Algorithmic Methods to Infer the Evolutionary Trajectories in Cancer Progression

Giulio Caravagna[1,*] Alex Graudenzi[1] Daniele Ramazzotti[1]
Rebeca Sanz-Pamplona[2] Luca De Sano[1] Giancarlo Mauri[1]
Victor Moreno [2,3] Marco Antoniotti[1,4,†] Bud Mishra[5,†]

[1] Dept. of Informatics, Systems and Communication, University of Milan-Bicocca, Italy

[2] Unit of Biomarkers and Susceptibility, Cancer Prevention and Control Program, Catalan Institute of Oncology (ICO), IDIBELL and CIBERESP. Hospitalet de Llobregat, Barcelona, Spain

[3] Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain

[4] Milan Center for Neuroscience, University of Milan-Bicocca, Italy

[5] Courant Institute of Mathematical Sciences, New York University, USA

Abstract

The evolutionary nature of cancer relates directly to a renewed focus on the voluminous NGS (next generation sequencing) data, aiming at the identification of explanatory models of how the (epi)genomic events are choreographed in cancer initiation and development. However, despite the increasing availability of multiple additional -omics data, this quest has been frustrated by various theoretical and technical hurdles, mostly related to the dramatic heterogeneity and temporality of the disease. In this paper, we build on our recent works on "selectivity" relation among driver mutations in cancer progression and investigate their applicability to the modeling problem – both at the population and individual levels. On one hand, we devise an optimal, versatile and modular pipeline to extract ensemble-level progression models from cross-sectional sequenced cancer genomes. The pipeline combines state-of-the-art techniques for sample stratification, driver selection, identification of fitness-equivalent exclusive alterations and progression model inference. We demonstrate this pipeline's ability to reproduce much of the current knowledge on colorectal cancer progression, as well as to suggest novel experimentally verifiable hypotheses. On the other hand, we prove that our framework can be applied, *mutatis mutandis*, in reconstructing the evolutionary history of cancer clones in single patients, as illustrated by an example with multiple biopsy data from clear cell renal carcinomas.

Introduction

Since the late seventies evolutionary dynamics, with its interplay between variation and selection, has progressively provided the widely-accepted paradigm for the interpretation of cancer emergence

[*] To whom the correspondence should be addressed at giulio.caravagna@unimib.it.

[†] Co-senior authors.

and development [1–3]. Random alterations of an organism's (epi)genome can sometimes confer a functional *selective advantage* to certain cells, in terms of adaptability and ability to survive and proliferate. Since the consequent *clonal expansions* are naturally constrained by the availability of resources (metabolites, oxygen, etc.), further mutations in the emerging heterogeneous tumor populations are necessary to provide additional *fitness* of different kinds that allow survival and proliferation in the unstable micro environment. Such further advantageous mutations will eventually allow some of their sub-clones to outgrow the competing cells, thus enhancing tumor's heterogeneity as well as its ability to overcome future limitations imposed by the rapidly exhausting resources. Competition, predation, parasitism and cooperation are indeed often observed in co-existing cancer clones [4].

In the well-known vision of Hanahan and Weinberg [5, 6], the phenotypic stages that characterize this multistep evolutionary process are called *hallmarks*. These can be acquired by cancer cells in many possible alternative ways, as a result of a complex biological interplay at several spatio-temporal scales that is still only partially deciphered [7]. In this framework, we distinguish alterations driving the hallmark acquisition process (i.e., *drivers*) by activating *oncogenes* or inactivating *tumor suppressor genes*, from those that are transferred to sub-clones without increasing their fitness (i.e., *passengers*) [8]. Driver identification is a modern challenge of cancer biology, as distinct cancer types exhibit very different combinations of drivers, some cancers display mutations in hundreds of genes [9], and the majority of drivers is mutated at low frequencies ("long tail" distribution), not allowing their detection by examining the recurrence at the population-level [10]. One can also use the evolutionary models to characterize, what may be called, *anti-hallmarks* – the phenotypes that are possible by the variational processes, but rarely found to be selected [11]. For instance, certain collections of driver mutations, whose individual members are often present in the patient genomes, are never seen jointly. These anti-hallmarks point to tumors' vulnerabilities, and thus, novel targets for therapeutic interventions.

Cancer clones harbour distinct types of "alterations". The *somatic* ones involve either few nucleotides or larger chromosomal regions, and are usually catalogued as mutations - i.e., *Single Nucleotide Variants* (SNVs) and *Structural Variants* (SVs) at multiple scales (insertions, deletions, inversions, translocations) – of which only some are detectable as *Copy Number Alterations* (CNAs), which appear to be most prevalent in many tumor types [12]. Also *epigenetic alterations*, such as DNA methylation and chromatin reorganization, play a key role in the process [13]. The overall picture is confounded by factors such as *genetic instability* [14], *aneuploidy* and *tumor-microenvironment* interplay [15], the latter involving stromal and immune-system cells with strong influence on the final effect of mutations [16]. Furthermore, spatial organization and tissue specificity play an essential role on tumor progression as well [17][1].

In this scenario, genomic alterations are related to the phenotypic properties of tumor cells via the structure and dynamics of *functional pathways*, in a process which has been only partially characterized [18–21]. In general, in fact, as there exist many equivalent ways to disrupt signaling and regulatory pathways, many mutations can provide equivalent fitness to cancer cells, leading to *alternative routes* to selective advantage across a population of tumors [22]. Practically, if multiple genes are equally functional for the same biological process, when any of those is altered the selection pressure on the others is diminished or even nullified [23]. Such genes, e.g., APC/CTNNB1 in colorectal cancer [24], therefore show a trend of *exclusivity* across a cohort – with few cases of

[1] We mention that much attention has been recently casted on newly discovered cancer genes affecting global processes that are apparently not directly related to cancer development, such as cell signaling, chromatin and epigenomic regulation, RNA splicing, protein homeostasis, metabolism and lineage maturation [10].

co-occurrent alterations. The same applies when disruptive alterations hit on the same gene, e.g., PTEN's mutations and deletions in prostate cancer [25].

An immediate consequence of this state of affair is the dramatic *heterogeneity* and *temporality* of cancer, both at the *inter-tumor* and at the *intra-tumor* levels [26]. The former manifests as different patients with the same cancer type can display few common alterations. This led to the development of techniques to stratify tumors into *subtypes* with different genomic signatures, prognoses and response to therapy [27].The latter refers to the noteworthy genotypic and phenotypic variability among the cancer cells within a single neoplastic lesion, characterized by the coexistence of more than one cancer clones with distinct evolutionary histories [28].

Cancer heterogeneity poses a serious problem from the diagnostic and therapeutic perspective as, for instance, it is now acknowledged that a single biopsy might not be representative of other parts of the tumor, hindering the problem of devising effective treatment strategies [4]. Therefore, the quest for an extensive etiology of cancer heterogeneity and for the identification of cancer evolutionary trajectories is nowadays central to cancer research, and attempt to exploit the massive amount of sequencing data available through public projects such as *The Cancer Genome Atlas* (TCGA) [29].

Such projects involve an increasing number of *cross-sectional* (epi)genomic profiles collected via single biopsies of patients affected by various cancer types, which might be used to extract trends of cancer evolution across a population of samples. Higher resolution data such as *multiple samples* collected from the same tumor [28], as well as *single-cell* sequencing data [30], might be complementarily used to face the same problem within a specific patient. However, either the lack of public data or problems of accuracy and reliability, currently prevent a straightforward application [31].

These different perspectives lead to the different mathematical formulations of the problem of *inferring a cancer progression model* from genomic data, which we shall examine at length in this paper [32]. Indeed, such models can either be focused to describe trends characteristics of a population, i.e. *ensemble-level*, or clonal progression in a *single-patient*. In general, both problems deal with understanding the *temporal ordering of somatic alterations* accumulating during cancer evolution, but use orthogonal perspectives and different input data – see Figure 1.

Ensemble-level cancer evolution. It may seem desirable to extract a *probabilistic graphical model* (PGM) explaining the statistical trend of accumulation of somatic alterations in a population of n cross-sectional samples collected from patients affected by a specific cancer. To make this problem independent of the experimental conditions in which tumors are gathered, we only consider the *list of alterations detected per sample* – thus, as 0/1 Bernoulli variables.

Much of the difficulty lies in estimating the true and unknown trends of *selective advantage* among genomic alterations in the data, from such observations. This hurdle is not unsurmountable, if we constrain the scope to only those alterations that are *persistent across tumor evolution in all sub-clonal populations*, since it yields a consistent model of a temporal ordering of mutations. Therefore, epigenetic and trascriptomic states, such as hyper and hypo-methylations or over and under expression, could only be used, provided that they are persistent thorough tumor development [34].

Historically, the linear colorectal progression by Vogelstein is an instance of a solution to the cancer progression modeling problem [35]. That approach was later generalized to accommodate *tree-models of branched evolution* [36–39] and, later, further generalized to the inference of directed acyclic graph (DAG) models by Beerenwinkel and others [40–42]. We contributed to this research

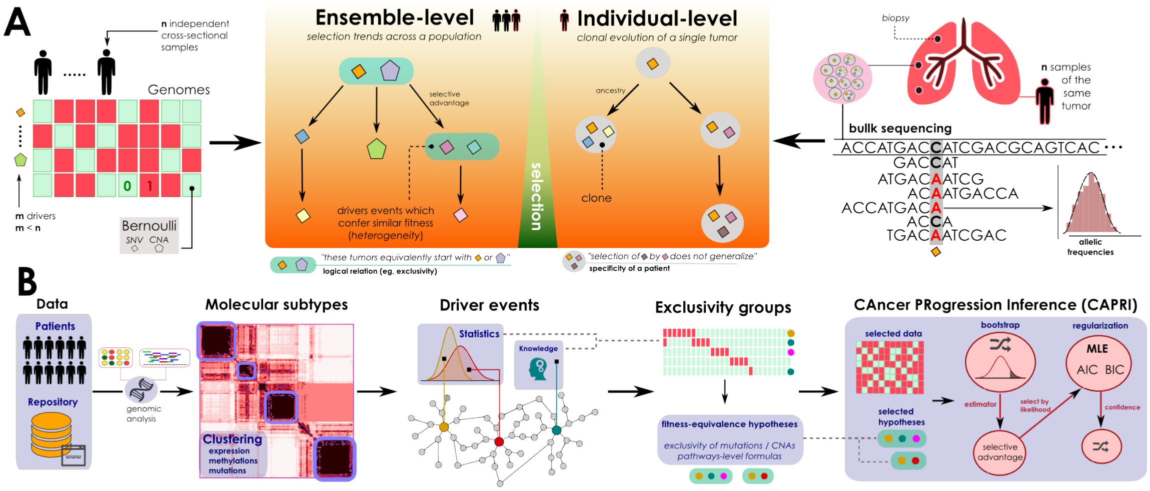

Figure 1: **(A) Problem statement.** (left) Inference of ensemble-level cancer progression models from a cohort of n independent patients (cross-sectional). By examining a list of somatic mutations or CNAs per patient (0/1 variables) we infer a probabilistic graphical model of the temporal ordering of fixation and accumulation of such alterations in the input cohort. Sample size and tumor heterogeneity harden the problem of extracting population-level trends, as this requires to account for patients' specificities such as multiple starting events. (right) For an individual tumor, its clonal phylogeny and prevalence is usually inferred from multiple biopsies or single-cell sequencing data. Phylogeny-tree reconstruction from an underlying statistical model of reads coverage or depths estimates alterations' prevalence in each clone, as well as ancestry relations. This problem is mostly worsened by the high intra-tumor heterogeneity and sequencing issues. **(B) A pipeline for ensemble-level inference.** The optimal pipeline includes several sequential steps to reduce tumor heterogeneity, before applying the CAPRI [33] algorithm. Available mutation, expression or methylation data are first used to stratify patients into distinct tumor molecular subtypes, usually by exploiting clustering tools. Then, subtype-specific alterations driving cancer initiation and progression are identified with statistical tools and on the basis of prior knowledge. Next is the identification of the fitness-equivalent groups of mutually exclusive alterations across the input population, again done with computational tools or biological priors. Finally, CAPRI processes a set of relevant alterations and such groups. Via bootstrap and hypothesis-testing, CAPRI extracts a set of "selective advantage relations" among them, which is eventually narrowed down via maximum likelihood estimation with regularization (with various scores). The ensemble-level progression model is obtained by combining such relations in a graph, and its confidence is assessed via bootstrap (see Online Methods).

prgram with two related algorithms: *CAncer PRogression Extraction with Single Edges* (CAPRESE, [43]) and *CAncer PRogression Inference* (CAPRI, [33]), which are currently implemented in TRONCO (TRanslational ONCOlogy), an open source R package available in standard repositories [44]. Both techniques rely on Suppes' theory of probabilistic causation to define estimators of selective advantage [45], are robust to the presence of noise in the data and perform well even with limited sample sizes. The former algorithm exploits shrinkage-like statistics to extract a tree model of progression, the latter combines bootstrap and maximum likelihood estimation with regularization to extract general directed acyclic graphs that capture branched, independent and confluent evolution. Both algorithms represent the current state-of-the-art to approach this problem, as they outperform others in speed, scale and predictive accuracy.

Clonal architecture in individual patients. At the time of this writing, technical and economical limitations of single-cell sequencing prevent a straightforward application of phylogeny inference algorithms to the reconstruction of the clonal evolutionary history of genomic alterations within a single tumor [46, 47]. Conversely, samples of cells collected from a single bulk tumor do not define an isogenic lineage [48] and most likely contain a large number of cells belonging to a collection of sub-clones resulting from the complex evolutionary history of the tumor, where the prevalence of a particular clone in time and its spatial distribution reflect its growth and proliferative fitness. To overcome hurdles such as this, many recent efforts have aimed at inferring the clonal signatures and prevalence in individual patients from sequencing data [28, 49].

The majority of attempts employ different strategies, usually based on Bayesian inference, to relate *allelic imbalance to cellular prevalence*, and benefit from multiple sample per patient, taken across time or space. In particular, most tools usually process a set of read counts from a high-coverage sequencing experiment to estimate *Variant Allele Frequency* (VAF). Some of them are based on the VAF analysis of specific SNVs [50, 51]. Recent algorithms attempt to minimize the error between the observed and inferred mutation frequencies with distinct optimization procedures [52–54]. Other approaches support explicitly short-read data and different types of data, such as CNAs, SNVs and B-allele fractions [55]. Distinct techniques, instead, use genome-wide segmented read-depth information to determine mixtures of subclonal CNA profiles [56, 57], while others use a generative approach to deconvolve sequencing data to clonal architectures [58]. Clearly, any of these approaches gains precision from high-coverage sequencing data, since high read counts yield high confidence estimate of allele frequency.

Results

Here, we report on the design, development and evaluation of an optimal, versatile and modular pipeline which exploits state-of-the-art tools to extract ensemble-level cancer progression models from cross-sectional data. We also show its applications in interpreting colorectal cancer data which, because of its high levels of heterogeneity, may be thought of as one of the most challenging case studies. Here, we are able to show that, in general, tools to detect cancer evolution at the ensemble-level can be effective even on single-patient data.

A pipeline to infer ensemble-level progression models

We have devised a customizable pipeline to infer ensemble-level cancer progression models from cross-sectional data with CAPRI [33]. To increase the statistical quality of its predictions our

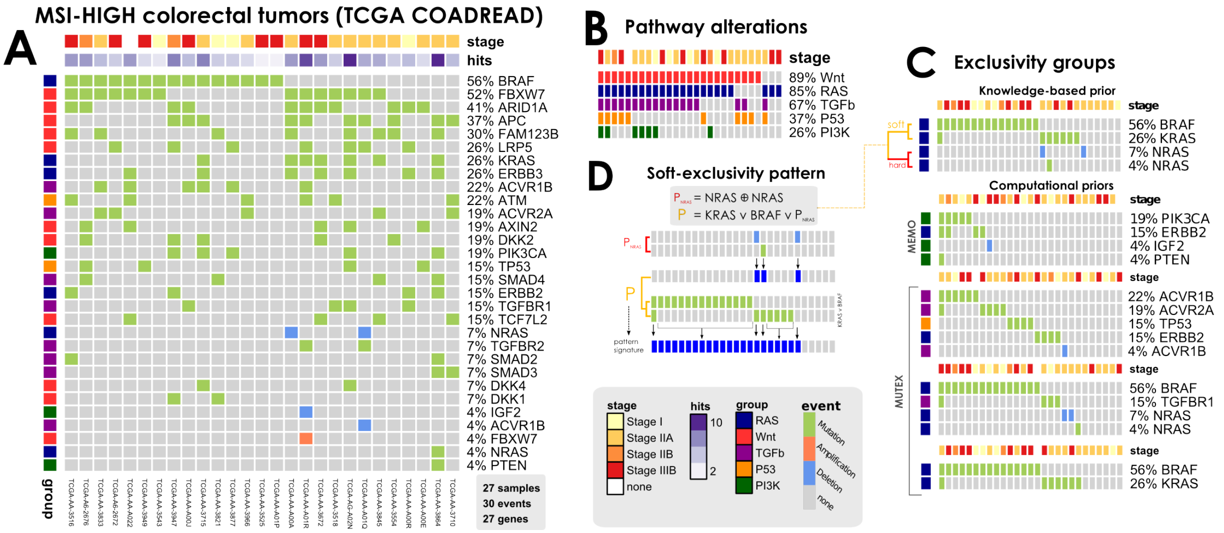

Figure 2: .**(A) Selected MSI-HIGH colorectal tumors used for inference.** Data from the TCGA COADREAD project [59], restricted to 27 samples with both somatic mutations and high-resolution CNA data available and a selection out of 33 driver genes annotated to WNT, RAS, PI3K, TGF-β and P53 pathways. This dataset is used to infer the model in Figure 4. **(B) Altered pathways.** Mutations and CNAs in these tumors mapped to pathways confirm heterogeneity even at the pathway-level. **(C) Mutually exclusive alterations.** Groups were obtained from [59] - which run the MEMO [60] tool - and by MUTEX [23] tool. Plus, previous knowledge about exclusivity among genes in the RAS pathway was exploited. **(D) Construction of a formula.** A Boolean formula inputed to CAPRI to test the hypothesis that alterations the RAF genes KRAS, NRAS and BRAF confer equivalent selective advantage. The formula accounts for hard exclusivity of alterations in NRAS mutations and deletions, jointly with soft exclusivity with KRAS and NRAS alterations.

pipeline pre-processes data to diminish the confounding effects of inter and intra-tumor heterogeneity. At a high-level, we shall thus identify: (*i*) biologically meaningful subtypes with similar molecular profiles via tumor stratification, (*ii*) the set of driver alterations and (*iii*) the groups of fitness-equivalent (i.e., exclusive) alterations.

Thus, this pipeline, which is briefly sketched in Figure 1 and detailed as Online Methods, is similar in spirit to those implemented by consortia such as TCGA to analyze huge populations of cancer samples [59, 61]. One of the main novelties of our approach, which is only possible by the specific features of hypothesis-testing provided by CAPRI [33], is the exploitation of groups of exclusive alterations as a proxy to detect fitness-equivalent routes of cancer progression. Thus, CAPRI may be thought of as an ideal tool for efficient and theoretically-grounded investigations in population-based studies on cancer genomics.

Our approach allows one to produce a progression model for virtually every cancer subtype identified in the input cohort, which shall be characteristic of the population trends of cancer initiation and progression. In the following, we empirically characterize the efficacy of our approach in processing colorectal cancer data from TCGA project [59], demonstrating that we were able to re-discover most of the existing body of knowledge about colorectal tumor progression or to propose further experimentally verifiable hypotheses[2].

Evolution in a population of MSI/MSS colorectal tumors with CAPRI

It is common knowledge that *colorectal cancer* (CRC) is a heterogeneous disease comprising different molecular entities. Since similar tumors are most likely to behave in a similar way, grouping tumors with homogeneous characteristics may be useful to define personalized therapies. Indeed, it is currently accepted that colon tumors can be classified according to their global genomic status into two main types: *microsatellite instable tumors* (MSI), further classified as high or low, and *microsatellite stable* (MSS) tumors (also known as tumors with chromosomal instability). This taxonomy plays a significant role in determining pathologic, clinical and biological characteristics of CRC tumors [62]. Thus, MSS tumors are characterized by changes in chromosomal copy number and show worse prognosis [63, 64]. On the contrary, the less common MSI tumors (about 15% of sporadic CRC) are characterized by the accumulation of a high number of mutations and show predominance in females, proximal colonic localization, poor differentiation, tumor-infiltrating lymphocytes and a better prognosis [65]. In addition, MSS and MSI tumors exhibit different responses to chemotherapeutic agents [66,67]. Regarding molecular progression, it is also well established that each subtype arises from a distinctive molecular mechanism. While MSS tumors generally follow the classical adenoma-to-carcinoma progression (sequential APC-KRAS-TP53 mutations) described in the seminal work by Vogelstein and Fearon [68], MSI tumors results from the inactivation of DNA mismatch repair genes like MLH-1 [65].

We instantiated the pipeline discussed as Online Methods to process MSI-HIGH - hereby shortly denoted as MSI - and MSS colorectal tumors collected from the The Cancer Genome Atlas project *"Human Colon and Rectal Cancer"* (COADREAD, [59]) – see Supplementary Figure S1. Details on the implementation are available as Supplementary Material, as well as source code to replicate this study. COADREAD has enough samples to implement a *training/test* statistical validation of our findings - see Supplementary Table S1 and Supplementary Figure S2. In brief, we split subtypes by the microsatellite status of each tumor, and select somatic mutations and focal CNAs in 33 driver genes manually annotated to 5 pathways in [59] - WNT, RAF, TGF-β, PI3K and P53. Groups of exclusive alterations were scanned by MUTEX [23] (Supplementary Table S2), and fetched by [59] using the MEMO [60] tool; groups were used to create CAPRI's formulas, see Supplementary Table S3. Data for MSI tumors are shown in Figure 2, for MSS tumors are shown in Supplementary Figure S3 and S4. CAPRI was run, on each subtype, by selecting recurrent alterations from the pool of 33 pathway genes and using both AIC/BIC regularizators.

The model inferred for MSS tumors is in Figure 3, the model for MSI-HIGH ones is in Figure 4. Each edge in the graph mirrors selective advantage among the upstream and downstream nodes, as estimated by CAPRI from training datasets (statistics: $p < 0.05$, 100 non-parametric bootstraps); only the minimum amount of edges is selected to maximize the likelihood of data (see Online Methods). As statistical validation of these models, we mark those relations that display significant p-values in the test datasets, and rank them if they contribute (or otherwise) to max-likelihood. For some edges it is not possible to provide a validation, as some upstream or downstream event may be missing in the test dataset, while other edges do not show statistical evidence in the test datasets.

[2] We remark that in-vitro and in-vivo experiments could provide an optimal validation for the newly suggested selective advantage relations and hypotheses, yet this is out of the scope of the current work.

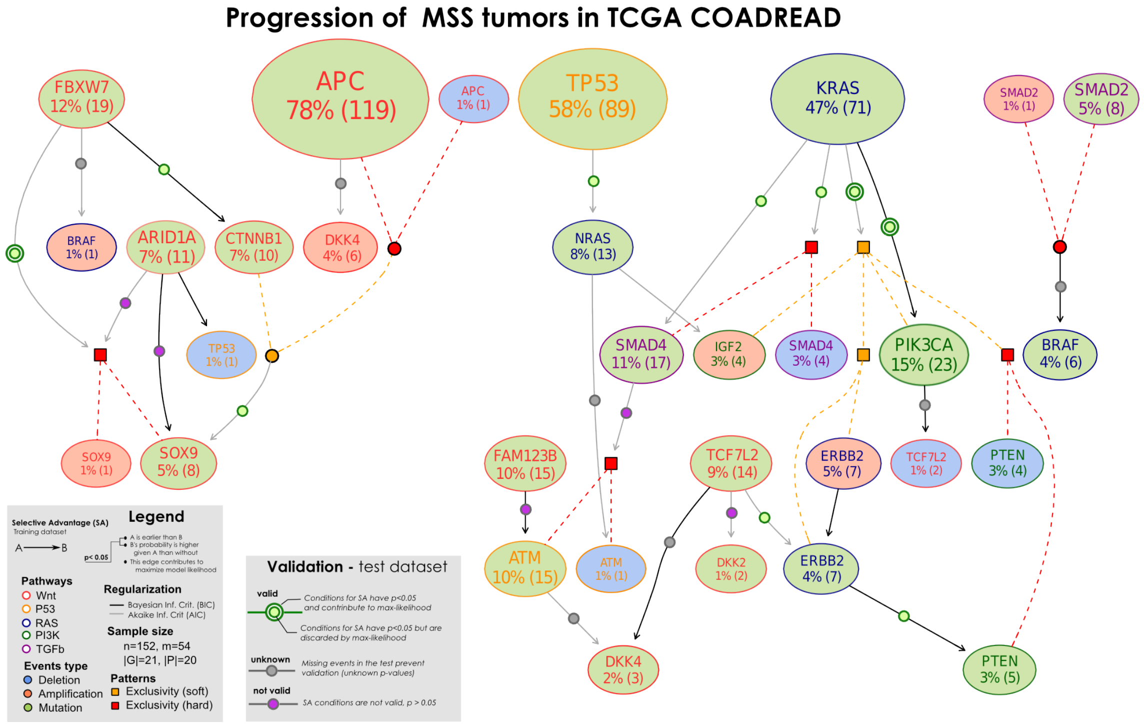

Figure 3: **Progression model of MSS colorectal tumors.** Selective advantage relations inferred by CAPRI constitute MSS progression; input dataset in Supplementary Figure S3 and S4. Formulas written on groups of exclusive alterations are expanded with colors representing the type of exclusivity (red for hard, orange for soft). We mark also those relations that display significant p-values in the test dataset, and rank them if they contribute (or otherwise) to max-likelihood. For all MSS tumors in COADREAD, we find at high-confidence selection of SOX9 alterations by FBXW7 mutations (with AIC), as well as selection of alterations in PI3K genes by the KRAS mutations (direct, with BIC, and via the MEMO group, with AIC).

Interpretation of the models. Our models capture the well-known features distinguishing MSS and MSI tumors, e.g., APC-KRAS-TP53 primary events and chromosomal aberrations in MSS, versus BRAF mutations in MSI, which lacks chromosomal alterations. Of all 33 driver genes, 15 are common to both models - e.g., APC, BRAF, KRAS, NRAS, TP53 and FAM123B among others (mapped to pathways like WNT, MAPK, apoptosis or activation of T-cell lymphocites), although in different relationships (position in the model), whereas new (previously un-implicated) genes stood out from our analysis and deserve further research.

MSS (Microsatellite Stable). In agreement with the known literature, we identify KRAS, TP53 and APC as primary and PTEN as late events in the carcinogenesis, as well as NRAS and KRAS determining two independent evolution branches, the former being "selected by" TP53 mutations, i.e. being a downstream event in the model, the latter "selecting for" PIK3CA mutations. The leftmost portion of the model links many WNT genes, in agreement with the observation that multiple concurrent lesions affecting such pathway confer selective advantage. In this

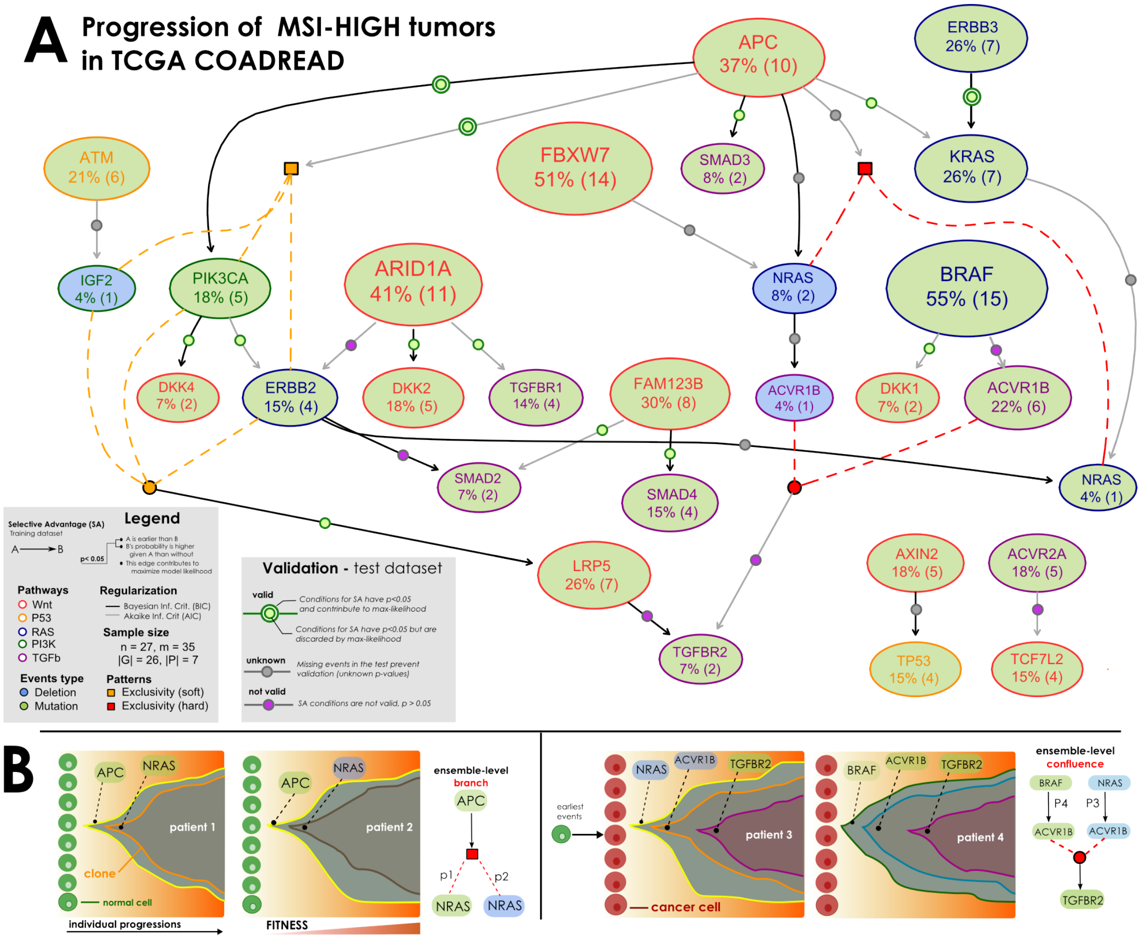

Figure 4: **(A) Progression model of MSI-HIGH colorectal tumors.** Selective advantage relations inferred by CAPRI constitute MSI-HIGH progression; input dataset in Figure 2. Formulas written on groups of exclusive alterations are expanded as in Figure 3. We note the high-confidence in APC mutations selecting for PIK3CA ones, both in training and test via BIC, as well as selection of the MEMO group (ERBB2/PIK3CA mutations or IGF2 deletions) predicted by AIC. Similarly, we find a strong selection trend among mutations in ERBB2 and KRAS. For each relation, confidence is annotated as in Figure 3. **(B) Predicting clonal expansion from the model.** Evolutionary trajectories from two example selective advantage relations. APC-mutated clones shall enjoy expansion, up to acquisition of further selective advantage via mutations or homozygous deletions in NRAS. These cases should be representative of different individuals in the population, and the ensemble-level interpretation should be that "APC mutations select for NRAS alterations, in hard exclusivity" as no sample harbour both alterations. A similar argument can show that the clones of patients harbouring distinct alterations in ACVR1B - and different upstream events - will enjoy further selective advantage by mutating TGFBR2 gene.

respect, our model predicts multiple routes for the selection of alterations in SOX9 gene, a transcription factor known to be active in colon mucosa [69]. Its mutations are indeed selected by APC/CTNNB1 alterations or by FBXW7, an early mutated gene that both directly, and in a redundant way via CTNNB1, relates to SOX9. The SOX family of transcription factors have emerged as modulators of canonical WNT/β-catenin signaling in many disease contexts, with evidences that multiple SOX proteins physically interact with β-catenin and modulate the transcription of WNT-target genes, as well as with evidences of regulating of SOX's expression by WNT, resulting in feedback regulatory loops that fine-tune cellular responses to β-catenin/TCF activity [70]. Also interestingly, FBXW7 has been previously reported to be involved in the malignant transformation from adenoma to carcinoma [71], and it was recently shown that SCFFbw7, a complex of ubiquitin ligase that contains such gene, targets several oncogenic proteins including SOX9 for degradation [72]; this relation has high-confidence also in the test dataset. The rightmost part of the model involves genes from other pathways, and outlines the relation between KRAS and the PI3K pathway. We indeed find, consistently in the training and test, selection of PIK3CA mutations by KRAS ones, as well as selection of the whole MEMO module, which is responsible for the activation of the PI3K pathway [59]. SMAD proteins relate either to KRAS or BRAF genes, and FAM123B, TCF7L2 converge in DKK2 or DKK4 which is interesting as these four genes are implicated in the WNT signalling pathway. It is also worth pointing that the model predicts a selection trend among SOX9/ARID1A and ATM/FAM123B; however, given that the within these couples the events have very similar frequencies, it is not possible to confidently assess the direction of the selectivity relations, which, in fact, are found to be reversed in the test dataset.

MSI (Microstaellite Instable). In agreement with the current literature, BRAF is the most commonly mutated gene in MSI tumors [73]. CAPRI predicted convergent evolution of tumors harbouring FBXW7 or APC mutations towards deletions of NRAS gene, as well as selection of SMAD2 or SMAD4 mutations by FAM123B mutations, for these tumors. Relevant to all MSI tumors seems again the role of the PI3K pathway. Indeed, a relation among APC and PIK3CA mutations was inferred with a high confidence in both training and test datasets, consistent with recent experimental evidences pointing at a synergistic role of these mutations, which co-occurr in the majority of human colorectal cancers [74]. Similarly, we find consistently a selection trend among APC and the whole MEMO module. Interestingly, both mutations in APC and ERBB3 select for KRAS mutations, which might point to interesting therapeutic implications (see Discussion). In contrast, mutations in BRAF mostly select for mutations in ACVR1B, a receptor that once activated phosphorylates SMAD proteins. It forms receptor complex with ACVR2A, a gene mutated in these tumors that selects for TCF7L2 mutations. Tumors harbouring TP53 mutations are those selected by exhibit mutations in AXIN2, a gene implicated in WNT signalling pathway, and related to instable gastric cancer development [75]. Inactivating mutations in this gene are important, as it provides serrated adenomas with a mutator phenotype in the MSI tumorigenic pathway [76]. Thus, our results reinforce its putative role as driver gene in these tumors.

By comparing these models we can find similarity in the prediction of a potential new early event for CRC formation, FBXW7, as other authors have recently described [71]. This tumor suppressor is frequently inactivated in human cancers, yet the molecular mechanism by which it exerts its antitumor activity remains unexplained [77], and our models provide a new hypothesis in this respect. We also note that genes involved in these models exhibit distinctive functional features, suggesting

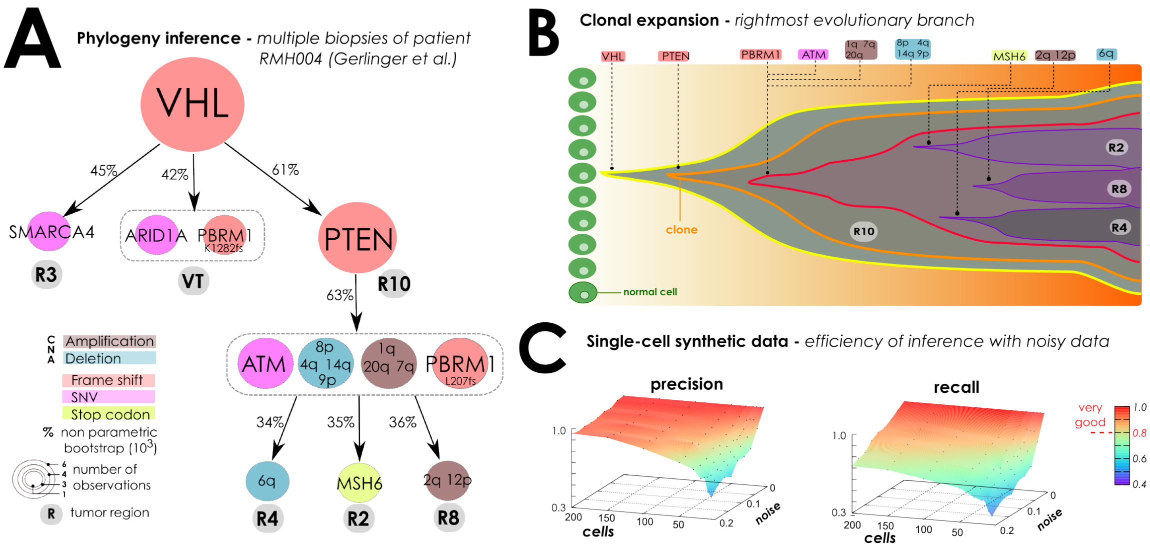

Figure 5: **(A) Application of ensemble-level algorithms to individual-patient data.** With data provided by Gerlinger *et al.* in [49], we infer a patient-specific clonal evolution from 6 biopsies of a clear cell renal carcinoma (5 primary tumor, 1 from the thrombus in the renal vein, VT). Validated non-synonymous mutations are selected for VHL, SMARCA4, PTEN, PBMR1 (p.Lys1282fs and p.Leu207fs), ARID1A, ATM and MSH6 genes. CNAs are detected on 12 chromosomes. For this patient, both region-specific allele frequencies and Bernoulli profiles are provided. Thus, we can extract a clonal tree, signature and diffusion of each clone, by the unsupervised CAPRESE algorithm [43]. **(B) Clonal expansion in patient RMH004.** The unsupervised model inferred by CAPRESE predicts an analogous clonal expansion observed in [49], and extracted with most parsimonious phylogeny tree reconstruction from allelic frequencies, and hand-curated for selection of the optimal model. For simplicity, we show only expansion of the sub-clones harbouring PTEN's frame shift mutation. **(C) Inference from single-cell data.** We estimate average precision and recall from single-cell sequencing data sampled from the phylogeny history of patient RMH004 (details as Supplementary Material). Sampled datasets vary for number of sequenced cells, $n \leq 200$, and noise in the data - as a model of potential experimental errors in data collection, manipulation and analysis.

that each one imparts alterations in different pathways in the early stages of carcinogenesis.

Private alterations of these tumors denote potential different progression mechanisms. Mutations or CNAs specific to MSS tumors involve intracellular genes like CTNNB1 or PTEN. In contrast, private MSI mutations appear in membrane receptors such as ACVR1B, ACVR2A, ERBB3, LRP5, TGFBR1 and TGFBR2; as well as in secreted proteins like IGF2.This suggests that MSI tumors need to disturb cell-cell and/or cell-microenvironment communication to grow, as their lesions accumulate in private pathways like cytokine-cytokine receptor, endocytosis and TGF-β signalling pathway. On the other hand, genes specific to MSS tumors are implicated in P53, MTOR, sodium transport and inositol phosphate metabolism.

Inference of patient-specific clonal evolution with CAPRESE

We also discovered that the CAPRESE [43] algorithm can be used to successfully reconstruct the clonal architecture in individual patients, an instance of tree-phylogeny of Figure 1. This result is indicative of the power of the selective advantage scores à-la-Suppes [45], even outside the scope of cross-sectional data. We performed our analysis on data from Gerlinger *et al.*, who have recently used multi-region targeted exome sequencing (> 70x coverage) to resolve the genetic architecture and evolutionary histories of ten *clear cell renal carcinomas* [49].

Besides quantification of intra-tumor heterogeneity, their work found that loss of the 3p arm and alterations of the Von Hippel-Lindau tumor suppressor gene VHL are the only events ubiquitous among their patients. In Figure 5 we show the clonal evolution estimated for one of those patients, RMH004, computed with CAPRESE (shrinkage coefficient $\lambda = 0.5$, time < 1 sec) from the Bernoulli 0/1 profiles provided in Supplementary Table 3 and Figure 4 of [49], with non-parametric bootstrap confidence (time < 6 sec). This model may be compared to the one inferred by processing the region-specific VAF with a max-mini optimization of most parsimonious evolutionary trees [78], and performing selection-by-consensus when multiple optimal solutions exist - Supplementary Figure 9 in [49]. CAPRESE requires no arbitrarily defined curation criteria to select the optimal tree, as it constructively searches for a solution which, in this case, is analogous in suggesting *parallel evolution of subclones* via deregulation of the SWI/SNF chromatin-remodeling complex – i.e., as may be noted from multiple clones with distinct PBMR1 mutations. Finally, the approach in [78], estimates also the number of non-synonymous mutations acquired on a certain edge of the tree. While our model is silent about this, it is very likely due to the limitations imposed by the lower-resolution and small sample size of the data – 9 events from 8 regions, and not the VAFs for all alleles.

Single-cell synthetic data. We estimate the efficiency of our approach to single-cell sequencing data, as if it was collected from patient RMH004 (synthetic data generated from the clonal phylogeny architecture of Figure 5). To mimic a poor reliability of this technology, to each sampled cell a noise model which accounts for false positives and negatives in the calls of their genomic alterations is applied. Performance is measured as the fraction of true-positive and negative ancestry relations inferred among cells (*precision* and *recall*), as a function of the number of sequenced cells and noise level. Results indicate a very good performance even with very small number of cells and reasonable noise levels, hinting at a promising application with this technology. Complete details for synthetic data generation and further performance measures are provided as Supplementary Material.

Discussion

In this paper, we have continued our exploration of the nature of somatic evolution in cancer, but with an emphasis on colorectal cancer and jointly with epidemiologists who study the disease. The nature of the proposed model of somatic evolution in cancer not only supports the heterogeneity and temporality seen in tumor population, but also suggests a selectivity/causality relation that can be used in analyzing (epi)genomic data and exploited in therapy design. We have shown in this paper that our approach can be effective in extracting evolutionary trajectories for cancer progression both at the level of populations and individual patients. In the former case we have set up a pipeline to minimize the confounding effects imputable to tumor heterogeneity, and we have applied it to a highly-heterogeneous cancer such as colorectal. In the latter we have have shown how our techniques can be readily applied to reconstruct clonal philogeny from multi-sample data, with an application to clear renal cell carcinoma.

Emphasis of this work is on the population-level inference of cancer progression. Our pipeline has been able to infer the role of many known events in colorectal cancer progression, and sheds light on the roles of new players such as FBXW7, SOX9 or AXIN2 which deserve further investigation.

In colon carcinogenesis, although each model identifies characteristic early mutations suggesting different initiation events, both model appear to be "converging" in common pathways and functions such as WNT or MAPK. However, each progression model recapitulates private functions related to microenvironment communication in the case of MSI tumors and with intracellular signalling in the case of MSS tumors.

Our models might have implications also for treatment strategies. For instance, some of the relations that we observed in our models might point to cancer hallmarks to be exploited for therapy design. As an example, the interesting relation between SOX9 and FBXW7 in microsatellite stable tumors, interpreted together with genes such as TP53, might point to a *DNA fragmentation and cell cycle arrest* hallmark as these genes are sensitive for cell-cycle regulation - via the p53 protein - and for degradation and senescence, via SOX9. This would also be supported by other cancer studies since transcription factor sox9 seems to play an important role in colon cancer development [79].

Personalized treatment strategies might also benefit from our analyses. In fact, KRAS status is currently used as a predictive biomarker for the selection of CRC patients susceptible to be treated with anti-EGFR targeted therapy [80]. However, resistance in KRAS wild-type tumors has been observed that could be caused by mutated genes in the same pathway other than KRAS [81]. Models could then be useful to detect these alternative mutated genes, like NRAS or ERBB family, which are characteristic of the population of observed tumors.

Remarkably, we could prove the effectiveness of our approach in inferring the clonal evolutionary history of single cancer patients as well, by showing a successful application to multiple-biopsy data on clear cell renal carcinoma. We also demonstrated that, in case of single-cell synthetic data generated by sampling a real clonal phylogenetic architecture, our inference techniques provide an excellent performance with a very limited number of samples and also in presence of a certain level of experimental noise. Even if further investigations on this topic are underway, these preliminary results point at the efficiency of our algorithmic framework in inferring the clonal architecture of single cancer patients, especially in anticipation of the expected increasing availability and reliability of single-cell sequencing data.

Authors contribution. The pipeline was discovered and realized by MA's Bioinformatics lab at University of Milan-Bicocca, within a project led and supervised by GC. GC, AG and DR designed the pipeline, GC, DR and LDS coded and executed it. Data gathering and models interpretation was done by GC, LDS, DR, AG together with VM and RSP. GM, MA, VM and BM provided overall organizational guidance and discussion. GC, AG and RSP wrote the original draft of the paper, which all authors reviewed and revised in the final form. This work follows up on an earlier project started by BM and involving BIMIB, subsequently including ICO after the 2014 School on Cancer, Systems and Complexity (CSAC).

Financial support. MA, GM, GC, AG, DR acknowledge Regione Lombardia (Italy) for the research projects RetroNet through the ASTIL Program [12-4-5148000-40]; U.A 053 and Network Enabled Drug Design project [ID14546A Rif SAL-7], Fondo Accordi Istituzionali 2009. BM acknowledges founding by the NSF grants CCF-0836649, CCF-0926166 and a NCI-PSOC grant. VM and RSP acknowledge the Instituto de Salud Carlos III supported by The European Regional Development Fund (ERDF) grants PI11-01439, PIE13/00022, the Spanish Association Against Cancer (AECC) Scientific Foundation, and the Catalan Government DURSI, grant 2014SGR647.

References

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

[2] Fidler, I. J. Tumor heterogeneity and the biology of cancer invasion and metastasis. *Cancer Research* **38**, 2651–2660 (1978).

[3] Dexter, D. L. *et al.* Heterogeneity of tumor cells from a single mouse mammary tumor. *Cancer Research* **38**, 3174–3181 (1978).

[4] Merlo, L. M., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer* **6**, 924–935 (2006).

[5] Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).

[6] Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

[7] Huang, S., Ernberg, I. & Kauffman, S. Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. In *Seminars in Cell & Developmental Biology*, 7, 869–876 (Elsevier, 2009).

[8] Futreal, P. A. *et al.* A census of human cancer genes. *Nature Reviews Cancer* **4**, 177–183 (2004).

[9] Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).

[10] Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).

[11] Loohuis, L. O., Witzel, A. & Mishra, B. Cancer hybrid automata: model, beliefs and therapy. *Information and Computation* **236**, 68–86 (2014).

[12] Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature Genetics* **45**, 1134–1140 (2013).

[13] Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome - biological and translational implications. *Nature Reviews Cancer* **11**, 726–734 (2011).

[14] Weinberg, R. *The Biology of Cancer* (Garland Science, 2013).

[15] Albini, A. & Sporn, M. B. The tumour microenvironment as a target for chemoprevention. *Nature Reviews Cancer* **7**, 139–147 (2007).

[16] Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).

[17] Nowak, M. A., Michor, F. & Iwasa, Y. The linear process of somatic evolution. *Proceedings of the National Academy of Sciences* **100**, 14966–14969 (2003).

[18] Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nature Medicine* **10**, 789–799 (2004).

[19] Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007).

[20] Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).

[21] Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).

[22] Nowak, M. A. *Evolutionary Dynamics* (Harvard University Press, 2006).

[23] Babur, Ö. *et al.* Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biology* **16** (2015).

[24] Gerstein, A. V. *et al.* APC/CTNNB1 (β-catenin) pathway alterations in human prostate cancers. *Genes, Chromosomes and Cancer* **34**, 9–16 (2002).

[25] Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).

[26] Fisher, R., Pusztai, L. & Swanton, C. Cancer heterogeneity: implications for targeted therapeutics. *British Journal of Cancer* **108**, 479–485 (2013).

[27] Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

[28] Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England Journal of Medicine* **366**, 883–892 (2012).

[29] The Cancer Genome Atlas (TCGA). URL `https://tcga-data.nci.nih.gov`.

[30] Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).

[31] Eberwine, J., Sul, J.-Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. *Nature Methods* **11**, 25–27 (2014).

[32] Beerenwinkel, N., Schwarz, R. F., Gerstung, M. & Markowetz, F. Cancer evolution: mathematical models and computational inference. *Systematic biology* **64**, e1–e25 (2015).

[33] Ramazzotti, D. *et al.* CAPRI: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics* **31**, 3016–3026 (2015).

[34] Ramchandani, S., Bhattacharya, S. K., Cervoni, N. & Szyf, M. DNA methylation is a reversible biological signal. *Proceedings of the National Academy of Sciences* **96**, 6107–6112 (1999).

[35] Vogelstein, B. *et al.* Genetic alterations during colorectal-tumor development. *The New England Journal of Medicine* **319**, 525–532 (1988).

[36] Desper, R. *et al.* Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of Computational Biology* **6**, 37–51 (1999).

[37] Desper, R. *et al.* Distance-based reconstruction of tree models for oncogenesis. *Journal of Computational Biology* **7**, 789–803 (2000).

[38] Szabo, A. & Boucher, K. Estimating an oncogenetic tree when false negatives and positives are present. *Mathematical Biosciences* **176**, 219–236 (2002).

[39] Beerenwinkel, N. *et al.* Learning multiple evolutionary pathways from cross-sectional data. *Journal of Computational Biology* **12**, 584–598 (2005).

[40] Beerenwinkel, N., Eriksson, N. & Sturmfels, B. Conjunctive Bayesian networks. *Bernoulli* 893–909 (2007).

[41] Gerstung, M., Baudis, M., Moch, H. & Beerenwinkel, N. Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics* **25**, 2809–2815 (2009).

[42] Misra, N., Szczurek, E. & Vingron, M. Inferring the paths of somatic evolution in cancer. *Bioinformatics* **17**, 245663 (2014).

[43] Olde Loohuis, L. *et al.* Inferring tree causal models of cancer progression with probability raising. *PLOS ONE* **9**, e115570 (2014).

[44] De Sano, L. *et al.* TRONCO: an R package for the inference of cancer progression models from heterogeneous genomic data. *Submitted* **archived at** `http://dx.doi.org/10.1101/027474` (2015). URL `http://bimib.disco.unimib.it/`.

[45] Suppes, P. *A Probabilistic Theory of Causality* (North-Holland Publishing Company Amsterdam, 1970).

[46] Navin, N. E. Cancer genomics: one cell at a time. *Genome Biology* **15**, 452 (2014).

[47] Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).

[48] Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).

[49] Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics* **46**, 225–233 (2014).

[50] Miller, C. A. *et al.* SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLOS Computational Biology* **8**, e1003665 (2014).

[51] Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nature Methods* **11**, 396–398 (2014).

[52] Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. & Morris, Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* **15**, 35 (2014).

[53] Malikic, S., McPherson, A. W., Donmez, N. & Sahinalp, C. S. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics* **31**, 1349–1356 (2015).

[54] 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).

[55] Fischer, A., Vázquez-García, I., Illingworth, C. J. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell Reports* **7**, 1740–1752 (2014).

[56] Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biol* **14**, R80 (2013).

[57] Oesper, L., Satas, G. & Raphael, B. J. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics* **30**, 3532–3540 (2014).

[58] Zare, H. *et al.* Inferring clonal composition from multiple sections of a breast cancer. *PLOS Computatioanl Biology* **7**, e1003703 (2014).

[59] The Cancer Genome Atlas Network *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

[60] Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research* **22**, 398–406 (2012).

[61] Network, C. G. A. R. *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England Journal of Medicine* **368**, 2059 (2013).

[62] Ogino, S. & Goel, A. Molecular classification and correlates in colorectal cancer. *The Journal of Molecular Diagnostics* **10**, 13–27 (2008).

[63] Markowitz, S. D. & Bertagnolli, M. M. Molecular basis of colorectal cancer. *The New England Journal of Medicine* **361**, 2449–2460 (2009).

[64] Walther, A. *et al.* Genetic prognostic and predictive markers in colorectal cancer. *Nature Reviews Cancer* **9**, 489–499 (2009).

[65] Vilar, E. & Gruber, S. B. Microsatellite instability in colorectal cancer - the stable evidence. *Nature reviews Clinical oncology* **7**, 153–162 (2010).

[66] Klingbiel, D. *et al.* Prognosis of stage II and III colon cancer treated with adjuvant 5-fluorouracil or FOLFIRI in relation to microsatellite status: results of the PETACC-3 trial. *Annals of Oncology* **26**, 126–132 (2015).

[67] Warusavitarne, J. & Schnitzler, M. The role of chemotherapy in microsatellite unstable (MSI-H) colorectal cancer. *International Journal of Colorectal Disease* **22**, 739–748 (2007).

[68] Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).

[69] Abdel-Samad, R. *et al.* MiniSOX9, a dominant-negative variant in colon cancer cells. *Oncogene* **30**, 2493–2503 (2011).

[70] Kormish, J. D., Sinner, D. & Zorn, A. M. Interactions between sox factors and wnt/β-catenin signaling in development and disease. *Developmental Dynamics* **239**, 56–68 (2010).

[71] Li, L. *et al.* Sequential expression of miR-182 and miR-503 cooperatively targets FBXW7, contributing to the malignant transformation of colon adenoma to adenocarcinoma. *The Journal of Pathology* **234**, 488–501 (2014).

[72] Hong, X., Liu, W., Inuzuka, H., Liu, L. & Pine, S. R. Negative regulation of Sox9 by glycogen synthase kinase 3 beta phosphorylation and SCFFbw7-dependent ubiquitination in cancer. *Cancer Research* **75**, 1957 (2015).

[73] Kim, J. H. & Kang, G. H. Molecular and prognostic heterogeneity of microsatellite-unstable colorectal cancer. *World Journal of Gastroenterology* **20**, 4230 (2014).

[74] Deming, D. A. *et al.* PIK3CA and APC mutations are synergistic in the development of intestinal cancers. *Oncogene* **33**, 2245–2254 (2014).

[75] Kim, M. S., Kim, S. S., Ahn, C. H., Yoo, N. J. & Lee, S. H. Frameshift mutations of Wnt pathway genes AXIN2 and TCF7L2 in gastric carcinomas with high microsatellite instability. *Human Pathology* **40**, 58–64 (2009).

[76] Muto, Y. *et al.* DNA methylation alterations of AXIN2 in serrated adenomas and colon carcinomas with microsatellite instability. *BMC Cancer* **14**, 466 (2014).

[77] Zhan, P. *et al.* FBXW7 negatively regulates ENO1 expression and function in colorectal cancer. *Laboratory Investigation* **9**, 9951004 (2015).

[78] Purdom Jr, P. W., Bradford, P. G., Tamura, K. & Kumar, S. Single column discrepancy and dynamic max-mini optimizations for quickly finding the most parsimonious evolutionary trees. *Bioinformatics* **16**, 140–151 (2000).

[79] Matheu, A. *et al.* Oncogenicity of the developmental transcription factor sox9. *Cancer research* **72**, 1301–1315 (2012).

[80] Bokemeyer, C. *et al.* Fluorouracil, leucovorin, and oxaliplatin with and without cetuximab in the first-line treatment of metastatic colorectal cancer. *Journal of Clinical Oncology* **27**, 663–671 (2009).

[81] Caiazza, F. *et al.* Targeting EGFR in metastatic colorectal cancer beyond the limitations of KRAS status: alternative biomarkers and therapeutic strategies. *Biomarkers in medicine* **9**, 363–375 (2015).

[82] Bennett, J. M. *et al.* Proposals for the classification of the acute leukaemias french-american-british (FAB) co-operative group. *British Journal of Haematology* **33**, 451–458 (1976).

[83] Lu, J. *et al.* MicroRNA expression profiles classify human cancers. *Nature* **435**, 834–838 (2005).

[84] Gao, Y. & Church, G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* **21**, 3970–3975 (2005).

[85] de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B. & Schliep, A. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics* **9**, 497 (2008).

[86] Network, C. G. A. R. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).

[87] Konstantinopoulos, P. A. *et al.* Gene expression profile of BRCAness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. *Journal of Clinical Oncology* **28**, 3555–3561 (2010).

[88] Reis-Filho, J. S. & Pusztai, L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet* **378**, 1812–1823 (2011).

[89] Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nature Methods* **10**, 1108–1115 (2013).

[90] Zhong, X., Yang, H., Zhao, S., Shyr, Y. & Li, B. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. *BMC Genomics* **16**, S7 (2015).

[91] Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

[92] Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Research* **21**, e169 (2012).

[93] Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).

[94] Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Research* **22**, 1589–1598 (2012).

[95] Tamborero, D., Lopez-Bigas, N. & Gonzalez-Perez, A. Oncodrive-CIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression. *PLOS ONE* **8**, e55489 (2013).

[96] Gundem, G. *et al.* IntOGen: integration and data mining of multidimensional oncogenomic data. *Nature Methods* **7**, 92–93 (2010).

[97] Yeang, C.-H., McCormick, F. & Levine, A. Combinatorial patterns of somatic gene mutations in cancer. *The FASEB Journal* **22**, 2605–2622 (2008).

[98] Miller, C. A., Settle, S. H., Sulman, E. P., Aldape, K. D. & Milosavljevic, A. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Medical Genomics* **4**, 34 (2011).

[99] Vandin, F., Upfal, E. & Raphael, B. J. De novo discovery of mutated driver pathways in cancer. *Genome Research* **22**, 375–385 (2012).

[100] Zhao, J., Zhang, S., Wu, L.-Y. & Zhang, X.-S. Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* **28**, 2940–2947 (2012).

[101] Leiserson, M. D., Blokh, D., Sharan, R. & Raphael, B. J. Simultaneous identification of multiple driver pathways in cancer. *PLOS Computational Biology* **5**, e1003054 (2013).

[102] Leiserson, M. D., Wu, H.-T., Vandin, F. & Raphael, B. J. CoMEt: A statistical approach to identify combinations of mutually exclusive alterations in cancer. In *Research in Computational Molecular Biology*, 202–204 (Springer, 2015).

[103] Hua, X. *et al.* MEGSA: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations. *bioRxiv* (2015).

[104] Szczurek, E. & Beerenwinkel, N. Modeling mutual exclusivity of cancer mutations. In *Research in Computational Molecular Biology*, 307–308 (Springer, 2014).

[105] Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics* **47**, 106–114 (2015).

[106] Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology* **18**, 507–522 (2011).

[107] Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (CRC press, 1994).

Online Methods

Our cancer bioinformatics pipeline is versatile and can be easily customized for multiple purposes. Below, we review how its features may be selected according to the specific research goals, input data, and cancer type.

A general pipeline to infer ensemble-level progression models

For each of n tumors (n patients) we assume relevant (epi)genetic data to be available. We do not put constraints on data gathering and selection, leaving the user to decide the appropriate "resolution" of the input mutational data. For instance, one might decide whether somatic mutations should be classified by type, or aggregated. Or, one might decide to lift focal CNAs to the wider resolution of cytobands or full arms. These choices depend on data and on the overall understanding of such alterations and their functional effects for the cancer under study, and no single all-encompassing rationale may be provided.

Step 1: Reducing inter-tumor heterogeneity by cohort subtyping. We might wish to identify cancer subtypes in the *heterogeneous mixture* of input samples. In some cases the classification can benefit from clinical biomarkers, such as evidences of certain cell types [82], but in most cases we will have to rely on multiple *clustering* approaches at once, see, e.g., [59, 61].

Many common approaches cluster expression profiles [83], often relying on non-negative matrix factorization techniques [84] or earlier approaches such as k-means, Gaussians mixtures or hierarchical/spectral clustering - see the review in [85]. For glioblastoma and breast cancer, for instance, mRNA expression subtypes provides good correlation with clinical phenotypes [86–88]. However, this is not always the case as, e.g., in colorectal cancer such clusters mismatch with survival and chemotherapy response [86]. Clustering of *full exome* mutation profiles or smaller panels of genes might be an alternative as it was shown for ovarian, uterine and lung cancers [89, 90]

Step 2: selection of driver events. In subtypes detection, with more alterations available it becomes easier to find similarities across n samples, as features selection gains precision. In progression inference, instead, one wishes to focus on $m \ll n$ driver alterations, which ensure also an appropriate statistical ratio between sample size (n) and problem dimension (m).

Multiple tools filter out driver from passenger mutations. MutSigCV identifies drivers mutated more frequently than background mutation rate, [91]. OncodriveFM, avoids such estimation but looks for functional mutations [92]. OncodriveCLUST scans mutations clustering in small regions of the protein sequence [93]. MuSiC uses multiple types of clinical data to establish correlations among mutation sites, genes and pathways [94]. Some other tools search for driver CNAs that affect protein expression [95]. All these approaches use different statistics to estimate signs of positive selection, and we suggest using them in an orchestrated way, as done in some platforms [96]. Notice that driver genes will likely differ across subtypes, mimicking the different molecular properties of each group of samples.

Step 3: fitness equivalence of exclusive alterations. When working at the ensemble-level, identification of "groups of equivalent but alternative" mutually exclusivity alterations is crucial, prior to progression inference [33]. A plethora of tools can be used; greedy approaches [97, 98] or their optimizations, such as MEMO, which constrain search-space with network priors [60]. This strategy is further improved in MUTEX, which scans mutations and focal CNAs for genes with a common downstream effect in a curated signalling network, and selects only those genes that significantly contributes to the exclusivity pattern [23]. Other tools, instead, employ advanced statistics or generative approaches without priors [99–104].

In the fitness equivalent groups, we distinguish between *hard* and *soft* exclusivity, the former assuming strict exclusivity among events, with random errors accounting for possible overlaps, the latter admitting co-occurrences. [23]. CAPRI is the only algorithm where relations among group of genes can be input as *"testable hypotheses"* via *logical Boolean formulas.* In this case, we can use logical connectives such as $\oplus$ (the logical "xor") as a proxy for hard-exclusivity, and $\vee$ (the logical "disjunction") as a proxy for soft-exclusivity[3]. For example, these can be used to test wether colorectal tumors "start" prevalently from β-catenin deregulation, i.e., APC $\vee$ CTNNB1 , and if they further progress exclusively ($\oplus$) through KRAS or NRAS alterations. In general, as this testing-feature leaves the inference *unbiased* - see [33] - arbitrary hypotheses on significantly mutated subnetworks could be considered as well [105, 106].

Step 4: progression inference and confidence estimation. Finally, we use CAPRI to reconstruct cancer progression models of each identified molecular subtype, provided that there exist a reasonable list of driver events and the groups of fitness-equivalent exclusive alterations.

CAPRI's input is a binary $n \times (m + k)$ matrix $\mathbf{M}$ with n samples, m driver alteration events (Bernoulli 0/1 variables) and k testable formulas. CAPRI first *scans pairwise* $\mathbf{M}$ to identify a set of $\mathcal{S}$ plausible selective advantage relations, which then reduces to the most relevant ones, $\mathcal{S}^* \subset \mathcal{S}$.

Construction of $\mathcal{S}$ depends on the number of *non-parametric bootstrap* iterations and confidence p-values for estimating selective advantage among input events x and y. CAPRI postulates that "x selects for y" if it estimates that "x is earlier than y" and that "x's presence increases the probability of observing y" [45]. These conditions are implemented with these inequalities

$$p(x) > p(y) \qquad\qquad p(y \mid x) > p(y \mid \neg x) \tag{1}$$

for which we get p-values by Mann-Withney U Testing. Here, $p(\cdot)$ is an empirical marginal probability, $p(\cdot \mid \cdot)$ is a conditional, and $\neg x$ is the negation of x.

[3] Logical disjunction of a set of operands is true if and only if *one or more* of its operands is true. For this reason, if we shall use that as a model of soft-exclusivity, we shall also check that the majority of observations indeed shows an exclusivity trend, meaning that few cases of co-occurent observations happen.

Optimization of $\mathcal{S}$ is central to our tolerance to *false positives* and *negatives* in $\mathcal{S}^*$. CAPRI's implementation in TRONCO [44] selects from $\mathcal{S}$ a subset of relations by optimizing the *score with regularization*

$$\mathcal{S}^* = \arg\min_{\hat{\mathcal{S}} \subset \mathcal{S}} \left\{ -2\log[\mathcal{L}(\hat{\mathcal{S}} \mid \mathbf{M})] + \theta|\hat{\mathcal{S}}| \right\}\ , \tag{2}$$

where $\mathcal{L}(\cdot)$ is the *model likelihood*; the estimated optimal solution is $\mathcal{S}^*$.

Different values of θ lead to different tolerance to errors in $\mathcal{S}^*$, the *Akaike Information Criterion* (AIC) being for $\theta = 2$, the *Bayesian Information Criterion* (BIC) for $\theta = \log(n)$. Both scores are approximately correct; AIC is more prone to overfitting but likely to provide also good predictions from data and is better when false negatives are more misleading than positive ones. BIC is more prone to underfitting errors, thus is more parsimonious and better in opposite cases. As often done, we suggest to combine both approaches and distinguish which relations are selected by BIC/AIC.

Model confidence can be estimated with *non-parametric*, *parametric* or *statistical bootstrap* [107]. These procedures re-sample datasets to provide a confidence to every selective advantage relation and to the overall model. Bootstrapped datasets are randomly generated by re-shuffling data and seed (non-parametric), just seed (statistical) or by sampling from the model (parametric). CAPRI's other statistics include hypergeometric tests to assess how significant is the overlap between pairs of alterations.