National Cancer Institute
Center for Biomedical Informatics
& Information Technology

National Cancer Informatics Program

NCI BIOMEDICAL INFORMATICS BLOG

NCIP

The Real Value of an Atlas

Categories: genomics

October 31, 2016

By **Trey Ideker, Ph.D.**, Director, San Diego Center for Systems Biology, University of California San Diego, Department of Medicine

These days there seems to be a lot of talk about atlases for cancer. Most of us are familiar with The Cancer Genome Atlas (TCGA), the long-running effort which, over the past decade, sequenced genomes from thousands of tumor samples covering dozens of cancer types. TCGA catalogued the complex patterns of gene mutations underlying tumors, implicated numerous new cancer genes, and is generally viewed as a resounding success. Efforts on the project officially wrapped up last year, creating a buzz of discussion and effort towards what might be the "Next Big Thing." The White House then announced the Cancer Moonshot, raising the stakes as it begins efforts to identify what broad steps should be taken to maximize progress in cancer research in the next five years. Some, but not all, of these steps might involve creation of atlases. Surrounding these developments have been a plethora of news articles and press releases, Blue Ribbon Panel recommendations, creation of new atlas-oriented institutes, and so on.

So what are some of the cancer atlases and how do we make sense of them all?

Heterogeneity and the Cancer Genome

To better appreciate atlases, let me review some key discoveries in cancer genomics up to this point, ending in a major finding: rampant genetic heterogeneity. The basic idea goes like this: The Human Genome Project completed early this century, leading to the identification of 20,000-some-odd genes and numerous regulatory elements. This compelled us to understand how variations in these sequences cause disease, leading to efforts like TCGA. From TCGA's genomes, we find that most genomic variations and mutations are really quite rare — hardly ever the same from cancer patient-to-patient, except for those that fall in a few well-known cancer genes. Although at first it was thought that these rare mutations might be passive "passengers" in tumors, or just noise, increasing evidence shows that many of them are indeed functional.

The key word for this situation is *heterogeneity*. Every tumor is genetically different, like a snowflake. And recently, sequencing of individual tumor cells has shown that such heterogeneity exists not only from patient-to-patient, but even from cell-to-cell within a tumor. The problem is that, if every tumor genome is so different, how are we supposed to understand and treat cancer? Genome analysis has long worked according to the laws of statistical association. To firmly link a mutation to disease, we need to observe that mutation recurrently in a large enough number of patients to be considered "significant," i.e., a number that is larger than would be

expected by random chance. However, heterogeneity by definition means that recurrent patterns are not observed for most mutations. To make matters worse, patients afflicted by such unique patterns of mutations have been labeled "N-of-1s," to capture the idea that they cannot be joined together with any other individuals to be analyzed and treated as a larger cohort (i.e., of size N>1). Patients enduring this desultory fate stand alone, without a friend even in disease.

The Rise of the Atlases

So, what to do next that will help us make sense of all this rampant heterogeneity? It is not clear that just more sequencing will help — due to the sheer number of genomic patterns that may give rise to cancer, one might sequence every cancer patient on the planet and still not be able to detect the necessary recurrent mutations. Hence a surge of interest (or resurgence, in some cases) in further types of data and data analysis that might inform the heterogeneous biology of cancer. For instance, originally genes were implicated in cancer not because their mutations correlated with patients who had disease, but because introducing those gene mutations in cells resulted in transformation into cancer. A direct cause-and-effect was established. Then, the effects of these mutations on the corresponding protein, and ultimately the cell, could be studied using the tools of cell biology. Microscopy could be used to study the cellular location of that protein, both with and without the oncogenic mutation. Co-precipitation could be performed to directly identify other cellular factors, including other proteins, small molecules, and nucleic acids, that were associated with the protein, and how these interactions were modulated by mutation. The 3D configuration of the protein could be solved using crystallography, again with or without the relevant genetic variant. And so on. All of these techniques give us direct information about the structure and function of the normal or diseased cell, without the need to rely on associations over large populations of individuals. And increasingly, these other layers and types of cellular information can also be measured comprehensively, or at least systematically, in a similar way to how sequencing machines systematically measure the genome.

Hence the "Rise of the Atlases." Everyone who can measure anything has one, or at least has one in mind, as long as the measurements can be said to be pseudo-comprehensive. ("Pseudo" is a key qualifier here, as it admits inevitable shortcomings. These also apply, by the way, to my efforts to list relevant atlases in Table 1 — apologies to efforts I have inadvertently overlooked). For instance, there are protein atlases, holding collections of all proteins, protein structures and possible modifications; protein network maps, charting all of the physical and genetic interactions that occur among proteins; cell type atlases, enabled and defined by advances in single-cell RNA sequencing; protein localization atlases, tracking the subcellular location of each protein in heathy and disease states. All of this makes one wonder whether, in the end, are atlases just another name for… 'omics? Certainly the etymology is better: after all, Atlas was a Greek Titan. In contrast the word Omics, in the words of Sydney Brenner, usually just sounds COmic.

All of this is certainly overwhelming. As has always been the case with the various types of 'omics, one quickly gets lost in the what, how, and why. What is the precise definition of each of these 'omes or atlases? How does one technologically achieve coverage of that space? Most importantly, why would one be compelled to do so? Genomics itself has always had better answers to these questions. It is after all the Human Genome, up there on a pedestal, close to God. Easy to define and, nowadays, easier to sequence. Other 'omes / atlases face tougher questions. Why would one want to catalog all protein locations, all interactions, all cell types? "Because it is there" is actually not a bad answer — we should indeed be inherently interested in cataloging the components of life. On the other hand, given limited resources, perhaps the most pertinent question isn't whether we should generate these catalogs — the work certainly seems worthwhile — but which, out of the many possibilities, one might focus on first. Certainly before embarking on any new campaign to generate massive data, it seems reasonable that the

leaders of such campaigns should make strong, quantitative arguments for why any particular investment is warranted.

Table 1. An Atlas of Atlases.

Atlas	Description	Main Data Type (in some cases best guess)
The Cancer Genome Atlas	The "original" cancer atlas, vis-à-vis those covered in this blog.	Exome sequencing of matched tumor and normal tissue to define somatic mutations
Pan-Cancer Analysis of Whole Genomes	Direct successor to The Cancer Genome Atlas; extends gene-only (exome) sequencing to whole-genome sequencing.	Whole genome sequencing of matched tumor and normal tissue
Pre-Cancer Genome Atlas	Catalog germline influences that give rise to, and shape the genetic and epigenetic landscape of, cancer.	Whole genome sequencing of germline
Cancer Immunity Atlas	Map the 3D configuration of the tissue surrounding a tumor, including immune and other cells, that promotes its growth.	Whole genome sequencing, Deep phenotyping of immune activity
Cancer Cell Map	Map protein interactions, complexes and synthetic lethals underlying cancer; how these are reprogrammed by mutations.	Affinity purification mass spectrometry, targeted perturbation by CRISPR and drugs
Cancer Cell Atlas	Fine-resolution catalog of different cell types in cancer.	Single-cell mRNA sequencing to define types
Cancer Protein Atlas	Catalog of protein abundances and modification states in cancer.	Protein arrays, mass spectrometry
Cell Explorer	Catalog of 3D and 4D (dynamic) cell images and associated computational cell models, tracking locations of all proteins.	Advanced imaging technology

Atlases as Integrators and Filters of Heterogeneity

Certainly many people have preferred 'omes they are focusing on. However, rather than advocate for one atlas or another, I think it might be more productive to discuss general principles, or guidelines, for thinking about cell atlases now and in the future. Here I would like to highlight one such principle I view as key, because it gets back to the main reason for the recent calls for cancer cell atlases in the first place. This key principle, and I will bold it here for emphasis, is that **cell atlases should help homogenize heterogeneity**.

Although this requirement may seem quite out of left field (how indeed are cell atlases supposed to do that?) it is actually quite natural. The simplest way that cell atlases unify genetic

heterogeneity — wildly different genetic alterations across a patient population — is by *integrating* mutations into functional groups, a.k.a. "modules," of genes. Gene expression data define clusters of co-expressed genes. Protein interaction maps define clusters of interacting proteins. Imaging data identify punctae of co-localized proteins, and so on. These gene clusters, in turn, allow us to merge into a single event a seemingly diverse set of nucleotide mutations, all to different genes but nonetheless impacting the same genetic module. Rather than "such-and-such nucleotide is mutated" we can say "this membrane complex is mutated" or "that pathway is mutated" or, at a higher scale still, "that cell type is mutated." Indeed, this modularization process has recently been fairly successful at increasing the power to explain the genetic variants underlying cancer and other diseases.

The human genome, itself a fundamental type of cell atlas, unifies genetic heterogeneity in exactly this same way, although most of us probably take it for granted. By identifying genes and gene boundaries, the human genome allows us to group individual nucleotide variants within larger genetic regions. We can say "this gene is mutated" not just "this nucleotide is mutated." Without this principle, by the way, most of the cancer alterations reported by the TCGA would not have been found. Other cell atlases simply extend this integration step above the gene level, by grouping mutations that hit any gene encoding a sub-unit of a common protein complex, or any gene whose protein is localized to the mitochondrial membrane. The cell, and indeed the human body, is an exquisite hierarchy of biological systems functioning inside of systems inside of systems inside of systems…and so on, and somewhere near the bottom of this hierarchy lie the genes. In so far as cell atlases inform the hierarchy, they help unify heterogeneity.

This is not the only means by which cell atlases can help, by the way. Although I will not say too much about them here, there are also other methods by which cell atlases can make sense of heterogeneity, for instance by helping to *filter* which variants and mutations are likely to be functional. Along these lines, a cell type atlas might indicate which genes are expressed and required for maintenance of a particular cell type underlying cancer. A protein atlas might do the same at the protein level. Both types of information elevate mutations in certain genes to a higher level of scrutiny, while demoting others.

No more N-of-1s

So cancer cell atlases may be able to help clarify the genetic heterogeneity of cancer. Well, so what? For one, more and more patients could be placed into genetically similar groups. When viewed at the nucleotide level, or even at the level of the gene, we nowadays throw up our hands and write off this patient as an "N of 1." What a mistake. I predict that, when viewed in the light of cell atlases, these individuals will suddenly find themselves in a society of friends. We will be able to discover that these patients, although mutated at different genes, are in fact affected in the same transcriptional networks, protein complexes, pathways, or cell types.

You want to know my suspicion? There are no "N-of-1s."

— T.I.

Trey Ideker is a systems biologist working to elucidate and model the genetic networks inside cells. He has introduced a variety of influential approaches for mapping and analyzing networks, including Cytoscape, an open software platform cited more than 10,000 times. He holds positions as Professor of Genetics in the Department of Medicine at the University of California at San Diego and leads the Program in Genomes and Networks at UCSD Moores Cancer Center. Dr. Ideker serves on the Editorial Boards of *Cell*, *Cell Reports*, and *Molecular Systems Biology*; is a fellow of the American Association for the Advancement of Science; was named one of the Top 10 Innovators of 2006 by Technology Review magazine; and is the 2009 recipient of the Overton Prize from the International Society for Computational Biology.

This post has no tag

Permanent link to this article: https://ncip.nci.nih.gov/blog/real-value-atlas/