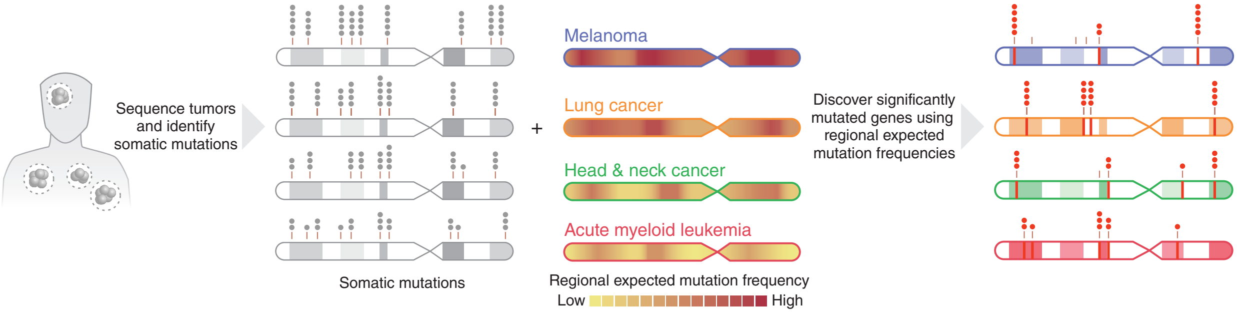
Sequence tumors and identify somatic mutations
Somatic mutations
+
Melanoma
Lung cancer
Head & neck cancer
Acute myeloid leukemia
Regional expected mutation frequency
Low
High
Discover significantly mutated genes using regional expected mutation frequencies

Figure 1 Genomic integrity is affected by external agents, including ultraviolet radiation, cigarette smoke and viral infection. Mutational heterogeneity is present in various cancer types, and mutation frequencies vary across the genome within individual cancer samples and types. The MutSigCV algorithm takes into account mutation frequency, mutation spectrum, replication timing and gene expression level to discover significantly mutated genes.

For example, with respect to regional heterogeneity, analysis of mutation data from published whole genome sequences (≥30× coverage) would provide even more substantial insights. The authors supported their analysis of exome data with an additional 126 whole genome sequences from ten cancer types. This provided additional evidence for the strong correlation of mutational frequencies with gene expression and replication timing.

Our own analysis of 50 acute myeloid leukemia (AML)[6,7], 46 breast cancer[8], 16 lung cancer[9] and 13 melanoma cases (unpublished data) also indicates that regional heterogeneity is cancer-type specific (e.g., moderately skewed distributions in AML and breast cancer, but heavily skewed distributions in lung cancer and melanoma with long tails) and, moreover, a function of sequence context. This would seem to preclude any kind of standard regional heterogeneity approach for pooled data from the Pan-Cancer project to determine common cancer genes. Instead, a more accurate estimation of regional heterogeneity for each cancer type and subtype will emerge as the community continues to generate high-coverage whole genome sequencing data for various cancer types and subtypes. Although regional effects such as replication timing and transcription-coupled repair are indeed factors in identifying significant cancer genes, there are many others—including GC content and carcinogens such as cigarette smoke and ultraviolet light exposure—whose influence on the mutational significance analysis remains to be investigated (**Fig. 1**). Our experience suggests that incorrect or biased annotation of mutations also contributes markedly to potential false positives in cancer gene analysis. For example, multiple open reading frames in genes like *TTN* or incomplete description of pseudogenes in olfactory receptors can lead to incorrect assignment and annotation of mutations resulting in false predictions.

The scientist's main quantitative handle in cancer genomics studies is the *P* value, that is, the probability of a given observation under the null hypothesis of random chance. The calculation of *P* values is affected by many factors, such as differences in gene size and the distribution of mutations among samples. Lawrence *et al.*[1] used the standard procedure of convolution to reconcile the probabilities of different categories of mutation within a sample. This is a mathematically approximate way of avoiding computationally intensive expansions by binning sufficiently similar events. Then they took another step of convoluting over all patient samples. Just as a single sample represents one test of the hypothesis that a given gene is cancer related, so many samples represent many independent instances of the same test. In most studies, sample-specific *P* values are combined in a systematic way: a meta-test on the distribution of the product of the sample-specific probabilities[10]. This procedure is typically carried out for samples of a single cancer type. However, meta-testing over multiple pan-cancer sets still appears to be an unsolved problem.

The methods presented by Lawrence *et al.*[1] will certainly facilitate the linking of cancers to their associated causative genes, but substantial work remains. Developing better estimates of regional differences in mutation rates will require more data as well as improvements in statistical tools for analyzing variable mutation frequencies and contexts, gene expression patterns and cancer type–specific heterogeneity. Improved mutation annotations will also be needed to better identify genes that have true biological relevance. Ultimately, conclusive proof that a given mutation has a causative role in cancer development will require functional validation and clinical assessment.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Lawrence, M.S. *et al. Nature* **499**, 214–218 (2013).
2. Cancer Genome Atlas Network. *Nature* **487**, 330–337 (2012).
3. Stephens, P.J. *et al. Nature* **486**, 400–404 (2012).
4. Cancer Genome Atlas Research Network. *Nature* **455**, 1061–1068 (2008).
5. Cancer Genome Atlas Network. *Nature* **490**, 61–70 (2012).
6. Cancer Genome Atlas Research Network. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
7. Welch, J.S. *et al. Cell* **150**, 264–278 (2012).
8. Ellis, M.J. *et al. Nature* **486**, 353–360 (2012).
9. Govindan, R. *et al. Cell* **150**, 1121–1134 (2012).
10. Fisher, R.A. *Am. Stat.* **2**, 2 (1948).

Research Highlights

Papers from the literature selected by the Nature Biotechnology editors. (Follow us on Twitter, @NatureBiotech #nbtHighlight)

Cerebral organoids model human brain development and microcephaly
Lancaster, M. *et al. Nature* doi:10.1038/nature12517.html (28 August 2013)

RNAi screens in mice identify physiological regulators of oncogenic growth
Beronja, S. *et al. Nature* doi:10.1038/nature12464.html (14 August 2013)

Design and characterization of synthetic fungal-bacterial consortia for direct production of isobutanol from cellulosic biomass
Minty, J. *et al. PNAS* **110**, 14592–14597 (2013)

Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity
Rann, F.A. *et al. Cell* **154**, 1380–1389 (2013)

A neo-substrate that amplifies catalytic activity of Parkinson's-disease-related kinase PINK1
Hertz, N.T. *et al. Cell* **154**, 737–747 (2013)