			3
CRPC	PC	15	13	11.5	11	6	3	3	4	4

b SCHISM Inferences

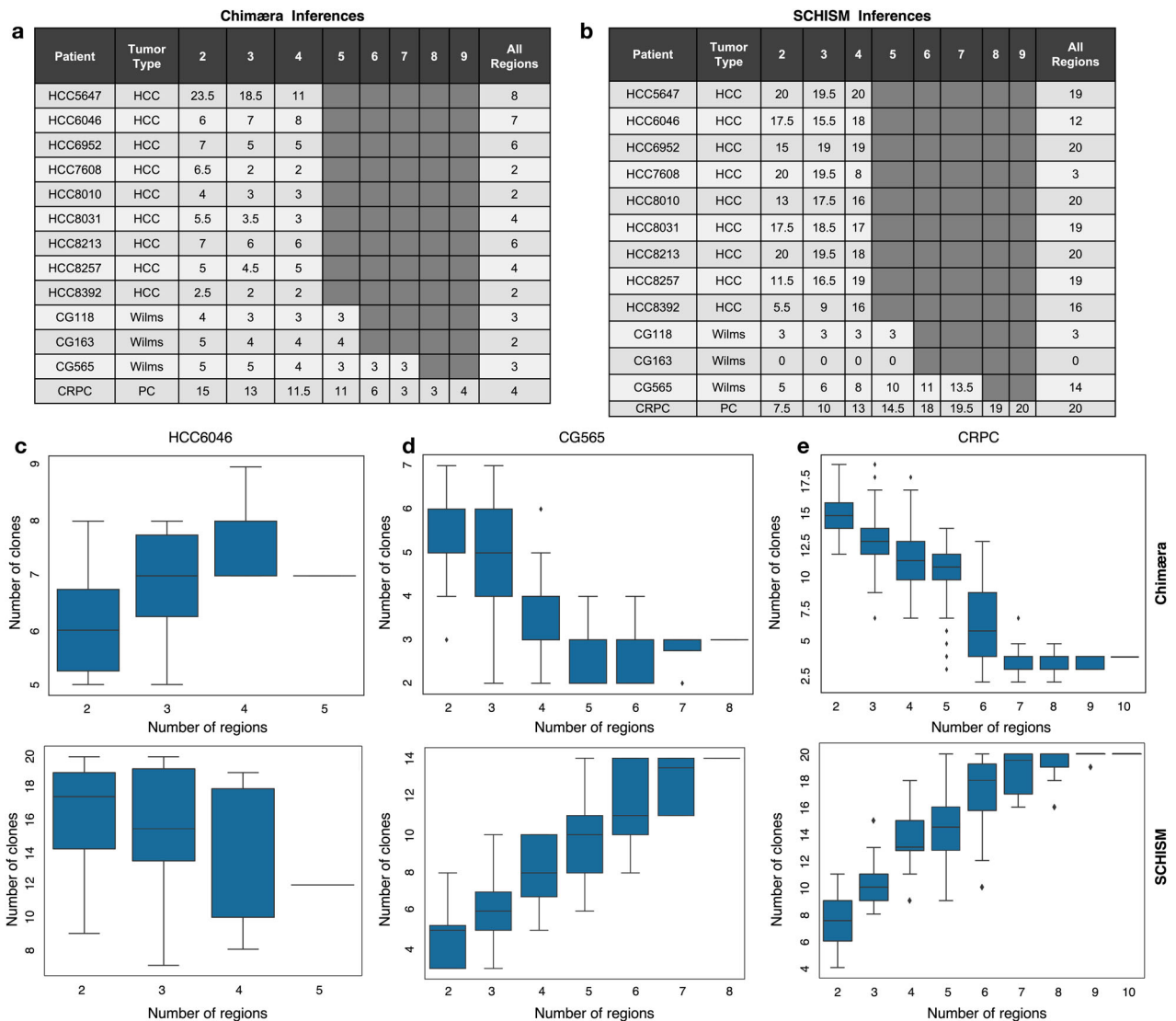| Patient	Tumor Type	2	3	4	5	6	7	8	9	All Regions
HCC5647	HCC	20	19.5	20						19
HCC6046	HCC	17.5	15.5	18						12
HCC6952	HCC	15	19	19						20
HCC7608	HCC	20	19.5	8						3
HCC8010	HCC	13	17.5	16						20
HCC8031	HCC	17.5	18.5	17						19
HCC8213	HCC	20	19.5	18						20
HCC8257	HCC	11.5	16.5	19						19
HCC8392	HCC	5.5	9	16						16
CG118	Wilms	3	3	3	3					3
CG163	Wilms	0	0	0	0					0
CG565	Wilms	5	6	8	10	11	13.5			14
CRPC	PC	7.5	10	13	14.5	18	19.5	19	20	20

Fig. 4 The number of predicted subclones depends on the number of profiled regions per tumor. a, b Lin et al. profiled five regions of each of nine HBV-positive HCCs. We profiled 6–8 regions of each of three Wilms' tumors and ten regions of a castrate-resistant prostate cancer (CRPC). For each tumor, we exhaustively selected all region subsets size-2 and up, and compared the number of predicted tumor subclones across subsets to those obtained using all available region profiles using **a** Chimæra and **b** SCHISM. Chimæra analysis of any 4-tumor regions resulted in a similar number of subclones as analysis of five regions, however, profiling two regions produced fewer predicted subclones. SCHISM performed better on stable genomes (CG118) than on unstable genomes (CG163 and CRPC). **c** Chimæra analyses (top) of HCC6046 profiles suggested convergence of subclone predictions using three profiled regions, while SCHISM analyses (bottom) produced a higher prediction variability across profiling subsets. **d** Similarly, Chimæra analyses of four regions of CG565 produced a similar number of clones as profiles of eight regions. **e** Chimæra analyses of seven regions of our CRPC tumor produced a similar number of clones as analysis of ten regions; however, analyses based on five or fewer tumor areas produced significantly more predicted tumor subclones due to reduced aggregating power. SCHISM analyses based on six or more regions produced consistent counts of predicted subclones, but this number was considerably greater than the number of subclones predicted by Chimæra.

for each pathway using the same test but after excluding WNT-signaling pathway genes. We note that MAPK-signaling and two other top-10 pathways were still enriched (Fig. 5e). Analyses products and input data are given in Table S3.

Phylogeny inference in Wilms' tumors
To test SCHISM's and Chimæra's predictive ability in tumors with a range of genomic instability, we selected three Wilms' tumors with low-genomic (CG118), intermediate-genomic (CG565), and high-genomic instability (CG163). Multiple regions of these tumors were profiled by WES, including six regions from each of CG118 and CG163 and eight regions from CG565 (Fig. 6a). SCHISM produced stable tumor subclone predictions that agreed with

predictions by Chimaera for CG118, but it did not converge on a set of subclones even when profiling was available for seven regions from CG565; it also was not able to predict any subclones for CG163 (Fig. 4b). Chimæra tumor subclone predictions converged with fewer profiled regions, and Chimæra predicted phylogenies for all three profiled tumors.

Chimæra analysis of CG118 profiles (Fig. 6b) suggested that the tumor was composed of primarily two types of cells: tumor subclones with a *CTNNB1* (S45F) mutation and those with a mutation in *WT1* (R445W) mutation; a predicted daughter clone of the *CTNNB1* mutation had no previously studied mutations. Both *CTNNB1* and *WT1* mutations have been previously implicated with Wilms' tumor genesis[34,35], and both clone types were present in
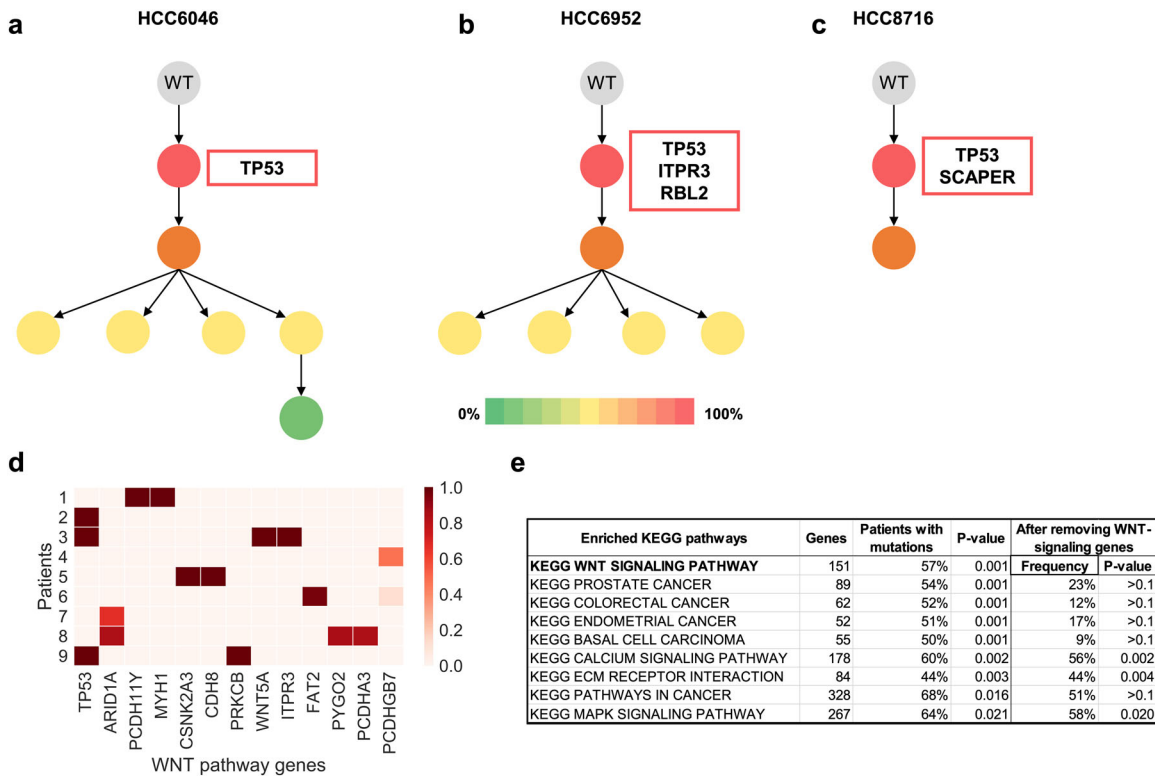
Fig. 5 Inferred tumor phylogenies for HBV-positive HCCs suggest that WNT-signaling pathway mutations play a key role in tumor initiation. a–c Mutations in *TP53* were inferred to initiate tumorigenesis in three of the nine tumors we studied; labels correspond to labeling by Lin et al. **d** All nine tumors had mutations in WNT-signaling pathway genes that were predicted in the initiating tumor subclone. **e** The majority of the 102 TCGA-profiled HBV-positive HCCs had mutations in WNT-signaling pathway genes, which was the most significantly mutated KEGG pathway in these patients. Most other significantly mutated pathways were no longer enriched for mutations after the exclusion of WNT-signaling pathway genes from the analysis.

every profiled region. However, the majority of cells—in all regions—were predicted to have *CTNNB1* mutations. To compare the effects of these mutations, we compared RNA-expression profiles in regions with the lowest and highest frequencies of *WT1*-mutated cells (Fig. 6c), 7% and 3%, respectively. Analysis of the CTNNB1-pathways and WT1-pathways[36,37] suggested that they are differentially regulated across these tumor regions (Fig. 6c). These data suggested that S45F activates CTNNB1 and that R445W inhibits WT1, as previously described[38,39].

Chimæra analysis of CG163 profiles suggested that the acquisition of missense mutation in *LIN28A* (p.R132H) was a key event early in the formation of this tumor. Chimæra identified a second event that produced a less frequent tumor clone with many coding mutations with unknown significance. RNA-expression profiles of regions with the lowest and highest concentration of *LIN28A* mutations suggested that genes downstream of LIN28 are significantly altered in mutated cells. Analysis of CG565 did not reveal any mutations with known significance in Wilms' tumors. Data and analyses are provided in Table S4.

Phylogeny inference in prostate tumors

To further test Chimæra's predictive ability, we studied ultra-deep profiles of multiple regions of a select set of prostate cancers at multiple time points. Ultra-deep targeted profiling allows for improved mutation identification and read-fraction estimation. A single region from each of these cancers was previously profiled and helped identify multiple predicted driver mutations[40,41]. However, because mutations detection by WES are not always reliable and often includes both false-positive and false-negative predictions[42], we selected three cancers and designed a mutation panel that targets mutated genes in these cancers, as well as other

known driver genes in prostate cancers[43] for ultra-deep sequencing. The identity of targeted genes is given in "Methods" section. This approach helped test Chimæra's performance in more restrictive assays, which are quickly becoming standard in oncology clinics. Controls and five areas of each cancer were profiled at 2, 3, and 5 time points per cancer using both our targeted sequencing panel and OncoScan arrays to estimate CNAs on genome scales; areas profiled from tumor PC1 at each of three time points are shown in Fig. 7a and given in Supplementary Table S5.

We recorded changes in treatment, identified mutations that may have been acquired following treatment, and predicted phylogenies for each tumor based on these mutations. We only considered mutations that were observed in multiple regions at the same time point, thus eliminating two-thirds of the candidates. Our results demonstrate the feasibility of phylogeny inference from targeted-panel profiling of multiple tumor regions and support phylogeny prediction efforts by suggesting that predicted subclones that are supported by multiple genetic variants may persist and accumulate additional genetic variants across time. We describe the results of Chimæra analysis below.

The tumor PC1 was diagnosed as Gleason $3 + 4$ and treated with an LHRH analog. It was profiled at 3-time points during follow-ups after treatment was started. Timepoint 1 in Fig. 7a was taken over a year after diagnosis and treatment, and Time points 2 and 3 followed 1.17 and 1.4 years after Time Point 1. These suggested increased risk, with Gleason $5 + 5$ and the discovery of a mutation in *AR* that has been associated with increased cancer cell proliferation and poor outcome. Chimæra analysis of profiles of PC1 suggested two predominant tumor subclones that are represented in Fig. 7b by predicted deleterious mutations in *EP300*
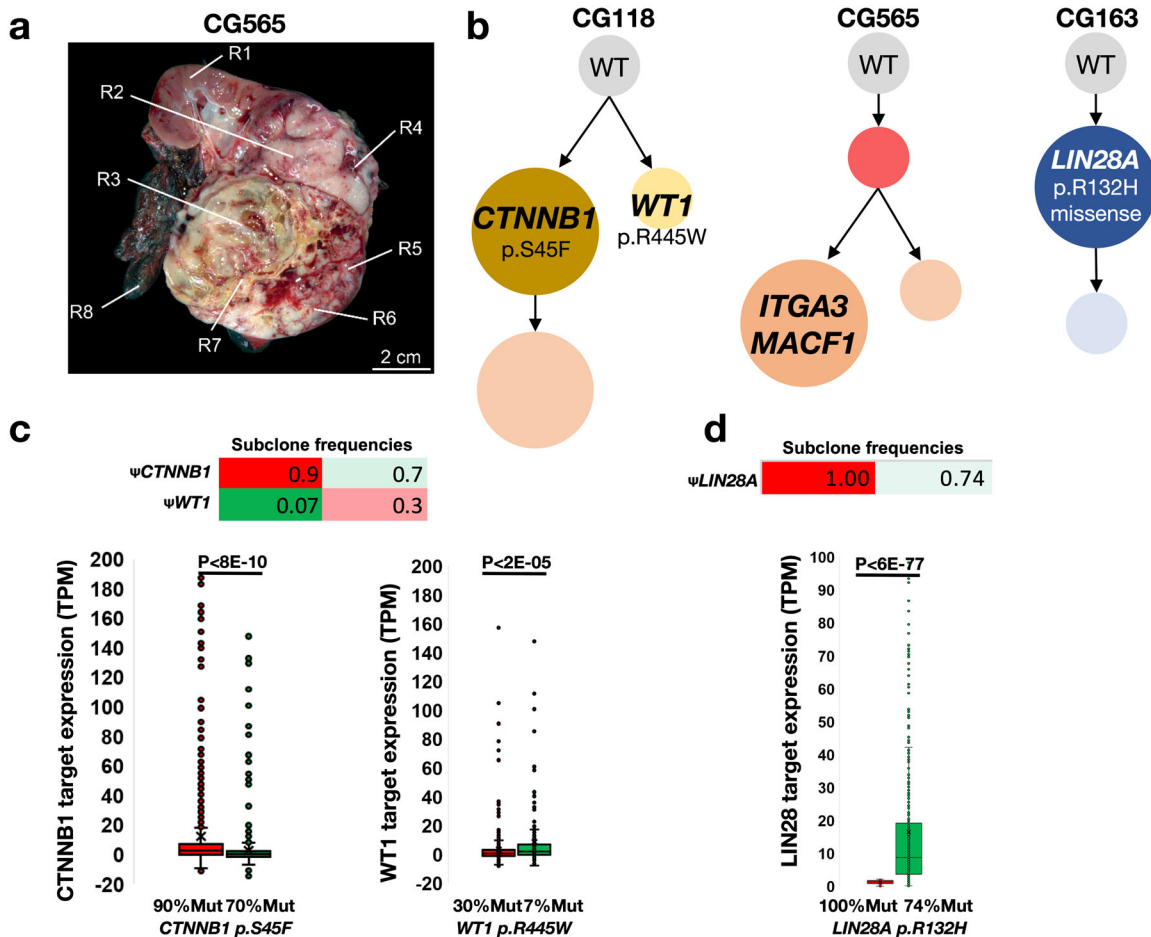
Fig. 6 Inferred tumor phylogenies for high-risk Wilms' tumors identified driving initiating mutations. a The location of eight CG565 regions selected for profiling. **b** The inferred phylogeny for CG118 suggested that it is composed of two major subclones driven by previously observed mutations in *CTNNB1* and *WT1*, with the *CTNNB1*-mutated clone accounting for a larger proportion of the tumor. CG565 was predicted to acquire mutations in *ITGA3* and *MACF1* that coincided with clonal expansion. The initiation of CG163 was predicted to include a mutation in *LIN28A*, which is sufficient to drive Wilms' tumor genesis. **c, d** RNA-expression profiles in **c** regions that were more abundant with each of the CG118 subclones—90% vs. 70%, and 30% vs. 7%, for the *CTNNB1*-mutated and *WT1*-mutated subclones, respectively—and **d** *LIN28A*-mutated subclones of CG163 (100% vs. 74%) suggested differential expression of the gene programs downstream from these predicted drivers.

(p.I997V) and *AR* (p.T878A); Fig. 7b. Tumor subclones with the *EP300* mutation were present in all regions profiled at Time Point 1. *AR* mutations were identified at high frequencies at Time Points 2 and 3 but only in regions that lacked the *EP300* mutation and had high Gleason scores. Together, these findings suggest that either the most aggressive tumor sections were not profiled in Time Point 1, or that subclones with the *AR* mutation have a proliferative advantage and have overtaken subclones with the *EP300* mutation. Protein profiling by Hyper Reaction Monitoring of regions with low-frequencies and high-frequencies for the subclones with the *AR* mutation confirmed that genes that have been shown to be downregulated by AR[44] are significantly upregulated in *AR*-mutated regions.

PC2 was diagnosed as Gleason 5 + 4 and biopsied before treatment. The patient was treated for 9 months with LHRH-Analogon, during which the cancer was biopsied two additional times and registered an increase in severity to Gleason 5 + 5 at Time Point 3. Following treatment, the tumor was biopsied a 4th time and showed no change in severity (Gleason 5 + 5). The patient was then treated with combined androgen blockage—LHRH and Casodex—for 2 months, followed by a Gleason 4 + 5 diagnosis at Time Point 5. Time Point 1 profiles revealed the loss of *RB1* and a commonly observed stop-gain mutation in *PTEN* (p.

R303X). Mutations in *BRCA2* and *EP300* were observed in Time Points 4 and 5. Chimæra analysis suggested that the RB1 locus deletion predates the *PTEN* mutation and that the BRCA2 and EP300 mutations were acquired in tumor subclones that following *RB1* loss, lacked the *PTEN* mutation (Fig. 7d).

The tumor PC3 was diagnosed as Gleason 5 + 4 using post-treatment profiles, which included androgen blockers and orchiectomy 9 years prior to Time Point 1 profiling. A second biopsy, taken 6 months after the first suggested increased severity and Gleason 5 + 5. Tumor profiles identified known deleterious mutations in *PTEN* (p.Q245) and *BRCA1* (p.E1038G), as well as a stop-gain mutation in BRIP1, and nonsynonymous mutations in *BRCA2* and *PALB2*. All mutations were identified at both time points and nearly all regions. Chimæra's analysis suggested that *RB1* loss was followed by the *BRCA1* mutation. This was followed by the acquisition of a mutation in *TP53* (p.V173G) or *PALB2* and *BRIP1*. These sister clones subsequently acquired mutations in *PTEN* and *BRCA2*, respectively (Fig. 7e). Comparing the expression of genes downstream from PTEN and TP53 by Hyper Reaction Monitoring in regions with the lowest-consternation and highest-consternation of subclones with *PTEN* and *TP53* mutations suggested that these mutations disrupt the associated pathway (Fig. 7f). Genes known to be downregulated by PTEN[45] were
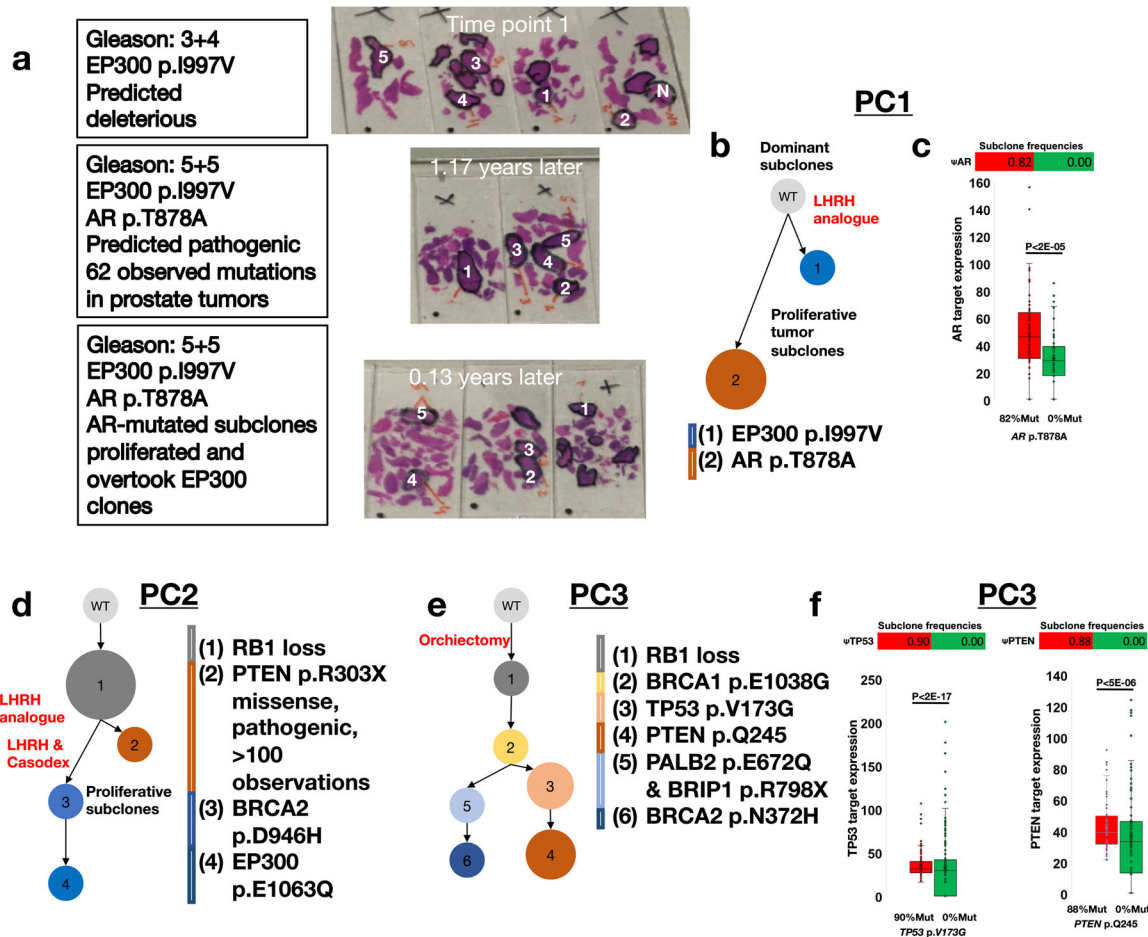
Fig. 7 Inferred phylogenies of tumors from three prostate cancer patients. a Five regions of Prostate Cancer 1 (PC1) were profiled at each of three time points, identifying potentially deleterious mutations at each time point. **b** PC1's inferred phylogeny suggested that *EP300* p.I997V mutation was present at Time Point 1, and that the tumor subclone with *EP300* p.I997V (Subclone 1) is distinct from the subclone with *AR* p. T878A (Subclone 2), which was observed only in the later time point and whose clonal frequency increased with time. **c** Differential protein expression analysis suggested that regions that were predicted to have high clonal composition of Subclone 2 (82% vs. 0%) had higher AR-target expression than regions without Subclone 2. **d** Inferred phylogeny for Prostate Cancer 2 (PC2) suggested that *RB1* loss was an initiating event, and that that majority of tumor cells are the results of divergent evolution following the acquisition of *PTEN* and *BRCA2* mutations, Subclones 2 and 3, respectively. In total, 5 regions of PC2 were profiled in each of the five Time Points; Subclone 1 was observed in all time points, and Subclones 3 and 4 in Time Points 4 and 5. **e** PC3's inferred phylogeny included *RB1* loss as an initiating event, followed by the acquisition of a *BRCA1* mutation. These tumor cells then either acquired *TP53* and *PTEN* mutations, or *PALB2*, *BRIP1*, and *BRCA2* mutations. All mutations and subclones were observed in each of the two time points profiled. The *PTEN*, *BRCA1*, and *BRCA2* mutations were previously observed in cancers. **f** A comparison of PC3 regions with low and high composition for Subclones 3 and 4, 90% vs. 0% and 88% vs. 0%, respectively, suggested significant differential protein expression for targets of TP53 and PTEN.

upregulated in regions rich with PTEN-mutated subclones as did targets of TP53[46]. All input data—including protein expression profiles—and analyses products are given in Table S5.

DISCUSSION

Bulk tissue DNA profiling by whole-genome and targeted sequencing can identify key DNA alterations that provide insight into the biology of tumors and indicate effective treatment options. Increasingly, these assays are used to help elucidate the clonal composition of heterogeneous tumors and even predict ancestral relationships between tumor subclones. The key to such efforts is the accurate estimation of mutation frequencies from both high-coverage and low-coverage DNA profiles. Our study suggested that such estimation efforts must explicitly account for the copy numbers of both the reference and alternative alleles and that only accurate mutation-frequency estimates can yield accurate tumor phylogenies.

Accordingly, we reported on methodology to improve the accuracy of tumor phylogeny reconstruction by improving mutation-frequency estimates from DNA profiles of multiple same-tumor biopsies. Our analysis suggested that mutation-frequency estimates are particularly challenging in the face of high genetic instability, which is characteristic of high-risk cancers, and that the accuracy of methods that rely on DNA profiles of a single biopsy of such cancers is poor. We also showed that even when profiles of multiple biopsies are available, methods that do not explicitly account for the full range of copy number variability produce inconsistent results and have poor accuracy.

We briefly outlined current methods to reconstruct tumor clonal evolution using DNA-profiling. These include methods that rely on single-cell profiles and methods that resolve subclone mixtures from profiles of bulk tissues. We note that single-cell DNA sequencing is expected to help produce more accurate mutation-frequency estimates, but this technology is yet to mature and does not produce accurate single-cell estimates for *FFPE* tumor samples. The greatest challenge in cancer genomics—and this is

not expected to change—is sample availability for patients with rich or specific clinical annotation, and FFPE is and will remain the primary preservation method for solid tumor resections. Moreover, frozen samples that could be used to generate cell suspensions that may be profiled using single-cell DNA sequencing cannot be easily used to evaluate the heterogeneity of tumors. We expect that whole-genome, whole-exome, and targeted-panel sequencing will produce the vast majority of tumor DNA profiles in the near future.

We expect that in the cases that single-cell DNA and RNA profiling is possible and pursued, whole-genome and whole-exome sequencing will be used to inform on the biology of clones. In these cases—where analyses of single-cell RNA, DNA, or protein expression will be integrated with analyses of bulk tumor profiles—methods that impute mutation frequencies and even tumor phylogeny from bulk tumor DNA profiles will play a key role in this integrated analysis. Most importantly, we argue that given the heterogeneity of solid tumors and the often observable and documented differences in the composition of regions on the same tumor, profiling multiple tumor regions will be required for both research and clinical efforts in the future.

Our analyses also suggested that Chimæra improves on mutation-frequency estimates by harnessing added information from multiple profiles and by directly accounting for the influence of CNAs on observations from DNA profiles. Chimæra's advantage was clearly observed in simulated data, where its performance was the most consistent and its accuracy the greatest. Interestingly, while Chimæra was able to estimate mutation frequencies with relatively high accuracy even at very high and very low copy numbers, its performance declined for the most unstable genomes where copy numbers for the same mutation varied widely across samples.

Using three tumor types—including Wilm's tumors, which is expected to have relatively low genetic instability and few mutations—and prostate cancers, where longitudinal biopsies from the same patients were available, we showed that when given a sufficient number of biopsies per tumor, Chimæra is able to address mutation-frequency estimate challenges arising from genetic instability. Our results suggested that Chimæra's mutation aggregation approach can help resolve issues arising from convergent evolution[20] and false-positive mutation calls[42]. We also showed how Chimæra could be used in conjunction with ultra-deep sequencing to improve on mutation calling accuracy. In conclusion, our results suggest that accurate mutation-frequency and cellularity inference are possible using profiles of multiple biopsies per tumor when coupled with analyses that account for the effects of CNAs on observed mutation read fractions.

METHODS

We formulated the phylogeny reconstruction problem in set-theoretic terms, which lead to a natural model for the effects of CNAs on mutated-read fractions in sequencing profiles. We describe our methodology for simulating WES tumor profiles, as well as our efforts to deconvolve mutation frequencies from simulated data using ABSOLUTE, AncesTree, EXPANDS, SCHISM, and Chimæra. Note that, to reduce analysis complexity, CNAs and mutation frequency simulations did not include alterations in sex chromosomes. Finally, to demonstrate that Chimæra can be effectively applied to clinical data, we described reconstructed phylogenies from WES profiles of ten same-tumor CRPC biopsies; a set of five same-tumor HCC biopsies from nine patients; two six same-tumor biopsies and an eight same-tumor biopsy of three Wilms' tumors; and three same-tumor prostate-cancer biopsies from three patients profiled at multiple time points.

Phylogeny reconstruction problem

Let $M = \{m : m \in \mathbb{N}, 1 \le m \le n\}$ denote the set of n mutations identified across a set of profiled biopsies S. The mutation burden in any given cell is given as a subset of M, $\gamma \subseteq M$, or as an element of the power set over M,

$\mathcal{P}(M)$; i.e., $\gamma \in \mathcal{P}(M)$ is a specific mutation ensemble that characterizes a tumor subclone. We denote the cellularity of γ and its corresponding subclone in biopsy $s \in S$ as ρ_γ^s, and the frequency of mutation $m \in \gamma$ in biopsy s as $\varphi_m^s = \sum_{\{\gamma:\gamma \in \mathcal{P}(M), m \in \gamma\}} \rho_\gamma^s$. Consequently, $\sum_{\gamma \in \mathcal{P}(M)} \rho_\gamma^s = 1$ and the assignment $A = \{\rho_\gamma^s : \gamma \in \mathcal{P}(M), s \in S\}$ produces a solution to our clonality reconstruction formulation.

Mutation frequencies

As defined above, for a mutation m in biopsy $s \in S$, φ_m^s denotes the frequency of cells in s with mutation m. The total copy number C^s of the allele targeted by the mutation can be estimated from WES data. C^s is composed by: the copy numbers of the allele in cells that lack mutation m, δ^s, the copy number of the wildtype allele in m-mutated cells, δ_w^s and the copy number of the mutated allele in m-mutated cells, δ_m^s (Fig. 2). Notice that if no copy number event has occurred at the locus: $\delta^s = 2$, $\delta_w^s = 1$ and $\delta_m^s = 1$. Adopting the infinite-sites assumption, we denote the mutated-read fraction—the fraction of reads reflecting the mutated versus wild-type allele in a WES profile—in sample s as f_m^s. Then, we can formulate the following equations (Eqs. 1, 2).

$$C^s = \delta^s(1 - \varphi_m^s) + (\delta_w^s + \delta_m^s)\varphi_m^s. \tag{1}$$

$$f_m^s = \frac{\varphi_m^s \delta_m^s}{C^s}. \tag{2}$$

Equation (1) provides a weighted sum of the copy number contribution from each allele type, and Eq. (2) gives the ratio of the number of reads from the mutated allele and the total number of reads.

Chimæra

Chimæra proceeds in three steps. First, mutation frequencies are approximated from sequencing and CNA data in each biopsy; then, mutations with similar frequency vectors (where each vector component gives the mutation frequency in each biopsy) are clustered together to form subclones; and finally, mutation frequencies and CNAs for these alleles are refined using an optimization process. The optimization assumes that clustered mutations that are associated with the same subclone have the same frequencies in each tumor biopsy and that δ_m^s—the average copy number of m (the mutated allele)—is unchanged across biopsies from the same tumor. This assumption can be relaxed in post-processing.

We first approximate the true frequency of the mutation φ_m^s by accounting for tumor purity, i.e., the fraction of tumor cells in biopsies that include non-tumor cells, and assuming that the allele's average copy number in tumor cells—whether mutated or not—is fixed. Let p^s be the purity of biopsy s, then Eq. (2) can be rewritten as follows:

$$f_m^s = \frac{\varphi_m^s \delta_m^s p^s}{2(1 - p^s) + C^s p^s}. \tag{3}$$

The experimentally observed copy number, C_{obs}^s, depends on the purity of the sample and the copy number of the sample tumor cells, C^s, as follows:

$$C_{obs}^s = 2(1 - p^s) + C^s p^s, \tag{4}$$

where C_{obs}^s can be estimated using additional biochemical assays, genetic sequencing, or through computational analysis of WES data[47], and the normal cells are assumed to have been corrected for germline copy number variants associated biases.

The simplifying assumption that the allele's average copy number—averaged across all profiled cells—of the mutated allele in tumor cells is constant across biopsies, i.e.: $\delta_m^s = \frac{C^s}{2}$. Under this approximation, we can use Eqs. (3) and (4) to eliminate C^s and obtain a first approximation of the mutation frequency $\widetilde{\varphi_m^s}$:

$$\widetilde{\varphi_m^s} = \min\left(\frac{2 f_m^s C_{obs}^s}{C_{obs}^s - 2(1 - p^s)}, 1\right). \tag{5}$$

This constraint will be later removed in the optimization process that follows but is necessary at this stage to obtain a first approximation that mutation frequencies that take into account the copy number influence from WES measurements. The minimization is necessary because of the interplay between the copy numbers at these alleles that may produce a first approximation above 1.

The approximate mutation frequency vectors (Eq. 5) are next clustered to identify candidate groups of mutations that form subclones. We

considered clustering algorithms with the robust treatment of outliers in order to ensure good clustering stability and quality. Specifically, we used a method modeled after *hdbscan*, a density-based hierarchical clustering method that aims at maximizing the stability of the obtained clustered against noise and requires minimal parameter selection. The number of clusters is determined automatically based on the minimal number of mutations that have to be considered to constitute a cluster. We also use tclust[29], a nonhierarchical robust clustering that trims outliers based on a probabilistic model. The number of clusters is selected by optimizing intracluster entropy or the sum of squared errors (SSE) and using a variety of optimization methods including the Elbow method, Gaussian mixture decomposition (GMD), and SD index[48–50]. The clustering based on *hdbscan* showed better performance on the generated synthetic data compared to others, especially when considering the number of mutations that could be assigned frequency estimates. Furthermore, it has the advantage of avoiding imposing a prior distribution on the mutations frequencies. Once the clusters are found, Chimæra assumes that each cluster represents a subclone and uses the mutation assignment to infer subclone frequencies and copy number estimates for each mutated allele in the final optimization step.

Focusing on subclone $\in \mathcal{P}(M)$, Eq. (3) describes a relationship between the frequencies and copy numbers of mutations in γ:

$$\varphi_m^s \delta_m^s = f_m^s \frac{C_{obs}^s}{p^s} \equiv \mathcal{B}_{ms}, \forall m \in \gamma, s \in S, \quad (6)$$

where $\mathcal{B}_{ms}$ is the entry of a matrix $\mathcal{B} \in \mathbb{R}^{|S|,|\gamma|}$ corresponding to mutation m and biopsy s. $\mathcal{B}$ is fully determined from analysis of sequencing assays, including purity, observed copy numbers, and observed mutated-read fractions of each mutation.

Unfortunately, the right-hand side of Eq. (6)—a multiplication of frequencies and copy numbers—can not be analytically decoupled. However, from our problem formulation, mutations from the same subclone are expected to have the same mutation frequencies, i.e., $\varphi_{m_i}^s = \varphi_{m_j}^s \equiv \varphi^s \, \forall \, m_i, m_j \in \gamma$. Further, we assume that the copy number of each mutation m is constant across biopsies, i.e., $\delta_m^{s_i} = \delta_m^{s_j} \equiv \delta_m \in [0, CN] \forall s_i, s_j \in S$, where CN is a fixed upper bound for the copy number; CN = 15 in our simulations. While we expect that this assumption will introduce some errors to the approximation of δ_m^s, it will have limited effects on the selection of optimal mutation frequencies because the variability of copy number averages for the mutated allele across biopsy is expected to be low for most mutated loci. We also note that we have not assumed stable genomes in our simulated data, i.e., the generated data displays variable copy numbers for the same mutated allele across biopsies in order to have an accurate estimate of the committed error.

After these assumptions, the optimization problem for each subclone $\in \mathcal{P}(M)$, based on Eq. (6), can be formulated as:

$$\min \| \vec{\varphi^s} \otimes \vec{\delta_m} - \mathcal{B} \|_2; 0 \le \delta_m \le CN, 0 \le \varphi^s \le 1, \forall m \in \gamma, \forall s \in S, \quad (7)$$

where $\vec{\varphi^s}$ is the mutation frequency vector across biopsies for all mutations in γ; $\vec{\delta_m}$ is the copy-number vector for each mutation in γ; $\mathcal{B}$ is as defined in Eq. (6); and $\varphi^s \otimes \delta_m$ denotes the outer product of vectors $\varphi^s \in \mathbb{R}^{|S|}$ and $\delta_m \in \mathbb{R}^{|\gamma|}$. We used the sequential least-squares programming (SLSQP) optimization[51] to find an optimal solution of Eq. (7). To avoid being trapped in local optima, multiple runs with different random initializations for the mutations frequencies and unitary variant allele copy numbers are performed.

Profiling and analysis of Wilms' tumors

DNA libraries were constructed and sequenced as previously described[52]. Briefly, after QC, high molecular weight double-strand genomic DNA samples are constructed into Illumina PairEnd pre-capture libraries according to the manufacturer's protocol (Illumina Inc.) with modifications. 300 base-pair fragments were checked using a 2.2% Flash Gel DNA Cassette 5 (Lonza, Cat. No.57023). The Fragmented DNA were End-Repaired, incubated, A-tailed, and ligated. Agencourt® XP® Beads (Beckman Coulter Genomics, Inc.; Cat. No. A63882) were used to purify DNA after each enzymatic reaction. After Beads purification, PCR product quantification and size distributions were determined using the Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517). Pre-capture libraries (1 µg) were hybridized in solution to VCRome 2.1 (NimbleGen) targeting 43 Mb of sequence from ~30 K genes, according to the manufacturer's protocol. Sequencing was performed in paired-end mode with Illumina HiSeq 2000. Cluster formation and primer hybridization were performed on the flow cell with Illumina's cBot cluster generation system. On average, about 80–100 million successful reads, consisting of 2 × 100 bp, were generated on each lane of a flow cell. CNAs and mutation frequencies in sex chromosomes were estimated using Genome Analysis Toolkit pipelines.

RNA-seq profiles of Wilms' tumors were aligned using STAR v2.3.0e[32] to an index of GRCh37 that included GENCODE v16 gene annotation. Alignment files were processed using Picard tools v1.54, and the final BAM files indexed using SAMtools index v0.1.11[33]. RNA-seq run quality was assessed using the RNA-SeQC package[34] using the same GENCODE 16.gtf file. Transcript quantification was performed using Cufflinks[18] v2.02 running in quantification mode against the GENCODE v16.gtf file. FPKM values were used for relative abundance estimation.

Profiling and analysis of WES of prostate cancer biopsies

To test our ability to infer mutation frequencies and ancestral relations between subclones based on clinical profiles of four prostate cancers. The Specimen were collected at the Department of Pathology and Molecular Pathology, University Hospital Zurich, Switzerland as previously described[53] with the approval of Cantonal scientific ethics committee Zurich, approval number KEK-ZH-No. 2014-0007, and with informed consent by the patient. Tumor regions were selected for heterogeneous histological presentation by an experienced uropathologist (PJW). DNA from peripheral blood and FFPE punches was isolated with the Maxwell 16 LEV Blood DNA kit (Promega, AS1290) and Maxwell 16 FFPE Tissue LEV DNA Purification Kit (Promega AS1130), respectively, according to manufacturer's recommendations; 300 µl of blood collected in a BD Vacutainer K2 (EDTA 18.0 mg) tube was added to 30 µl of Proteinase K solution (final concentration 2 mg/ml) and subsequently mixed with 300 µl lysis buffer, vortexed, and incubated for 20 min at 56 °C. FFPE cylinders were deparaffinised with xylene, washed twice with ethanol, dried 10 min at 37 °C and re-suspended in 200 µl incubation buffer containing 2 mg/ml Proteinase K. Samples were incubated overnight at 70 °C and mixed with 400 µl lysis buffer. Lysates from both, blood and FFPE tissues, were transferred to well 1 of the supplied cartridge of the corresponding kit and DNA was automatically purified and eluted in 30 µl Tris-buffer, pH 8.0 by the Maxwell instrument.

WES samples were profiled using Agilent SureSelect Whole Exome Enrichment, v6 (58 Mbp) and 2 × 75 bp paired-end reads were used for optimal performance on a HiSeq 4000 (Illumina). Mutation calling was followed by protocols established by TCGA and ExAC[21,54]. Reads were aligned to hg19 using BWA[55], and variants were called with GenomeAnalysisTK, MuTect[56], Picard MarkDuplicates, and additional post-processing utilities from GATK including BaseRecalibrator. FastQ files were deposited in EBI's ENA project PRJEB19193. Predicted mutations are given in Table S5; mutations were annotated with estimated read fractions and estimated CNAs by VarScan using default parameters and after setting the maximum amplification to 15×[47].

Profiling and analysis of targeted-panel sequencing of prostate cancer biopsies

We selected a total of 36 genes whose mutations are enriched in prostate cancers for ultra-deep sequencing[40,41]. ImmQuant was used to capture their coding regions and the completeness of the capture was verified against the human reference genome GRCh38. Target genes are given below.

AKT1	CDH1	MED12	PMS2
AR	CDKN1B	MRE11A	PTEN
ATM	CHEK2	MSH2	RAD51C
ATR	EP300	MSH6	RAD52D
AURKA	ERG	MYC	RB1
BARD1	EZH2	MYCN	SPOP
BRACA1	FOXA1	NBN	TMPRSS
BRACA2	GEN1	PALB2	TP53
BRIP1	HOBX13	PIK3CA	ZNF595

The samples were taken from the Prostate Cancer Outcomes Cohort Study from UZH (ProCOC) and Metastatic Prostate Cancer Biobank from UZH (metaProC). Patients and target genes were selected for ultra-deep

sequencing based on WES profiling and using our predictive panel. Their profiles were implemented in three batches and sequenced using Illumina HiSeq using 150 bp pair-end reads by Sophia Genetics. Analysis of DNA profiles mirrored the steps described for WES analysis. BAM files are freely available on ENA project PRJEB19193.

Copy numbers were estimated using Affymetrix OncoScan arrays and the OncoScan FFPE Assay Kit for detection of genome-wide copy number changes and loss of heterozygosity in FFPE samples. Oncoscan uses molecular inversion probe technology to query over 220,000 SNPs at carefully selected genomic locations, evenly distributed across the genome and with increased density within approximately 900 cancer or cancer-related genes. All samples underwent array hybridization and analysis and passed gel QC, as established during validation of the assay for clinical genetic analysis. Data were analyzed using Nexus Express.

Protein-expression profiling and analysis of prostate cancer biopsies

For the comprehensive quantitation of proteins in high throughput manner, we employed a data-independent acquisition (DIA) workflow compiled of the commercially available and standardized iST sample preparation kit (PreOmics GmbH), a Q Exactive HF (Thermo Fisher Scientific Inc.) and Spectronaut Pulsar data analysis software[57]. We have applied a combination of published and specifically generated spectral libraries. We have prepared and analyzed a selected five tissue samples from human prostate adenocarcinomas from the retrospective FFPE PC cohort from the University Hospital Zurich as described above[41]. Each specimen was analysed once.

First, the FFPE fixed material was deparaffinised and peptides were generated by using the iST 96× kit from PreOmics following the optimized protocol for FFPE punches. This was followed with DIA-MS analysis of 1 μg of total peptide mixture for each patient sample including iRT standard peptides for non-linear calibration, batch correction and large-scale data set merge. Then, we generated and applied spectral libraries through the merge of an in house generated library (Spectronaut Pulsar) and a commercial organ-specific library 311 FFPE prostate library (311_FFPE_prostate). The resulting library contained 8200 protein groups with a size of 65.5 mb. For quality control, the peptide yield was determined after protein extraction, digestion, and peptide clean up. MS signal intensity was monitored (TIC) during entire MS run as well as the performance and stability of the liquid chromatography according to a reference peptide set.

Reporting summary

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

DATA AVAILABILITY

BAM files are freely available without restrictions on ENA project PRJEB19193. All supplementary tables, including synthetic data, processed data, and analyses used to produce Fig. 3 are also given in the GitHub repository https://github.com/drugilsberg/chimaera.

CODE AVAILABILITY

The Chimæra web server is available at https://ibm.biz/chimaera-aas and source code is available at https://github.com/drugilsberg/chimaera.

REFERENCES

1. Futreal, P. A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
2. Higgins, M. E., Claremont, M., Major, J. E., Sander, C. & Lash, A. E. CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res.* **35**, D721–D726 (2007).
3. Ding, L. et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell* **173**, 305–320. e310 (2018).
4. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
5. Fidler, I. J. & Hart, I. R. Biological diversity in metastatic neoplasms: origins and implications. *Science* **217**, 998–1003 (1982).
6. Boutros, P. C. et al. Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.* **47**, 736–745 (2015).
7. Wang, Y. et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
8. Suzuki, H. et al. Mutational landscape and clonal architecture in grade II and III gliomas. *Nat. Genet.* **47**, 458–468 (2015).
9. Mann, K. M. et al. Analyzing tumor heterogeneity and driver genes in single myeloid leukemia cells with SBCapSeq. *Nat. Biotechnol.* 34, 962–972 (2016).
10. Gao, R. et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* **48**, 1119–1130 (2016).
11. Andor, N. et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22**, 105–113 (2016).
12. El-Kebir, M., Oesper, L., Acheson-Field, H. & Raphael, B. J. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* **31**, i62–i70 (2015).
13. Niknafs, N., Beleva-Guthrie, V., Naiman, D. Q. & Karchin, R. SubClonal hierarchy inference from somatic mutations: automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing. *PLoS Comput. Biol.* **11**, e1004416 (2015).
14. Espiritu, S. M. G. et al. The evolutionary landscape of localized prostate cancers drives clinical aggression. *Cell* **173**, 1003–1013 (2018).
15. Chu, W. K. et al. Ultraaccurate genome sequencing and haplotyping of single human cells. *Proc. Natl Acad. Sci.* **114**, 201707609 (2017).
16. Getz, G. & Ardlie, K. in *TCGA Second Annual Scientific Symposium* (eds Meyerson, M. & Shmulevich, I.) (National Institutes of Health, Crystal City, Virgina, 2012).
17. Cieslik, M. et al. The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res.* **25**, 1372–1381 (2015).
18. Wang, J. et al. Tumor evolutionary directed graphs and the history of chronic lymphocytic leukemia. *eLife* **3**, e02869 (2014).
19. Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
20. Kuipers, J., Jahn, K., Raphael, B. J. & Beerenwinkel, N. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res.* **27**, 1885–1894 (2017).
21. The Cancer Genome Atlas. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
22. The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
23. TCGA. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169**, 1327–1341 (2017).
24. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
25. Andor, N., Harness, J. V., Muller, S., Mewes, H. W. & Petritsch, C. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* **30**, 50–60 (2014).
26. Deshwar, A. G. et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16**, 35 (2015).
27. Malikic, S., McPherson, A. W., Donmez, N. & Sahinalp, C. S. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics* **31**, 1349–1356 (2015).
28. McInnes, L., Healy, J. & Astels, S. hdbscan: hierarchical density based clustering. *J. Open Source Softw.* **2**, 205 (2017).
29. Fritz, H., García-Escudero, L. A. & Mayo-Iscar, A. tclust: an r package for a trimming approach to cluster analysis. *J. Stat. Softw.* **47**, 1–26 (2012).
30. Fraley, C. & Raftery, A. E. *MCLUST Version 3: An R Package for Normal Mixture Modeling and Model-based Clustering* (DTIC Document, 2006).
31. Zhao, X. & Sandelin, A. GMD: measuring the distance between histograms with applications on high-throughput sequencing reads. *Bioinformatics* **28**, 1164–1165 (2012).
32. Lin, D.-C. et al. Genomic and epigenomic heterogeneity of hepatocellular carcinoma. *Cancer Res.* **77**, 2255–2265 (2017).
33. Roy, A. et al. Integration of whole transcriptome sequencing into the genomic analysis of pediatric solid tumors: early experience and challenges. *J. Mol. Diagn.* **16**, 754–755 (2014).
34. Li, C.-M. et al. CTNNB1 mutations and overexpression of Wnt/β-catenin target genes in WT1-mutant Wilms' tumors. *Am. J. Pathol.* **165**, 1943–1953 (2004).
35. Pelletier, J. et al. WT1 mutations contribute to abnormal genital system development and hereditary Wilms' tumour. *Nature* **353**, 431 (1991).
36. Liu, C. et al. Control of β-catenin phosphorylation/degradation by a dual-kinase mechanism. *Cell* **108**, 837–847 (2002).
37. Kim, H.-S. et al. Identification of novel Wilms' tumor suppressor gene target genes implicated in kidney development. *J. Biol. Chem.* **282**, 16278–16287 (2007).
38. Rubinfeld, B. et al. Stabilization of β-catenin by genetic defects in melanoma cell lines. *Science* **275**, 1790–1792 (1997).

39. Wang, Y. et al. Mutation spectrum of genes associated with steroid-resistant nephrotic syndrome in Chinese children. *Gene* **625**, 15–20 (2017).

40. Guo, T. et al. Multi-region proteome analysis quantifies spatial heterogeneity of prostate tissue biomarkers. *Life Sci. Alliance* **1**, e201800042 (2018).

41. Zhong, Q. et al. A curated collection of tissue microarray images and clinical outcome data of prostate cancer patients. *Sci. Data* **4**, 170014 (2017).

42. Shi, W. et al. Reliability of whole-exome sequencing for assessing intratumor genetic heterogeneity. *Cell Rep.* **25**, 1446–1457 (2018).

43. Robinson, D. et al. Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).

44. Doane, A. et al. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. *Oncogene* **25**, 3994 (2006).

45. Iwanaga, K. et al. Pten inactivation accelerates oncogenic K-ras–initiated tumorigenesis in a mouse model of lung cancer. *Cancer Res.* **68**, 1119–1127 (2008).

46. Bruins, W. et al. The absence of Ser389 phosphorylation in p53 affects the basal gene expression level of many p53-dependent genes and alters the biphasic response to UV exposure in mouse embryonic fibroblasts. *Mol. Cell. Biol.* **28**, 1974–1987 (2008).

47. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

48. Krzanowski, W. J. & Lai, Y. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* **44**, 23–34 (1988).

49. Kovács, F., Legány, C. & Babos, A. in *6th International symposium of hungarian researchers on computational intelligence*. (Citeseer, 2005).

50. Celeux, G. & Govaert, G. Gaussian parsimonious clustering models. *Pattern Recognit.* **28**, 781–793 (1995).

51. Sheppard, D., Terrell, R. & Henkelman, G. Optimization methods for finding minimum energy paths. *J. Chem. Phys.* **128**, 134106 (2008).

52. Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

53. Mortezavi, A. et al. KPNA2 expression is an independent adverse predictor of biochemical recurrence after radical prostatectomy. *Clin. Cancer Res.* **17**, 1111–1121 (2011).

54. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

55. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

56. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

57. Bruderer, R. et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteom.* **14**, 1400–1410 (2015).

AUTHOR CONTRIBUTIONS
M.M., D.W.P., P.J.W., M.R.M., and P.S. concieved and designed the project. M.M., M.R. M., and P.S. lead the project's implementation. M.M., H.R.M., R.M., P.C., A.P., and P.S. performed data analyses and designed methods. D.R., L.D.V.R., U.W., K.O., K.S., and P.J. W. lead efforts of collecting, profiling, and analyzing prostate cancer samples. A.R. and D.W.P. lead efforts of collecting, profiling and analyzing Wilms' tumor samples, and M.M., B.S., J.S.Z., and P.S. analyzed regulatory networks and altered pathways.

COMPETING INTERESTS
The authors declare no competing interests.

ADDITIONAL INFORMATION
Supplementary information is available for this paper at https://doi.org/10.1038/s41540-020-00147-5.

Correspondence and requests for materials should be addressed to P.J.W., M.R.M. or P.S.

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

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