



ARTICLE

DOI: 10.1038/s41467-018-06867-x

OPEN

# Mutational interactions define novel cancer subgroups

Jack Kuipers<sup>1,2</sup>, Thomas Thurnherr<sup>1</sup>, Giusi Moffa <sup>3,4</sup>, Polina Suter<sup>1,2</sup>, Jonas Behr<sup>1</sup>, Ryan Goosen<sup>5</sup>, Gerhard Christofori<sup>5</sup> & Niko Beerenwinkel <sup>1,2</sup>

Large-scale genomic data highlight the complexity and diversity of the molecular changes that drive cancer progression. Statistical analysis of cancer data from different tissues can guide drug repositioning as well as the design of targeted treatments. Here, we develop an improved Bayesian network model for tumour mutational profiles and apply it to 8198 patient samples across 22 cancer types from TCGA. For each cancer type, we identify the interactions between mutated genes, capturing signatures beyond mere mutational frequencies. When comparing mutation networks, we find genes which interact both within and across cancer types. To detach cancer classification from the tissue type we perform de novo clustering of the pancancer mutational profiles based on the Bayesian network models. We find 22 novel clusters which significantly improve survival prediction beyond clinical information. The models highlight key gene interactions for each cluster potentially allowing genomic stratification for clinical trials and identifying drug targets.

<sup>1</sup>Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland. <sup>2</sup>SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. <sup>3</sup>Division of Psychiatry, University College London, London WC1E 6BT, UK. <sup>4</sup>Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, 4031 Basel, Switzerland. <sup>5</sup>Department of Biomedicine, University of Basel, 4058 Basel, Switzerland. Correspondence and requests for materials should be addressed to J.K. (email: [jack.kuipers@bsse.ethz.ch](mailto:jack.kuipers@bsse.ethz.ch)) or to N.B. (email: [niko.beerenwinkel@bsse.ethz.ch](mailto:niko.beerenwinkel@bsse.ethz.ch))

The past years have seen great progress towards a deeper understanding of the molecular changes underpinning cancer progression. Identification and characterisation of molecular subtypes within and across different cancer types has emerged as a promising approach for the development of targeted therapies<sup>1–3</sup>. Nevertheless, cancer treatment is far from optimal. The approval of the limited number of available cancer drugs is often limited to a specific cancer type or subtype, preventing widespread use of targeted therapies. Moreover, cancers may develop resistance against these therapies, rendering them ineffective<sup>4</sup>.

Pancancer analyses enabled by large datasets such as The Cancer Genome Atlas (TCGA)<sup>5</sup> or the International Cancer Genome Consortium (ICGC)<sup>6</sup> may aid a better understanding of the disease biology across different tissues, and can, for example, identify mutational hotspots in tumours<sup>7</sup>. Genes strongly associated with a cancer type or pancancer subgroups provide insights into the molecular mechanisms, which are key for pinpointing novel therapeutic opportunities and improving current treatment strategies. For example, analysis of endometrial carcinoma showed genetic similarities to certain types of breast and ovarian cancer<sup>8</sup> while olaparib, approved for *BRCA*-mutated ovarian cancer, provided a good response in metastatic prostate cancer patients with DNA repair mutations<sup>9</sup>. Pancancer, or basket, clinical trials<sup>10</sup> could extend targeted treatments beyond their current indication, like testing the *BRAF* inhibitor vemurafenib outside of metastatic melanoma<sup>11</sup>, or test novel agents, like Loxo Oncology's trial of larotrectinib for patients with a *TRK* gene fusion mutation. Recently, the FDA granted its first approval for a drug (pembrolizumab) based on genetic markers, regardless of the tissue type.

Cancer is known to be a disease characterised by a progression of molecular changes leading to malignant features and activities<sup>12,13</sup>. Alongside stratifying tumours based on static molecular profiles<sup>14,15</sup>, investigating their development<sup>16–22</sup> may offer a new perspective on pancancer analyses with the potential to identify key drivers and provide benefits on multiple levels, including (1) prioritisation of mutation-based biomarkers; (2) uncovering previously unknown mutational dependencies; (3) identification of biomarkers of progression; and (4) biological insight into the genetic progression of cancer.

The facets of clustering patient samples, inferring their genetic tumour progression and mutational interactions are highly inter-related (Supplementary Section A). Here, we introduce a unified statistical framework to combine them by modelling the mutations as a Bayesian network. The probability of observing each mutation depends on the state of its parents in the network, thereby accounting for mutational interactions such as co-occurrence or mutual exclusivity, as well as more complex relationships. The directions of the connections may be suggestive of causal relationships<sup>23–25</sup>, though they may not be fully resolved from the data.

We develop efficient methods to infer the dependency structure of the mutations and performed fully Bayesian inference (Methods) to capture the uncertainty in the network structure learned from mutational profile data. Characterising mutational data through Bayesian networks provides useful insights through the analysis of the mutational interactions encoded by the network, beyond just analysing mutational frequencies. We employ our Bayesian network modelling to cluster patient samples into groups, with different interactions among mutated genes. The key interactions within and across novel subgroups may uncover common mechanistic insights, potential therapeutic targets, and prognostic and predictive biomarkers.

## Results

**Analysis overview.** We performed two distinct analyses of non-silent mutation data, summarised at the gene level for 201 genes, from 8198 patient samples across 22 cancer types (Supplementary Table 1) from TCGA. Initially, in a supervised analysis, we built cancer-specific probabilistic models to explore type-specific mutational interactions and pancancer heterogeneity. Then, in an unsupervised analysis, we proceeded to cluster the samples into novel mutational subgroups (Fig. 1).

**Cancer-specific Bayesian networks.** Stratifying the TCGA mutation data by tissue of origin, we built Bayesian networks (Methods) separately for each cancer type. We obtained an alternative representation of the mutational landscape that goes beyond mutation frequencies by highlighting the inter-dependencies between genes. Edges show co-occurrence or mutual exclusivity, or higher order correlations between sets of genes, offering a systematic visualisation of the most important mutational interactions in 22 different cancer types (Fig. 2).

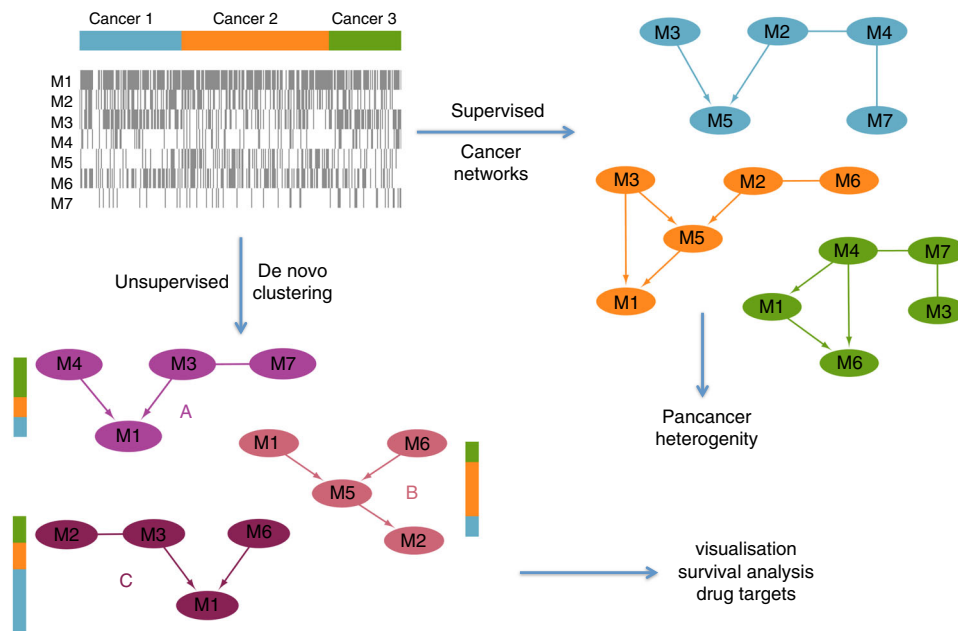
The network parameters estimated from the data capture information both about the mutational frequencies and about their interactions, allowing us to evaluate how well each model explains the mutational status of each patient sample (Methods), including those from other cancer type. Simulations show that our approach performs notably better than alternatives in learning the network structure (Supplementary Section B). It also improves on the potential to effectively characterise different cancer types with respect to simple distance-based measures (Supplementary Section C). The network structure inference is informed (Methods) by using the STRING protein–protein interaction network<sup>26</sup> as a prior, but even without this information a comparison (Methods) reveals a significant overlap of the edges (permutation test;  $p = 4.1 \times 10^5$ ), suggesting the inferred network of mutational interactions is biologically relevant as known functional interactions coincide.

We rediscovered many key genes that have previously been associated with their respective cancer types. For example, genes *ATM*, *PIK3CA* and *PTEN* with a lot of connections in colorectal cancer (Fig. 2) have previously been reported as highly mutated<sup>27</sup>. In lung adenocarcinomas, we recapitulate the mutual exclusivity between *KRAS* and *EGFR* as well as the importance of *TP53*<sup>28</sup>, for which our model suggests that it is frequently co-mutated with both *KRAS* and others like *MLL3*, as well as *KRAS* and *STK11* are frequently co-mutated.

*TP53* is a major hub with the most interactions across multiple cancer types (67 in total) and with multiple dependencies in brain cancer and especially lower-grade glioma. *TP53* mutations in lower-grade glioma have previously been associated with the disease, along with *IDH1*, *FUBP1*, *ATRX*, *CIC*, *NOTCH1*, *EGFR*, and *PIK3CA*<sup>29</sup> among which we see some interactions. Similarly in glioblastoma, where in addition to *TP53*, we observe interactions involving *PTEN*, *IDH1* and *ATRX* which were previously reported to be mutated<sup>30</sup>.

Other hubs with high connectivity like *TP53* include *MLL2* (65 interactions), *MLL3* (58), *XYLT2* (56) and *FAT1* (55). Mutations in these genes have previously been associated with cancer<sup>31–33</sup>. *MLL3* and *FAT1* share connections across several cancer types, with both having several interaction in uterine cancer, while *MLL2* exhibits many connections for stomach, and *XYLT2* for oesophageal cancer.

Strong edges can hint at common mechanistic causes or fitness effects. For the example of *FAT1*, our study finds relevant interacting mutations in breast, colorectal, endometrial, kidney, lung, liver, stomach and head and neck cancer, with mutation rates between 2 and 24%. Although *FAT1* was reported to be



**Fig. 1** Overview of the analyses. Starting from the mutation data, we perform two types of analysis. Supervised learning of the Bayesian network structure for each known cancer type, allowing us to uncover mutational interactions and visualise pancancer heterogeneity. Unsupervised clustering of the mutation data into components with common interactions to uncover a novel stratification of the patient samples

recurrently mutated in several cancer types<sup>32,34,35</sup>, our results suggest that the gene correlates with a large number of mutations in different cancer types, including *FAT1* in breast cancer, *ATM* in colorectal cancer, and *APC*, *MTOR* and *MLL3* in endometrial cancer. On the pathway level, we found highly connected genes across cancer types to be significantly involved in many signal transduction pathways, as well as cellular processes and DNA damage repair (Supplementary Table 8).

Several genes with mutational interactions are putatively actionable, with drugs either approved or currently tested in a clinical study (labelled with black diamonds in Fig. 2). In general, approval of targeted therapies is limited to one or several cancer types or cancer subtypes. Strong dependencies in the Bayesian network potentially indicate effectiveness of targeted therapies against dependent gene mutations, particularly if it is in the same pathway. Moreover, the network potentially expands the group of tumours responsive to a specific targeted therapy, within the same tumour types and between different tumour types. Although the drugs are only targeting the mutation itself, we see for example several interactions of *KRAS* in lung adenocarcinoma (with *TP53*, *EGRF* and *STK11*) and there are currently several clinical trials investigating the effectiveness and safety of targeted therapies against *KRAS* for this cancer type (NCT02642042, NCT01912625 or NCT02079740).

A two-dimensional visualisation based on multiscale projections (Methods) highlights the differences and similarities of patient samples as measured over the Bayesian networks, within and across cancer types (Fig. 3). Some, like colorectal, thyroid and lower grade glioma, are well separated and hence well defined by the mutation profiles of their patient samples; others are much more similar, for example those within overlapping groups. Even for the better separated cancer types, we observe substantial heterogeneity with some patient samples closer to those from other cancers.

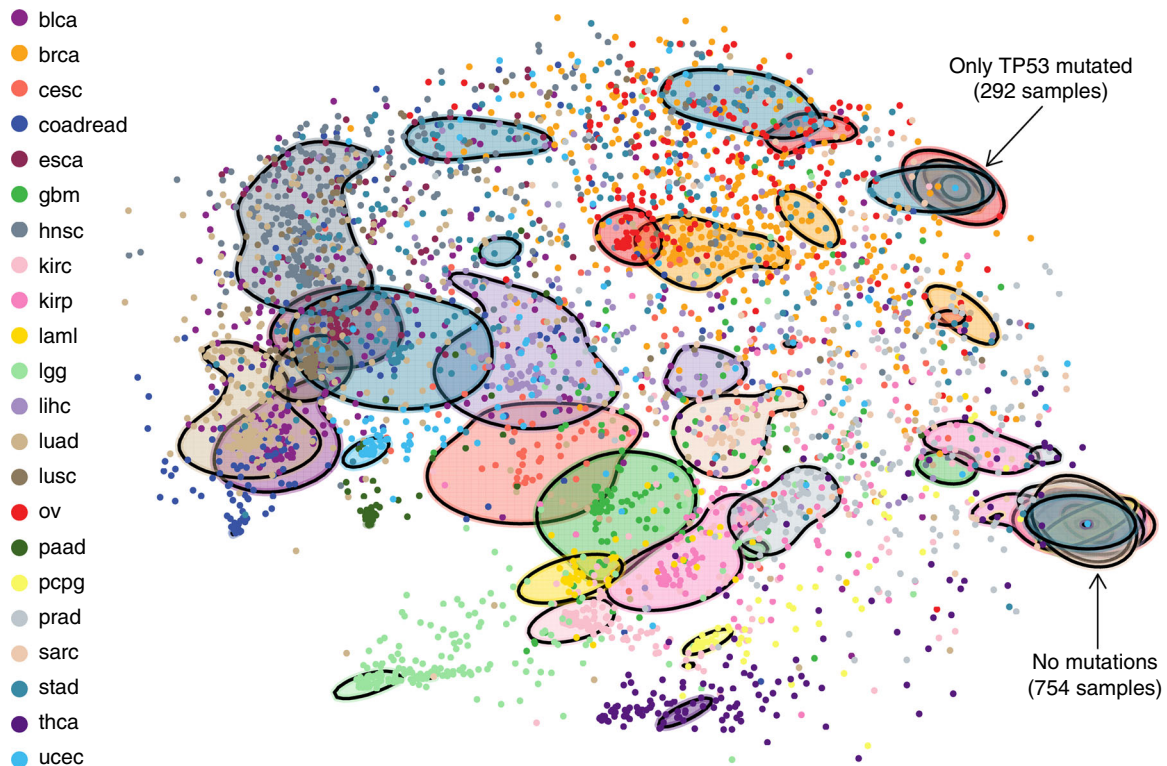
As well as being applicable to pancancer cohorts, these methods naturally apply when focussing on individual cancer types and their subtypes (as we discuss for breast cancer in Supplementary Section D).

**De novo clustering based on Bayesian networks.** Given the considerable heterogeneity across cancer types, we asked ourselves whether the mutation profiles themselves can be re-clustered on the basis of Bayesian network models without knowing the cancer types. This model-based de novo clustering of the binary mutation data (Methods and Supplementary Section F) identified 22 groups, coincidentally the same number as the original cancer types (which was not imposed), but distinct in their composition. Each cluster is defined by a Bayesian network which constitutes a generative model of its patient data. These models fit the data much better than the partitioning by cancer type—the average cluster assignment is 91.5% compared to 71.9% for the cancer-specific models.

The composition of the clusters (Fig. 4) shows that the more differentiated cancer types, like colorectal, thyroid and lower grade glioma from the 2D projection (Fig. 3) initiate new clusters. Clusters G and I are mostly composed of glioma and colorectal samples respectively. Almost all thyroid cancer samples belong to cluster V, along with samples from other cancer types, and this cluster exhibits a strong enrichment in *BRAF* mutations (17% compared to an overall rate of 7%, Supplementary Table 9). Cluster K, where *TP53* plays a strong role, mostly consists of colorectal samples, and cluster L of leukaemia with *IDH* mutations being prominent. The ovarian cancer samples belong almost entirely to cluster U (where all samples exhibit a *TP53* mutation). Samples with no mutations among the 201 genes are assigned to cluster V.

The de novo clustering also split patient samples from the same cancer type, like the glioma samples into clusters G and N: cluster N has elevated mutation rates in *TP53* and *ATRX* which are largely absent from cluster G which instead has elevated rates of *CIC* as well as even higher rates of *IDH1* than N (98% compared to 86%). Patient samples from certain cancer types may actually be more similar to tumours in other cancers. Cluster S, for example, groups samples of several different cancer types, including bladder, breast, liver and lung cancers and possesses elevated mutation rates in several genes, like *TP53*, *MLL*, *ARID* and *CTNNB1*.





**Fig. 3** Visualisation of pancancer heterogeneity. 2D visualisation of the similarity between patient samples based on their fit to each cancer-specific Bayesian network, highlighting the heterogeneity within and across cancer types. For example, stomach, breast and liver cancer show high inter-tumour heterogeneity with a high spread across the plot, whereas pancreatic cancer shows low inter-tumour heterogeneity and is much more localised. Ovarian and breast cancers as well as bladder cancer and lung adenocarcinomas show similar mutational profiles, while lower grade glioma is rather distinct from other cancer types, as is thyroid cancer. The solid shapes are based on contours that together contain a total of 50% of the respective cancer types. The group of samples on the lower right possess no mutations among the 201 genes while those exhibiting a mutation only in *TP53* are also indicated. Versions highlighting certain cancer types are displayed in Supplementary Fig. 5

main groups. Clusters J, with a large group of lung adenocarcinomas, and U have significantly poorer prognosis.

Some cancer types are split by the clustering, with one portion forming the bulk of a cluster, and show differences in survival between the patients assigned to different clusters. For example, leukaemia samples in cluster L show a significantly lower lethal risk compared to samples in cluster V (hazard ratio = 0.51;  $p = 0.012$ ). Lower grade glioma samples in cluster N have a non-significantly higher risk than those in cluster G (hazard ratio = 1.64;  $p = 0.056$ ). Similarly colorectal samples in cluster K show non-significantly higher lethal risk than colorectal samples in cluster I (hazard ratio = 1.67;  $p = 0.063$ ), in line with the fact that cluster K contains a large number of samples from the MSS subtype which has the worst outcome<sup>36</sup>.

The probabilistic model describing each cluster can be utilised to visualise the important mutational interactions characterising the mutation profiles of its samples (Fig. 5). The cluster composition is distinct from the grouping by cancer (Fig. 2), even for clusters dominated by one cancer type such as cluster B, G, I or J, so that the key interactions mostly differ, especially when focussing on 20 genes per cluster for clarity.

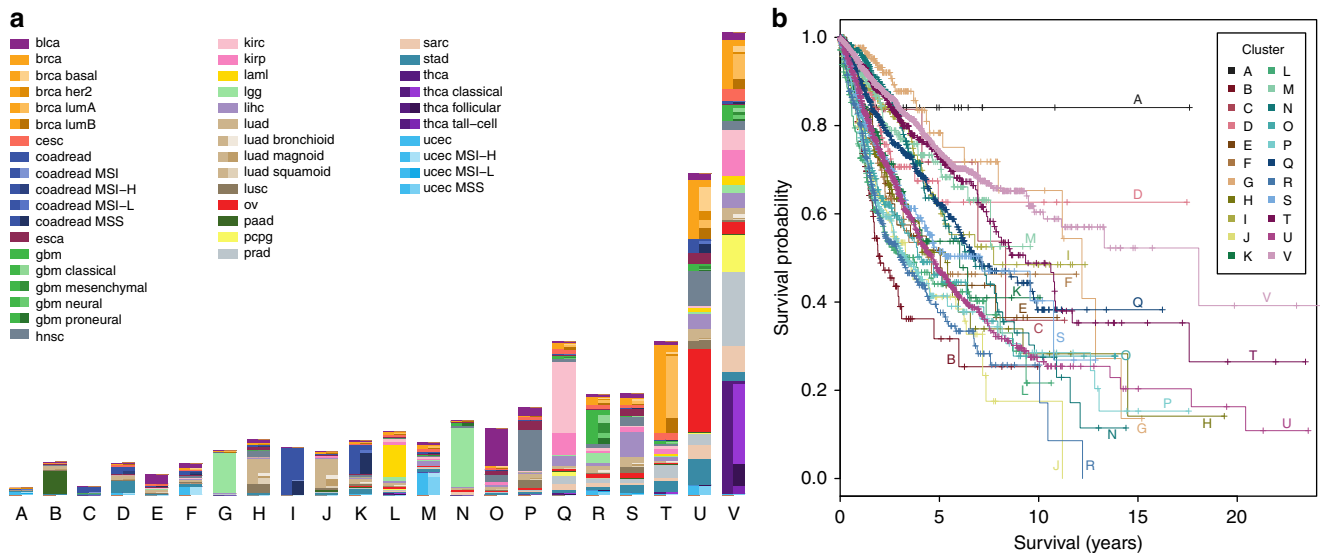
*TP53*, as the most frequently mutated gene across all cancer types, remains fairly prominent when clustering samples by their mutation profile. Looking at its connections among all genes, these differ substantially across the clusters even when the marginal frequency is similar (Fig. 6), suggesting that it plays different roles in different clusters. Our approach emphasises mutational interaction patterns which are specific to a cluster. Along with its interactions, the prevalence of *TP53* in each cluster

is also used in assigning patient samples. For example all members of cluster U possess a *TP53* mutations while none of cluster V do.

Among the clusters, genes such as *ERBB2* (37 interactions), *MSH3* (36) or *CDKN1B* (33) have more interactions than *TP53* (32) in total. Of the selected interactions (Fig. 5), *ERBB2* has connections particularly for clusters T and M, and for cluster M interacts with *CTNNB1*. *PIK3CA* has several interactions, including three for cluster C (with *TP53*, *CTNNB1* and *VHL*), a cluster dominated by colorectal cancer samples.

We found several cancer-associated signalling pathways significantly enriched in highly connected genes across clusters, including HIF-1, Jak-STAT, p53, Toll-like receptor, and TNF signalling (Supplementary Table 10). These results suggest that genes involved in signalling pathways are not only highly connected, with a large number of mutational interactions with other genes, but that they also play an important role in characterising the clusters.

Clustering samples by their mutational profile, independent of their cancer type, provides an alternative patient stratification which may inform targeted treatment (putatively actionable genes are marked by black diamonds in Fig. 5). For example, *BRAF* inhibitors are being currently tested or have already been approved for multiple cancer types, including lung cancer, ovarian cancer, and thyroid cancer<sup>37</sup>. Substantial fractions of these cancer types are grouped in the single cluster V (for which *BRAF* exhibits three interactions in Fig. 5). The similarity of mutational profiles to other cancer types in cluster V suggests that they may be responsive to the same targeted treatment, such as



**Fig. 4** De novo clustering. **a** Assignment of the 8198 patient samples to the 22 clusters, labelled A–V, based on Bayesian network clustering of their mutation profiles. The left-hand side of the bar for each cluster indicates the number of patient samples with a given cancer type, while the right-hand side indicates the breakdown into the known subtypes. **b** Survival probabilities of the 22 clusters

*BRAF* inhibitors, provided the targeted gene or pathway is mutated.

## Discussion

We modelled mutational profiles with Bayesian networks, which capture the interactions between mutations, in a pancancer setting across 22 cancer types. Clustering of pancancer data can be highly insightful into the molecular similarities across cancer types<sup>14,15</sup>. In order to concurrently model cancer heterogeneity and mutational interactions we combined the Bayesian network approach with clustering in an integrated framework. The challenge for network clustering of large datasets resides in achieving reliable inference of networks with many nodes. Therefore we also developed methods for fast network learning for large data (Methods), so to extend the analysis to a couple of hundred important genes. Larger networks than those analysed may not lead to any further benefit, since many cancer mutations are quite rare and therefore unlikely to show strong interactions detectable in network modelling.

Mutations can be modelled at different levels, from the finer scale of individual mutation sites or hotspots<sup>7</sup> up to the pathway level. Higher resolution comes at the cost of lower frequencies, but this could be balanced out by combining aberrations by their molecular or pathway functions. A powerful alternative could be to use diffusion algorithms to condense aberrations to their affected subnetworks<sup>38</sup> or mediator genes<sup>39</sup>. Here we focussed on mutation data summarised at the gene level, which contain a large amount of information. Mutational profiles are one genomic lens through which to identify molecular subtypes which can then complement other views based on, for example, copy number, expression and methylation profiles<sup>15</sup> and there may also be interactions across data modalities. Cancers can be highly heterogeneous, potentially harbouring distinct clones with different mutational profiles and prognostic signatures. Finer modelling of clonal structure and its impact on patient stratification present substantial challenges, as well as opportunities for more precise treatment options.

Furthermore, we account for uncertainty in the network structure through a fully Bayesian approach (Methods). The networks learnt provide the key gene interactions characterising each cancer type (Fig. 2) and a significantly extended view over

just mutational frequencies. Along with the edges, the networks also utilise the frequency of each mutation to quantify how similar or disparate the mutation patterns of patient samples are within and across cancer types (Fig. 3).

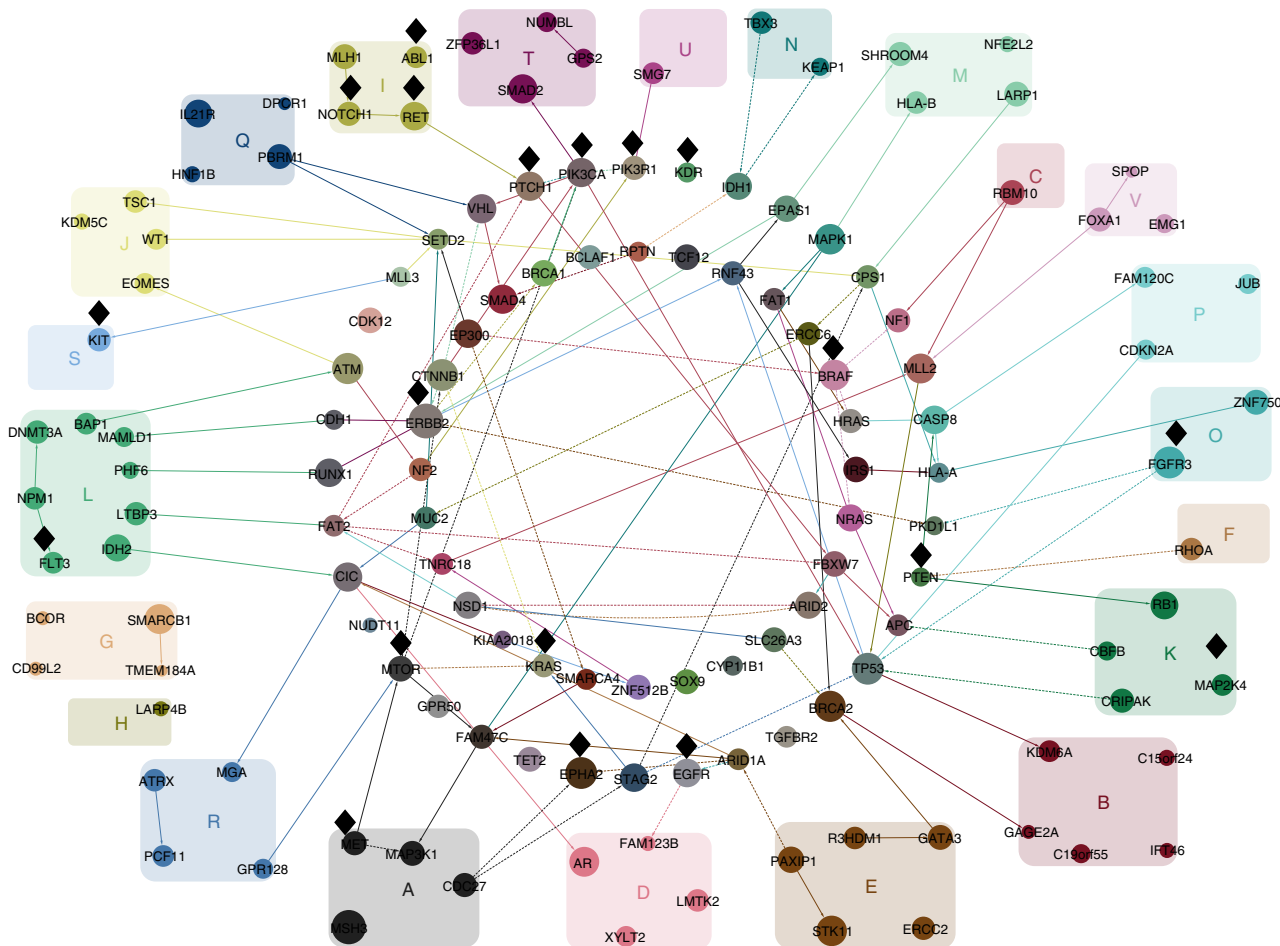
Because the networks we infer are related to a class of clustering methods, we directly employed our models to re-cluster the TCGA data. We discovered and characterised 22 clusters in the data, distinct from the original cancer types but happening to match in number. New patient samples, including mutation profiles from different tumour types, can also be classified into the clusters, for which we provide a web application (Software). The new clusters significantly improve survival prediction, over and above that from the clinical and histopathological status of each patient sample. Integrating this information along with the mutational profiles during the clustering could further improve the survival prediction. Different types of mutations can occur and mutations also have different functional impact, which may affect their potential interactions, and may have different effects in different tissues. Accounting for such interactions could offer further refinements.

The Bayesian network modelling and clustering developed here provides us with insights into the mutation events and specifically their dependencies in cancer types and in novel cancer subgroups. These can be used as biomarkers which may be explored experimentally for therapeutic intervention.

## Methods

**Cancer mutation data.** Mutation annotation files were obtained through GDAC Firehose<sup>40</sup> for the 22 cancer types with sequenced primary tumour samples from more than 100 patients (see Supplementary Table 1), giving a total of 8198 patient samples. The 16 most significantly mutated genes for each cancer type were collated to give a total of 201 genes considered. The different sample sizes and potentially different effect sizes makes comparing the significance of gene mutations challenging across cancer types. We found such comparisons led to gene lists dominated by a small number of cancer types, with correspondingly fewer discriminatory pancancer markers. Instead, a fixed number per cancer type was employed to ensure markers were retained for each cancer type to better characterise inter-tumour heterogeneity. A binary matrix of the non-silent mutations in these genes across all patient samples was generated.

Putatively actionable targets are identified based on approved drugs or drugs currently undergoing clinical studies as collected by MyCancerGenome<sup>37</sup>, a manually curated precision cancer medicine knowledge resource. MyCancerGenome was queried through rDGIdb<sup>41,42</sup>. Functional annotation of genes was performed using the KEGGREST package<sup>43</sup>.



**Fig. 5** The cluster-specific connections between the genes. The connections between the 20 most frequent and connected genes per cluster (rather than per cancer type, as in Fig. 2). Black diamonds near nodes indicate putatively actionable genes. Edges highlight interdependencies between the selected mutations: solid for positive and dashed for negative correlation. Directed edges point in the inferred direction of dependency, whereas undirected edges are where the direction cannot be inferred. Node size reflects the total number of edges, including edges not shown. Nodes are coloured by combining the colours of their edges from the different clusters

**Graph inference.** To quantify the extent to which a Bayesian network can explain a set of binary data, we use the BDe score<sup>44</sup> where the state of a node  $X$  is determined by a different parameter  $\theta_Y$  for each configuration of its  $m$  parents  $Y: P(X = 1 | Y) = \theta_Y$  with a beta prior on each  $\theta_Y$  with hyperparameters  $\alpha = \beta = \frac{\chi}{2^{m+1}}$  defined in terms of a single parameter  $\chi$  which represents the number of pseudocounts added.

To make inference of the Bayesian network structure, which is a directed acyclic graph (DAG), we adopt a modified version of order MCMC<sup>45,46</sup>. For large networks we search over a reduced skeleton, initially found via the PC algorithm<sup>47</sup> and then expanded by MCMC search using the software package **BiDAG**<sup>48</sup>. From the final skeleton, a sample from the posterior is obtained using partition MCMC<sup>46</sup> optimised for such a skeleton<sup>48</sup>.

**Prior edge knowledge.** We obtained all human functional interactions from STRING<sup>26</sup>, selecting those between the 201 genes included in this study ( $\approx 7000$  interactions). Edges that were not in this STRING network were penalised by a factor 2 for the graph inference.

**Additional edge penalisation.** When examining the networks learned from binary data, as in Figs. 2 and 5 we additionally penalise all edges. Edge penalisation (by a factor of 2) has previously been examined with the marginal uniform prior<sup>49</sup> and shown to improve network reconstruction. Based on simulation studies (Supplementary Section B) we find stronger regularisation further improves the accuracy for data mimicking the TCGA and employ a factor of 16 to regularise the network. We sample 100 DAGs from the posterior. Edges are only displayed if they appear in at least half of the posterior sample.

To select genes to display per cancer type or per cluster in Figs. 2 and 5, for each gene we multiply its number of connections by its frequency, and choose the 20 genes with the largest product.

**Scoring samples against DAGs.** For a given DAG the posterior distribution of each node  $X$  given its parent state  $Y$  is again a beta distribution with updated parameters  $\alpha + \tilde{\alpha}_Y$  and  $\beta + \tilde{\beta}_Y$ , where  $\tilde{\alpha}_Y$  is the number of times  $X$  takes the value 1 when the parents are  $Y$  and  $\tilde{\beta}_Y$  is the number of times it takes the value 0 in the data. Hence the likelihood

$$P(X = 1 | G, Y) = \hat{\theta}_Y = \frac{\alpha + \tilde{\alpha}_Y}{\alpha + \tilde{\alpha}_Y + \beta + \tilde{\beta}_Y} \tag{1}$$

can be evaluated for each node of an arbitrary observed binary vector  $X$ , providing a measure of fit of the observed vector to the DAG.

Given a sample of  $M$  DAGs from the posterior distribution  $P(G | k)$  of different cancer types (indexed by  $k$ ), we can score each patient sample  $D_i$  against each DAG  $G_j$  (dropping the index for the cluster  $k$ ). From the sample we can build the Monte Carlo approximation to the likelihood of the data for a given cluster  $k$

$$P(D_i | k) \approx \frac{1}{M} \sum_{j=1}^M P(D_i | G_j) \tag{2}$$

using the likelihoods in Eq. (1). Under a uniform prior over cancer types, the likelihoods in Eq. (2) are normalised to probability vectors over the collection of cancer types according to  $P(k | D_i) \propto P(D_i | k)P(k)$ . Similarities between the probability vectors of different patient samples are computed as their Jensen–Shannon divergence. Calculating this divergence between all pairs of patient samples provides a distance matrix between patient samples which we project into 2D with multidimensional scaling using the **cmdscale** command in R, as in Fig. 3.

**Clustering with frequency information.** We cluster the data with a mixture model. Given  $K$  graphs and parameter sets  $(G_k, \theta_k)$ , we assume that each patient sample  $D_i$  is generated from one of  $K$  models depending on the value of a latent





was evaluated by the likelihood ratio test, based on the asymptotic  $\chi^2$  distribution.

The possible values for stage were I, II, III, IV, along with unclassified values of ‘[Not Applicable]’, ‘[Not Available]’, ‘[Discrepancy]’, ‘[Unknown]’, and ‘Stage X’. The ‘[Not Applicable]’ category is entirely determined by the tissue type and hence does not affect the regression when adjusting for tissue type. The remaining categories other than ‘[Not Available]’ consisted of only 56 patient samples. Performing a Cox regression with adjustment for age and tissue type indicated that keeping those categories separate did not lead to significantly better predictions (likelihood ratio = 1.2;  $p = 0.13$ ) due to their small sizes. All the unclassified values were hence combined into a single stage X. Keeping stage I and stage II separate led to significantly better survival prediction compared to their combination into a single category, and similarly for stage III and IV, and hence all stages were kept separate. Therefore five levels (whose size is summarised in Supplementary Table 5) were retained for the stage covariate in the Cox regression.

Modelling survival on the basis of the clinical data available from TCGA is challenging, due to intrinsic limitations in the data<sup>51</sup>, including, for example, relatively short follow-up times as well as heterogeneity in data collection and cohort selection across the different cancer types. Therefore, although we adjust for cancer type in our Cox regressions and use the same survival model to compare different clustering methods, there may be other effects like cancer-type-specific confounding not accounted for.

**Overlap with functional network.** To compare the STRING<sup>26</sup> network to the Bayesian networks inferred in this study we re-ran the analysis without using the functional STRING network as a prior. We then created the union of all edges in the Bayesian networks over all cancer types. We performed a permutation test on the overlap between these two networks by generating one million random permutations of the gene labels in the functional network. For each permuted network, we computed the mean overlap to derive the empirical  $p$  value.

A web interface to classify new patient samples is at <https://cbg.bsse.ethz.ch/pancancer/>. The package BiDAG for Bayesian network inference is at <https://CRAN.R-project.org/package=BiDAG>. R code for the Bayesian network clustering and survival analysis is available at <https://github.com/cbgeth/pancancer-clustering>.

**Code availability.** The network inference code is available at <https://CRAN.R-project.org/package=BiDAG>. The clustering and survival analysis code is available at <https://github.com/cbg-ethz/pancancer-clustering>.

## Data availability

The mutational profiles and clinical information of the patient samples used in the study are available at <https://github.com/cbg-ethz/pancancer-clustering>, along with their cluster assignments.

Received: 20 September 2017 Accepted: 1 October 2018

Published online: 19 October 2018

## References

- Sun, S., Schiller, J. H., Spinola, M. & Minna, J. D. New molecularly targeted therapies for lung cancer. *J. Clin. Investig.* **117**, 2740–2750 (2007).
- Higgins, M. J. & Baselga, J. Targeted therapies for breast cancer. *J. Clin. Investig.* **121**, 3797–3803 (2011).
- Roock, W. D., Vriendt, V. D., Normanno, N., Ciardiello, F. & Tejpar, S. KRAS, BRAF, PIK3CA, and PTEN mutations: implications for targeted therapies in metastatic colorectal cancer. *Lancet Oncol.* **12**, 594–603 (2011).
- Groenendijk, F. H. & Bernards, R. Drug resistance to targeted therapies: déjà vu all over again. *Mol. Oncol.* **8**, 1067–1083 (2014).
- McLendon, R. et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
- Hudson, T. J. et al. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
- Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).
- TCGA Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
- Mateo, J. et al. DNA-repair defects and olaparib in metastatic prostate cancer. *N. Engl. J. Med.* **2015**, 1697–1708 (2015).
- Cunanan, K. M. et al. Basket trials in oncology: a trade-off between complexity and efficiency. *J. Clin. Oncol.* **35**, 271–273 (2017).
- Hyman, D. M. et al. Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. *N. Engl. J. Med.* **373**, 726–736 (2015).
- Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
- Vogelstein, B. et al. Genetic alterations during colorectal tumor development. *N. Engl. J. Med.* **319**, 525–532 (1988).
- Ciriello, G. et al. Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).
- Hoadley, K. A. et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues-of-origin. *Cell* **158**, 929–944 (2014).
- Gerstung, M., Baudis, M., Moch, H. & Beerenwinkel, N. Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics* **25**, 2809–2815 (2009).
- Attolini, C. S.-O. et al. A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc. Natl. Acad. Sci. USA* **107**, 17604–17609 (2010).
- Gerstung, M., Eriksson, N., Lin, J., Vogelstein, B. & Beerenwinkel, N. The temporal order of genetic and pathway alterations in tumorigenesis. *PLoS ONE* **6**, e27136 (2011).
- Farahani, H. S. & Lagergren, J. Learning oncogenetic networks by reducing to mixed integer linear programming. *PLoS ONE* **8**, e65773 (2013).
- Misra, N., Szczurek, E. & Vingron, M. Inferring the paths of somatic evolution in cancer. *Bioinformatics* **30**, 2456–2463 (2014).
- Ramazotti, D. et al. CAPRI: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics* **31**, 3016–3026 (2015).
- Cristea, S., Kuipers, J. & Beerenwinkel, N. pathTiME: joint inference of mutually exclusive cancer pathways and their progression dynamics. *J. Comput. Biol.* **24**, 603–615 (2017).
- Pearl, J. & Verma, T. S. A theory of inferred causation. (eds. Allen, J. F., Fikes, R. & Sandewall, E.) In *Second International Conference on Principles of Knowledge Representation and Reasoning* 441–452, (Morgan Kaufmann Publishers, San Francisco, CA, USA, 1991).
- Pearl, J. *Causality: Models, Reasoning and Inference* (MIT Press, Cambridge, MA, 2000).
- Dawid, A. P. Beware of the DAG! *J. Mach. Learn. Res. Workshop Conf. Proc.* **6**, 59–86 (2010).
- Szklarczyk, D. et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2014).
- TCGA Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
- TCGA Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
- TCGA Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* **372**, 2481–2498 (2015).
- Brennan, C. W. et al. The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
- Olivier, M., Hollstein, M. & Hainaut, P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.* **2**, a001008 (2009).
- Morris, L. G. et al. Recurrent somatic mutation of FAT1 in multiple human cancers leads to aberrant wnt activation. *Nat. Genet.* **45**, 253–261 (2013).
- Kantidakis, T. et al. Mutation of cancer driver MLL2 results in transcription stress and genome instability. *Genes Dev.* **30**, 408–420 (2016).
- Katoh, M. Function and cancer genomics of FAT family genes (review). *Int. J. Oncol.* **41**, 1913–1918 (2012).
- Garg, M. et al. Profiling of somatic mutations in acute myeloid leukemia with FLT3-ITD at diagnosis and relapse. *Blood* **126**, 2491–2501 (2015).
- Phipps, A. I. et al. Association between molecular subtypes of colorectal cancer and patient survival. *Gastroenterology* **148**, 77–87 (2015).
- Taylor, A. D., Micheel, C. M., Anderson, I. A., Levy, M. A. & Lovly, C. M. The path(way) less traveled: a pathway-oriented approach to providing information about precision cancer medicine on my cancer genome. *Transl. Oncol.* **9**, 163–165 (2016).
- Leiserson, M. D. et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106 (2015).
- Dimitrakopoulos, C. et al. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* **34**, 2441–2448 (2018).
- Broad Institute TCGA Genome Data Analysis Center. Analysis-ready standardized TCGA data from Broad GDAC Firehose stddata\_\_2015\_08\_21 run (2016).
- Wagner, A. H. et al. DGIdb 2.0: mining clinically relevant drug–gene interactions. *Nucleic Acids Res.* **44**, D1036–D1044 (2015).
- Thurnherr, T., Singer, F., Stekhoven, D. J. & Beerenwinkel, N. Genomic variant annotation workflow for clinical applications. *F1000Res.* **5**, 1963 (2016).

43. Tenenbaum, D. *KEGGREST: Client-Side REST Access to KEGG*. R package version 1.14.0 (2016).
44. Heckerman, D. & Geiger, D. Learning Bayesian networks: a unification for discrete and Gaussian domains. (eds. Besnard, P. & Hanks, S.) In *Eleventh Conference on Uncertainty in Artificial Intelligence* 274–284 (Morgan Kaufmann Publishers, San Francisco, CA, USA, 1995).
45. Friedman, N. & Koller, D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Mach. Learn.* **50**, 95–125 (2003).
46. Kuipers, J. & Moffa, G. Partition MCMC for inference on acyclic digraphs. *J. Am. Stat. Assoc.* **112**, 282–299 (2017).
47. Spirtes, P., Glymour, C. N. & Scheines, R. *Causation, Prediction, and Search* (MIT Press, Cambridge, MA, 2000).
48. Suter, P. & Kuipers, J. BiDAG: Software for the efficient inference and sampling of Bayesian networks. <https://CRAN.R-project.org/package=BiDAG> (2017).
49. Scutari, M. An empirical-Bayes score for discrete Bayesian networks. *J. Mach. Learn. Res.* **52**, 438–448 (2016).
50. Dean, N. & Raftery, A. E. Latent class analysis variable selection. *Ann. Inst. Stat. Math.* **62**, 11–35 (2010).
51. Liu, J. et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416 (2018).
52. Wu, R.-C., Wang, T.-L. & Shih, I.-M. The emerging roles of ARID1a in tumor suppression. *Cancer Biol. Ther.* **15**, 655–664 (2014).
53. Longo, T. et al. Targeted exome sequencing of the cancer genome in patients with very high-risk bladder cancer. *Eur. Urol.* **70**, 714–717 (2016).

### Acknowledgements

J.K. was supported by ERC Synergy Grant 609883 (<http://erc.europa.eu/>). T.T. was supported by EU Horizon 2020 PHC Grant 633974 (SOUND).

### Author contributions

J.K., G.M., J.B., and N.B. designed the study. J.K., G.M., and P.S. developed the methodology. T.T. collated the data. J.K., T.T., and R.G. performed the analyses. J.K., T.T., G.

M., and R.G. drafted the article. G.C. and N.B. critically reviewed the manuscript. All authors approved the final version.

### Additional information

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-018-06867-x>.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018