Title

Network-based stratification of tumor mutations.

Journal

Nature methods, 10(11)

ISSN

1548-7091

Authors

Hofree, Matan

Shen, John P

Carter, Hannah

et al.

Publication Date

2013-11-01

DOI

10.1038/nmeth.2651

Peer reviewed

Network-based stratification of tumor mutations

Matan Hofree[1], John P Shen[2], Hannah Carter[2], Andrew Gross[3] & Trey Ideker[1–3]

Many forms of cancer have multiple subtypes with different causes and clinical outcomes. Somatic tumor genome sequences provide a rich new source of data for uncovering these subtypes but have proven difficult to compare, as two tumors rarely share the same mutations. Here we introduce network-based stratification (NBS), a method to integrate somatic tumor genomes with gene networks. This approach allows for stratification of cancer into informative subtypes by clustering together patients with mutations in similar network regions. We demonstrate NBS in ovarian, uterine and lung cancer cohorts from The Cancer Genome Atlas. For each tissue, NBS identifies subtypes that are predictive of clinical outcomes such as patient survival, response to therapy or tumor histology. We identify network regions characteristic of each subtype and show how mutation-derived subtypes can be used to train an mRNA expression signature, which provides similar information in the absence of DNA sequence.

Cancer is a disease that is not only complex, i.e., driven by a combination of genes, but also wildly heterogeneous, in that gene combinations can vary greatly between patients. To gain a better understanding of these complexities, researchers involved in projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) are systematically profiling thousands of tumors at multiple layers of genome-scale information, including mRNA and microRNA expression, DNA copy number and methylation, and DNA sequence[1–3]. There is now a strong need for informatics methods that can integrate and interpret genome-scale molecular information to provide insight into the molecular processes driving tumor progression. Such methods are also of pressing need in the clinic, where the impact of genome-scale tumor profiling has been limited by the inability to derive clinically relevant conclusions from the data[4,5].

One of the fundamental goals of cancer informatics is tumor stratification, whereby a heterogeneous population of tumors is divided into clinically and biologically meaningful subtypes as determined by similarity of molecular profiles. Most prior attempts to stratify tumors with molecular profiles have used mRNA expression data[2,6–9], resulting in the discovery of informative subtypes in diseases such as glioblastoma and breast cancer. On the other hand, in TCGA cohorts including colorectal adenocarcinoma and small-cell lung cancer, subtypes derived from expression profiles do not correlate with any clinical phenotype including patient survival and response to chemotherapy[2,10]. These results might be due to limitations of expression-based analysis[11] such as issues with RNA sample quality, lack of reproducibility between biological replicates and ample opportunities for overfitting of data.

A promising new source of data for tumor stratification is the somatic mutation profile, in which high-throughput sequencing is used to compare the genome or exome of a patient's tumor to that of the germ line to identify mutations that have become enriched in the tumor cell population[12]. As this set of mutations is presumed to contain the causal drivers of tumor progression[13], similarities and differences in mutations across patients could provide invaluable information for stratification. Although individual mutations in cancer genes have long been used to stratify patients[14–17], stratification based on the entire mutation profile has been more challenging. Somatic mutations are fundamentally unlike other data types such as expression or methylation, in which nearly all genes or markers are assigned a quantitative value in every patient. Instead, somatic mutation profiles are extremely sparse, with typically fewer than 100 mutated bases in an entire exome (**Supplementary Fig. 1**). They are also remarkably heterogeneous, such that it is very common for clinically identical patients to share no more than a single mutation[2,18,19].

Here we report that these problems can be largely overcome by integrating somatic mutation profiles with knowledge of the molecular network architecture of human cells. It is widely appreciated that cancer is a disease not of individual mutations, nor of genes, but of combinations of genes acting in molecular networks corresponding to hallmark processes such as cell proliferation and apoptosis[20,21]. We postulated that, although two tumors may not have any mutations in common, they may share the networks affected by these mutations (as per Waddington's original theory of 'genetic canalization'[22]). Although current cancer pathway maps are incomplete, much relevant information is available in public databases of human protein-protein, functional and pathway interactions. An increasing number of studies have successfully integrated these network databases with tumor molecular profiles to map the molecular pathways of cancer[23–27]. Here we focus on the orthogonal problem of using network knowledge to stratify a cohort into meaningful subsets. Using this

[1]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA. [2]Department of Medicine, University of California, San Diego, La Jolla, California, USA. [3]Department of Bioengineering, University of California, San Diego, La Jolla, California, USA. Correspondence should be addressed to T.I. (tideker@ucsd.edu).

RECEIVED 14 FEBRUARY; ACCEPTED 12 AUGUST; PUBLISHED ONLINE 15 SEPTEMBER 2013; DOI:10.1038/NMETH.2651

knowledge, we were able to cluster somatic mutation profiles into robust tumor subtypes that are biologically informative and have a strong association to clinical outcomes such as patient survival time and emergence of drug resistance. As a proof of principle, we applied this method to stratify the somatic mutation profiles of three major cancers cataloged in TCGA: ovarian, uterine and lung adenocarcinoma.

RESULTS

Overview of network-based stratification

NBS combines genome-scale somatic mutation profiles with a gene interaction network to produce a robust subdivision of patients into subtypes (**Fig. 1a**). Briefly, somatic mutations for each patient are represented as a profile of binary (1, 0) states on genes, in which a '1' indicates a gene for which mutation (a single-nucleotide base change or the insertion or deletion of bases) has occurred in the tumor relative to germ line. For each patient, we project the mutation profile onto a human gene interaction network obtained from public databases[28–30]. Next we apply network propagation[31] to spread the influence of each mutation over its network neighborhood (**Fig. 1b**). The resulting matrix of 'network-smoothed' patient profiles is clustered into a predefined number of subtypes (k = 2, 3, … 12) via non-negative matrix factorization[32] (NMF, **Fig. 1c**), an unsupervised technique. Finally, to promote robust cluster assignments, we use consensus clustering[33], aggregating the results of 1,000 different subsamples from the entire data set into a single clustering result (**Fig. 1d**). For further details, see Online Methods. To evaluate the impact of different sources of network data, we used three interaction databases for this analysis: search tool for the retrieval of interacting genes (STRING)[29], HumanNet[28] or PathwayCommons[30]. **Supplementary Table 1** summarizes the number of genes and interactions used in our analysis from each of these three networks. Our implementation of NBS is available as **Supplementary Software**; for updated versions, NBS may be downloaded from http://idekerlab.ucsd.edu/software/NBS/.

Benchmarking and performance analysis

In an initial exploration of NBS, we simulated a somatic mutation data set using the structure of the TCGA ovarian tumor mutation data and the STRING gene interaction network (**Fig. 2a**). Mutation profiles were permuted, and patients were divided randomly and uniformly into a predefined number of subtypes (k = 4). Next we reassigned a fraction of mutations in each patient to fall within genes of a single 'network module' characteristic of that patient's subtype (the 'driver' mutation frequency f, varied from 0% to 15%); the remaining mutations were left to occur randomly. We selected the network modules randomly from the set of all network modules in STRING, defined as sets of densely interacting genes with size range s = 10–250 (see Online Methods for details and justification for the ranges of k, f and s). Although it is unknown whether these assumptions completely mirror the biology of cancer, they provide a reasonable model of a pathway-based genetic disease that is (i) driven by genetic circuits corresponding to a molecular network whose activity can be altered by mutations at multiple genes and (ii) characterized by many additional mutations that are noncausal 'passengers'.

Using this simulation framework, we measured the ability of NBS to recover the correct subtype assignments in comparison to a standard consensus clustering approach not based on network knowledge (Online Methods). NBS showed a striking improvement in performance, especially for large network modules, as these can be associated with any of numerous different mutations across the patient population (**Fig. 2b**). As module size decreased, the chance of observing the same mutated gene in patients of the same subtype increased, and the standard clustering algorithm performed increasingly well. We found that the high performance of NBS depended not only on network smoothing but also on the

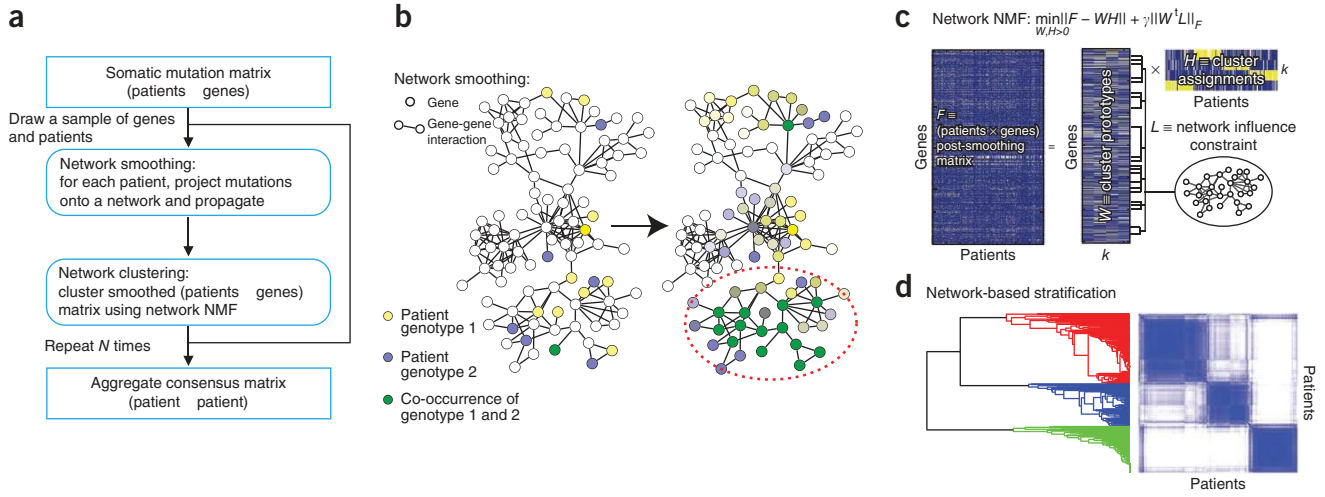

Figure 1 | Overview of network-based stratification (NBS). (**a**) Flowchart of the approach. (**b**) Example illustrating smoothing of patient somatic mutation profiles over a molecular interaction network. Mutated genes are shown in yellow (patient 1) and blue (patient 2) in the context of a gene interaction network. Following smoothing, the mutational activity of a gene is a continuous value reflected in the intensity of yellow or blue; genes with high scores in both patients appear in green (dashed oval). (**c**) Clustering mutation profiles using non-negative matrix factorization (NMF) regularized by a network. The input data matrix (F) is decomposed into the product of two matrices: one of subtype prototypes (W) and the other of assignments of each mutation profile to the prototypes (H). The decomposition attempts to minimize the objective function shown, which includes a network influence constraint L on the subtype prototypes. k, predefined number of subtypes. (**d**) The final tumor subtypes are obtained from the consensus (majority) assignments of each tumor after 1,000 applications of the procedures in **b** and **c** to samples of the original data set. A darker blue color in the matrix coincides with higher co-clustering for pairs of patients.

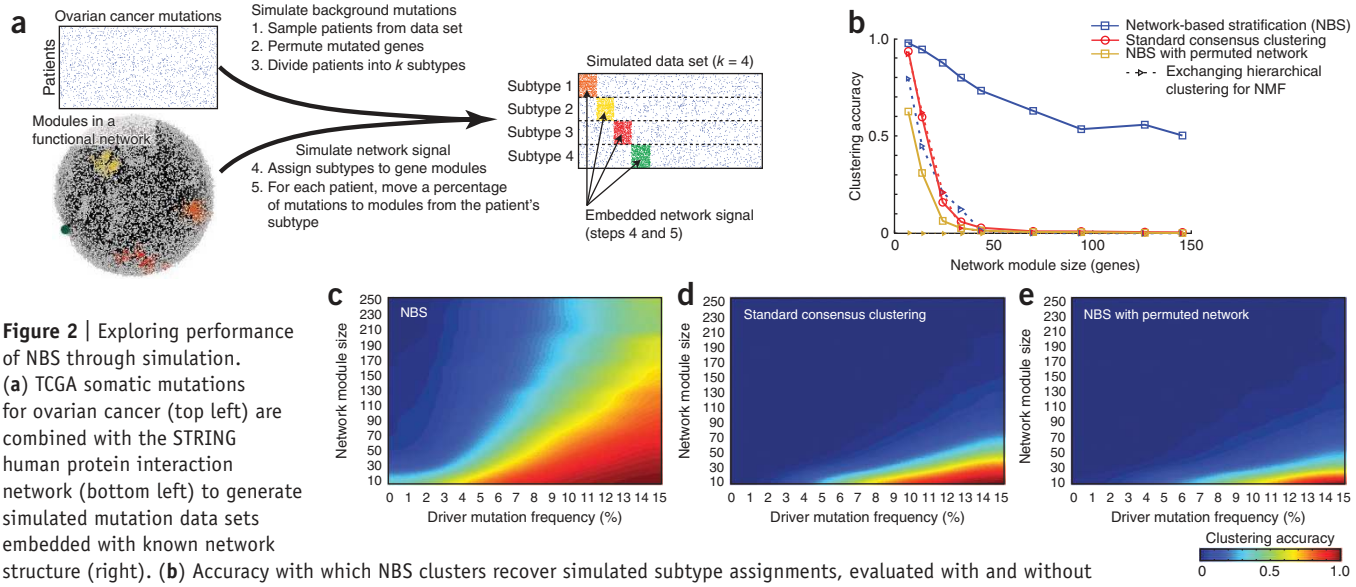

Figure 2 | Exploring performance of NBS through simulation. (**a**) TCGA somatic mutations for ovarian cancer (top left) are combined with the STRING human protein interaction network (bottom left) to generate simulated mutation data sets embedded with known network structure (right). (**b**) Accuracy with which NBS clusters recover simulated subtype assignments, evaluated with and without network smoothing and using non-negative matrix factorization (NMF) versus hierarchical clustering. Accuracy is calculated as the Adjusted Rand Index[57] of overlap between the clusters and correct subtype assignments, for which a score of 0 represents random overlap and 1 represents perfect overlap. Simulation was performed with a driver mutation frequency $f = 7.5\%$ with a single network module assigned to each subtype. (**c**) Accuracy landscape of NBS across varying driver mutation frequency and module size. (**d**) As in **c**, for a standard non-network–based clustering approach. (**e**) As in **c**, using a permuted network.

NMF clustering approach; substitution of NMF with an alternative method such as hierarchical clustering resulted in relatively poor performance (**Fig. 2b**).

Next we investigated how NBS performance was affected as a function of mutation frequency (**Fig. 2c**). Standard consensus clustering was sufficient for stratification at high mutation frequencies and for small modules, for which there is substantial overlap in mutations among patients of the same subtype (**Fig. 2d**); however, NBS was able to accurately recover the correct subtypes for a much larger range of both variables. Applying NBS on a permuted network resulted in poor performance (**Fig. 2e**), which is on par with that observed with standard consensus clustering. These results were qualitatively similar when we used multiple network modules per patient (2–6) and/or a different network (**Supplementary Fig. 2**).

Network-based stratification of tumor mutations

We next sought to apply NBS to stratify patients profiled by TCGA full-exome sequencing for uterine, ovarian and lung cancers (see Online Methods for further details). In each of the three cancers, we observed that NBS resulted in robust subtype structure, whereas standard consensus clustering was unable to stratify the patient cohort (**Fig. 3a** for uterine cancer; **Supplementary Figs. 3a** and **4a** for ovarian and lung cancers, respectively). Similar results were obtained when we used any of the three human networks (STRING, HumanNet and PathwayCommons).

To determine the biological importance of the identified subtypes, we investigated whether they were predictive of observed clinical data. In uterine cancer, NBS subtypes (**Supplementary Table 2**) were closely associated with the recorded subtype on a histological basis (**Fig. 3b**,**c** and **Supplementary Fig. 5**). Survival analysis was not possible owing to low mortality rates for this cohort. In ovarian cancer, the identified subtypes (**Supplementary Table 3**) were significant predictors of patient survival time (log-rank $P = 1.59 \times 10^{-5}$; **Fig. 3d**,**e** and **Supplementary Fig. 3b**,**c**). Patients with the most aggressive ovarian tumor NBS subtype had a mean survival of approximately 32 months, compared to more than 80 months for those with the least aggressive NBS subtype (**Supplementary Fig. 3d**,**e**). Moreover, the NBS subtypes were predictive of survival independently of clinical covariates including tumor stage, age, mutation rate and residual tumor presence after surgery (**Supplementary Fig. 6**; likelihood ratio test, $P = 3.75 \times 10^{-5}$) and were also predictive of time to relapse after treatment with platinum chemotherapy ('platinum-free interval') (**Supplementary Fig. 3f**), as measured using a Kaplan-Meier analysis of platinum-free survival[34]. Finally, in lung cancer the identified NBS subtypes (**Supplementary Table 4**) were also significant predictors of patient survival (log-rank $P = 1.95 \times 10^{-6}$, **Fig. 3f**,**g**; median survival of 12 months versus approximately 50 months for the best-surviving subtype, **Supplementary Fig. 4**), with predictive value beyond known clinical covariates such as tumor stage, grade, mutation frequency, age at diagnosis and smoking status (likelihood ratio test, $P = 3.3 \times 10^{-4}$). Stratification using a network in which the mapping between mutated genes and the network was permuted, which disrupted the relationship between mutations and network structure, resulted in degraded predictive performance (**Fig. 3b**,**d**,**f**).

We compared these results to subtypes derived from other data types in the TCGA, including copy-number variation (CNV), methylation, mRNA expression, microRNA expression and protein profiles. For ovarian cancer, all other data types had inferior ability to predict survival beyond what could be predicted from clinical covariates (**Fig. 4a**) and led to different subtype assignments than NBS (**Fig. 4b**). In lung cancer, both NBS subtypes and those based on RNA-seq had good predictive power (**Fig. 4c**) and had some overlap in terms of patient assignments (**Fig. 4d**), whereas other data types were not predictive of survival. In uterine cancer, subtypes derived from all data types were highly predictive of histology

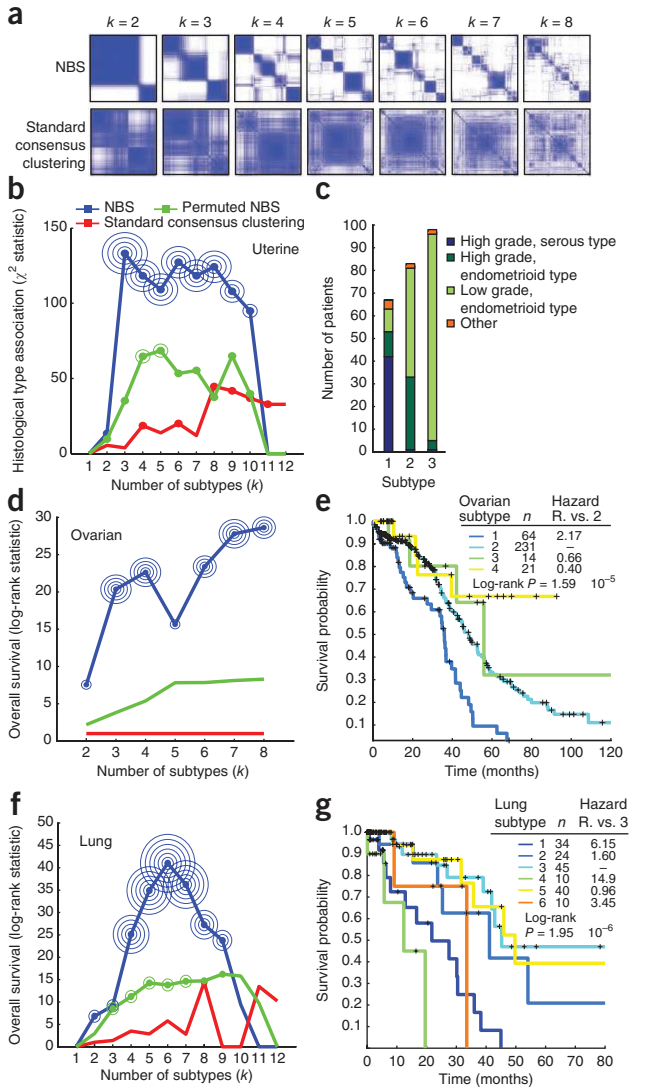

Figure 3 | NBS of somatic tumor mutations. (**a**) Co-clustering matrices for uterine cancer patients, comparing NBS (STRING) (top) to standard consensus clustering (bottom). (**b**,**c**) Association of NBS subtypes with histology (**b**) and composition of NBS subtypes in terms of histological type and tumor grade (**c**) for uterine cancer. (**d**,**e**) Association of NBS subtypes (HumanNet) with patient survival time (**d**) and Kaplan-Meier survival plots for NBS subtypes (**e**) for ovarian cancer. (**f**,**g**) Association of NBS subtypes (HumanNet) with patient survival time (**f**) and Kaplan-Meier survival plots for NBS subtypes (**g**) for lung cancer. (**b**,**d**,**f**) *P* value of significance of 10^{-k} is indicated by *k* concentric circles surrounding a data point (for example, three concentric circles indicate $P < 0.001$); in the case of uterine a significance of 10^{-5k} is indicated by *k* concentric circles (for example, one circle indicates $P < 10^{-5}$). Hazard R., hazard ratio, the ratio of fatalities between the two indicated subtypes over the studied time interval.

(**Fig. 4e**; CNVs had highest predictive power overall) and also had very high overlap with NBS subtype assignments (**Fig. 4f**).

Distinct network modules associate with each tumor subtype

We next sought to identify the regions of the network that are most responsible for discriminating the somatic mutation profiles of tumors of different subtypes. Focusing on ovarian cancer as a proof of principle, for each subtype we identified genes for which the network-smoothed mutation state differs significantly for patients of that subtype versus the others (false discovery rate <0.05; Online Methods). This set of genes was projected onto the HumanNet network and visualized using Cytoscape[35]. The network for subtype 1 (**Fig. 5**), which had the worst overall survival and shortest platinum-free interval, contained over 20 genes in the fibroblast growth factor (FGF) signaling pathway, which has previously been implicated as a driver of tumor progression and associated with resistance to platinum and anti-VEGF therapy[36]. The network for subtype 2 was enriched in DNA damage–response genes including *ATM*, *ATR*, *BRCA1*, *BRCA2*, *RAD51* and *CHEK2* (**Supplementary Fig. 7**). Collectively these highlighted pathways are characteristic of a functional deficit in response to DNA damage, which has been referred to as 'BRCAness'[7,37]. Consistent with this finding, this subtype also included the vast majority of patients with *BRCA1* and *BRCA2* germ-line mutations (15 of 20 and 5 of 6 patients in the cohort, respectively). The network for subtype 3 was enriched for genes in the NF-κB pathway (**Supplementary Fig. 8**), whereas subtype 4 was enriched for genes involved in cholesterol transport and fat and glycogen metabolism (**Supplementary Fig. 9**). A similar analysis in uterine and lung cancers produced other subnetworks with unique characteristics, including enrichments for DNA-damage response, WNT signaling and histone modification (**Supplementary Figs. 10–16**). Thus, the NBS approach not only can stratify patients into clinically informative subtypes but may help identify the molecular network regions commonly mutated in each subtype.

Translation to predictive signatures

For NBS to be applicable to new patients not in the TCGA, it is necessary to complement it with a procedure for assigning a patient to one of the existing NBS subtypes. For this purpose, we explored the nearest shrunken centroid approach[38], a standard method for sample classification that summarizes each subtype with a class 'centroid' and assigns new samples to the subtype with closest centroid. We found that this method was able to classify the network-smoothed mutation profile of an individual patient with over 95% accuracy (**Fig. 6a**; tenfold cross-validation).

However, mRNA expression data are presently much more widely available than are full genome or exome sequences: there are numerous existing cohorts of cancer patients that have been profiled in mRNA expression but not in somatic mutations[7,39–42]. We therefore sought to test whether, having used NBS to define subtypes within TCGA somatic mutation data, we could assign a new patient to these subtypes using an expression signature. To explore this idea, we used the mRNA expression profiles available for the TCGA ovarian tumor cohort to learn an expression signature for each subtype defined earlier by NBS, again using the nearest shrunken centroid approach[38]. We found that expression performed as an adequate surrogate for mutation profile, albeit at a reduced accuracy (**Fig. 6a**; >95% for mutations, ~60% for expression and ~30% at random). This expression signature was nonetheless able to recover stratification predictive of survival (**Fig. 6b**).

We examined the predictive value of this gene expression signature in two independent studies of serous ovarian tumors by Tothill *et al.*[40] and Bonome *et al.*[42] as well as in a meta-analysis including over 1,000 patients, which subsumes Tothill, Bonome and TCGA samples that included expression profiles but lacked somatic mutation profiles[41] (**Fig. 6c** and **Supplementary Fig. 17**) and incorporates an unknown number of nonserous ovarian cancer samples. Using the expression signature we had learned from NBS

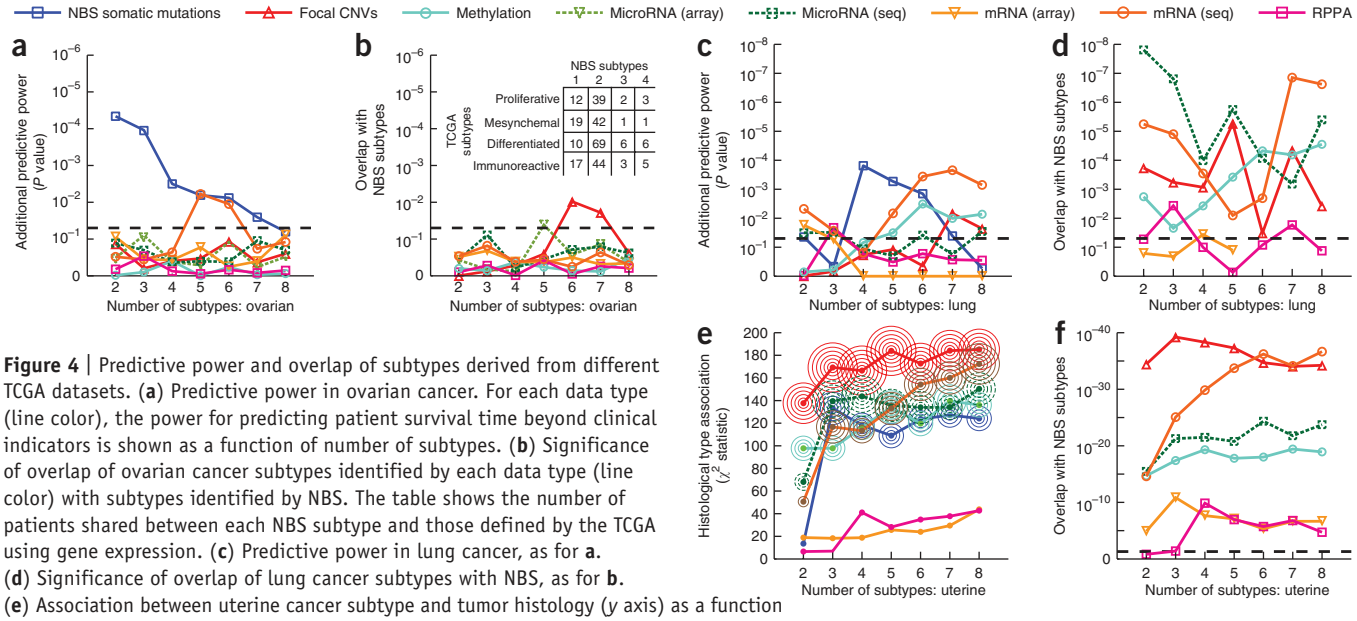

Figure 4 | Predictive power and overlap of subtypes derived from different TCGA datasets. (**a**) Predictive power in ovarian cancer. For each data type (line color), the power for predicting patient survival time beyond clinical indicators is shown as a function of number of subtypes. (**b**) Significance of overlap of ovarian cancer subtypes identified by each data type (line color) with subtypes identified by NBS. The table shows the number of patients shared between each NBS subtype and those defined by the TCGA using gene expression. (**c**) Predictive power in lung cancer, as for **a**. (**d**) Significance of overlap of lung cancer subtypes with NBS, as for **b**. (**e**) Association between uterine cancer subtype and tumor histology (*y* axis) as a function of the number of subtypes. *P* value of significance is indicated by concentric circles as in **Figure 3**. Colors are as in other panels, symbols have been omitted for clarity. (**f**) Significance of overlap of uterine cancer subtypes with NBS, as for **b**. Dashed horizontal lines indicate the $P = 0.05$ threshold of significance.

analysis of TCGA data, all patients could be assigned to one of the four NBS subtypes. In the Tothill data set, the subtype assignments were found to be significantly predictive of patient survival and platinum drug resistance (log-rank $P = 6.1 \times 10^{-3}$ and 1.65×10^{-6} respectively; **Fig. 6c** and **Supplementary Fig. 17**), following the same trends observed in the original TCGA cohort. In the Bonome and the meta-analysis data sets, the recovered subtypes were again significantly associated with patient survival (log-rank $P = 1.40 \times 10^{-3}$ and 1.22×10^{-4}, respectively; **Supplementary Fig. 17**). We note that the proportions of the recovered subtypes in each of the three independent expression cohorts appeared to be different (**Supplementary Table 2**), a phenomenon possibly due to different

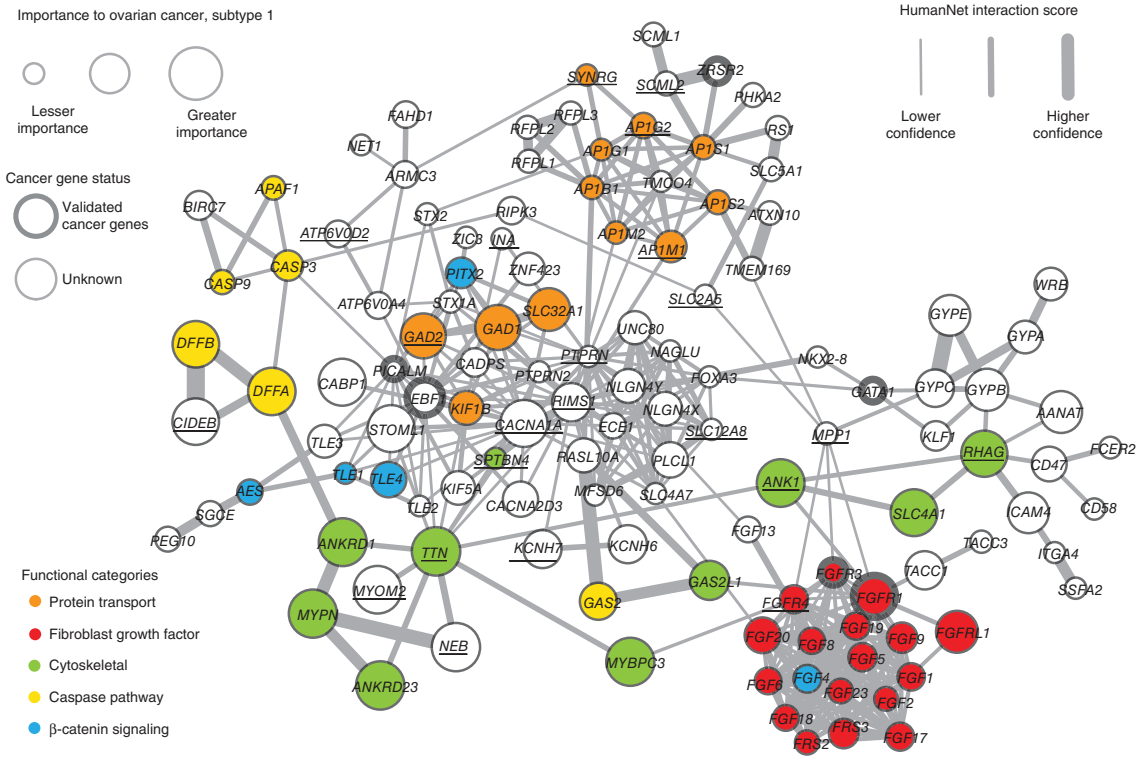

Figure 5 | Network view of genes with high network-smoothed mutation scores in HumanNet ovarian cancer subtype 1 (relative to scores of other subtypes). Subtype 1 had the lowest survival and highest platinum-resistance rates amongst the four recovered subtypes. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plug-in. Thickened node outlines indicate genes that are known cancer genes included in the COSMIC cancer-gene census. An underlined gene symbol in the network indicates that somatic mutations were found for that gene in the examined cohort.

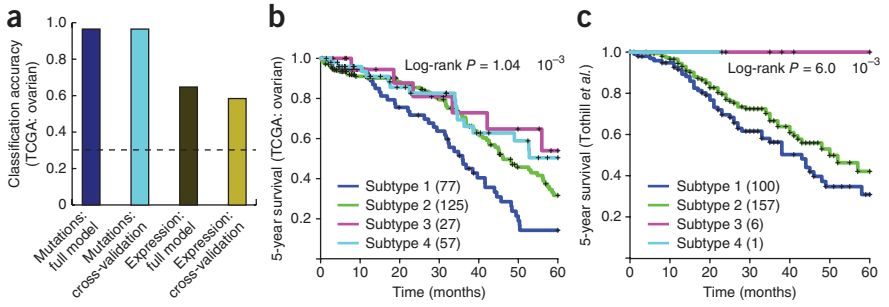

Figure 6 | From mutation-derived subtypes to expression signatures. (**a**) Classification accuracy (fraction of correctly classified patients) when using a supervised learning method trained to learn a signature on the basis of either somatic mutation profiles or gene expression, showing training error and cross-validation error. Dashed line shows the accuracy for a random predictor. (**b**) Kaplan-Meier survival plots for the TCGA ovarian cancer patients using a classifier trained on subtypes from NBS of mutation data in TCGA. (**c**) Results of the same classifier applied to serous ovarian cancer samples from an independent data set (Tothill *et al.*[40]).

criteria for inclusion in each study (for example: the TCGA ovarian cohort is primarily composed of high-grade, late-stage patients) or possibly differences due to population substructure. As a final control, we performed clustering of the Tothill expression profiles independent of NBS subtypes; this resulted in a different set of subtypes that associated with survival to a more limited extent ($P = 0.01$, **Supplementary Fig. 18**). These results show that tumor subtypes defined by NBS can be identified in independent data sets when gene expression is used as a surrogate biomarker.

Effects of different classes of mutation on stratification

We studied the impacts of different classes of somatic mutation on the NBS approach. We first tested the effect on NBS of disrupting synonymous mutations by reassigning them to new randomly chosen gene locations. For uterine and lung cancers (**Fig. 7a** and **Supplementary Fig. 19**, respectively), disruption of synonymous mutations had little effect on NBS performance. In sharp contrast, disruption of nonsynonymous mutations or of all mutations greatly affected stratification performance. Interestingly, in the ovarian cancer cohort (**Fig. 7b**), disruption of either synonymous or nonsynonymous mutations was detrimental to performance.

We also studied the effect of removing mutations judged to be nonfunctional in cancer by methods such as MutationAssessor[43], cancer-specific high-throughput annotation of somatic mutations (CHASM)[13] and the variant effect scoring tool (VEST)[44], which use features such as sequence conservation and protein structural information to assess the likely impact of mutations. Filtering mutations with these tools resulted in decreased association of NBS subtypes with patient survival in all three cancers (**Fig. 7c–e**, with the possible exception of VEST for ovarian tumors: **Fig. 7d**). Finally, we studied the effect of removing genes with long sequences or late cell-cycle replication times: both of these characteristics have been postulated to accrue high numbers of mutations that may be unrelated to tumor progression[45]. We found that removal of long genes substantially degraded the ability to identify ovarian and lung subtypes predictive of survival (**Fig. 7d,e**). However, removal of late-replicating genes had little effect and, in the case of the lung tumor cohort, actually increased predictive power (**Fig. 7e**).

DISCUSSION

Here we have reported the discovery that, through the use of prior knowledge captured in molecular networks, a set of tumor mutation profiles can be stratified into subtypes that are both biologically and clinically informative. These subtypes are distinct from those recovered through stratification of other types of data and are independent of other clinical markers known to be associated with survival. We can identify network modules characteristic of each subtype, which may provide new insight into the biological mechanisms driving tumor progression. To our knowledge, this is the first time that somatic mutation profiles have been used to stratify patients in an unsupervised fashion.

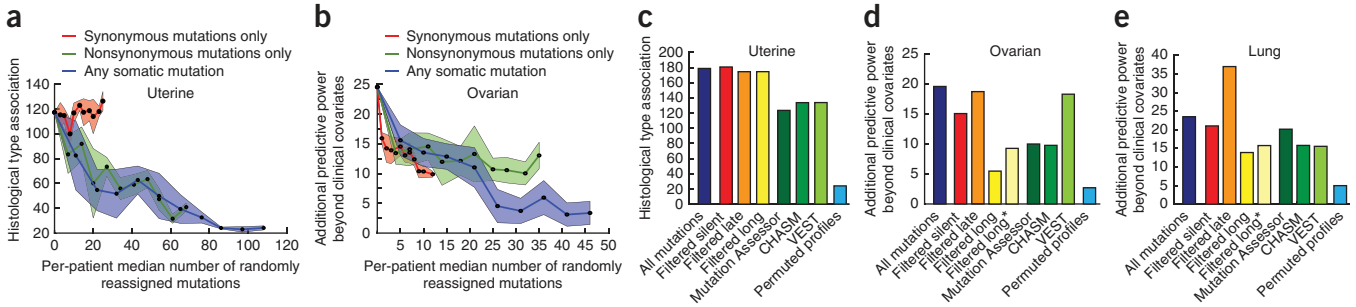

Figure 7 | Effects of different types of mutations on stratification. (**a**,**b**) Effects of permuting a progressively larger fraction of mutations per patient for different types of somatic mutation, for the uterine (**a**) and ovarian (**b**) tumor cohorts. Lines show the median performance, and colored regions represent the median absolute deviation. (**c–e**) Different types of filters were applied as a preprocessing step before NBS was run on the uterine (**c**), ovarian (**d**) and lung (**e**) cohorts. In blue is the full data set; in red we filter all synonymous mutations; in orange and yellow we filter the top 2% late-to-replicate and long genes, respectively (long*: top 2% long genes, with any COSMIC cancer gene census genes included in the analysis). In green are three types of filters based on predictors of the functional effect of mutation; in light blue is the performance we observed after permuting all mutations within each patient separately as a control. (**a–e**) For uterine cancer, we report the median χ^2 statistic; for ovarian and lung cancer, we report the median likelihood difference of a full model to a base model including just clinical covariates (age, grade, stage, mutation rate and residual tumor after surgery).

One might consider at least three potential reasons for the good performance of NBS. First, somatic mutations represent a digital signal in that a given gene can be considered either mutated or not, whereas most other data layers are analog signals representing measurements of continuous values. In general, digital systems have improved accuracy and reproducibility and are more robust to noise[46]. Second, somatic mutation profiles are differential measurements between tumor and normal tissue, whereas expression and other 'omics profiles are absolute measurements in each patient. The differential analysis filters out mutations or variants present in the patient's germ line, leaving only tumor-specific changes. In contrast, it has been difficult to identify a true 'baseline' gene expression state for a tissue, as these measurements are dynamic and highly context specific. Finally, the somatic mutation profile captures the causal genetic events underlying tumor progression, whereas mRNA or protein expression profiles are a functional readout of the current cell state and are influenced by external factors that may be unrelated to tumor biology.

The network modules we identified as characteristic for each tumor subtype provide new insights into the biology of cancer and raise many new questions. One particularly promising finding was the prominence of the *FGF* pathway in ovarian tumor subtype 1 (**Fig. 5**). This pathway has been implicated in tumor proliferation and angiogenesis, and many inhibitors for this pathway are in clinical development[47]. Specifically, it has been shown that increased expression of *FGF1* is associated with poor survival in ovarian cancer[48], and inhibition of *FGFR1* and *FGFR2* increases sensitivity to cisplatin in ovarian cancer cell lines[36]. An intriguing question for future work is whether subtype 1 patients are particularly responsive to therapy directed at network-identified targets, such as treatment with inhibitors of FGFR1.

Another interesting observation is that several network modules are enriched for long genes. For example, for ovarian tumor subtype 2, a total of 12 of 176 genes in the module are in the top 2% by length ($P = 2.3 \times 10^{-4}$). One prominent example is *TTN*, the longest known coding gene. Although prominent 'gold-standard' catalogs of cancer genes—such as the Catalogue of Somatic Mutations in Cancer (COSMIC) cancer gene census[49] and the list of Vogelstein *et al.*[50]—are also enriched for long genes (for example, 17 of 125 in the Vogelstein list, $P = 5.11 \times 10^{-10}$), there remains some controversy about the roles these genes may play in cancer. On the one hand, it is possible that long genes are highly mutated not because they are drivers of cancer but simply owing to chance because they are a bigger 'target' to hit. On the other hand, there is no definitive evidence that mutations in long genes are not functional or do not contribute to tumor progression. Our analysis provides some evidence that these long genes should not be ignored. In the molecular network, long mutated genes were highly interconnected to other functionally related genes of all lengths, which are also found to be mutated in patients of that subtype. For example, the network region for ovarian tumor subtype 1 (**Fig. 5**) showed *TTN* interconnected to genes such as *NEB*, *ANK1* and *MYOM2*, all of which are also mutated in patients of this subtype. These genes encode components of the cytoskeleton thought to have both structural and signaling roles[51]. Although *TTN* is a long gene and thus might accrue mutations by chance, it is striking that other members of the same protein interaction neighborhood are also found to be mutated in tumors of the same subtype. Using permutation analysis, we estimated that the chance of *TTN* having an immediate network neighborhood with this same number of mutations is roughly $P < 0.0001$. Thus, one possibility is that the TTN and other cytoskeletal components are required for platinum-induced, P53-independent apoptosis, and that mutation in either structural or signaling proteins in this pathway leads to platinum resistance. In support of this theory is prior work demonstrating that cell shape is associated with chemotherapy response in ovarian cancer[52].

Another interesting observation is that synonymous mutations, though dispensable for stratification of uterine and lung tumors, appear to have some predictive power in stratification of ovarian tumors. In support of this finding, a number of high-profile studies have suggested that synonymous mutations may indeed play a causal role in cancer progression[53–56]. Further study is needed to understand whether ovarian cancer is indeed the outlier in this respect and whether and how synonymous mutations truly function in this disease.

Finally, we see many opportunities to improve upon the basic concept of NBS in future work. First, integrating multiple layers of information beyond somatic mutations (for example: CNVs, epigenome, transcriptome, etc.) into a composite stratification method might further expand our ability to identify subtypes with clinically relevant differences. Second, although we have shown the utility of three sources of gene-gene interactions, there are other types of networks worth exploring, such as those involved in signaling, metabolism or transcription. Although this study focused on uterine, ovarian and lung cancers, the NBS method is broadly applicable to any cohort of cancer patients for which somatic mutations are known. Finally, analyzing NBS subtypes across all cancers simultaneously (i.e., a pan-cancer analysis) will offer the intriguing opportunity to explore whether the genes and networks underlying the progression of a tumor are more informative of clinical outcome than its tissue of origin.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank all members of the Ideker lab, specifically J. Dutkowski, R. Srivas, G. Bean, M. Yu and M. Choueiri, for many fruitful discussions during various stages of this project. We also thank G. Hofree for her input, patience and support. J.P.S. is supported in part by grants from the Marsha Rivkin Center for Ovarian Cancer Research and the Conquer Cancer Foundation of the American Society of Clinical Oncology. This work was supported by US National Institutes of Health grants P41 GM103504 and P50 GM085764.

AUTHOR CONTRIBUTIONS

M.H. and T.I. conceived of and designed the approach. M.H. performed the data analysis, implemented the method and performed all computational experiments. J.P.S. assisted in characterizing clinical aspects of the cohorts and interpreting the biological implication of the resulting subtypes. H.C. assisted in analyzing the contributions of different mutation types and the functional scoring of mutations. A.G. compared resulting subtypes to other data layers. M.H., J.P.S. and T.I. wrote the manuscript. All authors approved the final version of this manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. The International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–996 (2010).
2. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
3. The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
4. Brunham, L.R. & Hayden, M.R. Whole-genome sequencing: the new standard of care? *Science* **336**, 1112–1113 (2012).
5. Chin, L., Andersen, J.N. & Futreal, P.A. Cancer genomics: from discovery science to personalized medicine. *Nat. Med.* **17**, 297–303 (2011).
6. Konstantinopoulos, P.A., Spentzos, D. & Cannistra, S.A. Gene-expression profiling in epithelial ovarian cancer. *Nat. Clin. Pract. Oncol.* **5**, 577–587 (2008).
7. 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. *J. Clin. Oncol.* **28**, 3555–3561 (2010).
8. Reis-Filho, J.S. & Pusztai, L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet* **378**, 1812–1823 (2011).
9. Esteva, F.J. *et al.* Prognostic role of a multigene reverse transcriptase-PCR assay in patients with node-negative breast cancer not receiving adjuvant systemic therapy. *Clin. Cancer Res.* **11**, 3315–3319 (2005).
10. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
11. Raspe, E., Decraene, C. & Berx, G. Gene expression profiling to dissect the complexity of cancer biology: pitfalls and promise. *Semin. Cancer Biol.* **22**, 250–260 (2012).
12. Mardis, E.R. Genome sequencing and cancer. *Curr. Opin. Genet. Dev.* **22**, 245–250 (2012).
13. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.* **69**, 6660–6667 (2009).
14. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
15. Wang, K. *et al.* Exome sequencing identifies frequent mutation of *ARID1A* in molecular subtypes of gastric cancer. *Nat. Genet.* **43**, 1219–1223 (2011).
16. Dulak, A.M. *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat. Genet.* **45**, 478–486 (2013).
17. Allegra, C.J. *et al.* American Society of Clinical Oncology provisional clinical opinion: testing for *KRAS* gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy. *J. Clin. Oncol.* **27**, 2091–2096 (2009).
18. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
19. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
20. Kreeger, P.K. & Lauffenburger, D.A. Cancer systems biology: a network modeling perspective. *Carcinogenesis* **31**, 2–8 (2010).
21. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
22. Waddington, C.H. Canalization of development and the inheritance of acquired characters. *Nature* **150**, 563–565 (1942).
23. Vandin, F., Upfal, E. & Raphael, B.J. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* **18**, 507–522 (2011).
24. Vaske, C.J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (2010).
25. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22**, 398–406 (2012).
26. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 140 (2007).
27. Dutkowski, J. & Ideker, T. Protein networks as logic functions in development and cancer. *PLoS Comput. Biol.* **7**, e1002180 (2011).
28. Lee, I., Blom, U.M., Wang, P.I., Shim, J.E. & Marcotte, E.M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**, 1109–1121 (2011).
29. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D561–D568 (2011).
30. Cerami, E.G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685–D690 (2011).
31. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* **6**, e1000641 (2010).
32. Lee, D.D. & Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
33. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52**, 91–118 (2003).
34. Yang, D. *et al.* Association of *BRCA1* and *BRCA2* mutations with survival, chemotherapy sensitivity, and gene mutator phenotype in patients with ovarian cancer. *J. Am. Med. Assoc.* **306**, 1557–1565 (2011).
35. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431–432 (2011).
36. Cole, C. *et al.* Inhibition of FGFR2 and FGFR1 increases cisplatin sensitivity in ovarian cancer. *Cancer Biol. Ther.* **10**, 495–504 (2010).
37. Wysham, W.Z. *et al.* BRCAness profile of sporadic ovarian cancer predicts disease recurrence. *PLoS ONE* **7**, e30042 (2012).
38. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 6567–6572 (2002).
39. Le Page, C. *et al.* Gene expression profiling of primary cultures of ovarian epithelial cells identifies novel molecular classifiers of ovarian cancer. *Br. J. Cancer* **94**, 436–445 (2006).
40. Tothill, R.W. *et al.* Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.* **14**, 5198–5208 (2008).
41. Győrffy, B., Lánczky, A. & Szállási, Z. Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. *Endocr. Relat. Cancer* **19**, 197–208 (2012).
42. Bonome, T. *et al.* A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer Res.* **68**, 5478–5486 (2008).
43. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
44. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14** (suppl. 3), s3 (2013).
45. Stamatoyannopoulos, J.A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).
46. Rabiner, L.R. & Gold, B. *Theory and Application of Digital Signal Processing* (Prentice Hall, 1975).
47. Turner, N. & Grose, R. Fibroblast growth factor signalling: from development to cancer. *Nat. Rev. Cancer* **10**, 116–129 (2010).
48. Birrer, M.J. *et al.* Whole genome oligonucleotide-based array comparative genomic hybridization analysis identified fibroblast growth factor 1 as a prognostic marker for advanced-stage serous ovarian adenocarcinomas. *J. Clin. Oncol.* **25**, 2281–2287 (2007).
49. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
50. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
51. Simon, D.N. & Wilson, K.L. The nucleoskeleton as a genome-associated dynamic 'network of networks'. *Nat. Rev. Mol. Cell Biol.* **12**, 695–708 (2011).
52. Liu, Y. *et al.* Integrated analysis of gene expression and tumor nuclear image profiles associated with chemotherapy response in serous ovarian carcinoma. *PLoS ONE* **7**, e36383 (2012).
53. Strauss, B.S. Role in tumorigenesis of silent mutations in the *TP53* gene. *Mutat. Res.* **457**, 93–104 (2000).
54. Kimchi-Sarfaty, C. *et al.* A "silent" polymorphism in the *MDR1* gene changes substrate specificity. *Science* **315**, 525–528 (2007).
55. Sauna, Z.E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* **12**, 683–691 (2011).
56. Salzman, D.W. & Weidhaas, J.B. miRNAs in the spotlight: making 'silent' mutations speak up. *Nat. Med.* **17**, 934–935 (2011).
57. Rand, W.M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).

ONLINE METHODS

Expanded overview of network-based stratification. The technique of network-based stratification (NBS) combines genome-scale somatic mutation profiles with a gene interaction network to produce a robust subdivision of patients into subtypes (**Fig. 1a**). Briefly, somatic mutations for each patient are represented as a profile of binary (1, 0) states on genes, in which a '1' indicates a gene for which mutation has occurred in the tumor relative to germ line (i.e., a single-nucleotide base change or the insertion or deletion of bases). For each patient independently we project the mutation profiles onto a human gene interaction network obtained from public databases[28–30]. Next, the technique of network propagation[31] is applied to spread the influence of each mutation profile over its network neighborhood (**Fig. 1b**). The result is a 'network-smoothed' profile in which the state of each gene is no longer binary but reflects its network proximity to the mutated genes in that patient along a continuous range [0, 1]. Following this 'network smoothing', patient profiles are clustered into a predefined number of subtypes (k = 2, 3, … 12) using the unsupervised technique of non-negative matrix factorization[32] (NMF; **Fig. 1c**). For NBS we use a variant of NMF that encourages the selection of gene sets supporting each subtype according to high network connectivity (NetNMF)[58]. Finally, to promote robust cluster assignments, we use the technique of consensus clustering[33], in which the above procedure is repeated for 1,000 different subsamples in which subsets of 80% of patients and genes are drawn randomly without replacement from the entire data set. The results of all 1,000 runs are aggregated into a (patient × patient) co-occurrence matrix, which summarizes the frequency with which each pair of patients has cosegregated into the same cluster. This co-occurrence matrix is then clustered a second time to recover a final stratification of the patients into clusters/subtypes (**Fig. 1d**). Our implementation of the NBS method is available for download as a Matlab package from http://idekerlab.ucsd.edu/software/NBS/ or as **Supplementary Software**. The former should be used for obtaining the most up-to-date versions.

Processing of patient mutation profiles. High-grade serous ovarian cancer, uterine endometrial carcinoma and lung adenocarcinoma somatic mutation data were downloaded from the TCGA data portal on 8 August 2012, 1 January 2013 and 1 January 2013, respectively. Only mutation data generated using the Illumina GAIIx platform were retained for subsequent analysis, and patients with fewer than 10 mutations were discarded. This left 356 patients with mutations in 9,850 genes for the TCGA ovarian cohort, 248 patients with mutations in 17,968 genes for the TCGA uterine endometrial cohort and 381 patients with mutations in 15,967 genes in the TCGA lung adenocarcinoma cohort. Patient mutation profiles were constructed as binary vectors such that a bit is set if the gene corresponding to that position in the vector harbors a mutation in that patient. Additional details on processing and organization of the data are available in a previous TCGA publication[2].

Sources of molecular network data. Patient mutation profiles were mapped onto gene interaction networks from three sources: STRING v.9 (ref. 29), HumanNet v.1 (ref. 28) and PathwayCommons[30] (**Supplementary Table 1**). STRING integrates protein-protein interactions from literature curation, computationally predicted interactions, and interactions transferred from model organisms based on orthology. HumanNet uses a naïve Bayes approach to weight different types of evidence together into a single interaction score focusing on data collected in humans, yeast, worms and flies. PathwayCommons aggregates interactions from several pathway and interaction databases, focused primarily on physical protein-protein interactions (PPIs) and functional relationships between genes in canonical regulatory, signaling and metabolic pathways (including hallmark pathways of cancer). **Supplementary Table 1** summarizes the number of genes and interactions used in our analysis from each of these three networks.

All network sources comprise a combination of interaction types, including direct protein-protein interactions between a pair of gene products and indirect genetic interactions representing regulatory relationships between pairs of genes (for example, coexpression or TF activation). The PathwayCommons network was filtered to remove any nonhuman genes and interactions, and all remaining interactions were used for subsequent analysis. Only the most confident 10% of interactions for both the STRING and HumanNet networks were used for this work, ordered according to the quantitative interaction score provided as part of both networks. This threshold was chosen using an independent ROC analysis with respect to a set of Gene Ontology–derived gold standards (data not shown). After filtering of edges, all networks were used as unweighted, undirected networks.

Network smoothing. After mapping a patient mutation profile onto a molecular network, network propagation[31] is applied to 'smooth' the mutation signal across the network. Network propagation uses a process that simulates a random walk on a network (with restarts) according to the function

$$F_{t+1} = \alpha F_t A + (1-\alpha)F_0$$

F_0 is a patient-by-gene matrix, and A is a degree-normalized adjacency matrix of the gene interaction network, created by multiplying the adjacency matrix by a diagonal matrix with the inverse of its row (or column) sums on the diagonal. α is a tuning parameter governing the distance that a mutation signal is allowed to diffuse through the network during propagation. The optimal value of α is network dependent (0.7, 0.5 and 0.7, for HumanNet, PathwayCommons and STRING, respectively), but the specific value seems to have only a minor effect on the results of NBS over a sizable range (for example, 0.5–0.8). The propagation function is run iteratively with t = [0, 1, 2, …] until F_{t+1} converges (the matrix norm of $F_{t+1} - F_t < 1 \times 10^{-6}$). Following propagation, the rows of the resultant matrix F_t are quantile normalized to ensure that the smoothed mutation profile for each patient follows the same distribution.

Network-regularized NMF. Network-regularized NMF is an extension that constrains NMF to respect the structure of an underlying gene interaction network. This is accomplished by minimizing the following objective function using an iterative method[32,58,59]:

$$\min_{W,H>0} ||F - WH||^2 + \text{trace}(W^t K W)$$

doi:10.1038/nmeth.2651

W and H form a decomposition of the patient × gene matrix F (resulting from network smoothing as described above) such that W is a collection of basis vectors, or 'metagenes', and H is the basis vector loadings. The trace(W^tKW) function constrains the basis vectors in W to respect local network neighborhoods. The term K is and adjacency matrix of a nearest neighbors network derived from the graph Laplacian of an influence distance matrix[23] that is derived from the original network. The degree to which local network topology versus global network topology constrains W is determined by the number of nearest neighbors. We experimented with neighbor counts ranging from 5 to 50 to include in the nearest network, and we observed only small changes in outcome (data not shown). For the work presented in this manuscript, the 11 most influential neighbors of each gene in the network as determined by network influence distance were used.

Consensus clustering. Clustering was performed with a standard consensus clustering framework, discussed in detail by Monti *et al.*[33] and used in previous TCGA publications[2,18,60]. Briefly, we used network-regularized NMF (see above) to derive a stratification of the input cohort. In order to ensure robust clustering, network-regularized NMF was performed 1,000 times on subsamples of the data set. In each subsample, we sampled 80% of the patients and 80% of the mutated genes at random without replacement. The set of clustering outcomes for the 1,000 samples was then transformed into a co-clustering matrix. This matrix records the frequency with which each patient pair was observed to have membership in the same subtype over all clustering iterations in which both patients of the pair were sampled. The result is a similarity matrix of patients, which we then used to stratify the patients by applying either average linkage hierarchical clustering or a second symmetric NMF step. Patients showing poor cluster association to a single subtype were excluded from further analysis.

Simulation of somatic mutation cohorts. We used simulations to determine the ability of NBS to recover subtypes from somatic mutation profiles. In order to quantify the performance of NBS, we needed a cohort with specified subtypes as a 'ground truth' reference and to be able to control the properties of the simulated signal determining the different subtypes. We simulated a somatic mutation cohort as follows. Patient mutation profiles were sampled with replacement from the TCGA ovarian data set. For each patient, the mutation profile was permuted, whereas the per-patient mutation frequency was kept invariant; this resulted in a background mutation matrix with no subtype signal. For simulation of an underlying network structure for NBS to detect, a network-based signal was added to the patient-by-mutation matrix as follows. First, we established a set of network communities (i.e., connected components enriched for edges shared within community members) in the input network (STRING, HumanNet or PathwayCommons) using the network community detection algorithm QCut[61]. Next, we divided the patient cohort randomly into four equal-sized subtypes (four was selected as reasonable owing to the four expression-based subtypes that have been identified for glioblastoma, ovarian and breast cancers[2,18,60,62]). Each subtype was assigned a small number (for example, 1–6) of network modules that together had a combined size s ranging from 10 to 250 genes. These network modules represent 'driver' subnetworks characterizing the subtype. For each patient, we reassigned a fraction of the patient's mutations f to genes covered by the driver modules for that patient's subtype. This procedure resulted in a patient × gene mutation matrix with underlying network structure while maintaining the per-patient mutation frequency.

A plausible range for the number of driver mutation in a tumor was recently proposed to be between 2 and 8 driver mutations[50]. We note that in our simulation framework, a 4% mutation rate corresponds to between 1 and 9 mutations with a median of 3, which is on par with the aforementioned estimate. In order to estimate the appropriate size of cancer pathways (s), we examined the known cancer pathways in the NCI-Nature pathway interaction database[63]. We observe that pathways in the database are of varying sizes, 2–139 genes, with a median size of 34, and over 23% of pathways include over 50 genes.

Identifying differentially mutated subnetworks. After applying NBS, we identified genes that were enriched for mutation in each of the subtypes relative to the whole cohort. To do this we applied the significance analysis of microarrays (SAM) method[64] on the network-smoothed mutation profiles. This is a nonparametric method developed for discovering differentially expressed genes in microarray experiments. We used a rank-based Wilcoxon-type statistic and compared each subtype against the remaining cohort. Significance was assessed using the SAM permutation scheme with 1,000 permutations. The resulting set of genes for each subtype was overlaid on the network used for network smoothing.

Survival analysis. Survival analysis was performed using the R "survival" package. We fit a Cox-proportional hazards model[65] to determine the relationship between the NBS-assigned subtypes and patient survival. A likelihood-ratio test and associated P value is calculated by comparing the full model, which includes subtypes and clinical covariates, against a baseline model that includes covariates only. Clinical covariates available in TCGA and included in the model were age, grade, stage, residual surgical resection and mutation rate, as well as cigarette smoking status for the lung cancer cohort.

Comparing predictive power and overlap with TCGA subtypes. Added predictive power is estimated using a likelihood-ratio test comparing the Cox proportional hazards model given subtypes and clinical covariates (age, stage, grade, mutation frequency and residual tumor presence after surgery) compared to a covariate-only model. Significance of overlap is assessed using a Pearson's χ^2 test of independence between NBS subtypes with a specific network and number of subtypes (ovarian, HumanNet, four subtypes; lung, HumanNet, six subtypes; uterine, STRING, three subtypes) and the different data types with varying number of subtypes reported in the TCGA and subtyped using consensus-clustering NMF. TCGA subtypes were downloaded from the Firehose run from 25 May 2012 (http://gdac.broadinstitute.org/runs/analyses__2012_05_25/reports/cancer/OV/).

Shrunken-centroid prediction on expression profiles. We used shrunken centroids to derive an expression signature equivalent to the somatic mutation-based NBS subtypes. Expression data

were provided by Győrffy *et al.*[41], who aggregated several expression data sets as part of a meta-analysis of ovarian cancer. In this analysis, all data were regularized using quantile and MAS5 normalization. We performed this analysis on the Tothill *et al.*[40] (ovarian serous samples only), Bonome *et al.*[42] and TCGA data sets, as well as across the full meta-analysis cohort. We used the "pamr" R package with default parameters to train a shrunken-centroid model[38] on mRNA expression levels for all genes in the TCGA ovarian data set with subtype assignment as the class label. The trained model was next used to predict subtype labels on the held-out Tothill *et al.* and Bonome *et al.* data or the full meta-analysis expression cohort (excluding any TCGA samples included in the training set).

We include a table of the class centroids for each of the three TCGA somatic mutation cohorts and the four expression cohorts of ovarian cancer included in this study (**Supplementary Table 5**).

Missense-mutation scoring. Missense mutations were scored using three methods: CHASM[13], VEST[44] and MutationAssessor[43]. CHASM and VEST use supervised machine learning to score mutations. The CHASM training set is composed of a positive class of driver mutations from the COSMIC database and a negative class of synthetic passenger mutations simulated according to the mutation spectrum observed in the tumor type under study. The VEST training set comprises a positive class of disease mutations from the Human Gene Mutation Database[66] and a negative class of variants detected in the ESP6500 (http://evs.gs.washington.edu/EVS/) cohort with an allele frequency of >1%. MutationAssessor uses patterns of conservation from protein alignments of large numbers of homologous sequences to assess the functional impact of missense mutations. CHASM and VEST scores were obtained from the CRAVAT webserver[44] (http://www.cravat.us/). MutationAssessor precomputed mutation scores were downloaded from http://mutationassessor.org/. After using each method to score all mutations across all patients, we picked a permissive threshold for retaining mutations to use for NBS (retaining the top 75% of mutations as scored by CHASM and VEST and using MutationAssessor with the "low threshold" setting).

Replication timing. RepliSeq[67] data for the GM12878 cell line were downloaded from the ENCODE project website (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/, downloaded March 2013). Summed normalized tag densities were used as a proxy for replication time (higher counts indicating that a transcript was replicated earlier in the cell cycle). Normalized tag densities for RefSeq protein coding regions were retrieved using bigWigAverageOverBed[68] with RefSeq gene sequence features in .gff3 format downloaded from http://www.yandell-lab.org/software/VAAST/data/hg19/Features/refGene_hg19.gff3. Tag densities were averaged for each transcript, and the longest transcript was selected to represent each gene.

58. Cai, D., He, X., Wu, X. & Han, J. Non-negative matrix factorization on manifold. in *8th IEEE Int. Conf. Data Mining* 63–72 (IEEE, 2008).
59. Brunet, J.P., Tamayo, P., Golub, T.R. & Mesirov, J.P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA* **101**, 4164–4169 (2004).
60. Verhaak, R.G. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).
61. Ruan, J. & Zhang, W. Identifying network communities with a high resolution. *Phys. Rev. E* **77**, 016104 (2008).
62. Verhaak, R.G. *et al.* Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J. Clin. Invest.* **123**, 517–525 (2013).
63. Schaefer, C.F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674–D679 (2009).
64. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
65. Andersen, P.K. & Gill, R.D. Cox's regression model for counting processes: a large sample study. *Ann. Stat.* **10**, 1100–1120 (1982).
66. Stenson, P.D. *et al.* The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr. Protoc. Bioinformatics* **39**, 1.13 (2012).
67. Hansen, R.S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. USA* **107**, 139–144 (2010).
68. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

doi:10.1038/nmeth.2651