BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 17 2014, pages 2456–2463
doi:10.1093/bioinformatics/btu319

Genetics and population analysis

Advance Access publication May 7, 2014

Inferring the paths of somatic evolution in cancer

Navodit Misra[1,*], Ewa Szczurek[1,2] and Martin Vingron[1]

[1]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, D-14195 Berlin, Germany and [2]Department of Biosystems Science and Engineering, ETH Zurich and Swiss Institute of Bioinformatics, CH-4058 Basel, Switzerland

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Cancer cell genomes acquire several genetic alterations during somatic evolution from a normal cell type. The relative order in which these mutations accumulate and contribute to cell fitness is affected by epistatic interactions. Inferring their evolutionary history is challenging because of the large number of mutations acquired by cancer cells as well as the presence of unknown epistatic interactions.

Results: We developed Bayesian Mutation Landscape (BML), a probabilistic approach for reconstructing ancestral genotypes from tumor samples for much larger sets of genes than previously feasible. BML infers the likely sequence of mutation accumulation for any set of genes that is recurrently mutated in tumor samples. When applied to tumor samples from colorectal, glioblastoma, lung and ovarian cancer patients, BML identifies the diverse evolutionary scenarios involved in tumor initiation and progression in greater detail, but broadly in agreement with prior results.

Availability and implementation: Source code and all datasets are freely available at bml.molgen.mpg.de

Contact: misra@molgen.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 4, 2013; revised on March 21, 2014; accepted on April 30, 2014

1 INTRODUCTION

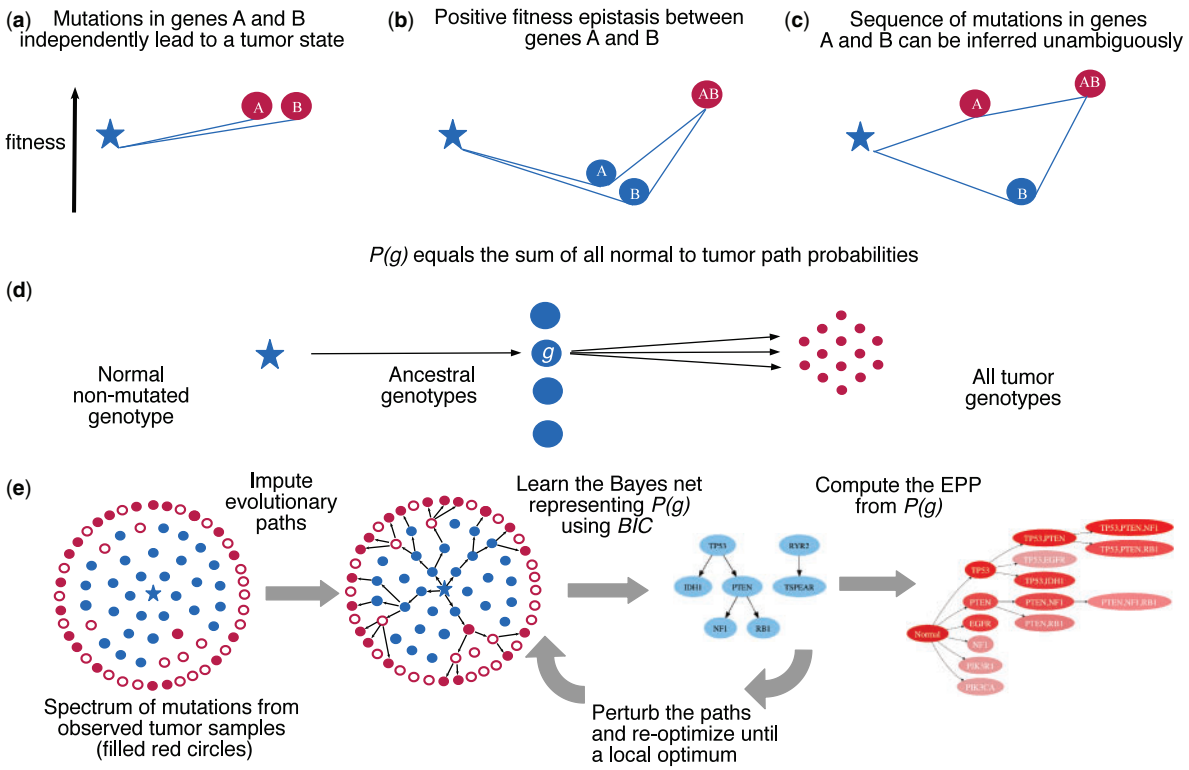
Tumor samples from cancer patients show a large variety of genetic abnormalities that have accumulated during somatic evolution from a normal cell state (Hanahan and Weinberg, 2011). Somatic mutations are continuously acquired in individual cells, but depending on, among other things, the fitness of the resultant genotype, only a small fraction may reach fixation within a cell population (Stratton *et al.*, 2009). Fitness change induced by a mutation can in turn depend on the genetic background, a phenomenon known as *epistasis* (Fisher, 1918). Epistasis has been known to play an important role in molecular evolution (Breen *et al.*, 2012; Kimura, 1985; Smith, 1970) and can constrain the sequence of mutation accumulation (Gong *et al.*, 2013; Weinreich *et al.*, 2005). The fitness function or landscape over the space of all genotypes depends both on the magnitude and sign of the epistatic interactions. Therefore, genotypes observed in tumor samples are likely the result of a diverse set of mutational paths evolving across a complex fitness landscape. Patterns of somatic mutations observed in tumor samples contain information, both about the evolutionary paths of cancer progression and the epistatic gene interactions that influence them. However, extracting this evolutionary information is challenging because the fitness landscapes are unknown, and analyzing large datasets with hundreds of recurrently mutated genes is computationally demanding. Owing to these difficulties, existing computational methods for cancer progression either constrain the set of possible evolutionary scenarios (Bozic *et al.*, 2010; Desper *et al.*, 1999) or are feasible for relatively small sets of genes (Attolini *et al.*, 2010; Gerstung *et al.*, 2009; Hjelm *et al.*, 2006).

2 APPROACH

Here, we report evolutionary progression paths (EPPs) for tumor samples from colorectal, glioblastoma, lung and ovarian cancer patients. The EPPs are estimated using a computational technique for reconstructing ancestral genotypes from observed tumor genotypes, called Bayesian Mutation Landscape (BML) (Fig. 1). The main novelty of BML is that it takes into account unobserved ancestral genotypes and unknown epistatic gene interactions, before inferring a probabilistic model for the accumulation of somatic mutations in a population of cancer cells. These unobserved precancer states present a systematic bias to all methods that attempt to compute EPPs directly from tumor samples. The nature and magnitude of epistatic interactions also influence EPPs and can distinguish between evolutionary scenarios with a clear sequence of genetic events from those with multiple parallel EPPs (Fig. 1a–c). Furthermore, unlike existing computational methods (Attolini *et al.*, 2010; Gerstung *et al.*, 2009; Hjelm *et al.*, 2006), BML incorporates several algorithmic improvements that allow, for the first time, to compute EPPs for some of the largest publicly available cancer datasets in their entirety.

BML is based on a probabilistic model where every evolutionary path (with irreversible mutations) from the normal genotype to any tumor genotype has a non-zero probability. BML first estimates the probability $P(g)$ that a particular combination of mutations (denoted by genotype g) reaches fixation in a cell population that has evolved from a normal cell genotype and will eventually attain a tumor cell genotype (Fig. 1d). We will refer to it as the *evolutionary probability* of genotype g. $P(g)$ equals the sum of path probabilities for every mutation path from the normal genotype that passes through g and ends as a tumor genotype. To get a better intuition as to what P represents, consider the following hypothetical scenario: assume we had a

*To whom correspondence should be addressed.



(f)

Data set	Samples	Recurrent genes	Cutoff	Total edges	Unpruned edges	Run time 100 restarts
Gliblastoma multiforme	194	22	4	231	9	2.1s
Lung adenocarcinoma	161	51	4	1275	52	29.7s
Ovarian cancer	326	192	5	18336	136	10m53.3s

Fig. 1. An overview of the BML model. Each circle represents a genotype that specifies the presence or absence of non-silent somatic mutation(s) in either copy of each gene. Blue star represents the normal non-mutated genotype, blue circles represent genotypes that may contain somatic mutations, but have not yet attained the phenotype of uncontrolled growth, and red circles represent cancer genotypes. Note that there may be additional unobserved tumor genotypes (empty red circles). **a**, **b** and **c** show the fitness profile and the distribution of observed samples (red circles) for three evolutionary scenarios. In a, mutations in genes *A* and *B* independently give rise to a genotype that is a fitter relative to the normal genotype. In contrast, b shows a scenario where there is positive fitness epistasis between genes *A* and *B*. However, there is no clear sequence of mutational events in *A* and *B*. c depicts a scenario where there is positive fitness epistasis between *A* and *B*, but the path through *B* encounters a fitness valley. This scenario is known as *sign epistasis* because the sign of fitness difference between *B* mutated and non-mutated states depends on the mutation state of *A*. In this case, an unambiguous sequence of mutations in *A* and *B* can be inferred (Supplementary Note S1). **d**, *Evolutionary probability* $P(g)$ for any ancestral genotype g contains information from all paths composed of sequential, irreversible mutations, from normal genotype to any tumor genotype that pass through g. Each possible path has a non-zero probability, and $P(g)$ equals the sum of all such path probabilities. **e**, Schematic for the algorithm. We use both observed tumor samples (filled red circles) and imputed evolutionary paths to infer the probabilities $P(g)$ of genotype g. $P(g)$ is represented by a Bayesian network, that is optimized for the best choice of imputed paths. Once a Bayesian network is selected, a recursive algorithm is used to infer the likely EPP. **f**, Efficiency of the pruning scheme and the BML algorithm for glioblastoma, lung and ovarian cancer datasets. Cutoff shows the minimum number of samples in which each retained gene was mutated to be considered *recurrent* (Section 3). Note that the search space is exponential in the number of unpruned edges in the worst case. Run time results are >100 random restarts of the algorithm on 2.4 GHz Intel Core i5 processor and 4 GB memory

database of tumor samples from large number of N cancer patients. In addition, assume we had perfect knowledge of the evolutionary paths followed by each tumor sample as it evolved from a normal cell state. If $n(g)$ was the number of samples in our database that had g as an ancestral or current cell state, then $P(g) \approx n(g)/N$. Because all tumor genotypes are assumed to have evolved from an initial normal state, $P(g_0) = 1$ for the non-mutated normal genotype g_0. Crucially, $P(g)$ is not simply the fitness of genotype g but depends on the fitness landscape over all ancestral cell states traversed during somatic evolution, as well as the details of evolutionary dynamics of cell populations. BML assumes that the mutation accumulation process is irreversible and sequential, proceeding one mutation at a time. Note that this assumption may not be valid for large-scale karyotypic and copy number changes that are frequently observed in tumor samples. We therefore restrict this approach to point mutations and small indels. We also ignore the effect of mutations already present in the germ line.

BML estimates the evolutionary probabilities using a graphical model known as a Bayesian network. Bayesian networks describe a large class of probability distributions that can be represented as directed acyclic graphs (DAGs). They have previously been applied to gene expression analysis (Friedman, 2000), as well as copy number variations in cancer (Bulashevska *et al*., 2004). Figure 1e provides a schematic for the algorithm. Inferring the distribution P over all genotypes is complicated because of a systematic bias, as the highest probability precancer genotypes (Fig. 1e, blue circles) are not present in the input, which consists of samples from diagnosed cancer patients (red circles). BML estimates P for these ancestral genotypes by imputing likely evolutionary paths (in the form of a bifurcating tree). The collection of paths connecting a set of vertices (observed tumor genotypes) to a common vertex (the normal genotype) can always be represented by a tree. The internal nodes of the tree represent ancestral genotypes and are treated as unobserved samples. These ancestral genotypes, along with the observed samples are then used to estimate a Bayesian network. Because we do not know the true paths followed by observed samples, we perform an additional optimization step, where we perturb the paths using a class of tree rearrangements known as nearest neighbor interchange (NNI) (Felsenstein, 2004) and repeat the process until the algorithm encounters a local optimum in tree space. The Bayesian network estimates P up to an overall normalizing factor, that is later set by requiring that the evolutionary probability for the non-mutated normal genotype is one. The inferred Bayesian network representation of $P(g)$ is then used to reconstruct the most likely EPP using a recursive algorithm (see Section 3).

Another advantage of using Bayesian networks is their ability of separating direct from indirect epistatic interactions, with network edges denoting direct epistatic interactions. Because negative epistatic interactions are difficult to separate from the scenario in Figure 1a (Supplementary Note S1), we restricted BML to model co-occurrence of mutations that provide a reliable signature of positive epistasis (Fig. 1b and c and Supplementary Note S1). Together with the algorithmic improvements introduced to BML modeling, based on pruning large regions of the search space (see Section 3 and Supplementary Note S2), this approximation of P allows for extremely efficient computations (Fig. 1f). As a result, BML can be used to perform comprehensive bootstrap analysis for tumor datasets with many more recurrently mutated genes than previously feasible.

3 METHODS

In this section, we discuss algorithms for learning the structure and parameters of the Bayesian network and for reconstructing the EPPs from observed tumor samples.

3.1 Datasets

We performed BML analysis for colorectal (Bamford *et al*., 2004; Sjöblom *et al*., 2006), glioblastoma (Parsons *et al*., 2008; TCGA consortium, 2008), lung (Ding *et al*., 2008) and ovarian cancer samples (TCGA consortium, 2011). The colorectal cancer dataset was obtained from the supplement to the paper by Attolini *et al*. (2010). Glioblastoma, lung and ovarian cancer datasets were all downloaded from publicly available databases maintained by The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) (Supplementary Table S1). Each dataset was preprocessed to retain non-silent mutations, identify recurrently mutated genes and coarse grain the data such that each gene can take two states, mutated and non-mutated. For glioblastoma, we combined the data from two sequencing studies (Parsons *et al*., 2008; TCGA consortium, 2008). We also removed one tumor sample in glioblastoma that was identified as hyper mutated in the original sequencing study (Parsons *et al*., 2008). We filtered genes that were mutated too infrequently by imposing a cutoff on the number of samples with mutations in a gene. For each dataset, this cutoff was chosen as the smallest number (greater than three) such that the number of retained genes was less than the number of available samples. The final input to our method is a matrix of genes versus tumor samples with 0/1 entries indicating the absence/presence of a non-silent somatic mutation in a gene for each tumor sample (Fig. 1f). We should point out that the set of genes used as an input to our method could also be restricted according to appropriate criterion (e.g. mRNA expression). This can be achieved via algorithms for restricting input gene sets such as the *MutSigCV* algorithm (Lawrence *et al*., 2013) or the somatic functional events in Ciriello *et al*. (2013). In the absence of such functional information, the events in the inferred paths must not be assigned functional importance and must be interpreted as the set of events that frequently occur during the process of cancer progression.

3.2 The BML model

BML models the evolutionary probabilities over the genotype space as a probability distribution that is represented by a Bayesian network, up to an overall normalizing factor. The normalizing factor is then obtained by imposing the constraint that $P(g_0) = 1$ for the normal genotype g_0 and appropriately scaling the Bayesian network probabilities. The network is defined on a set $\mathcal{C}$ of binary random variables that represent mutations of the genes, whereas edges represent direct epistatic interactions. Formally, a Bayesian network $B(G, \Theta)$ is specified by G, a DAG whose vertices are the genes in $\mathcal{C}$, and a set of parameters $\Theta = \{\theta_C | C \in \mathcal{C}\}$, representing conditional probabilities $\theta_C(c|\pi) \equiv Pr(C = c|\Pi_C = \pi)$ for each gene C given the state of its parents Π_C in G (Koller and Friedman, 2009). Let D denote a $m \times n$ data matrix with binary entries. The columns of D represent the set of genes $\mathcal{C} = \{C_1, \ldots C_m\}$, and rows represent the set of samples $\mathcal{S} = \{S_1, \ldots S_n\}$, such that $D_{ij} = 1$ if gene C_i is mutated in sample S_j (with respect to a reference state designated as normal) and 0 otherwise. The data matrix can be used to compute sufficient statistics for learning the network structure, in the form of counts $n_{c,\pi}$ for the number of samples where gene $C = c$ when its parents $\Pi = \pi$. We use the Bayesian information criterion (BIC) for selecting candidate structures:

$$\log Pr(D|B) \equiv \sum_{C \in \mathcal{C}} Fam(C; \Pi_C) \tag{1}$$

where $Fam(C; \Pi_C)$ is the BIC score for a family $\{C, \Pi_C\}$ consisting of each gene and its parents and is given by

$$Fam(C; \Pi_C) = \max_{\theta_C} \sum_{\pi} \left\{ \sum_{c} n_{c,\pi} \log[\theta_C(c|\pi)] - \frac{\log n}{2} \right\} \tag{2}$$

The BIC score is known to be statistically consistent in the sense that given sufficiently many samples from an underlying Bayesian network, we can learn the true structure by maximizing the BIC score.

3.2.1 Learning the evolutionary probability distribution To correctly learn the evolutionary probability P, we need to consider a dataset containing both the given cancer genotypes and the unobserved precancer ones. If the samples are lacking data from certain regions of the state space, the inferred network parameters will be biased accordingly. To account for this problem, we construct bifurcating trees, with normal

genotype at the root (a node of degree one), tumor samples at the leaves and all other internal nodes as degree three. If O is the input dataset of observed tumor samples and T denotes the degree three internal nodes and the root of the bifurcating tree, then the complete data $D = O \cup T$, and this is used for estimating the Bayesian network in Equation (1).

We also make the following simplifications in selecting the trees and model parameters for reasons of computational efficiency. First, we assume that the accumulation of mutations in a sample is irreversible, with 0–1 transitions from the root to the leaf. The root node in our problem is the normal state with all genes in state 0. We choose the state of any gene at any internal vertex as 1 only if all its descendant leaves are in state 1.

The second restriction we make is motivated by our model choice. Because the probabilities that we infer represent the chance of a combination of mutations reaching fixation in a cell population, as it evolves from a normal state, the probabilities for the mutated states must be smaller than those of the normal state. We use a simple and computationally efficient heuristic criterion to incorporate this feature of BML by requiring that the number of samples in D with a mutation in any gene should not be more than half the total number of samples. Note that this condition can always be satisfied by choosing an appropriate labeling of the internal nodes. This holds because n, the total number of nodes, equals the sum of s observed samples, one normal and $s+1-2$ degree three internal nodes, which yields $n = 2s$.

Third, we restrict the parameters of our model such that given any genotype, the probability of accumulating a mutation in a gene does not decrease on acquiring a mutation in another gene (see Supplementary Methods for details). Formally, we require the conditional probabilities for a gene $C \in \mathcal{C}$ to be mutated, given the state of its parents $\Pi_C \subset \mathcal{C}$, to obey the following constraints.

$$Pr(C=1|\Pi_C=\pi_A) \geq Pr(C=1|\Pi_C=\pi_B) \forall \pi_B \subset \pi_A \quad (3)$$

where $\pi_B \subset \pi_A$ means that all genes in Π_C that are mutated in state π_B are also mutated in π_A. Prior attempts at modeling the dynamics of cancer progression have included similar parameter constraints (Gerstung *et al*., 2009; Hjelm *et al*., 2006). One justification for this constraint is that co-occurrence of mutations in any pair of genes is unlikely by chance and serves as a reliable test for positive epistasis, whereas mutations that show a tendency to be mutually exclusive are not necessarily because of epistatic interactions (Supplementary Note S1) and represent a weaker signal to be resolved at the small sample sizes of available datasets. Note that this constraint does not imply a monotonically increasing fitness landscape.

The structure learning problem with these simplifications is to estimate the tree T_* and Bayes net B_* that maximize log $Pr(D|B)$. With $D = T \cup O$, we can formally write our objective as

$$(T_*, B_*) = \arg\max_{T,B} \log Pr(D|B) \quad (4)$$

3.2.2 Bayesian network structure and parameter learning algorithm In this section, we describe a heuristic for efficiently learning the distribution P. Given the leaves and internal nodes of the tree, we search for the optimal DAG using the method of ordering-based search (OBS) (Teyssier and Koller, 2005). OBS initializes an ordering on the variables and constrains each variable to choose parents exclusively from the set of its predecessors in the ordering. The algorithm then searches the space of all orderings by flipping the order of any pair of variables adjacent in the ordering. Note that this ordering is not the same as the ordering of mutations in genes during somatic evolution. The search over tree space was performed by a class of local moves known as NNI (Felsenstein, 2004). We use two asymptotic pruning results that allow us to greatly restrict the search space (Supplementary Note S2).

Algorithm 1: BML Structure learning.

1. Perform a global pairwise pruning for each pair of genes (Supplementary Note S2).
2. Randomly initialize a bifurcating tree with observed samples as leaves and normal state as root and assign internal node labels.
3. Perform pairwise local pruning (Supplementary Note S2).
4. Find the DAG that maximizes BIC score and obeys Equation (3) using OBS.
5. Perturb the tree using NNI and repeat the search steps 3 and 4 until local optimum.

After performing structure learning using BIC, we used an empirical Dirichlet prior for learning the parameters of the Bayes Net. For each gene C, the parameters were chosen as $\theta(C=c|\Pi_C=\pi) = (n_{c\pi} + \alpha_c)/(n_{0\pi} + n_{1\pi} + 1)$, where the hyper parameter α_c denotes the fraction of samples in D that have $C = c$, and $n_{c\pi}$ is the number of samples where $C = c$ and $\Pi_C = \pi$.

3.3 Reconstructing the most likely EPP

The analysis performed in Supplementary Note S1 suggests that the most probable ancestor for a given genotype is the one with highest evolutionary probability. We use this observation to reconstruct the most likely EPPs, presented in Figure 2, using a recursive algorithm. Briefly, the algorithm starts with a set of most likely states with three mutations and retraces their mutational history by connecting each genotype to their most likely ancestral state (i.e. the ancestral state with largest P). The algorithm takes as input the inferred P and parameters $k > 1$ and $c < 1$. Paths are initialized starting from all genotypes representing combinations of k mutations, which are present in the observed data and have a probability larger than a cutoff $c * m_k$, where m_k equals the largest probability of a genotype with k mutations. The user can vary the level of detail in the reconstructed paths by varying c, and the size of the paths with k. At each subsequent step $i < k$ of the algorithm, for each genotype, the algorithm identifies the most likely ancestral state with $i-1$ mutations, by choosing the one with the highest P. The algorithm then adds a set of nodes with $i-1$ mutations that are either identical to the genotype of at least one observed sample and have a probability larger than $c * m_k$, or were identified as the likely ancestral state for a node retained at the previous step $i + 1$. This process is repeated all the way up to the node representing the normal genotype. Figure 2 shows the likely paths with $c = 0.3$ and $k = 3$.

4 RESULTS

BML analysis for each dataset was accompanied with 1000 parametric bootstrap replicates to assess the robustness of the inferred Bayesian network. Figure 2 shows the highest probability genotypes and the most likely paths of progression for each dataset. Note that the trees shown in Figure 2 are not the same as the full bifurcating tree used by the algorithm, but only the high probability genotypes traversed by tumor samples (Fig. 1e, see Section 3).

4.1 Most likely paths of progression

The temporal order of mutations has perhaps been best studied in colorectal cancer (Fearon and Vogelstein, 1990). The temporal order of *APC*, *KRAS* and *TP53* mutations was also investigated computationally in Attolini *et al.* (2010). Therefore, we first present the results of BML analysis for the colorectal cancer dataset analyzed by Attolini *et al.* (2010) for comparison. Their results support the hypothesis that *APC* mutations are more likely to

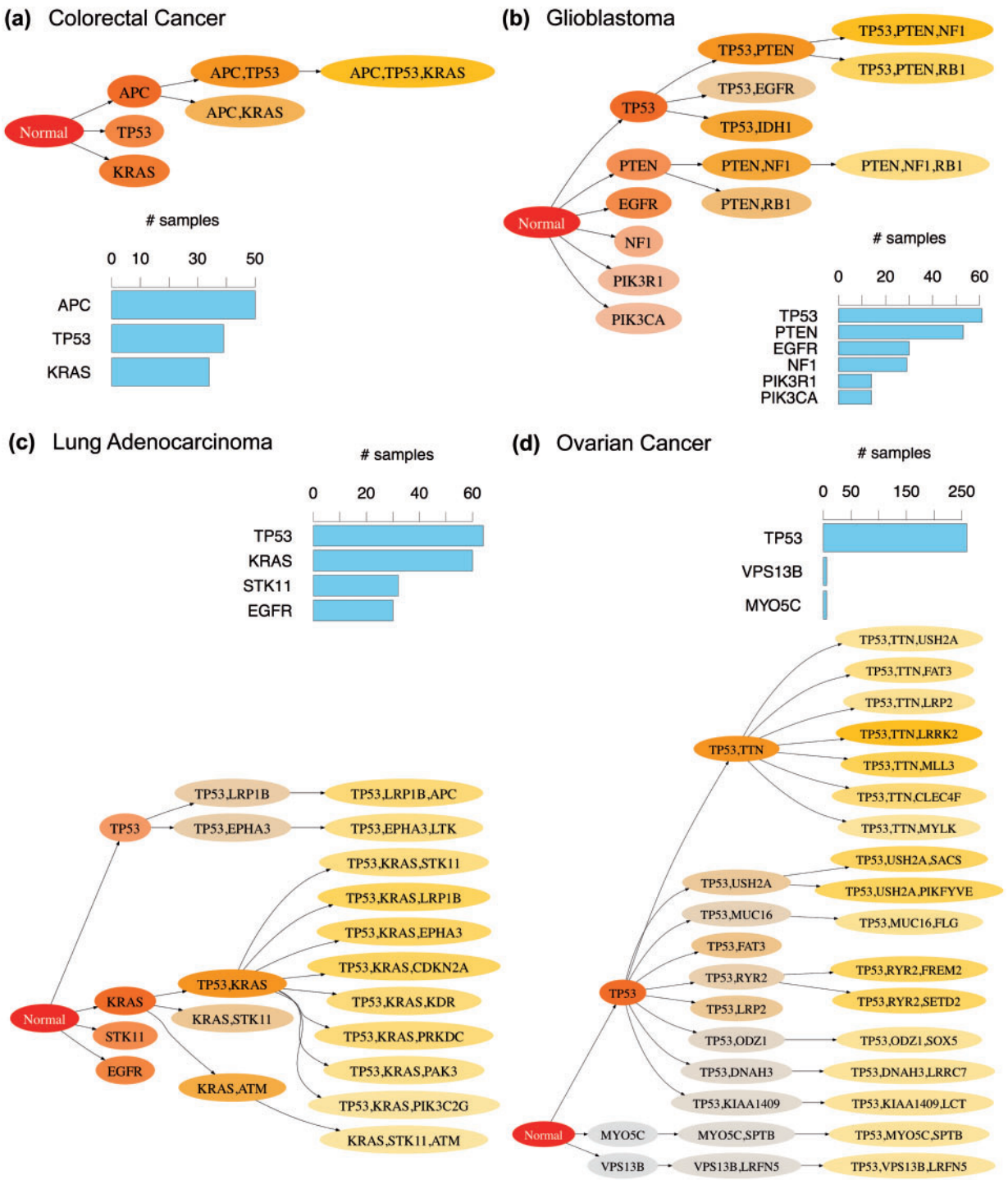


Fig. 2. Most likely paths followed by (a) colorectal, (b) glioblastoma, (c) lung adenocarcinoma and (d) ovarian cancer tumor samples. In general, there are several other low-probability events that may occur, but have been left out for clarity. Color for a genotype g with k mutations is scaled according to its relative probability $P(g)/m_k$ (decreasing from darker shade to light), where m_k is the maximum probability for a node with k mutations (Section 3). The bar plots (blue) show the number of samples with mutations in a given gene for the initial events shown in the paths

initiate tumorigenesis than *KRAS* mutations, which in turn are more likely than *TP53* mutations. BML estimates for the evolutionary probabilities of initial mutations (Fig. 2a) agree with these conclusions of Attolini *et al.* (2010). However, BML detects a robust positive epistatic interaction (bootstrap confidence >98%) between *APC* and *TP53*. As a consequence, conditional on *APC* being the initial mutation, a *TP53* mutation is more likely than a *KRAS* mutation. Therefore, in tumor samples

that contain mutations in each of the three genes, the most likely sequence is an *APC* mutation followed by a *TP53* mutation, which is then followed by *KRAS*.

We next performed BML analysis for a set of 22 recurrently mutated genes in 194 glioblastoma samples. A model proposed by Ohgaki *et al*. (2004) and Ohgaki (2007) indicated that a *TP53* mutation is the initiating event in secondary glioblastomas, followed most commonly by *EGFR* and *PTEN* mutations. In the case of primary glioblastomas, *TP53*, *EGFR* and *PTEN* mutations are present in roughly equal frequencies and provide alternative paths of tumor initiation. BML recapitulates these findings and also identifies alternative lower probability paths that are initiated by *NF1*, *PIK3R1* and *PIK3CA* mutations (Fig. 2b). The BML prediction that an initiating mutation in *NF1* is less likely than *TP53* also agrees with the computational analysis of Attolini *et al*. (2010). For lung cancer, BML analysis of 51 recurrently mutated genes in 161 adenocarcinoma samples inferred *KRAS*, *TP53*, *EGFR* and *STK11* mutations as likely early events in alternative paths during cancer progression (Fig. 2c). *TP53* and *KRAS* mutations tend to co-occur during the later steps of mutation accumulation. In contrast, *EGFR* and *KRAS* mutations are mutually exclusive and, as reported by (Ding *et al*., 2008), correlated with the smoking status of the patient, with EGFR mutations more common in non-smokers.

Applied to 192 recurrently mutated genes in 326 ovarian cancer samples, BML inferred *TP53* as the most likely initiator of tumor cells. *TTN* is the second most common mutation followed by several other recurrently mutated genes. BML predicts that *TTN* mutation is unlikely before *TP53*, but the *TP53-TTN* genotype is the most likely among states with two mutations in the observed tumor samples (Fig. 2d).

4.2 Fitness epistasis and sequence of genetic events

Fitness epistasis refers to a departure from additivity in the effect of mutation combinations with respect to their contribution to log fitness (Fisher, 1918). Epistatic interactions contribute both to the distribution of observed tumor samples in the genotype space as well as the evolutionary probability (Fig. 1). Even though BML is not constrained to any specific model of evolutionary dynamics, it is instructive to estimate and interpret the evolutionary probabilities for a population genetics model used in prior studies (Attolini *et al*., 2010; Komarova *et al*., 2003; Michor *et al*., 2004). The model is a stochastic process that describes the evolutionary dynamics of a population of cells as they randomly accumulate mutations (with gene-dependent mutation rates) during cell division, and compete for resources based on the fitness of the genotype (Supplementary Note S1). This model can be used to establish the following connection between epistatic interactions and the evolutionary probability for the scenarios depicted in Figure 1a–c (Supplementary Note S1):

(i) *Positive fitness epistasis* (in Fig. 1b and c) leads to a tendency for mutations in A and B to co-occur and implies that the double-mutant genotype satisfies $P(AB) \geq P(A)P(B)$.
(ii) *Sign epistasis* (Fig. 1c) implies $P(A) >> P(B)$, and an unambiguous ordering of mutations leading to the double mutant genotype can be inferred. Furthermore, if mutations in A occur at a sufficiently high frequency and/or the epistatic interaction is particularly strong, then, $P(A) \geq P(AB) \geq P(B)$.

These observations suggest that epistatic interactions can lead to scenarios where we can unambiguously infer the sequence of genetic events (Fig. 1c and Supplementary Note S1). We use BML estimates for P to detect such evolutionary scenarios by identifying pairs of genes where the double-mutant genotype has an evolutionary probability in between the two single-mutant genotypes. Figure 3a shows one such instance for each dataset analyzed.

The case of *TTN*, the gene that codes for the largest human protein and is frequently mutated across multiple cancer types (Balakrishnan *et al*., 2007; Greenman *et al*., 2007), highlights the utility of this criterion. In particular, tumor samples from ovarian cancer patients show a tendency of *TP53* and *TTN* mutations to co-occur, suggesting a possible epistatic interaction. However, BML analysis predicts that *TTN* mutations rarely precede *TP53* mutations (Fig. 3a) and are unlikely to initiate tumor formation. This conclusion is in agreement with the original TCGA publications that did not identify *TTN* mutations as significant in initiating tumor formation (TCGA consortium, 2011). Although frequent mutations in *TTN* may likely be due to its huge length and not entirely due to functional reasons, there have been prior experimental studies that have suggested a possible role for *TTN* during cell division (Machado *et al*., 1998; Machado and Andrew, 2000; Qi *et al*., 2008). A definite answer regarding the role of *TTN* mutations and their contribution to tumor cell fitness would require further experimental investigation.

4.3 Simulations validate the improvement in accuracy and robustness with BML

We performed a simulation-based parametric bootstrap (Friedman *et al*., 1999) for validating our method as well as demonstrating the effect of unobserved genotypes on the reconstruction algorithm. Parametric bootstrap involves learning a model from the given data and simulating the learnt model to generate new datasets for learning. This way we have access to ground truth for the simulated datasets and can estimate both the accuracy and robustness of the learning algorithm. We performed parametric bootstrap by simulating samples from the DAG learnt by BML on each of the datasets. We only retained those simulated samples that had at least one mutation and where all the mutations were present in at least one observed tumor sample. Because we wanted to assess the uncertainty in estimated evolutionary probabilities, our goal was to simulate a dataset where the region with unobserved precancer states closely mimics real data. This is important because the retained genotypes specify how many and which combination of mutations are needed before a cell population becomes capable of uncontrolled growth. The number of retained simulated samples was set equal to the number of observed tumor samples in each case.

Figure 3b–d shows the results from bootstrapping of 1000 simulated datasets for our approach with and without the tree estimation and parameter constraints. The former scenario reflects the DAG inferred by the full BML model, which takes into

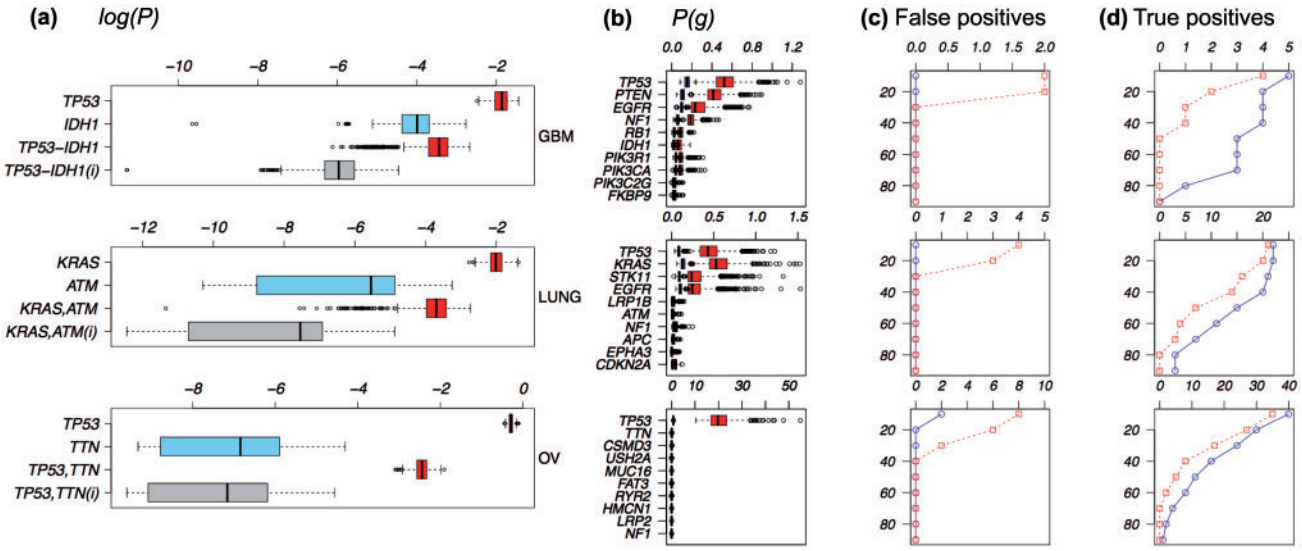


Fig. 3. (**a**) Sequence of genetic events. Box plots for 1000 bootstrap replicates showing examples of strong departures from additivity in $log(P)$ for each dataset. The third row (red) shows the computed $log(P)$ and the fourth row (gray) shows the sum of the two single mutated states. Mutations in these genes show a tendency to co-occur, suggesting that the departures from additivity are due to positive epistasis. Furthermore, a clear sequence of genetic events can be inferred (along the red box plots), suggesting the presence of a fitness valley along one (blue single mutant) of two possible paths to reach the double mutant genotype. The case of *TP53-TTN* genotype in ovarian cancer shows how *TTN* mutations, despite their recurrence, rarely initiate tumor progression. (**b–d**). Simulation-based parametric bootstrap for each of the datasets. Blue (box plots in b and curves in c and d) show the results for the BML model that uses inferred ancestral information, whereas the results for standard Bayesian network learning algorithm are in red. The box plots in b show $P(g)$ for single mutated states for 10 frequently mutated genes. These genes were selected based on an ordering that the algorithm automatically assigns to the genes (Section 3). Vertical axes in c and d represent the percentages of bootstrap confidence, whereas the horizontal axes represent the number of edges in the inferred networks that were false and true positives, respectively

account and aims to correct for the bias due to unobserved precancer genotypes. This is compared with standard Bayesian network learning where the unobserved states (represented by the internal nodes of the tree) are not included. Figure 3b shows the probability of mutations for 10 frequently mutated genes for the simulated Bayesian network in each case. These genes were selected based on an ordering that the algorithm automatically assigns to the genes (Section 3). The mutation probabilities for the BML model are consistently lower, as expected, because the standard algorithm does not take into account the probability mass from ancestral states that have fewer mutations than in the observed samples. As can be seen in Figure 3c and d, including the tree and parameter constraints leads to both fewer false positives and fewer false negatives in inferred edges, at a fixed confidence level.

By default, we constrain the parameters of our model to account for patterns of mutation co-occurrence (Section 3). To test the assumption in a model-based manner, we also implemented an alternative version of our method without any parameter constraints for glioblastoma and lung cancer. For glioblastoma, our method did not infer any additional edges. For lung cancer, this method inferred an edge between EGFR and KRAS, but with a low bootstrap confidence of 45%.

5 DISCUSSION

Modeling the evolutionary events leading to cancer and characterizing the fitness landscape of cancer cells promises innovative applications in clinical cancer research (Merlo *et al*., 2004). BML allows the reconstruction of likely ancestral genotypes and the paths of mutation accumulation in greater detail than existing methods. BML accomplishes these tasks owing to several algorithmic improvements that take into account the unobserved precancer genotypes that provide a systematic bias to EPP reconstruction, as well as the effects of unknown epistatic interactions.

We should emphasize that the goal of BML is not to classify somatic mutations as drivers or passengers; rather BML recapitulates the likely sequence of somatic mutation accumulation in recurrently mutated genes. It should also be noted that the bifurcating tree used by BML is simply an efficient data structure to represent paths and does not necessarily imply a hierarchical ordering of mutations. This distinction is important because somatic evolution occurred independently in each cancer patient and different tumor samples do not have a shared evolutionary history. Even though the genotypes at the internal nodes of the bifurcating tree allow us to correct for the systematic bias due to unobserved precancer genotypes, the estimated evolutionary probabilities are still only an approximation of the true distribution. However, since the estimated P are computed after taking into account the inferred precancer genotypes, they also incorporate the evolutionary aspect of the true evolutionary probabilities.

There are some obvious limitations of BML analysis because it does not include copy number and genomic rearrangements that likely provide alternative paths for tumor initiation and progression. Another source of complexity is the existence of genetic heterogeneity within individual tumor samples (Nik-Zainal

et al., 2012), as well as the role of tumor microenvironment during cancer progression (Bissell and Hines, 2011). BML ignores the possible cooperative interactions between subclonal cell populations within a tumor and between tumor and surrounding stromal cells. These are all important avenues that are left for further exploration.

Aside from these limitations, an extensive bootstrap analysis demonstrates that BML estimates of P are accurate and robust (Fig. 3). Simulations (Fig. 3c and d) also show that BML identifies epistatic interactions with greater accuracy than a naive network reconstruction algorithm. At the same time, BML is scalable for application to some of the largest available cancer datasets (Fig. 1f). Therefore, BML is an efficient and powerful tool that brings us a step closer to understanding the evolution of the cancer genome.

ACKNOWLEDGMENT

The authors thank Dmitri Petrov for discussion of fitness landscapes.

Funding: German Cancer Aid (109679).

Conflicts of Interest: none declared.

REFERENCES

Attolini,C.S. *et al.* (2010) A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc. Natl Acad. Sci. USA*, **107**, 17604–17609.

Balakrishnan,A. *et al.* (2007) Novel Somatic and germline mutations in cancer candidate genes in glioblastoma, melanoma, and pancreatic carcinoma. *Cancer Res.*, **67**, 3545.

Bamford,S. *et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer*, **91**, 355358.

Bissell,M.J. and Hines,W.J. (2011) Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression. *Nat. Med.*, **17**, 320–329.

Bozic,I. *et al.* (2010) Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl Acad. Sci. USA*, **107**, 18545–18550.

Breen,M.S. *et al.* (2012) Epistasis as the primary factor in molecular evolution. *Nature*, **490**, 535538.

Bulashevska,S. *et al.* (2004) Pathways of urothelial cancer progression suggested by Bayesian network analysis of allelotyping data. *Int. J. Cancer*, **110**, 850–856.

Ciriello,G. *et al.* (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, **45**, 1127–1133.

Desper,R. *et al.* (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comp. Biol.*, **6**, 37–51.

Ding,L. *et al.* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069–1075.

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

Felsenstein,J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Fisher,R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.*, **52**, 399–433.

Friedman,N. *et al.* (1999) Data analysis with bayesian networks: a bootstrap approach. In: *Proceedings of the fifteenth conference on Uncertainty in Artificial Intelligence (UAI)*. pp. 196–205.

Friedman,N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comp. Biol.*, **7**, 601–620.

Gerstung,M. *et al.* (2009) Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, **25**, 2809–2815.

Gong,L.I. *et al.* (2013) Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife*, **2**, e00631.

Greenman,C. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153–158.

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

Hjelm,M. *et al.* (2006) New probabilistic network models and algorithms for oncogenesis. *J. Comp. Biol.*, **13**, 853–865.

Kimura,M. (1985) The role of compensatory neutral mutations in molecular evolution. *J. Genet.*, **64**, 7–19.

Koller,D. and Friedman,N. (2009) *Probabilistic Graphical Models: Principles and Techniques*. MIT press, Cambridge, MA.

Komarova,N.L. *et al.* (2003) Mutation-selection networks of cancer initiation: tumor suppressor genes and chromosomal instability. *J. Theor. Biol.*, **223**, 433–450.

Lawrence,M.S. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214218.

Machado,C. *et al.* (1998) Human autoantibodies reveal titin as a chromosomal protein. *J. Cell Biol.*, **141**, 321–333.

Machado,C. and Andrew,D.J. (2000) D-Titin: a giant protein with dual roles in chromosomes and muscles. *J. Cell Biol.*, **151**, 639–652.

Merlo,L.M.F. *et al.* (2004) Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer*, **6**, 924–935.

Michor,F. *et al.* (2004) Dynamics of cancer progression. *Nat. Rev. Cancer*, **4**, 197–205.

Nik-Zainal,S. *et al.* (2012) The life history of 21 breast cancers. *Cell*, **49**, 994–1007.

Ohgaki,H. *et al.* (2004) Genetic pathways to glioblastoma: a population-based study. *Cancer Res.*, **64**, 6892–6899.

Ohgaki,H. and Kleihues,P. (2007) Genetic pathways to primary and secondary glioblastoma. *Am. J. Pathol.*, **170**, 1445–1453.

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

Qi,J. *et al.* (2008) Nuclear localization of the titin Z1Z2Zr domain and role in regulating cell proliferation. *Am. J. Cell Physiol.*, **295**, 975–985.

Sjöblom,T. *et al.* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268274.

Smith,J.M. (1970) Natural selection and the concept of a protein space. *Nature*, **225**, 563–564.

Stratton,M.R. *et al.* (2009) The cancer genome. *Nature*, **458**, 719–724.

Teyssier,M. and Koller,D. (2005) Ordering-based search: A simple and effective algorithm for learning Bayesian networks. In: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*. pp. 548–549.

The Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

The Cancer Genome Atlas Research Network. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 330–615.

Weinreich,D.M. *et al.* (2005) Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution*, **59**, 1165–1174.