Institute of Science and Technology

Reconstructing robust phylogenies of metastatic cancers

Johannes G Reiter and Alvin P Makohon-Moore and Jeffrey M Gerold and Ivana Bozic and Krishnendu Chatterjee and Christine A Iacobuzio-Donahue and Bert Vogelstein and Martin A Nowak

Technical Report No. IST-2015-399-v1+1
Deposited at 30 Dec 2015 16:09
http://repository.ist.ac.at/399/1/treeomics.pdf

IST Austria (Institute of Science and Technology Austria)
Am Campus 1
A-3400 Klosterneuburg, Austria

Reconstructing robust phylogenies of metastatic cancers

Johannes G. Reiter[1,2,3,4], Alvin P. Makohon-Moore[5,6], Jeffrey M. Gerold[1], Ivana Bozic[1,7], Krishnendu Chatterjee[2], Christine A. Iacobuzio-Donahue[5,6,8], Bert Vogelstein[9,10], Martin A. Nowak[1,7,11]

[1]Program for Evolutionary Dynamics, Harvard University, Cambridge, MA, USA. [2]IST (Institute of Science and Technology) Austria, Klosterneuburg, Austria. [3]Dana-Farber Cancer Institute, Boston, MA, USA. [4]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [5]The David M. Rubenstein Center for Pancreatic Cancer Research, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [6]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [7]Department of Mathematics, Harvard University, Cambridge, MA, USA. [8]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [9]The Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [10]The Ludwig Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [11]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA.

A comprehensive understanding of the clonal evolution of cancer is critical for understanding neoplasia. Genome-wide sequencing data enables evolutionary studies at unprecedented depth. However, classical phylogenetic methods often struggle with noisy sequencing data of impure DNA samples and fail to detect subclones that have different evolutionary trajectories. We have developed a tool, called *Treeomics*, that allows us to reconstruct the phylogeny of a cancer with commonly available sequencing technologies. Using Bayesian inference and Integer Linear Programming, robust phylogenies consistent with the biological processes underlying cancer evolution were obtained for pancreatic, ovarian, and prostate cancers. Furthermore, Treeomics correctly identified sequencing artifacts such as those resulting from low statistical power; nearly 7% of variants were misclassified by conventional statistical methods. These artifacts can skew phylogenies by creating illusory tumor heterogeneity among distinct samples. Importantly, we show that the evolutionary trees generated with Treeomics are mathematically optimal.

Genetic evolution underlies our current understanding of cancer[1–3] and the development of resistance to therapies[4–7]. The principles governing this evolution are still an active area of research, particularly for metastasis, the final biological stage of cancer that is responsible for the vast majority of deaths from the disease. Although many insights into the nature of metastasis have emerged[8,9], we do not yet know how malignant tumors evolve the potential to metastasize nor do we know the temporal or spatial rules governing the seeding of metastases at sites distant from the primary tumor[10–13].

In an effort to help understand this process, reconstructions of the temporal evolution of a patient's cancer from genome sequencing data have been reported[14–17]. But phylogenomic analysis has largely been focused on the subclonal structure and branching patterns of primary tumors[18–20]. The evolutionary relationships among metastases have less often been determined[21–25], despite their importance. Several factors complicate the determination of the evolutionary histories of metastatic cancers. First, comprehensive data sets of samples from spatially-distinct metastases in different organs are rarely available. Second, most advanced cancer samples are derived from patients who have been treated with toxic and mutagenic chemotherapies, imposing a variety of unknown constraints on genetic evolution and its interpretation. Third, tumors are composed of varying proportions of neoplastic and non-neoplastic cells, and inferring meaningful evolutionary patterns from such impure samples is challenging[26,27]. Moreover, the situation for solid tumors differs from that of “liquid tumors”, where mutant allele fractions are high and can be easily determined from cytological analysis. Fourth, chromosome-level changes, including losses, are frequently observed in cancers, and previously acquired variants can be lost (i.e., some variants are not "persistent"). Finally, even when performed at high depth, next-generation sequencing coverage is always non-uniform, resulting in different amounts of uncertainty among different loci within the same DNA sample as well as among different samples at the same locus.

The variety of methods that have recently been used to infer evolutionary relationships among tumors underscore these complicating factors and the need for a more robust phylogenomic approach. The methods include those based on genetic

distance[23,28–30], maximum parsimony[31,25], clonal ordering[3,17,24] and variant allele frequency[18,32,33]. Classical phylogenetics assumes that the individual traits are known with certainty[26]. Consequently these methods struggle with noisy high-throughput DNA sequencing data, possibly from very impure samples, and fail to exploit the full potential of these data due to the error-prone binary present/absent classification of variants. Modern phylogenomic methods[34–37] estimate variants from the observed variant allele frequencies (VAF). However, inaccurate VAFs resulting from insufficient sequencing depth or low sample purity introduce potential errors in the analysis. Furthermore, many of the methods used for inferring cancer evolutionary trees are based on those designed for more complex evolutionary processes involving sex and recombination[12].

Our current study was inspired by a different component of evolutionary biology, involving the analysis of asexual rather than sexual populations. One key conceptual difference between the new approach used here ("Treeomics") and previous ones is that we determined the probability that each variant was or was not found in each sequenced lesion rather than rely on a binary input ("present" or "absent"), as used in classical phylogenetic methods. This evolutionary approach results in multiple advantages: (i) it is amenable to low coverage sequencing data and impure samples, (ii) no constraints on tree topologies, substitution models or mutation rates are required, (iii) Mixed Integer Linear Programming[38] produces a single result without convergence or termination issues, and (iv) the obtained evolutionary tree is mathematically guaranteed to be optimal.

RESULTS

Evolutionarily incompatible mutation patterns

To illustrate our approach, we first focused on the data of a treatment-naïve pancreatic cancer patient Pam03[22] (Fig. 1). WGS (whole-genome sequencing; coverage: median 51x, mean 56x) as well as deep targeted sequencing (coverage: median 296x, mean 644x) was performed on ten spatially-distinct samples from the primary tumor and distinct liver and lung metastases (Online Methods and ref. 22). Estimated purities ranged from 16% to 43% per sample[22], typical for low-cellularity cancers. Founder variants (present in all

samples) and unique variants (present in exactly one sample) are parsimony-uninformative and hence irrelevant for the branching in an evolutionary tree. Parsimony-informative variants (variants present in some but not in all samples; depicted by black dots in Fig. 1) exhibited many evolutionary incompatibilities when we tried to reconstruct a phylogeny consistent with the evolutionary processes underlying tumor progression using conventional methods. In particular, evolutionary relationships could not be inferred based on standard present/absent classification of variants (Fig. S1).

The mutation pattern of a variant is denoted by the set of samples where the variant is present (Fig. S1). Two somatic variants α and β are evolutionarily incompatible if and only if samples with the following three patterns exist: (i) variant α is absent and β is present, (ii) α is present and β is absent, and (iii) both variants are present. Because somatic variants are by definition absent in the germline, α and β are evolutionarily incompatible, so no perfect (the same variant is not independently acquired twice; infinite sites model[39]) and persistent (acquired variants are not lost; no back mutation) phylogeny can be inferred (Fig S1).

A perfect and persistent tree consistent with the observed (noisy) data of Pam03 cannot be inferred and may not even exist[40–42]. Treeomics shows that such a phylogeny may indeed exist but that it is hidden behind technical and biological artifacts. Although the *median* coverage in the sequencing data from Pam03 was high, many of the identified variants had a coverage below 20x in at least one of the impure samples (purity <20%; Fig. S2), leading to potentially misleading evolutionary patterns with standard approaches, as shown below.

Identifying evolutionarily compatible mutation patterns

To account for inconclusive data, we developed a Bayesian inference model to calculate the probability that a variant is present in a sample (detailed in Online Methods). Using these probabilities for each individual variant, we calculated *reliability scores* for each possible mutation pattern. We constructed an evolutionary conflict graph where the nodes were determined through analysis of all mutation patterns, with the weights of each node

provided by the calculated reliability scores (Fig. S3). If two nodes (mutation patterns) were evolutionarily incompatible, an edge between the corresponding nodes was added. We aimed to identify the set of nodes that maximized the sum of the weights (reliability scores) when no pair of nodes was evolutionarily incompatible. This maximal set represents the most reliable and evolutionarily compatible mutation patterns (Supplementary Information). We modeled and solved the maximization challenge using a Mixed Integer Linear Program[38] (MILP; see Online Methods). Additionally, we proved via a reduction to the weighted minimum vertex cover problem that the decision version of finding the most reliable and evolutionarily compatible mutation patterns is *NP*-complete[43] (see Supplementary Information for mathematical proofs).

Predicting putative artifacts in sequencing data

The solution obtained with the MILP provided the most likely evolutionarily compatible mutation pattern for each variant. By comparing our inferred classifications to conventional binary classification, Treeomics predicted putative sequencing or biological artifacts in the data (Fig. 2). The conventional classifications differed in 8.8% of the variants in Pam03 (78 putative artifacts from 89 variants in 10 samples; Fig. 2). As expected, the majority (72) of the differences were caused by putative false-negatives in the binary classification that were inferred to be present by Treeomics (Table 1). Fifty-nine of these putative false-negatives had relatively low coverage, explaining how they could easily be misclassified as absent given the low neoplastic cell content in these samples. Accordingly, many of these under-powered false-negatives occurred in samples with the lowest coverage (LiM 5, LuM 2-3) or lowest neoplastic cell content (LuM 1). In LuM 2, the driver gene mutation *KRAS* was incorrectly classified as absent by conventional means though it is most likely a clonal founding mutation and was present at a VAF of 19% in the original WGS sample (Supplementary Table S1). Some variants contained false-negatives across many samples, indicating that these variants were generally difficult to call. Remarkably, 95% (56/59) of the predicted under-powered false-negatives were either significantly present in the WGS data (mostly at higher coverage than in the targeted sequencing data), or the genomic region of the variant possessed a low alignability score (Supplementary Table S1).

An additional 13 putative false-negatives were sequenced at relatively high coverage, but might be explained by loss of heterozygosity (LOH), which frequently occurs in pancreatic cancers. Of the 6 putative false-positives (purple squares in Fig. 2b; e.g., *abParts*, *MFN1*), 83% (5/6) were classified as absent in the original WGS data and all of them were in a genomic region with a low alignability score[44,45] (Supplementary Table S1). Hence, at least 6.9% (56 putative false-negatives + 5 putative false-positives) of the variants were misclassified by conventional binary classification. If a phylogenomic method does not account for sequencing artifacts, a large fraction of variants will often be inconsistent with any inferred evolutionary tree. In our case, at least 31.5% of the variants would be evolutionarily incompatible – independent of the inferred tree topology (Fig. 2a). These putative artifacts may also help to explain the observed high tumor heterogeneity in earlier studies and the recently reported tumor homogeneity when sequencing depth is increased[22,28].

Inferring evolutionary trees

From the identified mutation patterns, Treeomics inferred an evolutionary tree rooted at the germline DNA sequence of Pam03 (Fig. 3). We found strong support for two major evolutionary clusters among the geographically distinct lesions: (i) samples LiM 2-5 (liver mets) and PT 11 (primary tumor) and (ii) samples LiM 1, LuM 1 (lung met.) and PT 10. These results indicate that a recent parental clone of PT 11 seeded the liver metastases in cluster (i) and a recent parental clone of PT 10 seeded the lung and liver metastases in cluster (ii); perhaps the same clone also seeded LuM 2 and 3, however, the low neoplastic cell content and the low coverage of LuM 2 and 3 prevented a definite conclusion. We also reconstructed the same major clusters by using the low-coverage WGS data (Fig. S4) instead of the high coverage targeted sequencing data (Fig. 3). The inferred trees indicated that the lung metastases had been seeded before most of the liver metastases in patient Pam03 (Fig. 3). Furthermore, the results suggested that the liver metastasis LiM 1 was seeded from a genetically different subclone than all other liver metastases. In a different treatment-naïve pancreatic cancer patient (Pam02) we also found that liver metastases diverged late in the inferred evolutionary tree (Fig. S5).

Confirming robustness of the identified mutation patterns

We investigated the robustness of our results by determining whether Treeomics could identify the inferred mutation patterns and their evolutionary trajectories from a random subset of the given variants. Through this analysis, we found that only ~two thirds of the variants in Pam03 were sufficient to identify all major evolutionary relationships and clusters, despite the fact that only 34% of variants were identified as parsimony-informative (Fig. S6a). As expected, subclusters within the main clusters were less frequently reproduced as indicated by the lower bootstrapping values because of inadequate supporting sequencing data.

To further validate our approach, we reanalyzed data from high-grade serous ovarian cancers[23]. We were able to reproduce all phylogenetic trees of Bashashati et al.[23] except for Case 5. In this case, the authors reported an early divergence of sample 5c while Treeomics suggested a late divergence (Fig. S7c). Comprehensive analysis of their data (reinterpreted in Fig. S7a,b) revealed that their tree either required that several variants (including two driver gene mutations and multiple indels) occurred independently twice or that two mutations in the driver genes *ABL1* and *MDM4* were lost; both possibilities seem implausible (Fig. S7 and Fig. 1D in ref. 23). Treeomics did not require these implausible scenarios to construct an otherwise similar tree. We confirmed the robustness of our results via bootstrapping (Fig. S6b). Distance-based methods, such as those used by Bashashati et al., can be compromised by large differences in the number of acquired mutations among samples; sample 5c had twice as many mutations than most other samples.

We also reanalyzed a comprehensive data set from prostate cancers[21]. Treeomics generally confirmed the results and further refined others. For example, for patient A32, Gundem et al. (2015) reported an inconclusive evolutionary tree due to evolutionary incompatible subclones present at low frequencies. Our method used the strong evidence for mutation patterns C, E and D, F (see Extended Data Figure 3p,q in ref. 21) and was

thereby able to illuminate the evolutionary relationships among these samples in a conclusive fashion (Fig. S8).

Detecting subclones of distinct origin

If multiple subclones were represented in the same sample, conventional phylogenetic approaches would be unable to separate their evolutionary trajectories. In the cases where multiple subclones present at low frequencies were apparent, evolutionarily incompatible mutation patterns with high reliability scores were identified (Fig. S9b). By investigating the VAFs of the variants in these patterns, we could infer separate evolutionary histories for the subclones (Online Methods). For both the prostate cancer data of case A22[21] (Fig. S9) and of case 6[20] (Fig. S10), Treeomics identified subclonal structures and separated their evolutionary trajectories without requiring high purity samples or deep sequencing data as are required by previously used methods.

DISCUSSION

The new approach described here efficiently reconstructs the evolutionary history, detects potential artifacts in noisy sequencing data, and finds subclones of distinct origin. The evolutionary theory of asexually evolving populations combined with Bayesian inference and Integer Linear Programming enabled us to infer detailed phylogenomic trees. In contrast to other tools, Treeomics accounts for putative artifacts in sequencing data and can thereby infer the branches where somatic variants were acquired as well as where some may have been lost during evolution, presumably through losses of heterozygosity resulting from chromosomal instability[46]. The branching in the inferred trees sheds light on the seeding patterns (timing[47] and location) of particular metastatic lesions[11,12].

We have designed Treeomics from first principles to directly handle ambiguity in high-throughput sequencing data, including samples with low neoplastic cell content or coverage. The mutation patterns and their evolutionary conflict graph form a robust data structure and consequently the painful task of semi-automatic filtering becomes unnecessary. As a result of the Bayesian confidence estimates for the individual variants,

this method can infer more robust results than traditional phylogenetic methods, which employ a binary representation of sequencing data. Furthermore, as shown above, distance-based methods can produce results inconsistent with the evolutionary theory of cancer as they often ignore knowledge of biological phenomena specific to neoplasia (Fig. S7). We compared our results to another state-of-the-art method in cancer phylogenomics[37] (other methods were not applicable for multiple spatially-distinct samples with low neoplastic cell content). *AncesTree*[37] roughly identified one of the major evolutionary clusters in Pam03 but excluded 58% (37/89) of the variants (among them the driver gene mutation in *KRAS*) in the inferred phylogeny due to evolutionary incompatibilities (Fig. S11).

At present Treeomics only employs nucleotide substitutions and short insertions and deletions – a subset of the available information. Other types of data, such as copy number alterations, structural variations and DNA methylation, could be incorporated into Treeomics to further improve the accuracy of the inferred results[48–51]. Such analyses can benefit from analyzing all tumor samples from the same patient together (plus a matched normal sample) to account for the joint evolution of cancer cells, yielding more robust results[52].

The challenge in finding the most likely evolutionary trajectories is *NP*-complete. However, medium-sized instances of *NP*-complete problems are no longer intractable due to the enormous engineering and research effort that has been devoted to ILP solvers. The MILP formulation enables an efficient and robust analysis of large datasets (see Supplementary Information, Theorem 1, for more details about the theoretical limits). MILPs may also be useful in other areas of phylogenetic inference where methods with strong biological assumptions (e.g. constant mutation rates or specific substitution profiles) are not applicable or are computationally too expensive to obtain guaranteed optimal solutions.

ONLINE METHODS

DNA sequencing design and validation

As described in detail in ref. 22, sequencing data were generated in two stages. First, genomic DNA from 22 tumor samples (16 metastases and 6 primary tumor sections) was evaluated by 60x whole genome sequencing (WGS) using an Illumina Hi-Seq 2000. Importantly, genomic DNA from the normal tissues of each patient was used to facilitate identification of somatic variants. We obtained an average coverage of 69x with 97.5% of bases covered at >10x, revealing a total of 106,919 putative coding and noncoding somatic mutations, (average of 4,860 per sample). To limit the artifacts generated by WGS and alignment, we filtered the putative variants using several quality parameters, including read directionality, mutant allele frequency detected in the normal, known human SNPs, and the number of independent tags at each site.

Second, we utilized a targeted sequencing approach to independently screen every mutation that we observed to be of high quality in at least one WGS tumor sample. Briefly, probes for capture were designed to flank each potential mutant base (n = 960) and libraries were prepared for the original 22 WGS samples. Using an Illumina chip-based approach, we successfully aligned, processed, and validated 219 mutations (range 107-112 per patient) at an average sequencing depth of 772x (Supplementary Tables S2 and S3). In addition to the increased coverage and sensitivity of targeted sequencing, both sequencing approaches generated independent datasets in which we could directly compare putative variants *in silico* among many tumors within a patient. Additional details regarding patient selection, processing of tissue samples and DNA extraction and quantification can be found in ref. 22.

Bayesian inference model

To compute reliability scores for each mutation pattern, we first extract posterior probabilities for the presence and absence of a variant in a sample from a Bayesian model of error-prone sequencing. If f is the true fraction of variant reads in the sample, π is our

prior belief about *f,* and *e* is the sequencing error rate, the posterior distribution *P* of *f* given *N* total reads and *K* variant reads is

$$P(f|N,K) = \binom{N}{K} \cdot [f(1-e) + (1-f)e]^K \cdot [f \cdot e + (1-f)(1-e)]^{N-K} \cdot \pi(f) \cdot \frac{1}{Z} \quad (1)$$

where Z is a normalizing constant (see Supplementary Information). A priori, the variant allele frequency in a sample is exactly zero ($f = 0$) with some positive probability c_0. The prior π is then of the following form

$$\pi(f) = c_0 \cdot \delta(f) + (1 - c_0) \cdot g(f), \quad (2)$$

where $\delta(f)$ denotes the Dirac delta function and $g(f)$ denotes a prior given the variant is present (Supplementary Information). The prior can differ for each variant to account for sample purity and variant ploidy. The probability that a variant is absent, denoted by q, and the probability that a variant is present, denoted by p, are

$$q = P(f = 0|N,K), \quad p = 1 - q. \quad (3)$$

A variety of more sophisticated variant detection algorithms can be used here as long as the output can be converted to posterior probabilities of presence and absence. We calculate the probability of each mutation pattern for a particular variant by multiplying the corresponding posterior probabilities for each sample. Each mutation pattern has some positive probability, but those supported by the data are given much more weight. A mutation pattern ν is denoted as a binary vector of length $|S|$ (total number of samples) where ν_s is 1 if the variant is present in sample s and 0 if absent. The likelihood $L_\mu(\nu)$ that a variant μ exhibits pattern ν is

$$L_\mu(\nu) = \prod_{s \in S} {p_{\mu,s}}^{\nu_s} \cdot {q_{\mu,s}}^{1-\nu_s}. \quad (4)$$

The reliability score ω_ν of each mutation pattern ν (corresponding to a node in the evolutionary conflict graph; Fig. S3) is given by

$$\omega_\nu = -log \prod_{\mu} \Big(1 - L_\mu(\nu)\Big). \quad (5)$$

The argument of the logarithm denotes the probability that no mutation has pattern ν and hence leverages the full sequencing information from all variants. With these scores (weights), the minimum weight vertex cover of the evolutionary conflict graph

corresponds to identifying the most reliable and evolutionarily compatible mutation patterns (see Supplementary Information for further details).

Identifying reliable evolutionarily compatible mutation patterns

Given the calculated reliability scores, we efficiently find the most reliable and evolutionarily compatible mutation pattern for all variants via solving a mixed integer linear program[38] (MILP). In the Supplementary Information we prove that finding these mutation patterns is equivalent to solving the *Minimum Vertex Cover problem*; one of Karp's original 21 *NP*-complete problems[43]. In the Minimum Vertex Cover problem one wants to find the minimum set of nodes in an undirected graph such that each edge in the graph is adjacent to one of the nodes in the minimum set. Therefore, by definition all edges are covered by the nodes in the minimum set. Similarly, we try to find the weighted set of nodes (here mutation patterns) with the minimal sum of reliability scores such that no evolutionary incompatibilities in the conflict graph remain. After this minimal set of nodes and their adjacent edges have been removed from the graph, we can easily infer an evolutionary tree since evolutionary conflicts no longer exist (i.e., all edges were covered and removed with the minimal set). The remaining set of mutation patterns is by definition the maximal set of evolutionarily compatible patterns (see Supplementary Information for details).

In the evolutionary conflict graph $G = (V, E)$, each node $i \in V$ represents a different mutation pattern. For n samples, the number of nodes $|V|$ is given by 2^n. For each pair of evolutionarily incompatible mutation patterns i and j, there exists an edge $(i, j) \in E$. The weight (c_i) of each node i is given by the reliability scores ω_i described in the Bayesian inference model section (Fig. S3).

The MILP to find the minimal-weighted set of evolutionarily incompatible mutation patterns is defined by the following objective function and constraints:

$$\begin{aligned} &\text{(objective function)} && \text{minimize } \textstyle\sum_{i \in V} c_i \cdot x_i && && (6) \\ &\text{(constraints)} && \text{subject to } x_i + x_j \geq 1 && \text{for all } (i, j) \in E \\ & && x_i \in \{0,1\},\ c_i > 0 && \text{for all } i \in V \end{aligned}$$

This formulation guarantees that the MILP solver finds the minimal value of the objective function such that all constraints are met and hence the nodes in the selected set cover all edges. The evolutionarily compatible and most reliable mutation patterns $\{i \mid x_i = 0\}$ are given by the complement set of the optimal solution $\{i \mid x_i = 1\}$ to the MILP.

Inferring evolutionary trees

After the evolutionarily compatible mutation patterns $\{i \mid x_i = 1\}$ have been identified and variants are assigned to their most likely evolutionarily compatible pattern based on the maximum likelihood weights given by the Bayesian inference model, the derivation of an evolutionary tree is a trivial computational task. In quadratic time ($\mathcal{O}(n \cdot m)$) of the input size we construct a unique phylogeny where n is the number of samples and m is the total number of distinct variants[53]. The branches where the individual variants are acquired follow from the inferred tree.

Detecting subclones of distinct origin

Evolutionary incompatible mutation patterns with high reliability scores may indicate mixed subclones with distinct evolutionary trajectories (Fig. S9b, Fig. S10a). Recall that evolutionary incompatibility requires that the conflicting variants need to be present together in at least one sample. However, even if both variants are mutated in a statistically significant fraction in the same sample, these variants may not be present in the same cells and the evolutionary laws of an asexually evolving population may not be violated. If low VAFs of those variants support this hypothesis, Treeomics updates the corresponding mutation patterns and infers distinct evolutionary trajectories to these subclones. Low VAFs of the variants in descending (not necessarily evolutionary incompatible) mutation patterns of the putative subclone provide additional evidence for mixed subclones in a sample. As outlined for prostate cancer case A22, subsets (descendants) and supersets (ancestors) of the conflicting mutation pattern can simultaneously be identified and a comprehensive evolutionary tree inferred (Fig. S9c). This approach also worked well among samples from the same tissue. After two subclones were separated, 12643 (out of 12645) variants supported the inferred

evolutionary tree (Fig. S10b). The remaining two variants were predicted to be false-positives by Treeomics.

Binary present/absent classification

We perform conventional binary present/absent classification of each variant to allow a comparison to the inferred classification used in our new approach. We scored each variant by calculating a *p*-value in all samples (one-tailed binomial test): $\Pr(X \geq K | H_0, K, N) = 1 - \sum_{i=0}^{K-1} \binom{N}{i} \cdot {p_{fpr}}^{i} \cdot (1 - p_{fpr})^{N-i}$ where N denotes the coverage, K denotes the number of variant reads observed at this position, and X denotes the random number of false-positives. As null hypothesis H_0, we assume that the variant is absent. Similar to Gundem et al.[21], we assumed a false-positive rate (p_{fpr}) of 0.5% for the Illumina chip-based targeted deep sequencing. In the WGS data set we assumed a conservative false-positive rate of 1%[54]. We used the step-up method[55] to control for an average false discovery rate (FDR) of 5% in the combined set of *p*-values from all samples of a patient. Variants with a rejected null hypothesis were classified as present. The remaining variants were classified as absent.

Treeomics

The source code and manual for Treeomics, as well as multiple examples illustrating its usage, are provided at https://github.com/johannesreiter/treeomics. The tool is implemented in Python 3.4. The inputs to the tool are the called variants and the corresponding sequencing data, either in tab-separated-values format or as matched tumor-normal VCF files. As output, Treeomics produces a comprehensive HTML report (Supplementary File 1) including statistical analysis of the data, a mutation table plot and a list of putative artifacts (false-positives, well-powered and under-powered false-negatives). Additionally, Treeomics produces evolutionary trees in LaTeX/TikZ format for high-resolution plots in PDF format. If circos[56] is installed, Treeomics automatically creates the evolutionary conflict graph and adds it to the HTML report. Treeomics also supports various filtering (e.g., minimal sample median coverage, false-positive rate, false-discovery rate) for an extensive analysis of the sequencing data. Detailed instructions for the filtering and analysis are provided in the readme file in the

online repository. For solving the MILP, Treeomics makes use of the common CPLEX solver (v12.6) from IBM.

ACKNOLEDGEMENTS

This work was supported by the European Research Council (ERC) start grant 279307: Graph Games (J.G.R., C.K.), Austrian Science Fund (FWF) grant no P23499-N23 (J.G.R., C.K.), FWF NFN grant no S11407-N23 RiSE/SHiNE (J.G.R., C.K.), a Landry Cancer Biology Fellowship (J.M.G.), National Institutes of Health grants CA179991 (C.I.-D.), F31CA180682 (A.M.-M.), CA43460 (B.V.), the Lustgarten Foundation for Pancreatic Cancer Research, the The Sol Goldman Center for Pancreatic Cancer Research, the The Virginia and D.K. Ludwig Fund for Cancer Research, and the John Templeton Foundation. We thank Bashashati et al. (2013), Cooper et al. (2015), and Gundem et al. (2015) for sharing their comprehensive data sets.

AUTHOR CONTRIBUTIONS

C.I.-D. and A.M.-M. performed autopsies and experiments; all authors analyzed data; J.G.R., J.M.G., and K.C. performed mathematical analyses; J.G.R. developed algorithms and implemented the tool with input from J.M.G.; J.G.R., B.V., and M.A.N. wrote the manuscript with input from A.M.-M., J.M.G., I.B., K.C., C.I.-D.; all authors read and approved the final manuscript.

COMPETING FINANCIAL INTEREST

The authors declare no competing financial interests.

REFERENCES

1. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
2. Vogelstein, B. *et al.* Genetic Alterations during Colorectal-Tumor Development. *N. Engl. J. Med.* **319,** 525–532 (1988).
3. Merlo, L. M. F., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer* **6,** 924–935 (2006).
4. Engelman, J. A. *et al.* MET amplification leads to gefitinib resistance in lung

cancer by activating ERBB3 signaling. *Science* **316,** 1039–1043 (2007).

5. Diaz, L. A. *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486,** 537–40 (2012).
6. Bozic, I. *et al.* Evolutionary dynamics of cancer in response to targeted combination therapy. *eLife* **2,** e00747 (2013).
7. Wodarz, D. & Komarova, N. L. *Dynamics of cancer: mathematical foundations of oncology*. (World Scientific Publishing Co., Inc., Singapore, 2014).
8. Nguyen, D. X., Bos, P. D. & Massagué, J. Metastasis: from dissemination to organ-specific colonization. *Nat. Rev. Cancer* **9,** 274–284 (2009).
9. Talmadge, J. E. & Fidler, I. J. The Biology of Cancer Metastasis: Historical Perspective. *Cancer Res.* **70,** 5649–5669 (2010).
10. McGranahan, N. & Swanton, C. Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. *Cancer Cell* **27,** 15–26 (2015).
11. Naxerova, K. & Jain, R. K. Using tumour phylogenetics to identify the roots of metastasis in humans. *Nat. Rev. Clin. Oncol.* **12,** 258–272 (2015).
12. Hong, W. S., Shpak, M. & Townsend, J. P. Inferring the origin of metastases from cancer phylogenies. *Cancer Res.* **75,** 4021–4025 (2015).
13. Altrock, P. M., Liu, L. L. & Michor, F. The mathematics of cancer: integrating quantitative models. *Nat. Rev. Cancer* **15,** 730–745 (2015).
14. Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci USA* **105,** 13081–13086 (2008).
15. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467,** 1114–1117 (2010).
16. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467,** 1109–1113 (2010).
17. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366,** 883–892 (2012).
18. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150,** 264–278 (2012).
19. de Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346,** 251–256 (2014).
20. Cooper, C. S. *et al.* Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat Genet* **47,** 367–372 (2015).
21. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520,** 353–357 (2015).
22. Makohon-Moore, A. *et al.* The Natural History of Pancreatic Cancer Metastasis Is Dominated by Driver Gene Homogeneity. (2016).

23. Bashashati, A. *et al.* Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J. Pathol.* **231,** 21–34 (2013).

24. Sanborn, J. Z. *et al.* Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proc Natl Acad Sci USA* **112,** 10995–11000 (2015).

25. Brastianos, P. K. *et al.* Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets. *Cancer Discov.* (2015). doi:10.1158/2159-8290.CD-15-0369

26. Beerenwinkel, N., Schwarz, R. F., Gerstung, M. & Markowetz, F. Cancer evolution: mathematical models and computational inference. *Syst. Biol.* **64,** e1–e25 (2015).

27. Turajlic, S., McGranahan, N. & Swanton, C. Inferring mutational timing and reconstructing tumour evolutionary histories. *Biochim. Biophys. Acta (BBA)-Reviews Cancer* **1855,** 264–275 (2015).

28. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346,** 256–259 (2014).

29. Schwarz, R. F. *et al.* Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PLoS Med* **12,** e1001789 (2015).

30. Naxerova, K. *et al.* Hypermutable DNA chronicles the evolution of human colon cancer. *Proc Natl Acad Sci USA* **111,** E1889–E1898 (2014).

31. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46,** 225–233 (2014).

32. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).

33. Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152,** 714–726 (2013).

34. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11,** 396–401 (2014).

35. Miller, C. A. *et al.* SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* **10,** e1003665 (2014).

36. Deshwar, A. G. *et al.* PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16,** 35 (2015).

37. 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).

38. Nemhauser, G. L. & Wolsey, L. A. *Integer and combinatorial optimization*. **18,**

(Wiley New York, 1988).

39. Ma, J. *et al.* The infinite sites model of genome evolution. *Proc Natl Acad Sci USA* **105,** 14254–14261 (2008).
40. Fitch, W. M. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Biol.* **20,** 406–416 (1971).
41. Felsenstein, J. *Inferring phylogenies*. **2,** (Sinauer Associates, Sunderland, MA, 2004).
42. Zare, H. *et al.* Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.* **10,** e1003703 (2014).
43. Karp, R. M. in *Complexity of Computer Computations* 85–103 (Springer US, 1972).
44. Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS One* **7,** e30377–e30377 (2012).
45. Rosenbloom, K. R. *et al.* The UCSC genome browser database: 2015 update. *Nucleic Acids Res.* **43,** D670–D681 (2015).
46. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396,** 643–9 (1998).
47. Bauer, B., Siebert, R. & Traulsen, A. Cancer initiation with epistatic interactions between driver and passenger mutations. *J. Theor. Biol.* **358,** 52–60 (2014).
48. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.* **14,** R80 (2013).
49. Schwarz, R. F. *et al.* Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput. Biol.* **10,** e1003535 (2014).
50. Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24,** 1881–1893 (2014).
51. Prandi, D. *et al.* Unraveling the clonal hierarchy of somatic genomic aberrations. *Genome Biol.* **15,** 439 (2014).
52. Josephidou, M., Lynch, A. G. & Tavaré, S. multiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples. *Nucleic Acids Res.* **43,** e61 (2015).
53. Gusfield, D. Efficient algorithms for inferring evolutionary trees. *Networks* **21,** 19–28 (1991).
54. Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3,** 811 (2012).
55. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300

(1995).

56. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).

FIGURE LEGENDS

Fig. 1: Observed tumor heterogeneity across lesions of pancreatic cancer patient Pam03. Each variant found in any lesion is shown and its chromosomal position indicated in the outermost circle. Parsimony-informative variants (black dots; black gene names) are present in more than one but not in all samples and can provoke evolutionary incompatibilities. Founder variants (present in all samples; green squares; green gene names) and unique variants (present in a single sample; blue triangles; blue gene names) are parsimony-uninformative. The five innermost circles correspond to samples from five distinct liver metastases (LiM 1-5); the following three circles correspond to samples from three distinct lung metastases (LuM 1-3); the other circles correspond to different parts of the primary tumor (PT 10-11).

Fig. 2: Treeomics identifies evolutionarily compatible mutation patterns after recognizing potentially misleading artifacts in the sequencing data. Variants shown in Fig. 1 are organized as evolutionarily-defined groups ("nodes") rather than by chromosomal positions. The nodes are indicated in the outermost circle: blue colored nodes are evolutionarily compatible and red colored nodes are evolutionarily incompatible. **a** | Based on conventional present/absent classification, at least 31.5% of the variants were evolutionarily incompatible (depending on the inferred tree topology). The incompatibilities are demarcated by red lines ("edges") in the center of the circle that connect each pair of incompatible nodes. **b** | Based on a Bayesian inference model and a Mixed Integer Linear Program, Treeomics identified the most likely evolutionarily compatible mutation pattern for each variant (Online Methods). This method predicted that 8.8% (78/890) variants across all samples were misclassified and thereby caused the evolutionary incompatibilities shown in panel **a**. Putative false-negatives with low coverage sequencing data are depicted by unfilled purple triangles. Powered (coverage above 100) putative false-negatives are depicted by filled purple triangles. Putative false-positives are depicted by purple squares. The driver gene mutation in KRAS was among the putative false-negatives in one of the ten lesions.

Fig. 3: Reconstructed evolution of patient Pam03's cancer from targeted sequencing data. Lung metastases (LuM 1-3) are depicted in red; Liver metastases (LiM 1-5) are depicted in green; Primary tumor samples (PT 10-11) are depicted in black. SC indicate predicted subclones. Gray percentages indicate bootstrapping

values from 1000 samples. Based on the identified evolutionarily compatible mutation patterns in Fig. 2b, a unique evolutionary tree exists. LiM 1 was seeded from a different subclone than all other liver metastases. Due to the limited number of targeted resequenced variants, the support for some branches was relatively low, in particular within the identified main clusters (e.g. LiM 2-5). The majority of variants (55%) were already present in the founding clone.

TABLES

Table 1. Treeomics predicted putative artifacts in ten sequencing samples of pancreatic cancer patient Pam03. Many putative false-negatives with low statistical power occurred in samples with the lowest coverage (LiM 5, LuM 2-3) or lowest neoplastic cell content (LuM 1). Five distinct liver metastases (LiM 1-5), three distinct lung metastases (LuM 1-3), two different parts of the primary tumor (PT 10, 11).

Artifact type	LiM 1	LiM 2	LiM 3	LiM 4	LiM 5	LuM 1	LuM 2	LuM 3	PT 10	PT 11	Total
Under-powered false-negatives	7	2	5	1	13	12	5	5	1	8	59
Powered false-negatives	2	2	2	2	1	1	0	0	1	2	13
False-positives	0	1	0	0	0	3	0	1	0	1	6

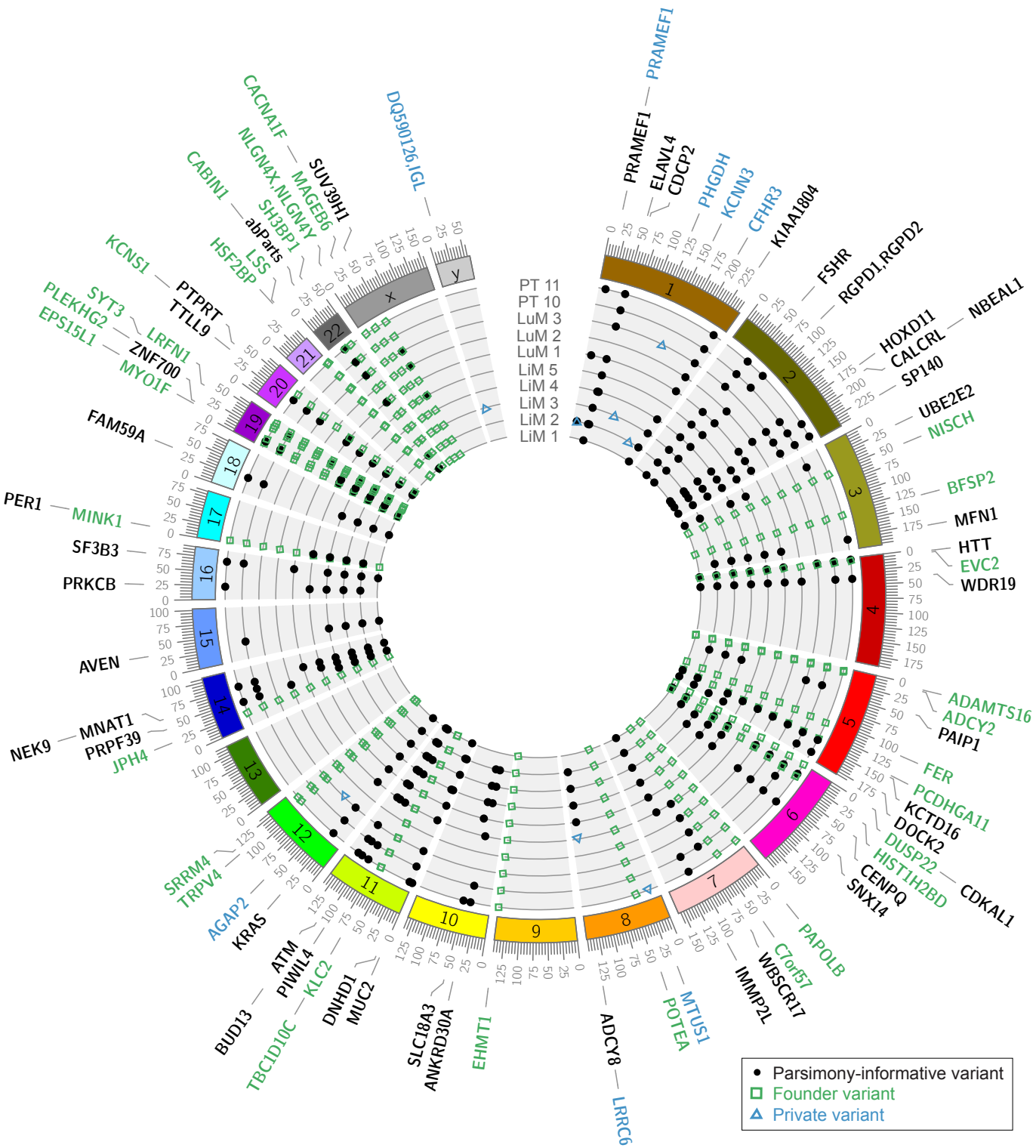
PT 11
PT 10
LuM 3
LuM 2
LuM 1
LiM 5
LiM 4
LiM 3
LiM 2
LiM 1
PRAMEF1
ELAVL4
CDCP2
PHGDH
KCNN3
CFHR3
KIAA1804
FSHR
RGPD1,RGPD2
HOXD11
CALCRL
SP140
NBEAL1
UBE2E2
NISCH
BFSP2
MFN1
HTT
EVC2
WDR19
ADAMTS16
ADCY2
PAIP1
FER
PCDHGA11
KCTD16
DOCK2
DUSP22
HIST1H2BD
CENPQ
SNX14
CDKAL1
PAPOLB
C7orf57
WBSCR17
IMMP2L
MTUS1
POTEA
ADCY8
LRRC6
EHMT1
SLC18A3
ANKRD30A
DNHD1
MUC2
KLC2
TBC1D10C
ATM
PIWIL4
BUD13
AGAP2
KRAS
SRRM4
TRPV4
NEK9
MNAT1
PRPF39
JPH4
AVEN
PRKCB
SF3B3
PER1
MINK1
FAM59A
EPS15L1
PLEKHG2
SYT3
LRFN1
ZNF700
MYO1F
KCNS1
PTPRT
TTLL9
HSF2BP
LSS
abParts
CABIN1
SH3BP1
NLGN4X,NLGN4Y
MAGEB6
CACNA1F
SUV39H1
DQ590126,IGL
Parsimony-informative variant
Founder variant
Private variant

Figure 1

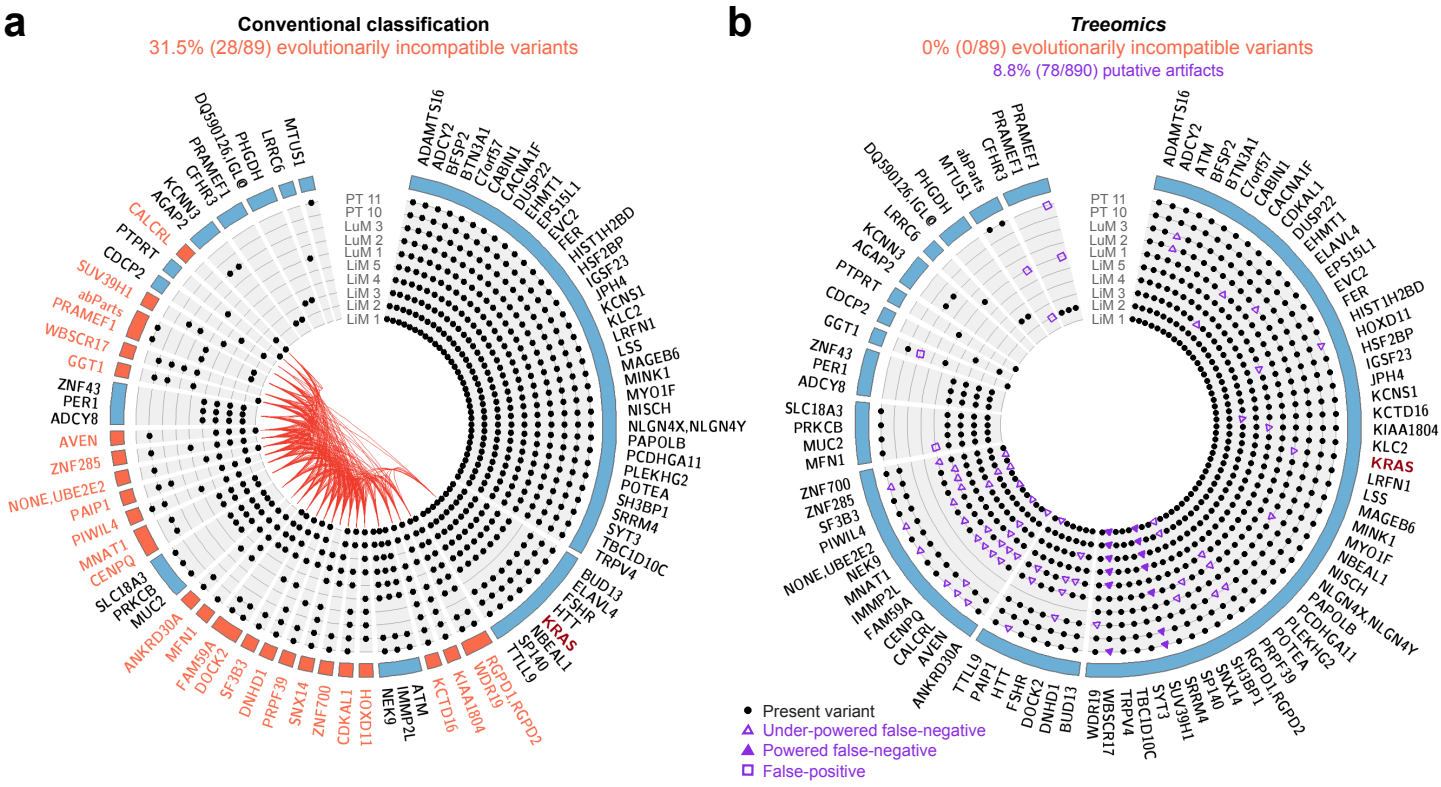
a
Conventional classification
31.5% (28/89) evolutionarily incompatible variants
b
Treeomics
0% (0/89) evolutionarily incompatible variants
8.8% (78/890) putative artifacts
PT 11
PT 10
LuM 3
LuM 2
LuM 1
LiM 5
LiM 4
LiM 3
LiM 2
LiM 1
KRAS
Present variant
Under-powered false-negative
Powered false-negative
False-positive

Figure 2

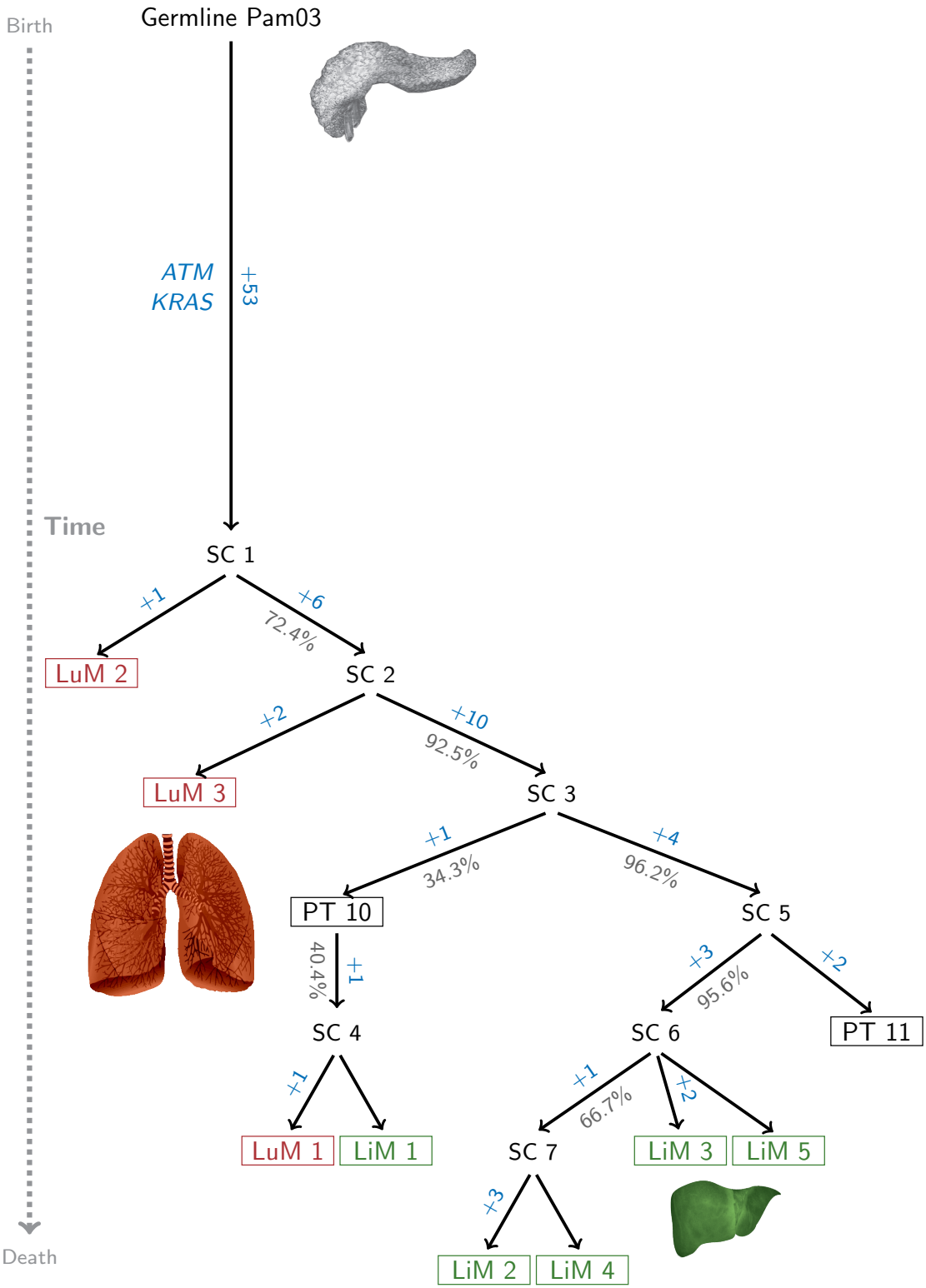
Birth
Germline Pam03
ATM
KRAS
+53
Time
SC 1
+1
+6
72.4%
LuM 2
SC 2
+2
+10
92.5%
LuM 3
SC 3
+1
34.3%
+4
96.2%
PT 10
SC 5
40.4%
+1
+3
95.6%
+2
SC 4
SC 6
PT 11
+1
+1
66.7%
+2
LuM 1
LiM 1
SC 7
LiM 3
LiM 5
+3
LiM 2
LiM 4
Death

Figure 3