PrECISE

D1.2

Final clone inference

Project number:	668858
Project acronym:	PrECISE
Project title:	PrECISE: Personalized Engine for Cancer Integrative Study and Evaluation
Start date of the project:	1st January, 2016
Duration:	36 months
Programme:	H2020-PHC-02-2015

Deliverable type:	OTHER
Deliverable reference number:	PHC-668858 / D1.2 / 1.0
Work package contributing to the deliverable:	WP 1
Due date:	December 2018 – M36
Actual submission date:	20th December, 2018

Responsible organisation:	BCM
Editor:	Pavel Sumazin
Dissemination level:	PU
Revision:	1.0

Abstract:	Final Clone Inference
Keywords:	copy number alterations, tumour phylogenies, clonal classification of tumours

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 668858.

This work was supported (in part) by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0324-2. The opinions expressed and arguments employed therein do not necessarily reflect the official views of the Swiss Government.

Editor

Sumazin, Pavel (BCM)

Contributors (ordered according to beneficiary numbers)

Dorothea Rutishauser (USZ)

Jelena Cuklina (ETH)

Peter Wild (UZH)

Dorothea Rutishauser (UZH)

Matteo Manica (IBM)

Executive Summary

Knowledge about tumour clonal evolution can help interpret the function of genetic alterations by pointing out initiating events and events that contribute to the selective advantage of proliferative, metastatic, and drug-resistant tumour subclones. Clonal evolution can be reconstructed from estimates of the relative abundance (*frequency*) of subclone-specific alterations in tumour biopsies, which, in turn, informs on the cellular composition of each tumour subclone. 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 dramatically alter mutation-frequency estimates and thus affect efforts to reconstruct tumour phylogenies. Our analysis suggests that a detailed accounting of CNAs is required for accurate mutation frequency estimates, and that such accounting is impossible for many cancer types using molecular profiling of one biopsy per tumour. Instead, we propose an optimization algorithm, *Chimaera*, to account for the effects of CNAs using profiles of multiple biopsies per tumour. Analyses of simulated data and profiles of a prostate cancer, hepatocellular carcinoma, and Wilms' tumours suggest that Chimaera estimates are consistently more accurate than previously proposed methods, resulting in improved phylogeny reconstructions, and the discovery of recurrent initiating mutations and key tumourigenesis events. We used Chimaera to analyse prostate cancer profiles, identifying mutations associated with initiating tumour subclones as well as tumour subclones that respond or are resistant to therapies.

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

Pan-cancer tumour profiling has identified recurrent alterations that are associated with tumour etiology at the loci of thousands of genes but the interpretation of genetic alterations remains a major challenge (Ding et al., 2018; Futreal et al., 2004; Higgins et al., 2007). Knowledge about the clonal evolution of tumours can point to genetic alterations that both contribute to tumorigenesis, indicate prognostically relevant intra-tumour variability, and point to refractory tumour subclones (Espiritu et al., 2018; Fidler et al., 1982; Nowell, 1976). Specifically, clonal evolution—depicted as a phylogenetic tree in Figure 1A—can help identify alterations that play a role in tumour initiation as well as those that confer a selective advantage to altered tumour cells. Moreover, information about its subclone composition is important for predicting the cancer's potential for drug resistance and metastasis, which vary across tumour subclones (Boutros et al., 2015) and are the key determinants of patient survival. Consequently, tumour-subclone characterization is essential for designing personalized therapies that target all tumour subclones and may hold the key to predicting tumour progression, drug sensitivity, and patient outcome.

Current methods that rely on DNA-profiling to reconstruct clonal evolution of tumours can be classified into two categories: methods that primarily rely on single-cell profiles (Gao et al., 2016; Mann et al., 2016; Suzuki et al., 2015; Wang et al., 2014b) and those that computationally resolve mixtures of subclones from molecular profiles of tumours, i.e. profiles of pools of cells that originate from a common malignant lesion (Andor et al., 2016; El-Kebir et al., 2015; Espiritu et al., 2018; Niknafs et al., 2015). Single-cell DNA sequencing can produce more definitive estimates of the proportion of tumour cells that contain each genetic alteration (*alteration frequencies*) and more complete profiles of tumour subclones, including information about the co-occurrence of alterations within each subclone. Its primary disadvantage is operational—the availability of high-quality tumour samples that permit single-cell isolation and profiling, and the accuracy and cost associated with parallel sequencing DNA from a multitude of cells per tumour. Alternatively, single-cell RNA sequencing or protein profiling can be used to define tumour subclones, but these do not directly point to key driving genetic alterations. Moreover, the accuracy of single-cell mutation profiling is an issue due to limited material availability in single cells(Chu et al., 2017), and this is not likely to improve as future sequencing technologies focus on profiling formalin-fixed paraffin-embedded (*FFPE*) tumour samples (Cieslik et al., 2015; Getz et al., 2012).

Focusing on single-nucleotide somatic variants (*SNV*s; or simply mutations), we sought to reconstruct clonal evolution from mutation profiles of genetically unstable cancers. This entails deconvolving mutation frequencies, alteration-subclone associations, and CNAs from molecular profiles—including whole-exome sequencing (*WES*) assays—that produce average estimates across cellular ensembles (Figure 1B). One approach to improving the accuracy of such deconvolutions is to profile multiple biopsies from the same tumour across time points (Wang et al., 2014a) or across regions (Boutros et al., 2015; Gundem et al., 2015). This approach relies on the assertions that genetic alterations that are specific to the same tumour subclone are expected to co-occur with the same frequency across biopsies and that the clonal composition across time or heterogeneous regions varies; i.e. multiple sampling will allow for the aggregation and deconvolution of the frequencies of most mutations with improved power. It's important to note that mutations that underwent convergent evolution (Kuipers et al., 2017) will not be aggregated with other mutations form the same tumour subclone because of differing frequency estimates across biopsies.

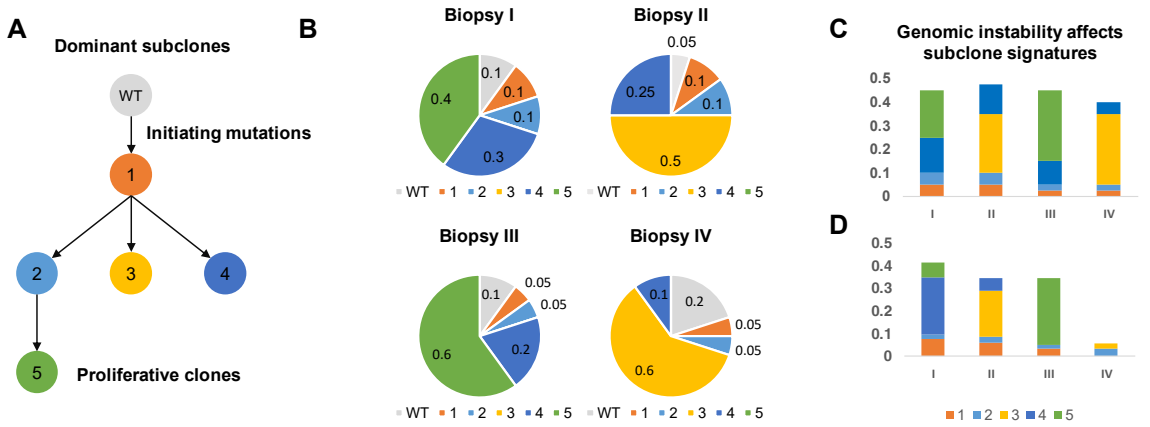

Figure 1. Footprint of clonal evolution across tumour biopsies.

(**A**) Tumour phylogeny composed of five dominant tumour subclones and wildtype (WT) cells—with no somatic mutations—that make up the cellular composition of four tumour biopsies (**B**). Subclones 3 and 5 were more proliferative—i.e. the proportion of these subclones (cellularity) in containing biopsies is greatest. (**C**) Failure to account for genomic instability can skew cellularity estimates because fractions of reads (mutated-read fractions) presenting each mutation in WES depend on the copy numbers of the alleles in both mutated and non-mutated cells. Consequently, in genomically-stable tumours, biopsies from (B) will have mutated-read fractions that differ from those of (**D**) genomically unstable tumours with the same mutation frequencies.

A central challenge for aggregating and estimating mutation frequencies in tumours with unstable genomes is accounting for the effects of CNAs that can alter *mutated-read fractions*—the frequencies of observed alternative alleles in the profiling assay—by altering contributions from reference alleles in mutation-free cells as well as both alternative and reference alleles in mutated cells (Figure 1C). In turn, errors in mutation-frequency estimates can prevent accurate phylogeny reconstructions (Figure 2). Our approach was to introduce a model for the effects of CNAs on mutated-read fractions. We use this model as a basis for simulations with CNA distributions that are compatible with observations from The Cancer Genome Atlas (TCGA)-profiled primary breast, HCCs, PCs, and Wilms' tumours (TCGA, 2017; The Cancer Genome Atlas, 2012; The Cancer Genome Atlas, 2015).

Data were simulated using synthetically generated phylogeny, including CITUP phylogenies (Malikic et al., 2015), followed by the duplication or loss of sequencing reads according to simulated effects of CNVs. Several methods are available in the literature to estimate mutation frequencies and clonal compositions. ABSOLUTE (Carter et al., 2012) infers tumour 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, with components centred at the discrete concentration-ratios implied by an initial frequentist estimation. AncesTree (El-Kebir et al., 2015) provides a combinatorial characterization of the clonal evolution of a tumour by assuming than in an error-free data mutations can be described by a perfect phylogeny matrix, which is found using integer linear programming; the problem is extended to real data using a probabilistic model for errors. EXPANDS (Andor et al., 2014) 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 (Deshwar et al., 2015) reconstructs phylogenies based on a model for simple somatic mutations in addition to a correction for CNAs, all based on a single biopsy per tumour. SCHISM (Niknafs et al., 2015) 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 algorithms is then used to build phylogenetic trees.

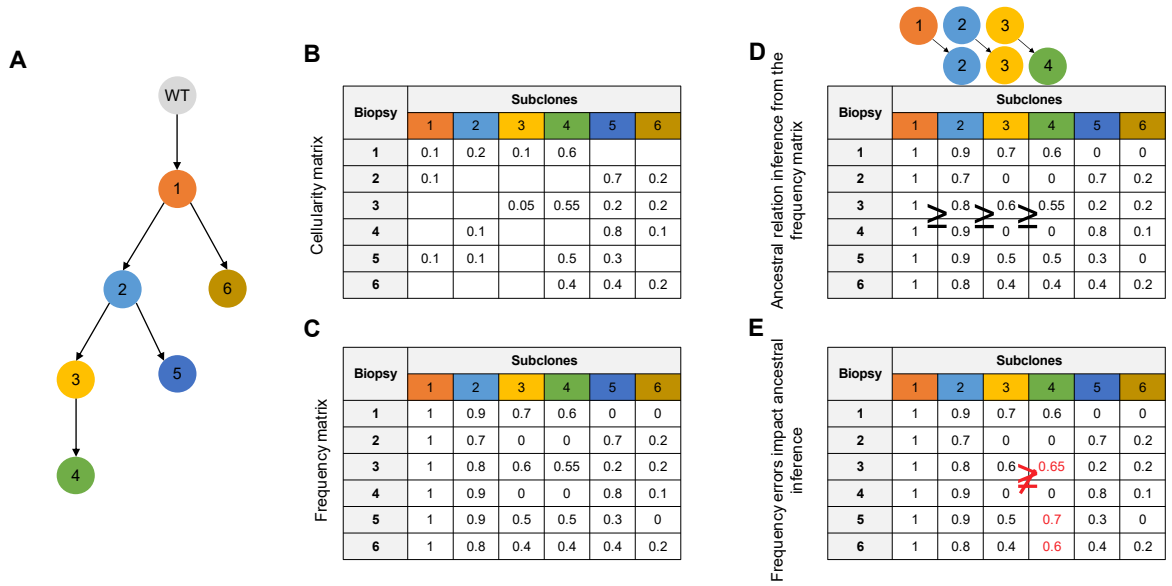

Figure 2. Small variations in mutation frequency estimates can impact the inference of ancestral relations.

(**A**) Simulated tumour phylogeny, (**B**) subclone cellularities, and (**C**) frequencies of subclonal mutations across biopsies. (**D**) Ancestral relations between subclones can be inferred from comparisons of their frequency vectors: Subclone 4 frequencies are greater than those of subclone 3 across all biopsies, but (**E**) errors in frequency estimates (red) can violate this relationship and complicate tumour-phylogeny reconstruction efforts.

Attempts to estimate the frequencies and cellularities of mutations and subclones using ABSOLUTE, AncesTree, EXPANDS, PhyloWGS, and SCHISM revealed variable success rates, with some methods showing consistently poor accuracy. EXPANDS and PhyloWGS, which were designed for phylogeny reconstruction using profiles of one biopsy per tumour, and ABSOLUTE, which is best known and most effective for estimating tumour purity, had consistently poor accuracy in our simulations. While SCHISM and AncesTree, which do not explicitly account for the full range of observed CNAs in tumours, were less accurate on simulations with CNAs. Like PhyloWGS, we concluded that explicit accounting for CNAs is required in order to approximate mutation frequencies accurately. However, more than one biopsy per tumour are required for accurately approximating mutation frequencies and CNAs at mutated loci.

To address this challenge and improve mutation-frequency and CNA estimations from WES of tumours with genetic instability, we developed Chimaera: clonality inference from mutations across biopsies. Chimaera relies on multiple biopsies for the same tumour to, first, approximate CNAs and mutation frequencies; then, identify mutations with similar approximate frequencies and associate them with subclones; and, finally, to estimate the true frequencies of these mutations and the associated subclones. As is the case for estimates made by SCHISM, ABSOLUTE and other methods, Chimaera is not able to produce frequency estimates for all mutations, but compared to existing methods is able to process and determine true frequencies for more variants, exhibiting more power in identifying potentially tumour initiating mutations and disease drivers. Finally, to demonstrate that Chimaera is able to reconstruct subclones from tumour profiles we produced Chimaera-inferred subclones and resulting phylogeny from profiles of 6 castration-resistant prostate cancer (CRPC) patients and a set of profiles extracted from 5 different tumour areas from 10 hepatocellular carcinoma (HCC) patients (Lin et al., 2017); 3 Wilm's tumour patients; and 5 CRPC patients that were profiled at multiple time points

using ultra-high resequencing. Results on our effort to identify prostate cancer subclones that are predictive of response to therapies are reported in WP2.

Focusing on single-nucleotide somatic variants (*SNV*s; or simply mutations), we describe the clonality problem as that of associating mutations with subclones and inferring ancestral relations between subclones. The goal of the resulting set-theoretic formulation, for each tumour, is 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 tumour subclone can be associated with a frequency vector that describes the proportion of cells containing its mutations in each biopsy. Establishing ancestral order between two subclones then depends on (probabilistic) comparisons between their corresponding mutation frequencies.

Our first challenge was to estimate mutation frequencies across biopsies. In cellular environments with stable genomes, where CNAs are few, accurate mutation-frequency estimation is a function of allele coverage. Mutation frequencies can be computed directly from sequencing evidence for the mutated allele—the fraction of reads (*mutated-read fraction*) that support the mutation as observed in sequencing data. However, CNAs can affect mutation frequency estimates because the mutated-read fraction is affected by contributions from alleles in mutation-free cells as well as both the mutated and wildtype forms of the allele in mutated cells. Changes to the copy number of one of these allele contributors can alter the mutated-read fraction dramatically, and, if not accounted for, will result in inaccurate mutation-frequency estimates (Figure 1C). These errors, in turn, prevent accurate phylogeny reconstructions (Figure 2). While our approach may not be feasible for all tumour types, it is a natural fit for high-risk patients with blood malignancies and some solid tumours, including hepatocellular carcinomas (*HCCs*), prostate carcinomas (*PCs*), and Wilm's tumours.

Chapter 2 Methods

We begin by formulating the phylogeny reconstruction problem in set-theoretic terms, which leads to a natural model for the effects of CNVs on mutated-read fractions in WES. We describe our methodology for simulating WES tumour profiles, as well as our efforts to deconvolve mutation frequencies from simulated data using ABSOLUTE, AncesTree, EXPANDS, SCHISM, and Chimaera. Finally, to demonstrate that Chimaera can be effectively applied to clinical data, we describe a reconstructed phylogeny from WES profiles of ten same-tumour CRPC biopsies and a set of five same-tumour HCC biopsies from nine patients.

2.1 Phylogeny reconstruction problem

Let $M = \{m: m \in \mathbb{N}, 1 \leq m \leq n\}$ 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 tumour 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.

2.2 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 (Figure 3). Notice that if no copy number event has occurred at the locus $m : \delta^s = 2$, $\delta_w^s = 1$ and $\delta_m^s = 1$. Adopting the infinite-sites assumption, we denote the mutated-read fraction—the fraction of reads reflecting the mutated versus wildtype allele in a WES profile—in sample s as f_m^s. Then, we can formulate the following equations (Eq. 1 and Eq. 2).

$$C^s = \delta^s\left(1 - \varphi_m^s\right) + (\delta_w^s + \delta_m^s)\varphi_m^s \qquad \text{Eq. 1}$$

$$f_m^s = \frac{\varphi_m^s \delta_m^s}{C^s} \qquad \text{Eq. 2}$$

Eq. 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 coming from the mutated allele to the total number of reads.

2.3 Chimaera

Chimaera proceeds in three steps. First, mutation frequencies are approximated from sequencing and CNV 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 CNVs for these alleles are refined using an optimization process. The optimization assumes that all clustered mutations that are associated with the same subclone have the same frequency in each tumour biopsy and that δ_m^s—the average copy number of the m-mutated allele—is the same across all biopsies from the same tumour.

A first approximation. We first approximate the true frequency of the mutation φ_m^s by accounting for tumour purity, i.e., the fraction of tumour cells in the biopsy versus normal cells,

and assuming that the allele's average copy number in tumour cells—whether mutated or not—is fixed. Let p^s be the purity of biopsy s, then Eq.2 can be rewritten as follows:

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

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

$$C_{obs}^s = 2(1-p^s) + C^s p^s \qquad \text{Eq. 4}$$

where C_{obs}^s can be estimated using additional biochemical assays, genetic sequencing, or through computational analysis of WES data (Koboldt et al., 2012), 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 of the mutated allele in tumour 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 $\widetilde{\varphi_m^s}$:

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

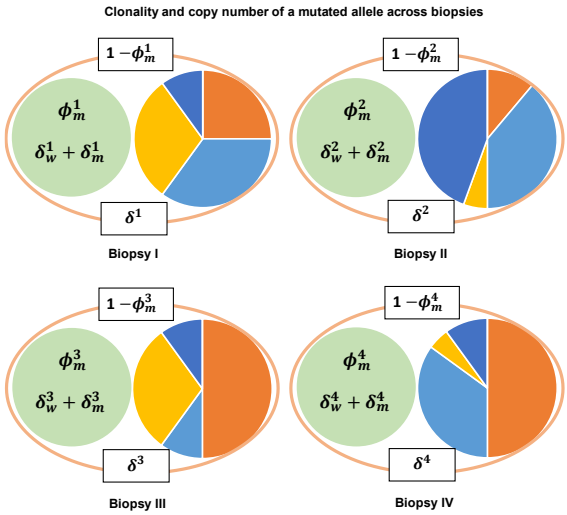

Figure 3. Our mutation-centric model for the effects of CNVs on mutated-read fractions in WES. In each biopsy s, the mutated-read fraction is a function of the true mutation frequency φ_m^s and (1) the copy numbers of the allele in all profiled cells—tumour and WT—that lack this mutation, δ^s, and (2) the copy numbers of the wildtype and the mutated allele in tumour cells with the mutation, δ_w^s, δ_m^s.

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 takes into account the copy number influence from WES measurements.

Subclone reconstruction. The approximate mutation frequency vectors (Eq. 5) are next clustered to identify candidate groups of mutations that form subclones. We consider clustering algorithms with robust treatment of outliers in order to ensure good clustering stability and quality. Specifically, we use *hdbscan* (McInnes, et al., 2017), a density-based hierarchical clustering method that aims at maximising 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 has to be considered to constitute a cluster. We also use tclust (Fritz et al., 2012), a non-hierarchical robust clustering that trims outliers based on a probabilistic model. The number of clusters is selected by optimizing intra-cluster entropy or the sum of square errors (SSE), and using a variety of optimization methods including the Elbow method, Gaussian mixture decomposition (GMD), and SD index (Celeux et al., 1995; Kovács et al., 2005; Krzanowski et al., 1988). The clustering based on *hdbscan* showed better performance on the generated synthetic data compared to others, especially when considering the number of mutations processed. Furthermore, it has the advantage of avoiding imposing a prior distribution on the mutations frequencies. Once the clusters are found, Chimaera 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 optimisation step.

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

$$\varphi_m^s \delta_m^s = f_m^s \frac{C_{obs}^s}{p^s} \equiv \mathcal{B}_{ms}, \qquad \forall\ m \in \gamma, s \in S. \qquad \text{Eq. 6}$$

where, $\mathcal{B}_{ms}$ is the entry of a matrix $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{|S|,|\gamma|}$ corresponding to mutation m and biopsy s. $\boldsymbol{\mathcal{B}}$ 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 –cannot be analytically decoupled. However, mutations from the same subclone occur in cells with shared evolutionary history, and thus are expected to show similar mutation frequencies, i.e. $\varphi_{m_i}^s = \varphi_{m_j}^s \equiv \varphi^s\ \forall\, m_i, m_j \in \gamma$. Notice that the same mutations m_i, m_j may have different frequencies in a different biopsy, as the subclones identified in different biopsies are not constraint to descend from the same ancestral parent. 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 and WES analysis. 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. 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 $\gamma \in \mathcal{P}(M)$, based on Eq. 6, can be formulated as:

$$min \left\| \overrightarrow{\varphi^s} \otimes \overrightarrow{\delta_m} - \boldsymbol{\mathcal{B}} \right\|_2 ; 0 \leq \delta_m \leq CN, 0 \leq \varphi^s \leq 1, \forall m \in \gamma, \forall s \in S. \quad \text{Eq. 7}$$

where $\overrightarrow{\varphi^s}$ is the mutation frequency vector across biopsies for all mutations in γ; $\overrightarrow{\delta_m}$ is the copy-number vector for each mutation in γ; $\boldsymbol{\mathcal{B}}$ is as defined in Eq. 6; and $\overrightarrow{\varphi^s} \otimes \overrightarrow{\delta_m}$ denotes the outer product of vectors $\overrightarrow{\varphi^s} \in \mathbb{R}^{|S|}$ and $\overrightarrow{\delta_m} \in \mathbb{R}^{|\gamma|}$. We used Sequential Least Squares Programming (SLSQP) optimization (Sheppard et al., 2008) to find an optimal solutions of Eq. 7, where multiple runs with multiple initializations are used to avoid being trapped in bad local optima.

Chapter 3 Simulation of WES data

WES simulations were based on phylogenies and associated cellularity matrices that describe ancestral relations between 6 to 12 subclones. These were either generated by us (see Figures 4A, B) or adapted from CITUP. Each subclone was associated with 20 to 50 somatic mutations, and each somatic mutation was associated with a trio of copy numbers—δ^s, δ_w^s, and δ_m^s—that were taken from truncated normal distributions with means μ∈{1,2,3}, where μ=1 corresponds to no copy number changes, and standard deviation σ∈{0,1,2,3}; σ=0 was used only when μ=1. The resulting copy numbers model a range of genetic instability conditions that was in line with observed copy number changes in PRAD and BRCA tumours (Figure 4C, D). We assumed no linkage between simulated CNVs of any mutations. In addition, we added up to 10% of wildtype reads for all simulated mutations to account for the potential inclusion of non-tumour cells in the assay (WT subclone in Figure 1A). Total coverage for each allele—i.e. the number of reads covering both wild-type and mutated genetic position —was taken by sampling mutation coverage values from our CRPC tumour biopsies. Finally, once idealized counts were available for both mutated and wild-type alleles, noise was added to simulate duplication or loss of up to 5% of the observations according to a uniform distribution. Each simulation was repeated to produce six biopsies per tumour using a distinct cellularity vector for each biopsy (as depicted in Figure 4A, B). The availability of six biopsies per tumour increases the likelihood that mutations can be aggregated and subclone mutation frequencies can be compared to infer ancestral relations. We note that while our CRPC tumour was profiled at ten regions, setting a six-biopsy minimum will exclude the profiling of many tumour types using our methods; this was a compromise between clinical feasibility and power to infer mutation frequencies and phylogenies.

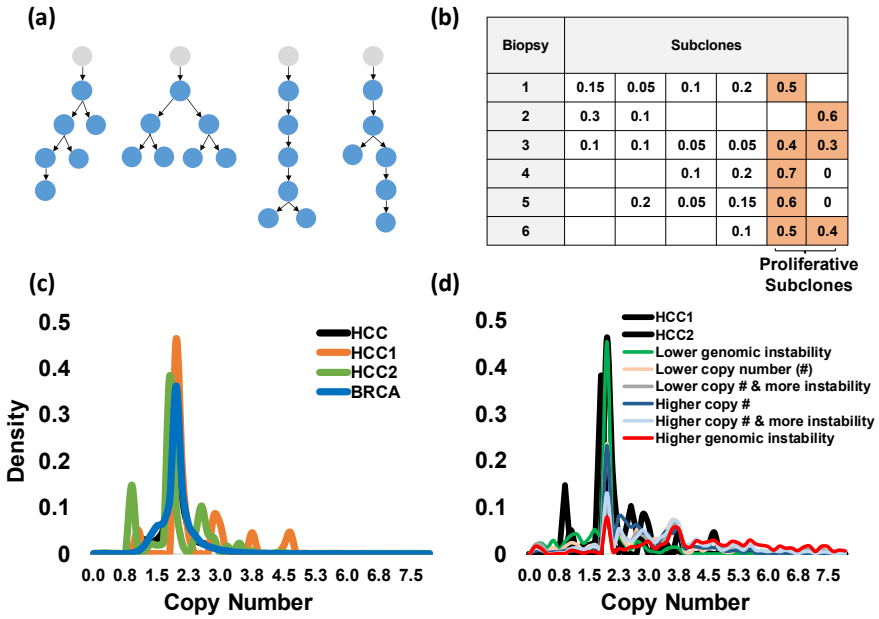

Figure 4. Our synthetic data generation and a comparison of simulated CNV distributions to those that were observed in tumours.

(A) Representative phylogenies and **(B)** a representative cellularity matrix. **(C)** Density plots of average copy numbers across profiles of TCGA prostate (PRAD) and breast (BRCA) tumours, and our CRPC tumour. PRAD1 and PRAD2 show genome-wide CNV distributions in each of two PRAD tumours, while PRAD and BRCA distributions are taken across genes and tumours; CNVs ranged from 0 to >260x. **(D)** Simulated CNVs ranged from 0 to 15x.

Chapter 4 Accuracy of mutation-frequency estimation on simulated data

We compared the accuracy of EXPANDS, ABSOLUTE, SCHISM, AncesTree, and Chimaera on simulated data, as described in Methods. Phylogeny reconstruction success and clonality-inference accuracy by EXPANDS and ABSOLUTE were the lowest. EXPANDS relies on single biopsies, and when evaluated on phylogenies that were composed of as few as 3 tumour subclones, EXPANDS-reconstructed phylogenies from profiles of same-tumour biopsies (both simulated and collected from the clinic including the CRPC reported on here) had few common ancestral inferences and performance was poor in every tested simulated instance. In contrast, SCHISM-reconstructed phylogenies from synthetic constructions with 3 tumour subclones were accurate in 100% of tested instances. ABSOLUTE can process profiles of multiple biopsies per tumour and has good accuracy for inferring tumour purity in our synthetic data. However, when using default parameters, errors in ABSOLUTE frequency-inferences were more than double those of SCHISM. Parameter optimization through human intervention consistently improved its accuracy, but it remained less accurate than SCHISM. Moreover, the degree of human intervention that this required was not compatible with large-scale benchmarking. Consequently, we focused on accuracy comparisons between inferences by SCHISM, AncesTree, and Chimaera (as given in Figure 5), and excluded EXPANDS and ABSOLUTE from further analyses.

AncesTree accepts no external input when estimating mutation frequencies, but SCHISM and can be guided by externally-inferred mutation clusters. SCHISM's implementation includes its own selected clustering methods, and these were also used to compare accuracy. We clustered mutations with tclust based on five optimization methods: ElbowSSE, Entropy, GMD, Mclust, and SDIndex. We compared the accuracy of methods and pipelines on 2000 simulated assays, including both simulated assays with and without modelled genetic instability (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 those included in SCHISM's implementation. Comparatively, Chimaera estimates were less dependent on clustering methods and significantly outperformed estimates by SCHISM with SDIndex ($p<1E-16$ by U test); Chimaera using *hdbscan* exhibited lower performance compared to other Chimaera runs (Figure 5A) but showed an increasing power in the percentage of mutations used to estimate the frequencies (Figure 5C).

Inference accuracy, for both SCHISM and Chimaera, was anti-correlated with the level of genetic instability, which followed truncated normal distributions with varying means and variances (Figure 5B). To better understand mutation-level behaviour, as opposed to the genome-level comparisons in Figure 5C, we rescued individual mutations from each simulation and compared accuracy, mutation by mutation, as a function of their simulated copy numbers (Figure 5D). The result suggests similar Chimaera accuracy across copy numbers. We note that many mutations were eliminated from the evaluation by both the SCHISM and Chimaera pipelines with *tclust* based methods. In total, only ~60% of mutations were assigned frequencies by Chimaera with *tclust* and SCHISM; on the contrary Chimaera is these proportion were independent of mutation copy numbers. While Chimaera assigned frequencies to all clustered mutations, SCHISM did not successfully estimate mutation frequencies for some simulated genomes. Accuracy comparisons in Figure 5 were made using only those mutations that had assigned frequencies by all methods.

In its totality, our analysis suggests that, at least under our model, mutation frequency estimation is more challenging for genomes with high copy-number variability. Chimaera inference accuracy for simulated genomes where all mutations had consistently low or consistently high copy numbers was relatively high. This is in part due to Chimaera's iterative process, where success in mutation clustering is followed by an optimization process that can correct for consistently high or consistently low mutation copy numbers.

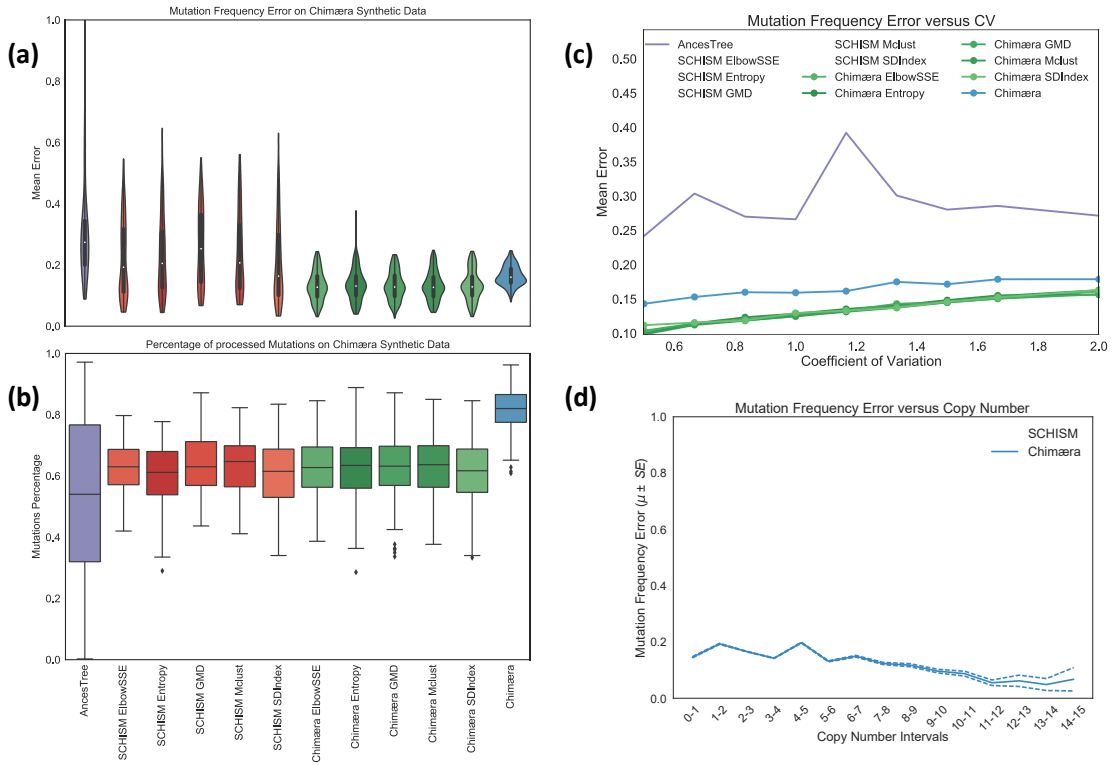

Figure 5. Accuracy on simulated data.

(**A**) Accuracy of mutation-frequency estimates by AncesTree (purple), SCHISM (red) and Chimaera (green and blue) on simulated WES data from genomes with varying mutation copy numbers; SCHISM and Chimaera were evaluated using multiple clustering methods with SDIndex (SCHISM) and ElbowSSE (Chimaera) producing top accuracy, respectively, in blue are reported estimates for Chimaera using *hdbscan*. (**C**) Percentage of mutations processed applying the three different algorithms. It is evident how Chimaera using *hdbscan* outperforms clearly other methods, being able of considering over 80% of the mutations considered. (**B**) Accuracy was inversely correlated with genetic instability, which was measured here as the coefficient of variation of the distributions used to simulate CNVs in each simulated WES profile; SCHISM with SDIndex clustering outperformed AncesTree inferences. (**D**) Evaluated independently, mutation copy numbers had relatively little effect on Chimaera accuracy. We report results for Chimera using *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.

Chapter 5 Clonality analysis

To test our methods, we analysed profiles from HCC and CRPC samples. We report on our HCC analysis below, and summarize analysis of our CRPC samples. More in depth treatment of CRPC subclones will be available in D1.3 and D2.4.

HCCs are high-risk liver tumours that are known to have high genetic instability (TCGA, 2017). To test our ability to infer mutation frequencies and ancestral relations between subclones using clinical data, we studied the profiles of nine HBV-positive HCC patients, with each tumour profiled in 5 areas (Lin et al., 2017). In total, we obtained mutated-read fractions and CNV estimates for 1,424 mutation candidates in 9 tumours and 43 tumour samples, while 7 tumours were profiled in 5 areas each, profiles from only 4 areas of tumours HCC5647 and HCC8716 passed quality control.

Chimaera inferred frequencies estimates for 60% (858/1424) of all mutations, reconstructing phylogenetic trees for each tumour sample and predicting initiating clones and clones that are associated with a proliferative advantage; see representative trees in Figure 6. To compare, SCHISM inferred mutations frequency for only 10% of mutation candidates. Interestingly, 78% (7/9) of the tumors included predicted initiating mutations in WNT-signalling pathway genes. An examination of 102 TCGA-profiled HBV-positive HCCs (TCGA, 2017) suggested that 74% (75/102) of samples carried mutations in WNT-signalling pathway genes, and the majority of these samples (76%) had WNT-signalling pathway mutations with mutated-read fractions above 25%—corresponding to mutations that are potentially present in the majority of cells. To test whether WNT-signalling pathway genes were enriched for mutations—and particularly mutations with mutated-read fractions above 25%—we calculated the proportion of tumours with such mutations in each of 186 KEGG pathways in MSigDB (Liberzon et al.). The top 10 pathways by p-value and mutated-sample fraction is given in Table 1. 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-singling was our top pathway for enrichment of mutations with mutated-read fractions above 25% or for any mutated-read fraction. To correct for the shadow effect (Roy et al., 2014), where pathways that overlap a pathway that is mutated in many samples are also significant, we recalculated enrichment significance for each pathway after excluding

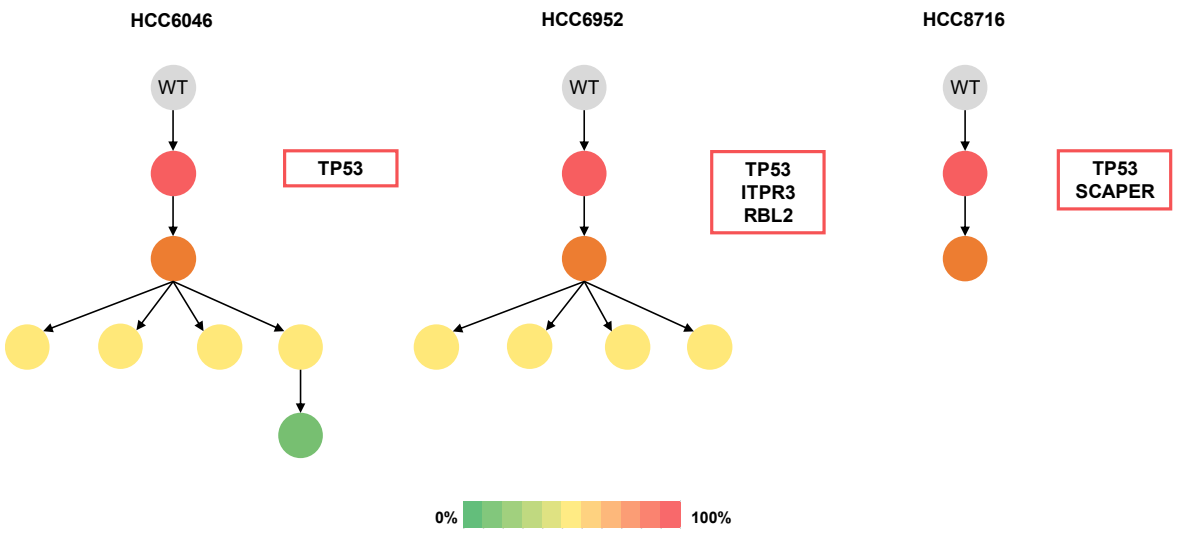

Figure 6: Predicted phylogenies in HBV-positive HCCs

Table 1: The 10 most enriched pathways for mutations with mutated-read fractions greater than 25% (high-frequency mutations) in TCGA-profiled virus-positive HCCs. Pathways were sorted by p-values followed by the proportion of patients with a high-frequency mutations in at least one pathway gene. P-values were estimated using permutation testing based on all expressed genes in 186 KEGG pathways; here, for each pathway and given the number of pathway genes (Genes, in the table below), each permutation test selected that number of genes uniformly at random and calculated the fraction of patients with a mutation in one of these genes. The same test was conducted after excluding WNT-signalling genes to establish independence from WNT-pathway signalling.

Enriched KEGG pathways	Genes	Patients with mutations	P-value	After removing WNT-signaling genes	
KEGG WNT SIGNALING PATHWAY	151	57%	0.001	**Frequency**	**P-value**
KEGG PROSTATE CANCER	89	54%	0.001	23%	>0.1
KEGG COLORECTAL CANCER	62	52%	0.001	12%	>0.1
KEGG ENDOMETRIAL CANCER	52	51%	0.001	17%	>0.1
KEGG BASAL CELL CARCINOMA	55	50%	0.001	9%	>0.1
KEGG CALCIUM SIGNALING PATHWAY	178	60%	0.002	56%	0.002
KEGG ECM RECEPTOR INTERACTION	84	44%	0.003	44%	0.004
KEGG PATHWAYS IN CANCER	328	68%	0.016	51%	>0.1
KEGG MAPK SIGNALING PATHWAY	267	64%	0.021	58%	0.020

WNT-signalling pathway genes, and note that MAPK-signalling and 2 other top-10 pathways were still enriched (Table 1).

CRPC analysis. In total, 72 CRPC samples corresponding to 5 CRPC patients were profiled, including a control for each patient. Profiled samples, including profiling time points, Gleason scores, and therapeutic history are given in Table S1[1]. Our analysis were used to infer tumour phylogenies, including the order of the emergence of dominant tumour subclones in these 5 patients (Patient 1-5).

Patient 1 was profiled at 3 time points across 2.3 years, and was assessed mutually disjoint mutations in *EP300* and *AR*, corresponding to 2 distinct proliferative tumour subclones. The clonal *EP300* mutation (p.I997V) is predicted to be deleterious and was inferred to be present in the majority of tumour cells at Time point 1. The well documented pathogenic mutation *AR* p.T878A was detected at Time point 2 and its detection coincided with a loss of the *EP300* mutation and an increase in the tumour's Gleason Score. At Time point 3, regions that were positive for the *AR* mutation tested negative for the *EP300* mutation.

Patient 2 was profiled at 5 time points across 1.8 years and was assessed the known pathogenic stop-gain mutation *PTEN* p.R303X at Time point 1 in addition to a heterozygous loss of *RB1*. At Time point 1, *PTEN* p.R303X was inferred in nearly all tumour cells. Following treatment with a luteinizing hormone releasing hormone (LHRH) analogue, this patient gained a mutation in *BRCA2* (p.N372H) and then in *EP300* (p.E1063Q); both were gained in Time point 4 and are predicted to be damaging. This coincided with an increase in Gleason score (5+4 to 5+5) and was followed by the introduction of the combination treatment LHRH and Casodex. Finally, Time point 5 profiles suggest that the *BRCA2-EP300* significantly increase in proliferation and this tumour subclone, which was infrequent at Time point 4, expands and accounts for most of the tumour cells.

Patient 3 was profiled at 2 time points and had an increase in Gleason score (5+4 to 5+5). This patient had a castrate resistant cancer and his therapy included orchiectomy 7 years prior to the first biopsy. His Chimaera inferred phylogeny suggested a dominant tumour subclone with a heterozygous loss of *RB1* together with previously-observed pathogenic *PTEN* nonsense (p.Q245) and *BRCA1* (p.E1038G) mutations. This subclone later acquired mutations in *PALB2*

[1] Separate confidential document: PrECISE-D1.2-M36-Table-S1-CO.xlsx

(p.E672Q) and *TP53* (p.V41G; predicted to be deleterious), followed by *BRCA2* (p.N372H; previously observed) and a stop-gain mutation in *BRIP1* (p.R798X). All of these mutations were present at Time point 1.

Phylogenies inferred for Patients 4 and 5 were simpler. Patient 4's phylogeny included a sequence of 5 intronic and synonymous mutations with unknown significance in *TMPRSS2*. While Patient 5's phylogeny included a predicted initiating mutation in *PIK3CA* (p.Y182H); see Table S1 for all mutation frequency data.

Chapter 6 Summary and Conclusion

We developed a method to infer mutation frequencies in genomically unstable cancers and used it to infer tumour subclones in prostate and other cancers. Our analysis of alterations in prostate cancers identified mutations associated with initiating tumour subclones as well as tumour subclones that respond or are resistant to therapies. Our study demonstrates how tumour profiles that consider multiple areas and time points during tumour evolution can help reveal the ordinal relationship between subclones and their responses to therapies.

Chapter 7 Bibliography

Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., Ji, H. P., and Maley, C. C. (2016). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nature medicine *22*, 105-113. PMC4830693

Andor, N., Harness, J. V., Muller, S., Mewes, H. W., and Petritsch, C. (2014). EXPANDS: expanding ploidy and allele frequency on nested subpopulations. Bioinformatics *30*, 50-60. PMC3866558

Boutros, P. C., Fraser, M., Harding, N. J., De Borja, R., Trudel, D., Lalonde, E., Meng, A., Hennings-Yeomans, P. H., McPherson, A., and Sabelnykova, V. Y. (2015). Spatial genomic heterogeneity within localized, multifocal prostate cancer. Nature genetics *47*, 736-745.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., and Weir, B. A. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nature biotechnology *30*, 413-421.

Celeux, G., and Govaert, G. (1995). Gaussian parsimonious clustering models. Pattern recognition *28*, 781-793.

Chu, W. K., Edge, P., Lee, H. S., Bansal, V., Bafna, V., Huang, X., and Zhang, K. (2017). Ultraaccurate genome sequencing and haplotyping of single human cells. Proceedings of the National Academy of Sciences, 201707609.

Cieslik, M., Chugh, R., Wu, Y.-M., Wu, M., Brennan, C., Lonigro, R., Su, F., Wang, R., Siddiqui, J., and Mehra, R. (2015). The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. Genome research *25*, 1372-1381. PMC4561495

Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., and Morris, Q. (2015). PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biology *16*, 35.

Ding, L., Bailey, M. H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D. L., Weerasinghe, A., Huang, K.-l., and Tokheim, C. (2018). Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell *173*, 305-320. e310.

El-Kebir, M., Oesper, L., Acheson-Field, H., and Raphael, B. J. (2015). Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics *31*, i62-i70. PMC4542783

Espiritu, S. M. G., Liu, L. Y., Rubanova, Y., Bhandari, V., Holgersen, E. M., Szyca, L. M., Fox, N. S., Chua, M. L., Yamaguchi, T. N., and Heisler, L. E. (2018). The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. Cell.

Fidler, I. J., and Hart, I. R. (1982). Biological diversity in metastatic neoplasms: origins and implications. Science *217*, 998-1003.

Fritz, H., Garcıa-Escudero, L. A., and Mayo-Iscar, A. (2012). tclust: An r package for a trimming approach to cluster analysis. Journal of Statistical Software *47*, 1-26.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. Nat Rev Cancer *4*, 177-183.

Gao, R., Davis, A., McDonald, T. O., Sei, E., Shi, X., Wang, Y., Tsai, P.-C., Casasent, A., Waters, J., and Zhang, H. (2016). Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nature Genetics *48*, 1119-1130. PMC5042845

Getz, G., and Ardlie, K. (2012). Mutation Analysis in Frozen and FFPE Tumor Samples. In TCGA Second Annual Scientific Symposium, M. Meyerson, and I. Shmulevich, eds. (Crystal City, Virgina: National Institutes of Health).

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L. B., Tubio, J. M., Papaemmanuil, E., Brewer, D. S., Kallio, H. M., Högnäs, G., and Annala, M. (2015). The evolutionary history of lethal metastatic prostate cancer. Nature *520*, 353-357.

Higgins, M. E., Claremont, M., Major, J. E., Sander, C., and Lash, A. E. (2007). CancerGenes: a gene selection resource for cancer genome projects. Nucleic acids research *35*, D721-726. PMC1781153

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research *22*, 568-576. PMC3290792

Kovács, F., Legány, C., and Babos, A. (2005). Cluster validity measurement techniques. Paper presented at: 6th International symposium of hungarian researchers on computational intelligence (Citeseer).

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

Kuipers, J., Jahn, K., Raphael, B. J., and Beerenwinkel, N. (2017). Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. Genome Research.

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J., and Tamayo, P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015; 1 (6): 417–25. In.

Lin, D.-C., Mayakonda, A., Dinh, H. Q., Huang, P., Lin, L., Liu, X., Ding, L.-w., Wang, J., Berman, B. P., and Song, E.-W. (2017). Genomic and epigenomic heterogeneity of hepatocellular carcinoma. Cancer Research *77*, 2255-2265.

Malikic, S., McPherson, A. W., Donmez, N., and Sahinalp, C. S. (2015). Clonality inference in multiple tumor samples using phylogeny. Bioinformatics *31*, 1349-1356.

Mann, K. M., Newberg, J. Y., Black, M. A., Jones, D. J., Amaya-Manzanares, F., Guzman-Rojas, L., Kodama, T., Ward, J. M., Rust, A. G., and van der Weyden, L. (2016). Analyzing tumor heterogeneity and driver genes in single myeloid leukemia cells with SBCapSeq. Nature biotechnology.

Niknafs, N., Beleva-Guthrie, V., Naiman, D. Q., and Karchin, R. (2015). SubClonal Hierarchy Inference from Somatic Mutations: automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing. PLoS Comput Biol *11*, e1004416. PMC4593588

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

Roy, A., Wang, L., Sumazin, P., Covington, K., Muzny, D., Kumar, V., Haines, K., Doddapaneni, H., White, S., and Chao, H. (2014). Integration of Whole Transcriptome Sequencing into the Genomic Analysis of Pediatric Solid Tumors: Early Experience and Challenges. Journal of Molecular Diagnostics *16*, 754-755.

Sheppard, D., Terrell, R., and Henkelman, G. (2008). Optimization methods for finding minimum energy paths. The Journal of chemical physics *128*, 134106.

Suzuki, H., Aoki, K., Chiba, K., Sato, Y., Shiozawa, Y., Shiraishi, Y., Shimamura, T., Niida, A., Motomura, K., and Ohka, F. (2015). Mutational landscape and clonal architecture in grade II and III gliomas. Nature genetics *47*, 458-468.

TCGA (2017). Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell *169*, 1327-1341. e1323.

The Cancer Genome Atlas (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70. PMC3465532

The Cancer Genome Atlas (2015). The molecular taxonomy of primary prostate cancer. Cell *163*, 1011-1025. PMC4695400

Wang, J., Khiabanian, H., Rossi, D., Fabbri, G., Gattei, V., Forconi, F., Laurenti, L., Marasca, R., Del Poeta, G., Foa, R., Pasqualucci, L., Gaidano, G., and Rabadan, R. (2014a). Tumor evolutionary directed graphs and the history of chronic lymphocytic leukemia. eLife *3*. PMC4308685

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., and Liang, H. (2014b). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature *512*, 155-160. PMC4158312