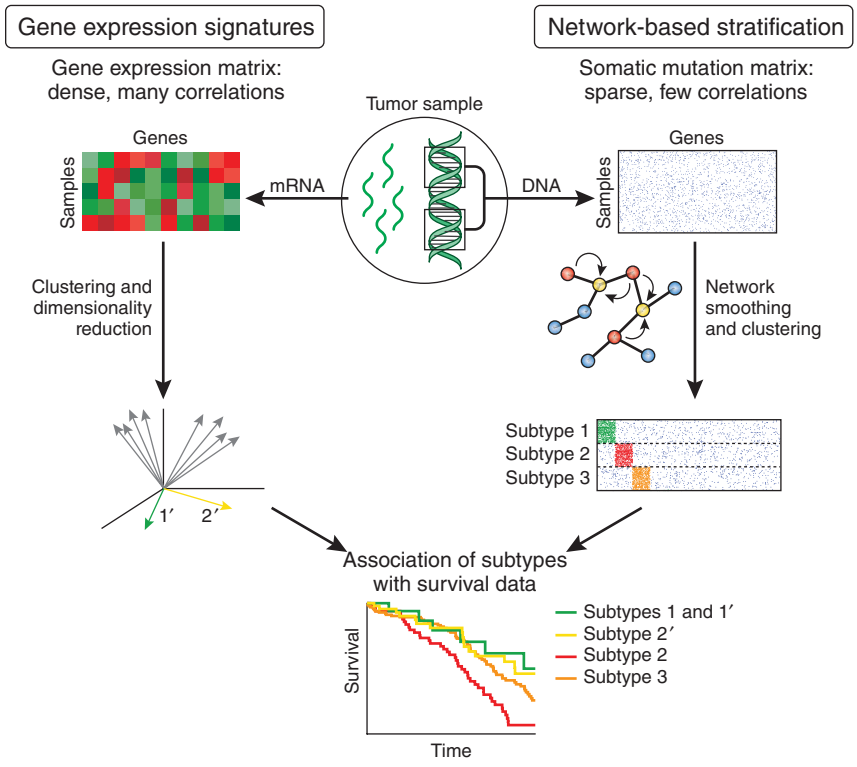
Making connections: using networks to stratify human tumors

Benjamin J Raphael

Network-based stratification (NBS) enables the subtyping of tumors on the basis of their mutational profiles, providing new avenues for cancer research and precision oncology.

Identifying molecular markers that stratify tumor samples into meaningful subtypes is an important goal in cancer genomics. Ideally, these subtypes correlate with clinical features, such as the aggressiveness of a tumor or response to drugs, and thus can be used to guide treatment. Early successes in defining such subtypes include the identification of translocations in leukemias, *ERRB2* (*HER2*) amplification in a subset of breast cancers, and others[1]. Since the introduction in the late 1990s of microarray techniques, there has been an explosion of studies (reviewed in ref. 2) to define subtypes according to gene expression signatures (**Fig. 1**). This work has led to some notable successes; but in many cancers, signatures or clinical correlations identified in one study were not reproduced in other studies.



Katie Vicari

Figure 1 | Network-based stratification and gene expression signatures are different approaches to stratify human tumors.

Benjamin J. Raphael is in the Department of Computer Science and Center for Computational Molecular Biology Brown University, Providence, Rhode Island, USA.
e-mail: braphael@brown.edu

In this issue, Hofree *et al.*[3] introduce a novel approach to stratify patients on the basis of the somatic mutations present in their tumors. Cancer is a disease driven by such somatic mutations, which accumulate in the genome during the lifetime of the individual. Recent advances in high-throughput DNA sequencing technologies now enable whole-genome or whole-exome measurement of somatic mutations. In particular, The Cancer Genome Atlas (TCGA) is using whole-exome sequencing to measure somatic mutations in protein-coding regions of genomes from ~500 samples from each of ~25 cancer types. Similar projects are underway by other groups, including dozens of national consortia under the umbrella of the International Cancer Genome Consortium.

The initial results from these large-scale sequencing studies demonstrated a major impediment to the use of somatic mutations for patient stratification, namely, cancers exhibit extensive mutational heterogeneity, with mutated genes varying widely across individuals. Moreover, an individual cancer sample may have somatic mutations in only a few to a few dozen of the ~21,000 human genes. In other words, if one builds a somatic mutation profile for a sample, where each gene is assigned a 1 or a 0 if the gene is mutated or not mutated, respectively, then the resulting profiles will be sparse, or nearly all 0s (**Fig. 1**). Consequently, comparison or clustering of such mutation profiles will not yield additional information beyond that revealed by direct examination of the handful of commonly mutated genes.

Although mutation profiles constructed from mutations at the gene level are sparse, it is widely reported that cancer-causing, or 'driver', mutations affect genes in a smaller number of signaling and regulatory pathways[4]. The innovation in Hofree *et al.*[3] is to use this pathway information, as represented in a protein-protein interaction network, to create a smoothed mutation profile for each sample. The smoothing process is analogous to heat diffusion: in brief, a mutation in a gene is a source of heat that diffuses to neighboring genes along the edges of the network. A similar diffusion was introduced in the HotNet algorithm[5] to identify

significantly mutated subnetworks in several TCGA publications. Hofree *et al.*[3] cleverly take an orthogonal approach with their NBS algorithm. They use the smoothed mutation profiles as the input to a clustering procedure, grouping the samples into a small number of subtypes according to the similarities between the smoothed mutation profiles.

Hofree *et al.*[3] apply NBS to somatic mutation data from TCGA studies of ovarian carcinoma, endometrial carcinoma and lung adenocarcinoma. On the ovarian and lung cancer data sets, NBS computes subtypes that discriminate the survival time of patients better than can subtypes derived from gene expression data. On the endometrial data set, NBS subtypes are closely associated with histological subtypes. Interestingly, although NBS significantly outperforms microarray-based gene expression for patient stratification, its gain over mRNA-Seq is smaller on the lung and endometrial data sets, suggesting an overall advantage for sequencing data (DNA or mRNA) over microarray data.

A second goal of tumor stratification is to examine whether the molecular markers in a signature are related to the disease mechanism. For gene expression signatures in cancer, the results of such analyses have been modest: many published signatures contain few genes whose aberrant expression plays a functional role in the pathogenesis of cancer. In some cases, the experiments necessary to demonstrate such a functional role have not yet been performed. However, the lack of a functional role for the genes in a signature is not surprising. Because gene expression data are high dimensional, it is typically necessary to use dimensionality-reduction techniques to find a smaller subset of genes that stratify tumors (**Fig. 1**). As the expression of some genes is highly correlated with the expression of others, there may be many possible selections of genes for the signature, with each selection nearly equal in its ability to discriminate samples[6,7]. This implies that functional interpretation of genes in an expression signature should be treated with caution.

Given that driver mutations are by definition directly responsible for cancer, one might anticipate that mutation profiles, or network-smoothed mutation profiles, would provide more functional insights than would gene expression signatures. Hofree *et al.*[3] determined the genes and associated subnetworks that distinguish individual tumor subtypes. On the ovarian cancer data set, the subnetwork for one subtype is enriched for DNA damage–response genes including *ATM*, *BRCA1* and *BRCA2*, all well-known cancer genes. A second subtype subnetwork contains multiple fibroblast growth factor (*FGF*) signaling genes, which were previously associated with resistance to therapy[8]. These results demonstrate that the advantages of NBS extend beyond patient stratification to include the identification of driver genes and networks in these subtypes.

Not all genes in the NBS subtype networks are well-known cancer genes: on the contrary, some are proposed to be genes containing an unusually high number of random, 'passenger' mutations. Prominent among these is titin, the largest protein (36,800 amino acids) in the human genome. Because of its length, titin will harbor passenger mutations in more samples than will shorter genes[9,10]. Additional factors such as replication timing and expression level also elevate the background mutation rate of genes, including that of titin[10]. NBS treats all mutations in all genes equally. However, unlike driver mutations, passenger mutations are not expected to cluster in subnetworks of a protein-protein interaction network. Thus, one might conjecture that the performance of NBS would improve if presumed passenger mutations were removed. Surprisingly, Hofree *et al.*[3] found the opposite: NBS performance typically degraded when long genes, late-replicating genes, or mutations with no predicted functional impact (including synonymous mutations in the ovarian data set) were removed.

Some of the presumed passenger mutations and genes appearing in the NBS subtype networks may not be passengers and might be incorrectly discarded by overly conservative statistical approaches for testing driver genes. Alternatively, these genes may in fact be passengers yet still be useful for patient stratification, perhaps owing to correlations with other features. For example, recent mathematical models propose that half or more of the passenger mutations in tumors from self-renewing tissues accumulate before tumorigenesis[11]. This discrepancy between the importance of certain mutations for NBS subtypes and their classification as passengers by other tests is worthy of further study.

NBS makes it possible to derive clinically and biologically meaningful subtypes directly from whole-exome and whole-genome cancer sequencing data sets. As these data sets continue to increase in size and scope, NBS may have a prominent role in cancer research and in precision oncology.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Chin, L., Andersen, J.N. & Futreal, P.A. *Nat. Med.* **17**, 297–303 (2011).
2. Sotiriou, C. & Piccart, M.J. *Nat. Rev. Cancer* **7**, 545–553 (2007).
3. Hofree, M., Shen, J.P., Carter, H., Gross, A. & Ideker, T. *Nat. Methods* **10**, 1108–1115 (2013).
4. Vogelstein, B. *et al. Science* **339**, 1546–1558 (2013).
5. Vandin, F., Upfal, E. & Raphael, B.J. *J. Comput. Biol.* **18**, 507–522 (2011).
6. Ein-Dor, L., Zuk, O. & Domany, E. *Proc. Natl. Acad. Sci. USA* **103**, 5923–5928 (2006).
7. Venet, D., Dumont, J.E. & Detours, V. *PLoS Comput. Biol.* **7**, e1002240 (2011).
8. Cole, C. *et al. Cancer Biol. Ther.* **10**, 495–504 (2010).
9. Greenman, C., Wooster, R., Futreal, P.A., Stratton, M.R. & Easton, D.F. *Genetics* **173**, 2187–2198 (2006).
10. Lawrence, M.S. *et al. Nature* **499**, 214–218 (2013).
11. Tomasetti, C., Vogelstein, B. & Parmigiani, G. *Proc. Natl. Acad. Sci. USA* **110**, 1999–2004 (2013).