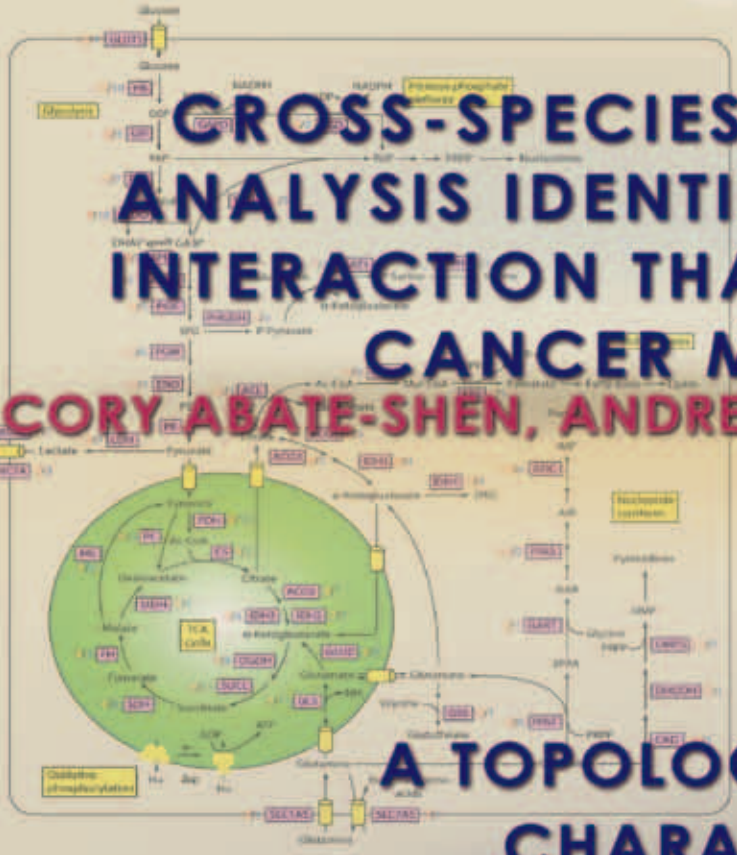
NATIONAL CENTER FOR THE MULTISCALE ANALYSIS OF GENOMIC AND CELLULAR NETWORKS

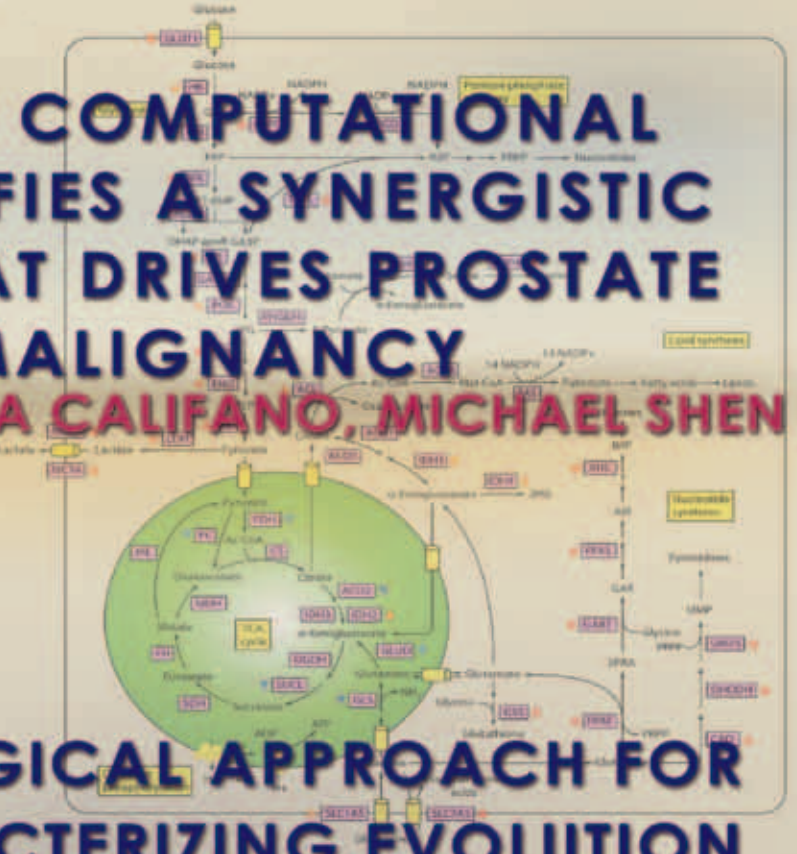
SPRING 2014 ISSUE NO.7

MAGNet NEWSLETTER



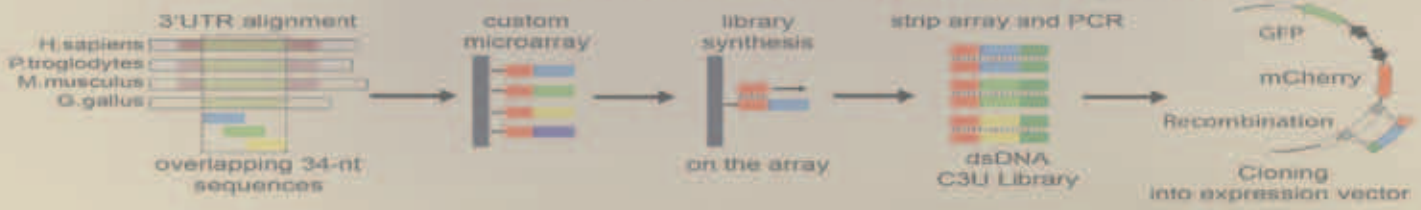
CROSS-SPECIES COMPUTATIONAL ANALYSIS IDENTIFIES A SYNERGISTIC INTERACTION THAT DRIVES PROSTATE CANCER MALIGNANCY

CORY ABATE-SHEN, ANDREA CALIFANO, MICHAEL SHEN



A TOPOLOGICAL APPROACH FOR CHARACTERIZING EVOLUTION

KEVIN EMMETT, RAUL RABADAN



A GLOBAL VIEW OF TUMOR-INDUCED GENE EXPRESSION CHANGES IN HUMAN METABOLISM

DENNIS VITKUP

SPRING 2014
ISSUE NO.7

MAGNet Newsletter
http://magnet.c2b2.columbia.edu

MAGNet NEWSLETTER

FEATURES

SECTIONS

This past year witnessed a milestone for biological research and education at Columbia University. On July 1, 2013, the University Trustees approved the creation of a new Department of Systems Biology (DSB), which now brings together a faculty of more than two-dozen principal investigators working at the intersection of the quantitative and the experimental biological sciences. In many ways, the founding of the DSB would not have been possible without the successes that have taken place over the last nine years at the Center for Multiscale Analysis of Genomic and Cellular Networks (MAGNet). The DSB embodies MAGNet's vision for using computational analysis and high-throughput experimentation to provide a systems-level understanding of the biological processes at the foundations of physiology and disease. It has done so by forging a dynamic, multidisciplinary research community and by coordinating the resources of the Columbia Center for Computational Biology and Bioinformatics (C2B2) — the original home of MAGNet — and the JP Sulzberger Columbia Genome Center. These measures have enabled the development of a state-of-the-art infrastructure for high-performance computing, next-generation sequencing, and high-throughput screening.

In parallel with these exciting developments, MAGNet continued to provide innovative new algorithms, computational tools, methodologies, and databases of biological information to the wider scientific community. Both in their own labs and as collaborators with colleagues at other institutions, MAGNet investigators produced more than 55 publications in 2013-2014, a large proportion of which appeared in high-impact journals. In collaboration with researchers in other departments across Columbia University Medical Center and at other universities, our Center's researchers are helping to catalyze clinical translation of basic science discoveries by supplying quantitative, predictive methods for addressing problems related to the diagnosis and treatment of human disease. The research profiled in this seventh MAGNet newsletter typifies the remarkable work undertaken by our investigators to develop quantitative and computational models that can lead to novel, validated insights into fundamental biological activity.

The first paper, recently published in *Cancer Cell*, represents a cross-disciplinary collaboration between the Abate-Shen, Califano, and Shen laboratories that grew from one of MAGNet's driving biological projects. The authors describe a new method for comparing human and mouse prostate cancer regulatory models (interactomes) that could lead to more useful genetically engineered mouse models for studying human biology. To generate gene expression profiles capable of capturing sufficient variability to assemble the mouse interactome, they perturbed 13 distinct transgenic mouse models of prostate cancer *in vivo*, using a repertoire of 13 small molecule compounds and DMSO as a control. Using MAGNet-developed tools including ARACNe (regulatory network reverse engineering) and a modified version of MARINa (master regulator analysis), the researchers then identified two synergistic master regulators of aggressive disease, *FOXM1* and *CENPF*. These genes were validated in mouse and human assays, both in vitro and in vivo. Although silencing either gene alone had a relatively minor effect, co-silencing both genes abrograted cancer growth and significantly reduced AKT and MAP kinase signals, a hallmark of aggressive tumors. Links between these two genes and poor clinical outcome were also identified and validated using a cohort of almost 900 patient TMAs. *FOXM1* and *CENPF* may thus provide a promising avenue for diagnosing and targeting aggressive prostate cancer.

In a report on a second paper that was published in the *Proceedings of the National Academy of Sciences*, Kevin Emmett and Raul Rabadan discuss a novel use of persistent homology, a technique from the mathematical field of algebraic topology, to model horizontal exchange of genetic information. They demonstrate how connectivity structures in topological spaces provide a robust representation of genetic exchanges, including events that cannot be represented by simple phylogenetic trees or networks. Applied to human H3N2 influenza virus, their method identified PB2 (polymerase basic 2) and HA (hemagglutinin) as being involved in horizontal exchanges. In H1N1pdm, the 2009 pandemic influenza strain, they found evidence of reassortment of its genome within humans. They also used the method to show a high recombination rate (22.16 reassortments/year) in avian influenza virus A, in contrast to very low rates for H1N1 swine flu and H3N2. In addition to their primary work on influenza, the authors also applied persistent homology to study the genetic exchange in HIV, hepatitis C, West Nile, and dengue viruses, finding evidence of horizontal exchange for the first two, but not for the latter two. These cases represent a truly exciting and successful application of the mathematics of topology to biological questions, allowing a unified representation of vertical (clonal) and horizontal (reticular) evolution.

In the third spotlighted article, published in *Nature Biotechnology*, Dennis Vitkup asks how similar various cancer types are in terms of their basic metabolic pathway activity. The analysis of a dataset comprising 2500 gene expression profiles from 22 diverse cancer types showed that tumors retain a metabolic imprint of their tissue of origin; i.e., tumor samples show more similarity in the expression of metabolic genes to normal samples of that tissue of origin than to tumors from other tissues. However, when the difference in expression between normal samples from different tissues was also considered, an intriguing finding emerged: the difference between distinct tissues was less pronounced in tumors than in normal samples, implying a process of cross-tissue metabolic convergence in the tumors. Much of the expression difference for major biochemical processes could be explained by just a few principal components, including glycolysis, nucleotide biosynthesis, and catabolic pathways, reflecting the changing environment of the cancer cell. Finally, the article reports hundreds of tumor-specific expression changes in isoenzymes, including previously unknown roles for succinate dehydrogenase and fumarate hydratase in colorectal cancer, a finding confirmed by measurements of specific metabolites from colon cancer patients.

These and other highlights captured in this newsletter show that MAGNet investigators continue their tradition of highly innovative, high-impact scientific research at the boundary of computational and experimental biology, providing new methods for gaining systems-level understanding of biological processes.

-Andrea Califano

CROSS-SPECIES COMPUTATIONAL ANALYSIS IDENTIFIES A SYNERGISTIC INTERACTION THAT DRIVES PROSTATE CANCER MALIGNANCY

CORY ABATE-SHEN, ANDREA CALIFANO, AND MICHAEL SHEN
DEPARTMENT OF SYSTEMS BIOLOGY
COLUMBIA UNIVERSITY

INTRODUCTION

Genetically engineered mouse models (GEMMs) play an important role in preclinical cancer research, and have been used to characterize disease-specific molecular pathways, find biomarkers of disease progression, and identify new strategies for disease prevention and treatment. In prostate cancer research, for example, there are now numerous GEMMs that model key molecular pathways implicated in cancer and specific stages of disease progression.[1,2] Nevertheless, mouse models have inherent limitations within the context of translational research. Although mice and humans share evolutionary roots and have extensive genomic overlap, differences in physiology mean that mouse models can at best approximate human biology, and all too often observations seen in mice turn out not to be replicated in humans. This poses particular problems in the study of complex human diseases such as prostate cancer, which is heterogeneous at the molecular level, and can assume a variety of phenotypes, from relatively benign to highly lethal.

In a Driving Biological Project at the Center for the Multiscale Center for Genomic and Cellular Networks (MAGNet), we set out to investigate whether methods based in systems biology could improve the effectiveness of animal models of cancer. Specifically, we hypothesized that differences in the organization of regulatory networks between mouse and humans is an important cause of phenotypic discrepancies seen in preclinical and clinical observations. By using computational methods to compare whole-genome networks of regulatory interactions in mice and humans (also called interactomes),[3] our goal was to identify master regulators of cancer aggressiveness that are conserved between species. (Master regulators are genes that function as key regulatory "bottlenecks", or points of convergence within the networks of molecular interactions that are essential for driving specific phenotypes.) Identifying similarities between species at the level of the regulatory network would therefore provide strong evidence that the conserved master regulators were actually important in human disease.

Because prostate cancer interactomes did not yet exist, our first step was to identify representative datasets from which to build one for each species. We assembled a human prostate cancer interactome based on the gene expression profile dataset reported in a 2010 paper by Taylor et al.,[4] which contained information about 185 patients ranging from those with benign tumors to those with highly aggressive tumors. The dataset also included gene expression profiles from a variety of sample types, including primary tumors, adjacent normal tissue, metastases, and cell lines, giving a high degree of variability. No analogous dataset existed for the mouse prostate cancer interactome, and so we undertook a series of assays to generate a collection of gene expression data that was both of appropriate size and represented a comparable degree of disease variability. With the cooperation of Charles Sawyers (Memorial Sloan-Kettering Cancer Center), Terry Van Dyke (National Cancer Institute), Bart Williams (Van Andel Institute), and Barbara Foster (Roswell Park Cancer Institute), all of whom maintain GEMMs, we chose 13 distinct genetically engineered mouse models that represent the full spectrum of prostate cancer phenotypes, including normal epithelium, low-grade disease, high-grade disease, castration-resistant disease, and metastatic disease. Each mouse model underwent *in vivo* administration of 13 distinct small-molecule compounds plus docetaxel, which were selected for their ability to modulate known prostate cancer molecular pathways. Gene expression signatures were recorded following each perturbation in each cell line, resulting in a large dataset comprising 384 gene expression profiles.

Once the human and mouse datasets were available, we used ARACNe (Algorithm for Reconstruction of Accurate Cellular Networks) to assemble a prostate cancer interactome for each species. Taking gene expression data as its input, ARACNe infers interactions among transcription factors and co-factors and their gene targets in an unbiased way. The algorithm's power stems from its ability to distinguish direct interactions from indirect interactions, resulting in a highly robust,

genome-wide model.[5,6] In this case, ARACNe generated a human interactome representing nearly 250,000 interactions between 2,681 transcriptional regulators and their target genes. The mouse interactome represented almost 223,000 interactions for 2,072 transcriptional regulators.

We still faced a challenge, however, because although we now had two interactomes, we needed a quantitative metric for determining conservation of regulatory networks between human and mouse prostate cancer. To do so, we modified another algorithm, called MARINa (Master Regulator Inference Algorithm),[7,8] to enable analysis of differences in the activity of the 2,028 transcriptional regulators that were present in both the human and mouse interactomes. We found that 70% of the transcriptional regulators in the human and mouse prostate cancer interactomes — including many genes known to play important roles in prostate cancer — regulate molecular programs that are highly conserved between species.

We then used MARINa to analyze the ARACNe-derived interactomes for master regulators (MRs) of malignant prostate cancer. When we interrogated the human prostate cancer genome, we identified 175 candidate master regulators, including 49 activated and 126 repressed MR's. In addition, when we performed MARINa analyses of 15 independent human interactomes — including prostate cancer interactomes as well as non-prostate cancer interactomes — we found a high degree of overlap between the MR's we had inferred and those in the 2 prostate cancer interactomes, but not those in the 13 non-prostate cancer interactomes. This gave further support to conclusions that we

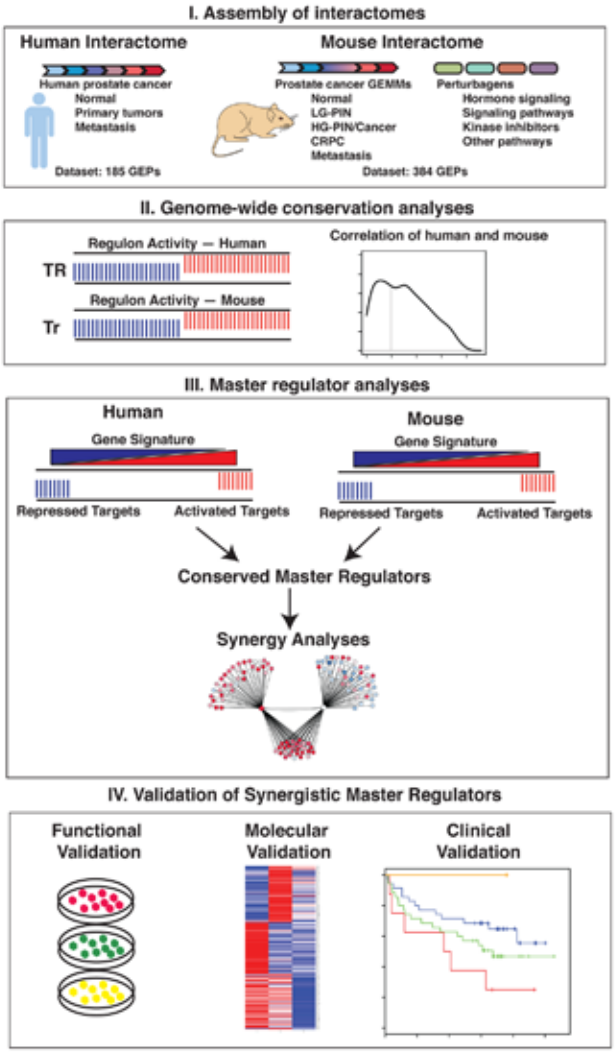


Figure 1. A strategy for genome-wide cross-species analyses of prostate cancer. Step I: Assembly of complementary human and mouse prostate cancer interactomes. Step II: Genome-wide computational analysis of conservation of transcriptional regulon activity. Step III: Master regulator analysis for identification of conserved master regulators and prediction of synergy. Step IV: Validation of master regulators using functional, molecular, and clinical analyses.

and others have drawn elsewhere; namely, that interactomes are highly context-specific, correlating closely with specific phenotypes. We also performed MARINa analysis on four distinct genetically engineered mouse model signatures. These signatures were selected because they are both associated with prostate cancer malignancy in mice and represent a range of cancer phenotypes. A meta-analysis identified 229 candidate master regulators of mouse prostate cancer, including 110 activated and 119 repressed MR's.

Integrating our findings from the mouse and human studies, we then produced a ranked list of 20 master regulators (7 activated, 13 repressed) that were conserved between species. We further prioritized these candidates by considering their potential for interacting synergistically to promote aggressive disease. We consider two master regulators to be synergistic if ARACNe determines that the co-regulated targets of two transcription factors are enriched in the malignancy signature to a higher degree than they would be by simply adding together the effects of each of the master regulators alone. When we considered all 21 possible synergistic relationships among our top-ranked master regulators, the only pair that was statistically significant and of clinical relevance was that of *FOXM1* and *CENPF*. Both of these genes had previously been associated with cancer-related biological processes — *FOXM1* in cell cycle progression and *CENPF* in mitosis — although this was the first time that a synergistic relationship between the two had been identified.

To test the accuracy of these computational predictions, we conducted a series of experiments in four distinct human prostate cancer cell lines using lentivirus vectors that, when induced by doxycycline, express shRNAs that silence *FOXM1* and *CENPF*. (We also used a control shRNA and tracked the activity of the lentiviruses using a fluorescent reporter gene.) We found that individual targeting of *FOXM1* or, to a lesser degree, *CENPF* resulted in reduced cell growth, but when the two were targeted synergistically, the combined effect was statistically greater than if the two were simply added together. Importantly, colony formation was nearly completely abrogated in each cell line following silencing of the two genes in combination.

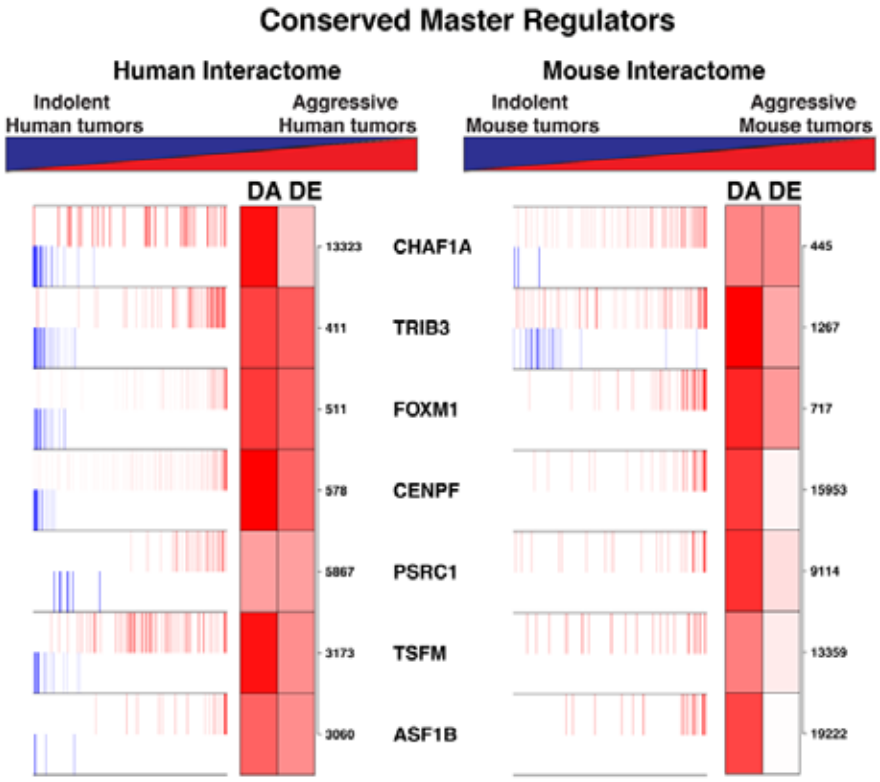


Figure 2. Conserved activated MRs are shown for the human (left) and mouse(right) malignancy signatures, depicting their positive (activated; red bars) and negative (repressed; blue bars) targets. The ranks of differential activity (DA) and differential expression (DE) are shown by the shaded boxes; the numbers indicate the rank of the DE in the signature.

We also investigated the effect of co-silencing *FOXM1* and *CENPF in vivo*. After engrafting DU145 cells (a population of prostate cancer metastatic cells) expressing silencing vectors for *FOXM1* and *CENPF* into immunodeficient mice, we saw consistent results: while individual silencing caused tumors to grow more slowly, co-silencing resulted in complete abrogation of tumor growth and a nearly 13-fold reduction in tumor weight. Once again, this synergistic effect was significantly greater than one would have expected to see if the effects of co-silencing were merely additive.

In a third experiment, we developed an *in vivo* competition assay in which DU145 cells were infected with silencing vectors expressing 1) a *FOXM1* shRNA and a red fluorescent reporter, 2) a *CENPF* shRNA and a green fluorescent reporter, or 3) both lentiviruses. As negative controls, we also infected DU145 cells with fluorescent markers but no shRNAs. We then implanted equal numbers of red, green, and yellow experimental and control cells into immunodeficient mice. After one month circulating *in vivo*, we isolated the resulting tumors and used FACS sorting to quantify the number of red, green, and yellow cells in the tumors. In the tumors derived from control cells, we found equivalent numbers of red, green, and yellow cells, suggesting that the corresponding lentiviruses had no effect on tumor growth. In contrast, tumors in the experimental group were made up primarily of red (*FOXM1*) or green (*CENPF*) cells, but virtually no yellow cells because co-silencing had eliminated nearly all of them. This finding supported our hypothesis that synergistic targeting of *FOXM1* and *CENPF* can dramatically interrupt prostate tumor growth.

With this strong evidence that the synergistic relationship between *FOXM1* and *CENPF* does actually take place, we undertook a number of studies to determine the molecular mechanisms that are responsible. We investigated whether *CENPF* was necessary for *FOXM1* to bind to known genomic binding sites on shared target genes. Experiments using chromatin immunoprecipitation (CHiP) followed by quantitative PCR showed that *FOXM1* binding was reduced when *CENPF* was silenced. These tests also suggested that *FOXM1* and *CENPF* both migrate to the nucleus in prostate cancer cells, and that this co-localization is mutually dependent. This is consistent with other findings, suggesting that the two transcription factors co-regulate expression of their shared target genes.

To identify the molecular pathways that are inhibited by co-silencing of *FOXM1* and *CENPF*, we also compared the gene expression profiles that were seen in cells after co-silencing with those from cells in which the genes were not silenced. The analysis showed that a majority of the targets of the two proteins were differentially expressed following silencing, giving us additional confidence that the ARACNe analysis was reliable. Co-silencing also revealed differential expression in a number of genes that were not previously known to be regulated by *FOXM1* and *CENPF*. This included important genes within several pathways associated with tumor growth. One of the most interesting findings was that the *PI3K*-kinase and *MAP* kinase signaling pathways were both enriched following co-silencing, particularly because these both constitute hallmarks of aggressive prostate cancer.[4,9] Western blot analysis showed that both pathways were completely abrogated during co-silencing of *FOXM1* and *CENPF*, suggesting that targeting these two genes could offer an effective strategy for inactivating them.

We also performed several retrospective studies looking at clinical data for correlation between *FOXM1/CENPF* co-expression, cancer progression, and disease outcome. Studying tissue microarrays from primary tumors at Memorial Sloan-Kettering Cancer Center and prostate cancer metastases at the University of Michigan, we found a number of strong associations between the expression of both of these genes and aggressive disease, and with poor clinical outcome. We also integrated statistical analyses of *FOXM1/CENPF* co-expression with Gleason scoring in the Memorial Sloan-Kettering prostate cancer tissue microarray data, and found that doing so dramatically improves the accuracy of prognoses over using Gleason scoring alone. This suggests that this two-gene signature could potentially be incorporated into prostate cancer diagnostics as a more effective predictor of poor disease outcome and metastasis.

This effort to bring the tools of systems biology to preclinical prostate cancer research offers a number of advances. For the first time, we have developed a method for performing cross-species interrogation of genome-wide, context-specific regulatory networks. Approaching genetically engineered mouse models from this perspective offers a higher-resolution view of the multitude of genetic, genomic, and epigenetic alterations that are associated with specific cancer phenotypes.

We anticipate that this strategy will improve researchers' ability to predict the applicability of observations seen in mouse models to human physiology, and will offer an efficient way to identify and focus investigators' attention on master regulators within interactomes that are critical for the generation of specific cancer phenotypes. Although *FOXM1* and *CENPF* have both been previously associated with various cancers, their synergistic ability to drive aggressive prostate cancer is a novel discovery. Considering the staggering number of possible synergies between molecular components within the prostate cancer cell, this finding is remarkable. A critical step in this approach was the

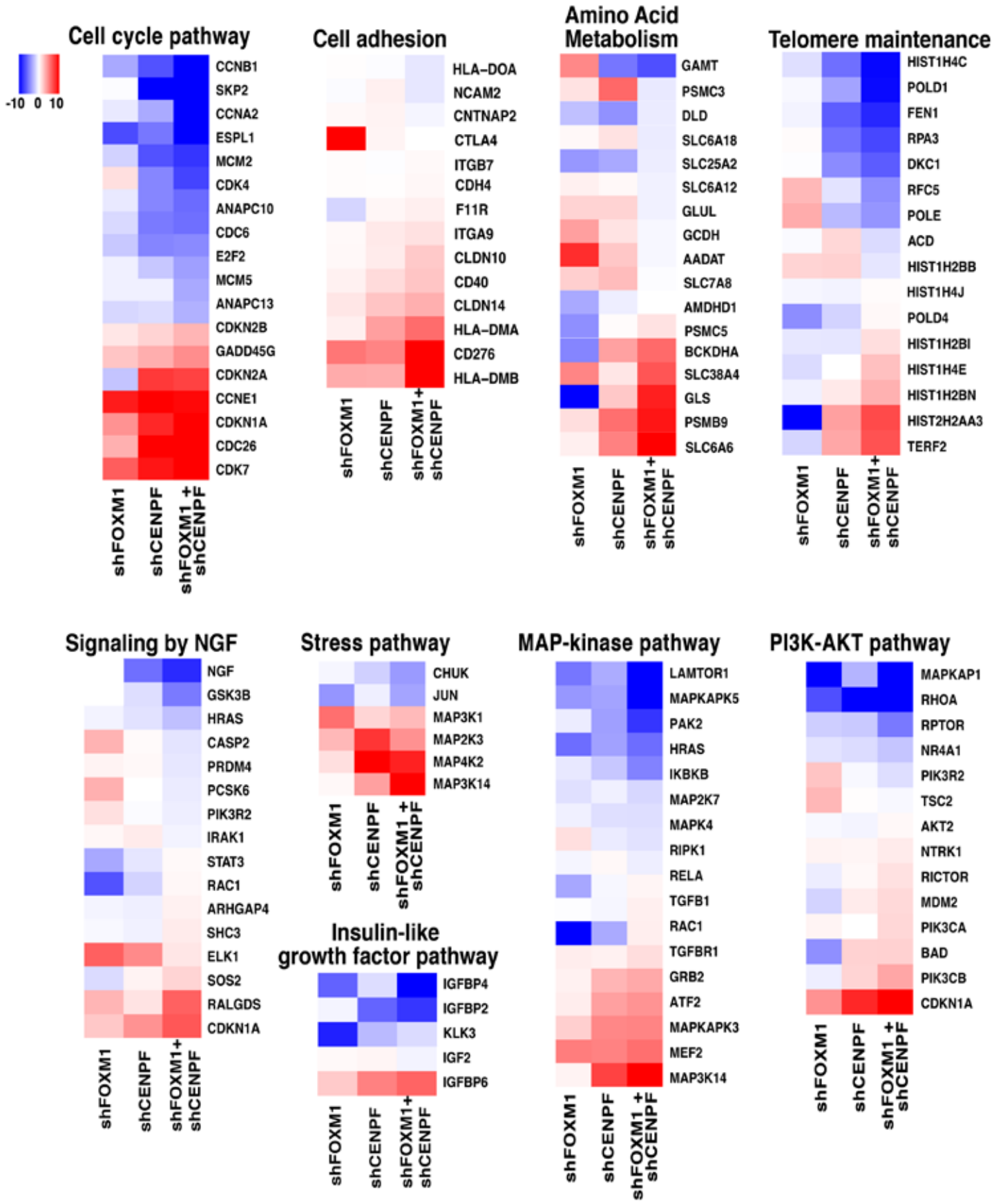


Figure 3. Heatmap of biological pathways. Pathway enrichment analysis was done by GSEA on the C2 pathway database. The heatmaps show differential expression of leading edge genes for selected pathways.

development of a new metric based on the MARINa algorithm that enables an assessment of the conservation of regulatory networks across species. When combined with the more conventional use of MARINa, the computational pipeline we developed was successful in identifying master regulators that are essential for driving aggressive prostate cancer. In addition, our use of ARACNe to infer the target genes for *FOXM1* and *CENPF* helped to characterize the molecular mechanism through which this synergy occurs. This computational result was confirmed experimentally, and suggests that inhibiting these two genes could overcome current challenges in targeting the *PI3K* and *MAPK* pathways, offering hope for new therapies against the most lethal form of prostate cancer. In addition, we anticipate that the recognition of the role of these two genes as master regulators of aggressive prostate cancer, combined with standard diagnostic approaches such as the Gleason score, could provide an improved method for differentiating tumors that pose a high risk to patients from those that are not life-threatening.

This new systems biology-based method for comparing cross species interactomes is not restricted to mice or prostate cancer, but could be used more broadly for the generation of interactomes for a wide array of physiologic and pathologic phenotypes in animal models. We expect that it will help to overcome some of the current limitations of genetically engineered mouse models and improve their usefulness in preclinical biological research.

References

1. Irshad S, Abate-Shen C. Modeling prostate cancer in mice: something old, something new, something premalignant, something metastatic. Cancer Metastasis Rev. 2013 Jun;32(1-2):109-22.

2. Ittmann M, Huang J, Radaelli E, Martin P, Signoretti S, Sullivan R, Simons BW, Ward JM, Robinson BD, Chu GC, Loda M, Thomas G, Borowsky A, Cardiff RD. Animal models of human prostate cancer: the consensus report of the New York meeting of the Mouse Models of Human Cancers Consortium Prostate Pathology Committee. Cancer Res. 2013 May 1;73(9):2718-36.

3. Lefebvre C, Rieckhof G, Califano A. Reverse-engineering human regulatory networks. Wiley Interdiscip Rev Syst Biol Med. 2012 Jul-Aug;4(4):311-25.

4. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, Antipin Y, Mitsiades N, Landers T, Dolgalev I, Major JE, Wilson M, Socci ND, Lash AE, Heguy A, Eastham JA, Scher HI, Reuter VE, Scardino PT, Sander C, Sawyers CL, Gerald WL. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010 Jul 13;18(1):11-22.

5. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet. 2005 Apr;37(4):382-90.

6. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006 Mar 20;7 Suppl 1:S7.

7. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H, Lasorella A, Aldape K, Califano A, Iavarone A. The transcriptional network for mesenchymal transformation of brain tumours. Nature. 2010 Jan 21;463(7279):318-25.

8. Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, Sato M, Wang K, Sumazin P, Kustagi M, Bisikirska BC, Basso K, Beltrao P, Krogan N, Gautier J, Dalla-Favera R, Califano A. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Mol Syst Biol. 2010 Jun 8;6:377.

9. Aytes A, Mitrofanova A, Kinkade CW, Lefebvre C, Lei M, Phelan V, LeKaye HC, Koutcher JA, Cardiff RD, Califano A, Shen MM, Abate-Shen C. ETV4 promotes metastasis in response to activation of PI3-kinase and Ras signaling in a mouse model of advanced prostate cancer. Proc Natl Acad Sci U S A. 2013 Sep 10;110(37):E3506-15.

A TOPOLOGICAL APPROACH FOR CHARACTERIZING EVOLUTION

KEVIN EMMETT[1] AND RAUL RABADAN[2]
[1]DEPARTMENT OF PHYSICS, COLUMBIA UNIVERSITY
[2]DEPARTMENT OF SYSTEMS BIOLOGY, COLUMBIA UNIVERSITY

The only image in Charles Darwin's seminal 1859 treatise, *On the Origin of Species*, visualizes the evolution of phenotypes in a branching pattern similar to that seen in a tree. Considered in the terms of modern genetics, this structure effectively models vertical, or clonal, evolution, which occurs because of random genomic mutations that are passed from one generation to the next over many generations. Since Darwin's time, however, it has become clear that evolution does not occur only in a branching, vertical pattern but can also be horizontal, or reticulate, when distinct clades of organisms exchange genetic information or merge together to form a new hybrid lineage. Examples of reticulate evolution include species hybridization in eukaryotes, lateral gene transfer in bacteria, recombination and reassortment in viruses, the incorporation of viral RNA into eukaryotic genomes, and genome fusions between symbiotic species. Phylogenetic trees lack expressiveness for capturing these important modes of genomic change.[1,2]

A number of techniques have been developed in an effort to account for reticulate evolution, but thus far have had only limited success. One approach that perhaps constitutes the greatest departure from phylogenetic trees has sought to explain reticulate evolution using the concept of the phylogenetic network, which allows for multiple paths between two lineages.[3-5] This approach has been limited in its effectiveness, however, because current implementations produce only one network that may be a suboptimal solution. In addition, this approach is computationally intensive, making it intractable to produce models for anything more than the smallest of datasets.[5-7]

This state of affairs begs the question of what other mathematical models might be capable of accounting for both clonal and reticulate evolution at the same time, in a way that is scalable to larger datasets in an efficient manner. In our lab, we have recently begun thinking of evolutionary events not as a tree or a network, but from the perspective of higher-dimensional objects with well-defined topological properties. Our approach is based on a branch of algebraic topology called *persistent homology*, which considers all possible topologies across a parameter space. By analyzing a variety of viral and simulated genomic datasets, we have shown that persistent homology can capture aspects of evolution that cannot be inferred using phylogeny. In addition, persistent homology can be used to quantify the rate of horizontal genomic events and statistical patterns of genomic cosegregation; i.e., the exchange of genes between individuals as a set.

Algebraic topology is a field of mathematics that characterizes global properties of a geometric object that do not change when the object is subjected to stretching or bending. To visualize this, consider what would happen if you could take a sphere and then push on opposite sides until the object flattened and ruptured to form a donut-like shape. This suggests concepts of connectedness between points on the surface of the object as well as holes that interrupt connectedness. Applied to evolution, our approach involves considering a traditional tree structure as a single point or connected component, whereas the presence of holes within a higher-dimensional object represents the occurrence of reticulate evolutionary events that confound phylogenetic evolution. We are particularly interested in holes that are "irreducible" cycles: cycles (holes) in dimension k that do not serve as the boundary of a $(k+1)$-dimensional object. We can define a topological invariant called the "homology group" H_k as an algebraic structure that encompasses all holes in dimension k, and the "Betti number" b_k as the count of these holes. By using algebraic topology to compute the number of holes in the evolutionary topological space, we should be able to gain insights into the occurrence and rate of reticulate events.

To do this, let us assume that evolution forms a topological space E. E is far too large a space to observe directly but we can observe a sample of data points within E, which corresponds to a set of genomic sequences separated from each other by some genetic distance. Although the set of these data points and space E do not share the same topology, we can estimate the topology of E by defining a function $B(x,\varepsilon)$ in which x is a data point at the center of a ball with a radius of genetic distance ε. We can show that for some value of ε, the union of all balls $B(x,\varepsilon)$ for all x shares the same topology as E. The topology of the union of balls can be estimated by constructing a corresponding

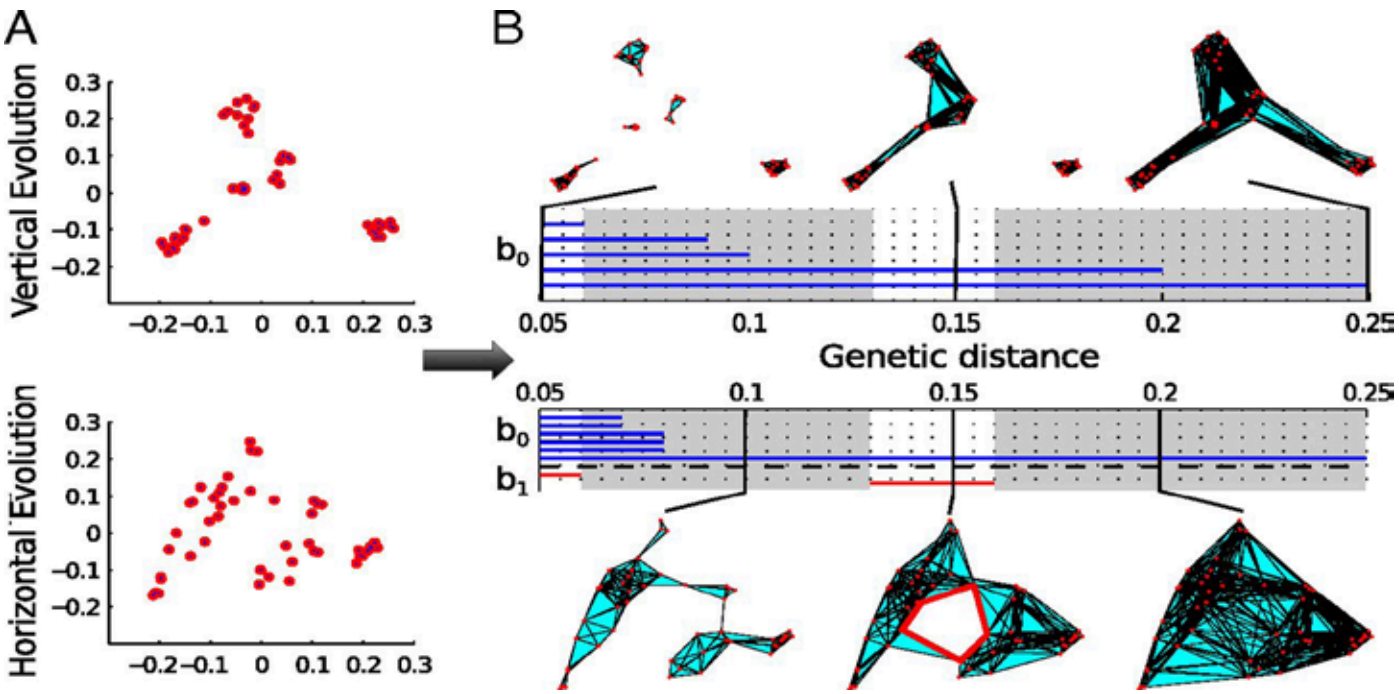


Figure 1: Persistent homology characterizes topological features of vertical and horizontal evolution. Evolution was simulated with and without reassortment. **(A)** A metric space of pairwise genetic distances *d(i,j)* can be calculated for a given population of genomic sequences $g_1,\ldots, g_n$. We visualize these data points using principal coordinate analysis (PCoA). **(B)** In the construction of simplicial complexes, two genomes are considered related (joined by a line) if their genetic distance is smaller than ε. Three genomes within ε of each other form a triangle, and so on. From there, we calculate the homology groups at different genetic scales. In the barcode, each bar in different dimensions represents a topological feature of a filtration of simplicial complexes persisting over an interval of ε. A one-dimensional cycle (red highlight) exists at ε = [0.13, 0.16 Hamming distance] and corresponds to a reticulate event. The evolutionary scales I where *b1* = 0 are highlighted in gray.

topological space called a *simplicial complex*, a set of points, lines, triangles, tetrahedrons and higher-dimensional "simplices" that are all within a distance ε of one another.[8-10] By varying ε across different scales, it becomes possible to create different simplicial complexes and reveal different irreducible cycles. In the context of our study of evolution, we consider a set of genomes as values of x within the topologic space. Using the pairwise distance matrix, we calculate the homology groups across all genetic distances ε in different dimensions. Zero-dimensional topologies represent vertical evolution while topologies of a dimension greater than zero result from horizontal exchanges of genomic information or complex reticulate events involving multiple parental strains. We call the set of sequences that represent a particular irreducible cycle the "*generator*" of that cycle. The generator describes the particular genomic mixture that the cycle is capturing.

Within the framework of persistent homology, then, we can mathematically formalize the role of phylogeny. By definition, a tree in which no reticulate evolution has occurred will have no holes and therefore a Betti number of zero. If a Betti number of dimension greater than zero exists, then there is by definition no "additive tree" in which genetic distance can be found simply by adding together genetic distance using phylogeny. Similarly, the disappearance of a nonzero topology reflects evolutionary scales at which no reticulate evolution has occurred and a tree model is sufficient. Using extensive coalescent simulations, we developed a maximum likelihood estimator for recombination rate that uses the topological information in the barcode diagram.

Persistent homology provides a lower bound for estimating recombination rates by considering independent irreducible cycles across all ε in a period of time. The irreducible cycle rate (ICR) is defined as the average number of one-dimensional irreducible cycles per unit of time. Simulations show that ICR is proportional to and provides a lower bound for the recombination/reassortment rate, adding a temporal dimension to our methodology that is not present in other methods.

To evaluate the ability of persistent homology to capture complex evolutionary process with

high sensitivity and specificity, we simulated three scenarios: clonal evolution, reassortment, and homologous recombination. Simulations showed that nontrivial homology appears when the recombination rate *r* is nonzero and that one-dimensional ICR increases proportionally to *r*. Reassortment events involving multiple loci can produce higher dimensional topology above 1-dimensional loops. In fact, the dimension of the topology is bounded by the number of loci involved in the reassortment.

To characterize influenza A evolution, we applied persistent homology to four influenza datasets from several hosts (avian, swine, and human), each numbering as many as 1,000 genomic sequences. When tested on single viral segments that had been unaffected by reassortment, higher-dimensional homology groups vanish, suggesting that no significant reticulate events have taken place. Phylogenic analysis in this situation is therefore appropriate for alignments of single segments. We can incorporate such examples of vertical evolution into our analysis by transforming a filtration of simplicial complexes of dimension zero into an equivalent distance-based dendrogram. From the bar sizes of the barcode, we can create a dendrogram that recapitulates classic phylogenetic analyses.[11,12]

Persistent homology also proved capable of identifying evidence of reassortment that could not be identified by phylogeny. We analyzed 1,000 genomes of the human H3N2 influenza virus and identified three generators of one-dimensional homology when joining two subunits called PB2 (polymerase basic 2) and HA (hemagglutinin). Observation of the resulting sequence alignment revealed two divergent allelic patterns between informative sites in PB2 and HA, as reflected in incongruent trees and reticulate cycles of the phylogenetic network. We also analyzed the concatenated H1N1pdm genome (the 2009 pandemic influenza strain) and identified two nontrivial cycles, suggesting that reassortment had taken place in humans. When we visually inspected informative sites, the results indicated potential reassortment of two viral strains, each contributing [PB2, M1, NS1] and [PB1, PA, HA]. Phylogenetic analysis supports these incompatibilities. Using one-dimensional ICR as a lower-bound estimate of reassortment rate, we calculated ICR < 1 event per year for classic H1N1 swine and H3N2 human influenza. In contrast, we calculated a high rate of 22.16 reassortments per year for avian influenza A. This difference could be explained by the high diversity and frequent co-infection of avian viruses and correlates with the high proportion of avian reassortants reported in previous studies.

Applying persistent homology to avian influenza, we also investigated whether gene segments cosegregate more than expected by chance. After considering all pairs of concatenated segments, we estimated the number of reassortments with a Betti number *b1*. We then tested for enriched reassortment between pairs of segments given the total estimate of reassortments in the concatenated genome. We found statistically significant configurations of four cosegregating segments: PB2, polymerase basic 1 (PB1), polymerase acidic (PA), and nucleoprotein (NP). This pattern is consistent with previous in vitro results suggesting that protein-protein interactions between the polymerase complex and NP protein constrain reassortment.

Persistent homology also offers an efficient method for real-time analysis of viral evolution. Following the April 2013 outbreak of H7N9 avian influenza in the Jiangsu province in China, Gao et al. constructed a series of trees per gene and observed conflicting structure by eye.[13,14] They determined that the novel virus was a triple reassortment of an H7N3 A/duck/Zhejiang/12/2011-like lineage, an H7N9 A/wild bird/Korea/A14/2011-like lineage, and an H9N2 A/brambling/Beijing/16/2012-like lineage donating HA, NA, and internal segments, respectively. Persistent homology of concatenated H7N9, H9N2, and H7N3 avian genomes identified a 2D irreducible cycle representing the H7N9 triple reassortment, representing it as a 2D cavity enclosed by a six-sided b2 polytope formed by joining two tetrahedra at the top and bottom. The top tetrahedron is formed at the apex by H7N9 reassortants and at the base by members of the three parental lineages, providing a visually interpretable representation of horizontal viral evolution.

Persistent homology is applicable not only to influenza, but also to a variety of other viruses that undergo reticulate evolution. The retrovirus HIV, for instance, has frustrated efforts to control it because of its high mutation rate and frequent homologous recombination, which leads to antiretroviral resistance.[15] This suggests the need for nonphylogenetic methods for characterizing the virus's evolution. When we applied our method to the independent and concatenated alignments of the HIV-1 genes gag, pol, and env, we identified one-dimensional topology, which suggests the presence of horizontal evolution. However, individual gene alignments also revealed one-dimensional homology groups, suggesting that recombination breakpoints exist within as well as between individual genes. In addition, persistent homology of the same concatenated segments produced 2D topology derived from complex recombination events. In studies of other viruses, we found high-dimensional topology for hepatitis C virus, but no high-dimensional structure for dengue or West Nile virus, suggesting a high rate of recombination in the first, but very low rates in the latter two, respectively.

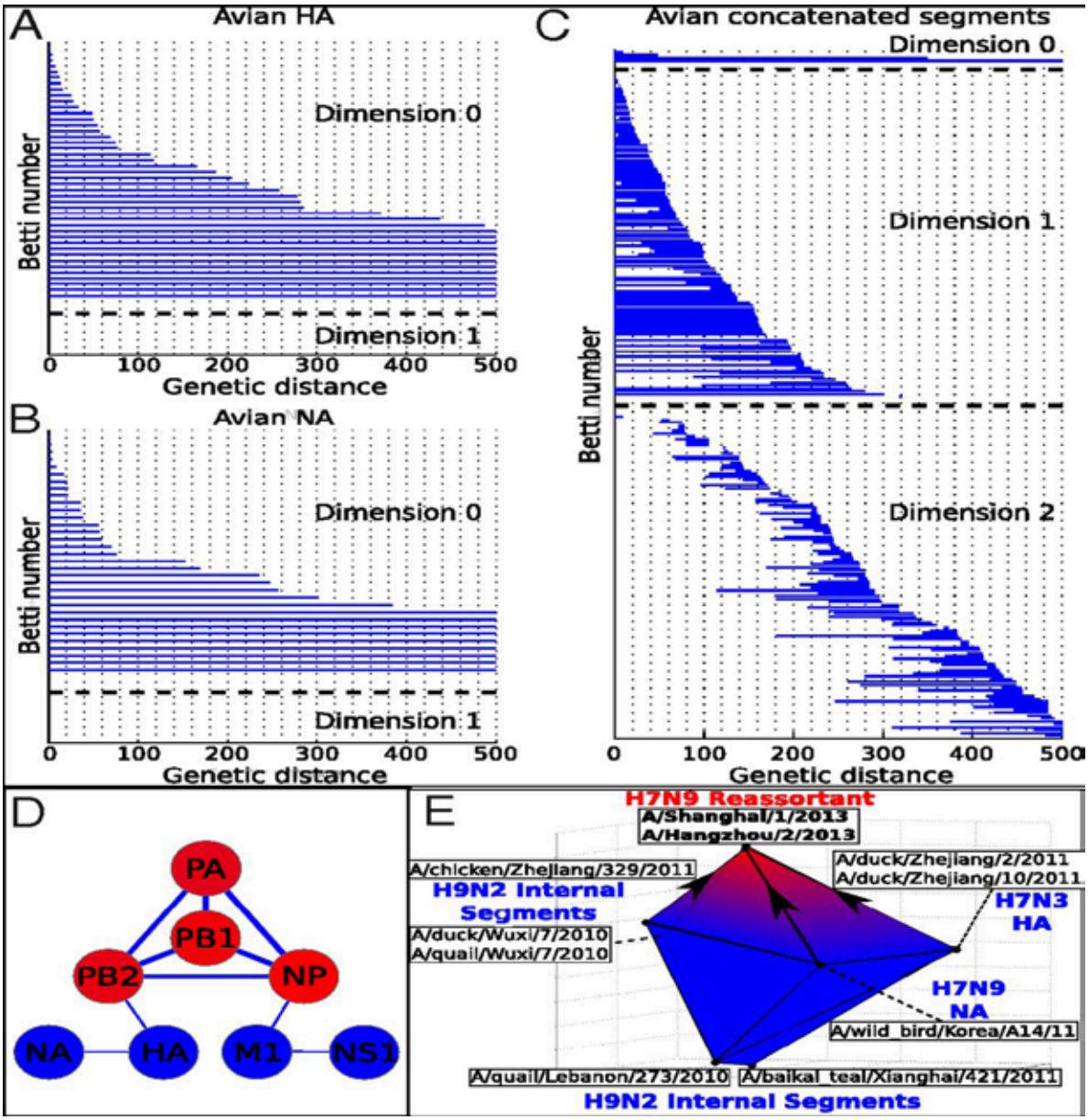


Figure 2: Persistent homology of reassortment in avian influenza. Analysis of **(A)** HA and **(B)** NA reveal no significant one-dimensional topological structure. **(C)** Concatenated segments reveal rich 1D and 2D topology, indicating reassortment. **(D)** Network representing the reassortment pattern of avian influenza deduced from high-dimensional topology. Line width is determined by the probability that two segments reassort together. Node color ranges from blue to red, correlating with the sum of connected line weights for a given node. **(E)** b2 polytope representing the triple reassortment of H7N9 avian influenza. Concatenated genomic sequences forming the polytope were transformed into 3D space using PCoA. Two-dimensional barcoding was performed using Vietoris–Rips complex and a maximum scale ε of 4,000 nucleotides.

We find these results very encouraging, suggesting that persistent homology can offer a fast and effective method for understanding properties of evolution from the perspective of topology. As these results show, this approach makes it possible to characterize both clonal and reticular evolution within one framework. It can account for horizontal events such as recombination and reassortment as well as higher-dimensional, complex evolutionary patterns such as segment cosegregation; and can provide a lower bound for the recombination/reassortment rate.

Persistent homology offers a number of advantages over other methods based on the tree paradigm of evolution. Whereas many phylogenetic methods produce a single, possibly suboptimal, tree or network, persistent homology can identify invariant topological characteristics of all simplicial complexes across the entire parameter space of genetic distance. Our methodology also shows stability to small fluctuations in the input data.

References

1. Nei M. Stochastic errors in DNA evolution and molecular phylogeny. Prog Clin Biol Res. 1986;218:133-47.

2. Doolittle WF. Phylogenetic classification and the universal tree. Science. 1999 Jun 25;284(5423):2124-9.

3. Fitch WM. Networks and viral evolution. J Mol Evol. 1997;44 Suppl 1:S65-75.

4. Holland BR, Huber KT, Moulton V, Lockhart PJ. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol Biol Evol. 2004 Jul;21(7):1459-61.

5. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006 Feb;23(2):254-67.

6. Morrison DA. Introduction to Phylogenetic Networks. Uppsala: RJR Productions; 2011. p. 216.

7. Kanj IA, Nakhleh L, Than C, Xia G. Seeing the trees and their branches in the network is hard. Theor Comput Sci. 2008 Jul 23;401(1-3):153-64.

8. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. Proceedings 41st Annual Symposium on Foundations of Computer Science; 2000; IEEE, Washington, DC; pp 454–463.

9. Zomorodian A, Carlsson G (2005) Computing persistent homology. Discrete Comput Geom. 2005;33(2):249–274.

10. Collins A, Zomorodian A, Carlsson G, Guibas LJ (2004) A barcode shape descriptor for curve point cloud data. Comput Graph-Uk 2004;28(6):881–894.

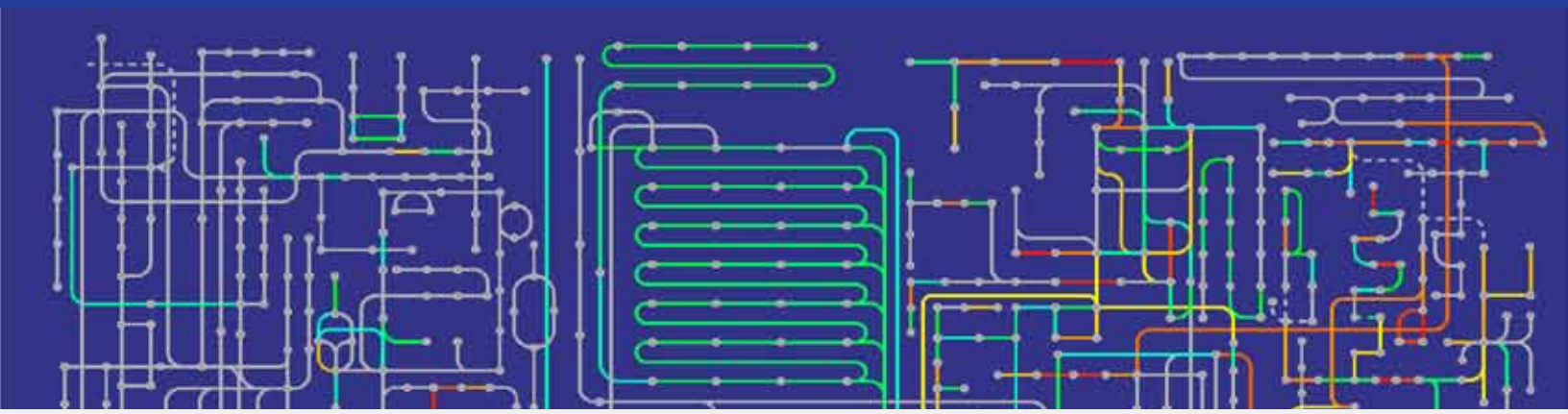
11. Fouchier RA, Munster V, Wallensten A, Bestebroer TM, Herfst S, Smith D, Rimmelzwaan GF, Olsen B, Osterhaus AD. Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. J Virol. 2005 Mar;79(5):2814-22.

12. Hifumi E, Fujimoto N, Ishida K, Kawawaki H, Uda T (2010) Characteristic features of InfA-15 monoclonal antibody recognizing H1, H3, and H5 subtypes of hemagglutinin of influenza virus A type. J Biosci Bioeng 109(6):598–608.

13. Hifumi E, Fujimoto N, Ishida K, Kawawaki H, Uda T (2010) Characteristic features of InfA-15 monoclonal antibody recognizing H1, H3, and H5 subtypes of hemagglutinin of influenza virus A type. J Biosci Bioeng. 2010 Jun;109(6):598-608.

14. Lycett SJ, Baillie G, Coulter E, Bhatt S, Kellam P, McCauley JW, Wood JL, Brown IH, Pybus OG, Leigh Brown AJ; Combating Swine Influenza Initiative-COSI Consortium. Combating Swine Influenza Initiative-COSI Consortium (2012) Estimating reassortment rates in co-circulating Eurasian swine influenza viruses. J Gen Virol. 2012 Nov;93(Pt 11):2326-36.

15. Nora T, Charpentier C, Tenaillon O, Hoede C, Clavel F, Hance AJ. Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment. J Virol. 2007 Jul;81(14):7620-8.

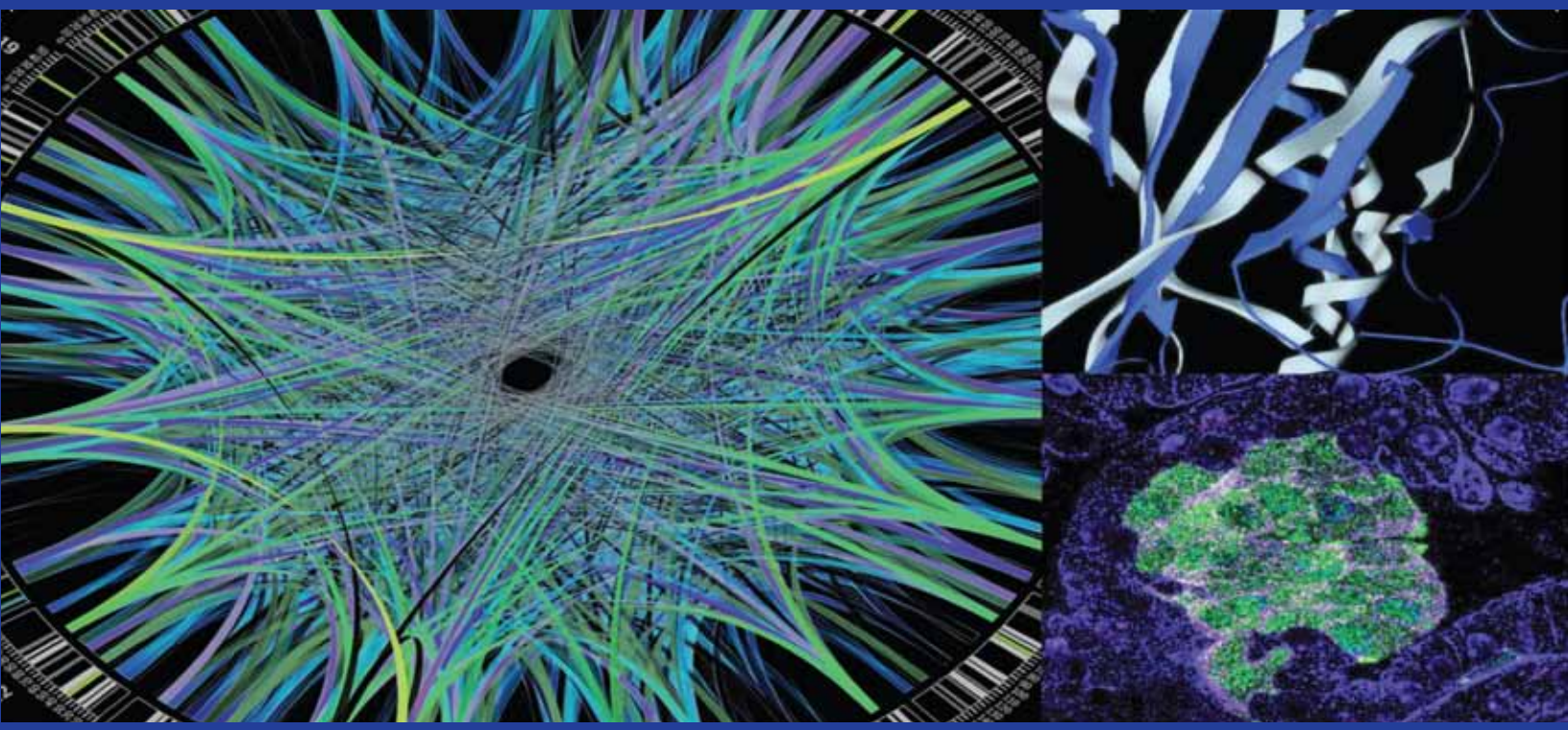


The Columbia University Department of Systems Biology

Research · Education · Technology

Moving toward the future of biomedical science

COLUMBIA UNIVERSITY MEDICAL CENTER



A Vision for the Future of Biomedical Research

The mission of the Columbia University Department of Systems Biology is to develop new methods for understanding the biological world from a systems perspective, particularly at the genomic and molecular levels. Our faculty shares a common interest in combining high-throughput experimentation, quantitative analysis, and innovative technology development. Researchers use computational approaches and laboratory experimentation in an iterative way, developing predictive models of biological systems, and then validating them in the laboratory.

Some facts about the Columbia University Department of Systems Biology

- With more than two-dozen faculty members, a community of more than 250 individuals working in department labs, and an outstanding record of publications, **the Department of Systems Biology is among the largest programs in this field**.

- **The Department is highly multidisciplinary**. Our faculty's research interests include bioinformatics, biophysics, cancer biology, chemical biology, genetics and genomics, machine learning, microbiology, molecular evolution, stem cell and developmental biology, structural biology, synthetic biology, and virology and immunology.

- **The Department is home to six centers of excellence in systems and computational biology**, including the Center for Multiscale Analysis of Genomic and Cellular Networks (MAGNet). MAGNet is one of eight National Centers for Biomedical Computing and one of 12 Integrated Cancer Biology Programs funded by the National Cancer Institute.

- Department members conduct collaborative research with investigators in the Herbert Irving Comprehensive Cancer Center, other institutes at Columbia University, and at other universities. In addition to producing new insights into basic biology, **many of these collaborations utilize approaches from systems and computational biology to study human diseases** such as cancer, infectious diseases, metabolic disorders, and psychiatric illnesses.

- The Department manages the JP Sulzberger Columbia Genome Center, which provides **a state-of-the art infrastructure for next-generation genome sequencing and high-throughput molecular screening**. The Genome Center supports researchers inside the department, across the Columbia University community, and at other institutions. The Genome Center is also the official high- throughput screening and chemistry core for NYSTEM, the New York State foundation for stem cell research.

- **We oversee graduate education and postdoctoral training in systems and computational biology to promote their use in biological research**. Our educational efforts include a training grant from the National Institutes of Health. Upon graduation, our alumni have a strong track record of launching successful careers in both academia and industry.

A GLOBAL VIEW OF TUMOR-INDUCED GENE EXPRESSION CHANGES IN HUMAN METABOLISM

DENNIS VITKUP
DEPARTMENT OF SYSTEMS BIOLOGY, COLUMBIA UNIVERSITY

One of the defining features of all cancers is the ability of tumor cells to divide and grow uncontrollably. In order to overcome the normal mechanisms that keep proliferation in check, cancer cells undergo a reconfiguration of the molecular networks that control metabolism. These changes are essential for synthesizing proteins that are necessary for cell survival and for generating the energy necessary for cell growth. More than 80 years ago Otto Warburg identified the importance of metabolism in cancer, finding that tumor cells often undergo a shift from oxidative to fermentative metabolism[1]. Recently, studies of signaling and gene regulatory pathways have shown that many key cancer signaling pathways are also key regulators of metabolic networks, prompting renewed interest in cancer metabolism.[2-4] This line of research is particularly enticing because identifying ways of selectively interrupting metabolism in cancer cells could lead to strategies for shutting down a tumor's energy supply and controlling disease.[5,6]

Exploring cancer metabolism from the perspective of metabolic networks is also timely because the accumulation of large collections of gene expression profiles from many tumor types over the last decade has produced a valuable resource for investigating this domain.[7,8] Recently, we undertook a comprehensive study of tumor-induced changes in mRNA expression of human metabolic genes, looking at 22 diverse cancer types. Using more than 2500 microarray measurements based on biopsies of primary tumors, our goal was to determine the range of variability in human metabolic networks at several layers of biochemical organization: at the global network level, at the level of individual biochemical pathways and networks, and at the level of single enzymatic reactions. Our findings provide the first genome-wide perspective on how metabolic networks become transformed when cells become cancerous, as well as a number of insights into specific tumor-induced enzymatic changes within metabolic networks.

Once our collection of gene expression data had been compiled, our first step was to investigate how metabolic gene expression changes in tumors, and how these changes compare with the corresponding gene expression patterns in normal cells of the same tissue. We found that differences in gene expression between normal and cancerous tissue are much greater than between different samples of the same normal tissue or between different samples of the same tumor types. However, differences in gene expression between tumor and normal cells of the same cell type are much smaller than differences in that seen between different tumor types. Moreover, differences in gene expression between different tumor types are smaller than differences between different types of normal tissue. All of this suggests that metabolic expression patterns are more similar between tumors than between normal tissues, but also that much of the expression pattern in normal tissue is retained even after a cell becomes cancerous.

To see how gene expression changes at the level of individual biochemical pathways, we compared expression signatures in tumors to those associated with biochemical pathways as defined in the KEGG database, a resource for the interpretation of large-scale biological datasets.[9] When we calculated the average fraction of tumor samples in which each metabolic pathway was significantly up-regulated and down-regulated across our 22 cancer types, we found that pathways responsible for the production of molecular components that are essential for cell division, such as pyrimidine and purine biosynthesis, are significantly up-regulated in many tumor samples. We also identified up-regulation in the glycolysis pathway, which reflects the enhanced glucose uptake often seen in tumors, and in pathways related to protein synthesis and glycoprotein biosynthesis. In contrast, pathways that are necessary for the degradation of essential amino acids, cofactors, and fatty acids are frequently and significantly down-regulated.

Analysis of biochemical pathways also produced a number of specific insights. For example, we saw down-regulation in the xenobiotic and drug metabolism pathways that might contribute to increased sensitivity of cancer cells to chemotherapy. We also saw heterogeneous behavior in the oxidative phosphorylation and TCA cycle pathways, which may reflect how different cancers adapt to tissue-specific physiological conditions such as hypoxia, nutrient availability, or the specific lesions driving a tumor type. In addition, oxidative phosphorylation was not only different between tumor

types, but also varied between tissue samples of the same tumor type. This suggests that this pathway is sensitive to specific physiological conditions and/or the genetic composition of specific tumors in

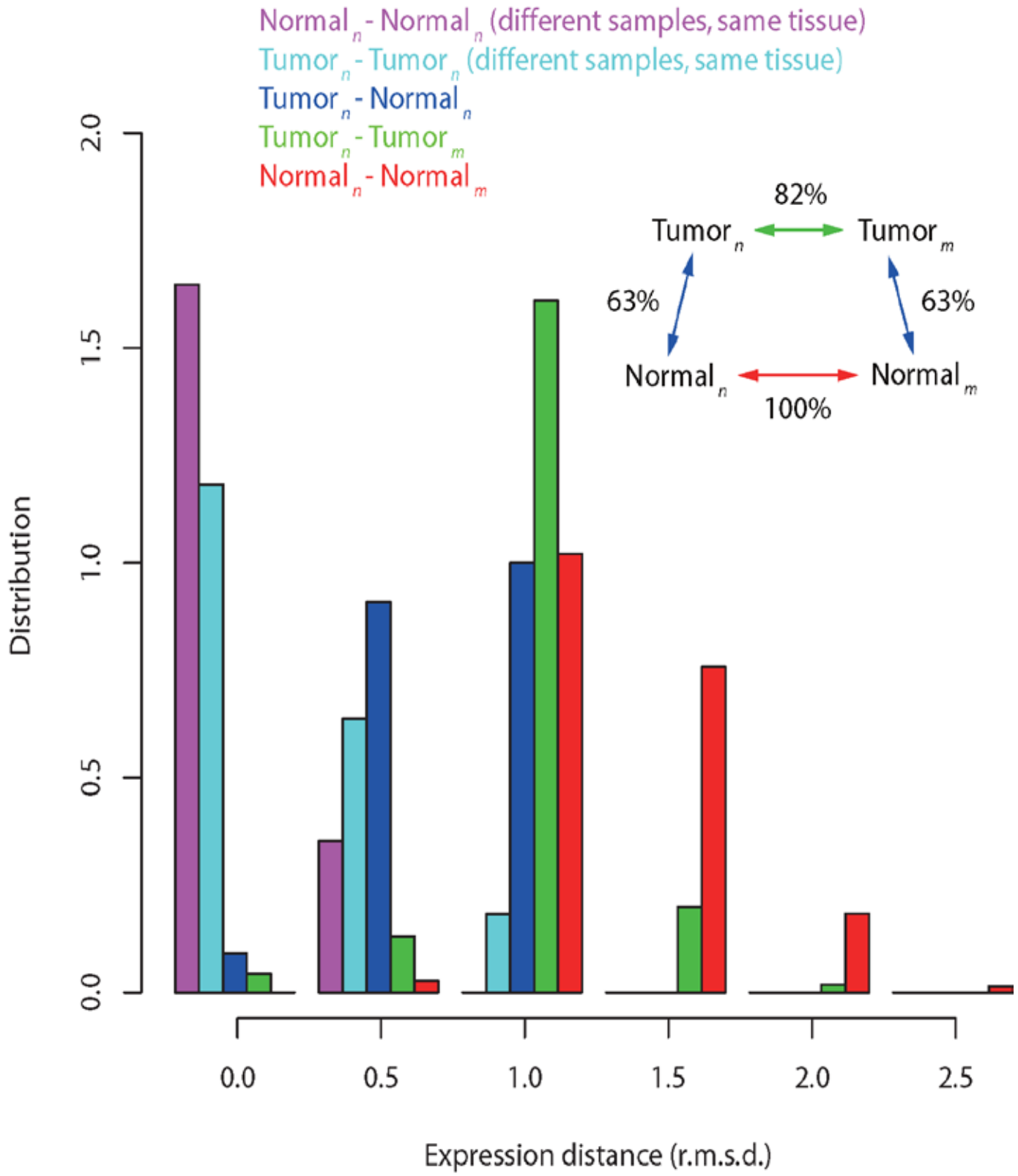


Figure 1. Colors represent distributions of the Euclidean expression distance (root mean squared deviation, r.m.s.d.) between different samples of identical normal tissues ($Normal_n$-$Normal_n$, magenta), different samples of identical tumors ($Tumor_n$-$Tumor_n$, cyan), tumors and corresponding normal tissues ($Tumor_n$-$Normal_n$, blue), different tumors ($Tumor_n$-$Tumor_m$, green) and different normal tissues ($Normal_n$-$Normal_m$, red). The distributions were binned for display purposes only. The inset summarizes the average distances between pairs of tissues as a percentage of the average distance between two different normal tissues.

individual patients.

We next investigated correlations between the expression of metabolic pathways and expression of signaling and regulatory genes frequently involved in tumorigenesis. This analysis revealed high mutual information between the oxidative phosphorylation pathway and the hypoxia-inducible factor (*HIF1A*) and its negative regulator *RBX1*. Notably, oxidative phosphorylation expression was correlated with expression of *RBX1*, but anti-correlated with *HIF1A* expression, suggesting that expression of oxidative phosphorylation genes is influenced by oxygen availability in the tumor.

Individual metabolic pathways are highly interdependent, and so we next used principal component analysis (PCA) to better understand correlations between them in cancer.[10] We considered expression changes in nine meta-pathways representing major metabolic processes, capturing 85% of the meta-pathway expression variance. The first principal component accounts for ~62% of the variance, representing an approximately uniform shift in the overall expression of metabolic genes. It is likely that these shifts reflect a loss of function in normal metabolic activity and a switch to a cancer metabolism program. Shifts along the second principal component, accounting for ~16% of the variance, involve a change in the expression of glycolysis and nucleotide biosynthesis. At the same time, there is an opposite change in the expression of three catabolic pathways, suggesting that dividing cells increasingly rely on glycolysis. Shifts along the third principal component (~7%) involve a strong change in the expression of oxidative phosphorylation with a concomitant opposite change in nucleotide biosynthesis.

Next we studied differences in expression at the level of individual biochemical reactions, focusing on the function of isoenzymes in the human metabolic network. Isoenzymes are enzymes that can be encoded by different genes or arise from splice variants of the same gene but play the same catalytic role in metabolic reactions. Because different kinetic and regulatory properties of isoenzymes are finely tuned to meet specific metabolic requirements of various human tissues, we investigated variation of isoenzyme expression in tumors. Using an information theoretic measurement

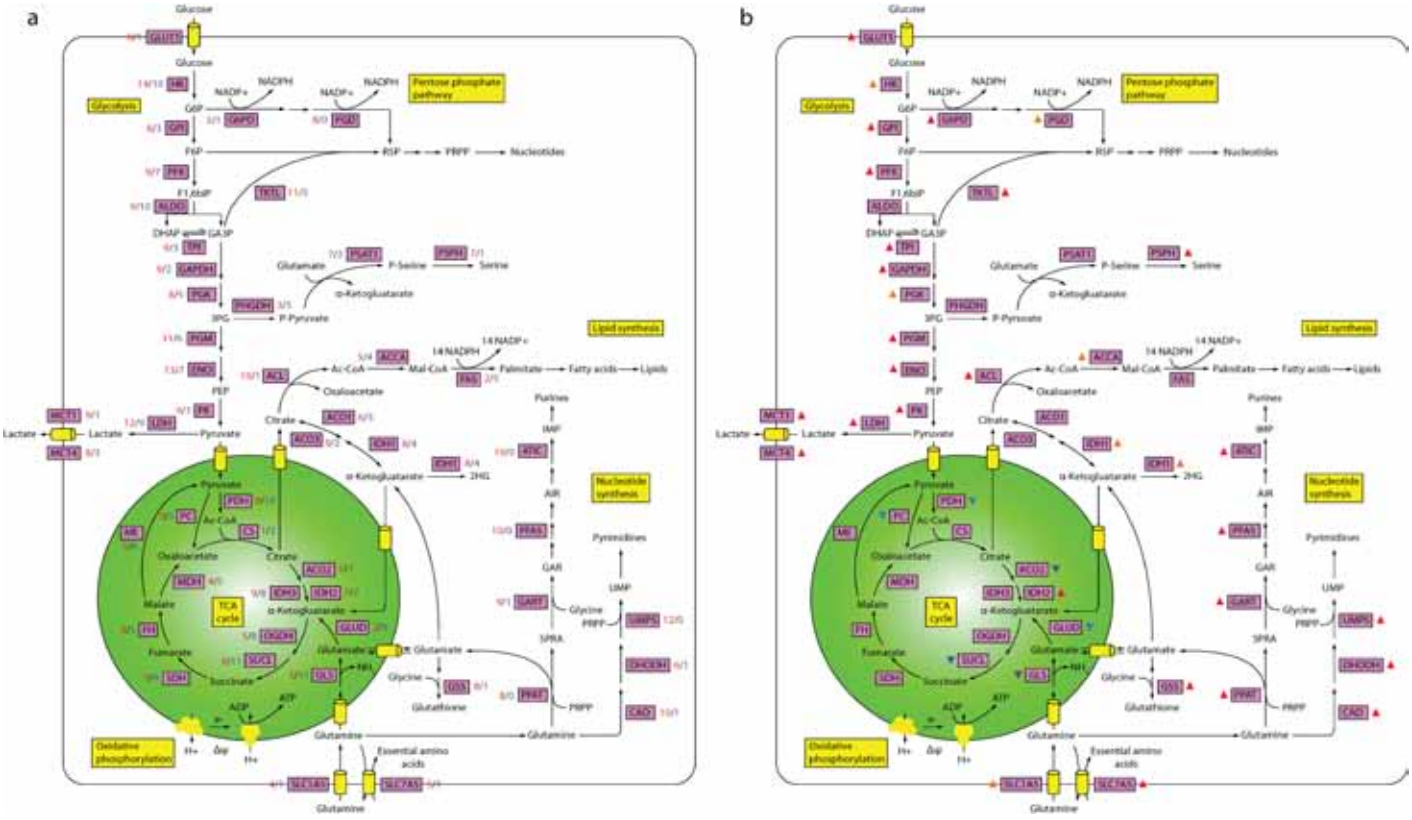


Figure 2. **(a)** Each metabolic reaction is marked with the number of tumors (out of 22 considered in our analysis) in which at least one isoenzyme catalyzing the corresponding reaction is significantly (FDR-corrected, $P < 0.05$) upregulated (red) and downregulated (blue). **(b)** Reactions that are significantly upregulated (red triangles) or downregulated (blue triangles) when all isoenzymes and members of the corresponding protein complexes are considered together across all tumors (deep red or deep blue, FDR-corrected, $P < 0.05$; orange or light blue, FDR-corrected, $P < 0.1$). If unmarked, no statistically significant change in mRNA expression was detected.

called the Kullback-Leibler divergence, we found that, on average, the relative expression patterns of isoenzymes are about two times more similar for different samples of identical normal tissues than for different samples of identical tumors. However, both of these distances are significantly smaller than the average distance between isoenzyme expression patterns in tumors and corresponding normal tissues, suggesting that for many biochemical reactions the transition of a cell to become cancerous leads to a significant shift in the relative expression of isoenzymes. To identify specific isoenzymes with frequently perturbed expression profiles, we calculated, for each isoenzyme in every biochemical reaction, the number of tumors in which the fractional expression of one isoenzyme among all isoenzymes catalyzing the same reaction is significantly up-regulated. We found 919 isoenzymes that were relatively up-regulated in at least one tumor type, and 322 that were up-regulated in more than 25% of the 22 tumor types we investigated.

Further investigation revealed that the isoenzymes *IDH1* and *IDH2* are frequently up-regulated in brain cancers and lymphoma, and that germline and somatic loss-of-function mutations in fumarate hydratase (*FH*) and three subunits of succinate dehydrogenase (*SDH*) are present in renal cell carcinoma and other cancers. Analysis also identified a previously unknown role of *SDH* and *FH* in colorectal cancer. We confirmed these computationally derived results by measuring and analyzing concentrations of specific metabolites from 10 colon cancer patients.

Notably, this wide range of findings provides strong evidence that there is no uniform change in cancer-induced changes in metabolic genes across all cancer types. At every level that we investigated, we found heterogeneous metabolic gene expression that is similar to the high variability between genetic and expression changes in signaling and regulatory networks. This said, we also found that changes in the expression of metabolic genes are not random in different samples of the same tumor, but are highly replicable. Indeed, we identified three general principles that characterize all cancer-induced changes in metabolic networks. First, tumors often retain a significant imprint of the metabolic expression patterns present in corresponding healthy tissues. Second, a large fraction of the variance in the expression of major biochemical processes can be rationalized in terms of several principal components, representing important expression modes for key metabolic processes. Third, we found many hundreds of isoenzymes that show significant and tumor-specific expression changes. Many of these are likely to be functionally important, and may offer potential drug targets. Taking advantage of these opportunities, however, will require additional analysis of the essential, context-specific metabolic transformations that take place in each specific cancer type

References

1. Warburg O, Posener K, Negelein E. On the metabolism of carcinoma cells. Biochem Z. 1924;152:309–344.

2. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science. 2009 May 22;324(5930):1029-33.

3. Deberardinis RJ, Sayed N, Ditsworth D, Thompson CB. Brick by brick: metabolism and tumor cell growth. Curr Opin Genet Dev. 2008 Feb;18(1):54-61.

4. Hsu PP, Sabatini DM. Cancer cell metabolism: Warburg and beyond. Cell. 2008 Sep 5;134(5):703-7.

5. Anastasiou D, et al. Pyruvate kinase M2 activators promote tetramer formation and suppress tumorigenesis. Nat Chem Biol. 2012 Oct;8(10):839-47.

6. Le A, et al. Glucose-independent glutamine metabolism via TCA cycling for proliferation and survival in B cells. Cell Metab. 2012 Jan 4;15(1):110-21.

7. Barrett T, et al. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res. 2011 Jan;39(Database issue):D1005-10.

8. Parkinson H, et al. ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res. 2009 Jan;37(Database issue):D868-72.

9. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010 Jan;38(Database issue):D355-60.

10. Jolliffe IT. Principal component analysis. 2nd Edn. New York: Springer. 2002

FEATURE SELECTOR FOR BREAST CANCER PROGNOSIS

DIMITRIS ANASTASSIOU LAB

The winning model of the Sage Bionetworks/DREAM Breast Cancer Prognosis Challenge [1,2] made use of several molecular features, called attractor metagenes, as well as another metagene defined by the average expression level of the two genes *FGD3* and *SUSD3*. As part of a follow-up study aimed at developing a breast cancer prognostic test derived from and improving upon that model, we designed a feature selector facility that calculates prognostic scores of combinations of features. These features included those that we had used earlier, as well as those used in existing breast cancer biomarker assays (http://www.ee.columbia.edu/~anastas/featureselector).

The prognostic score displayed for each combination of selected features was designed to be resistant to overfitting. It is evaluated as the asymptotic average of the concordance indices resulting from random 2-fold cross-validation experiments in the METABRIC data set. Each experiment uses the selected features as covariates to train a Cox proportional hazards model on half of the data set based on random splitting, and evaluates the corresponding concordance index of the fitted model on the other half. Our analysis leads to the unexpected and remarkable suggestion that *ER*, *PR*, and *HER2* status or molecular subtype classification do not provide additional prognostic value in breast cancer when the values of the *FGD3, SUSD3* and attractor metagenes are taken into account.

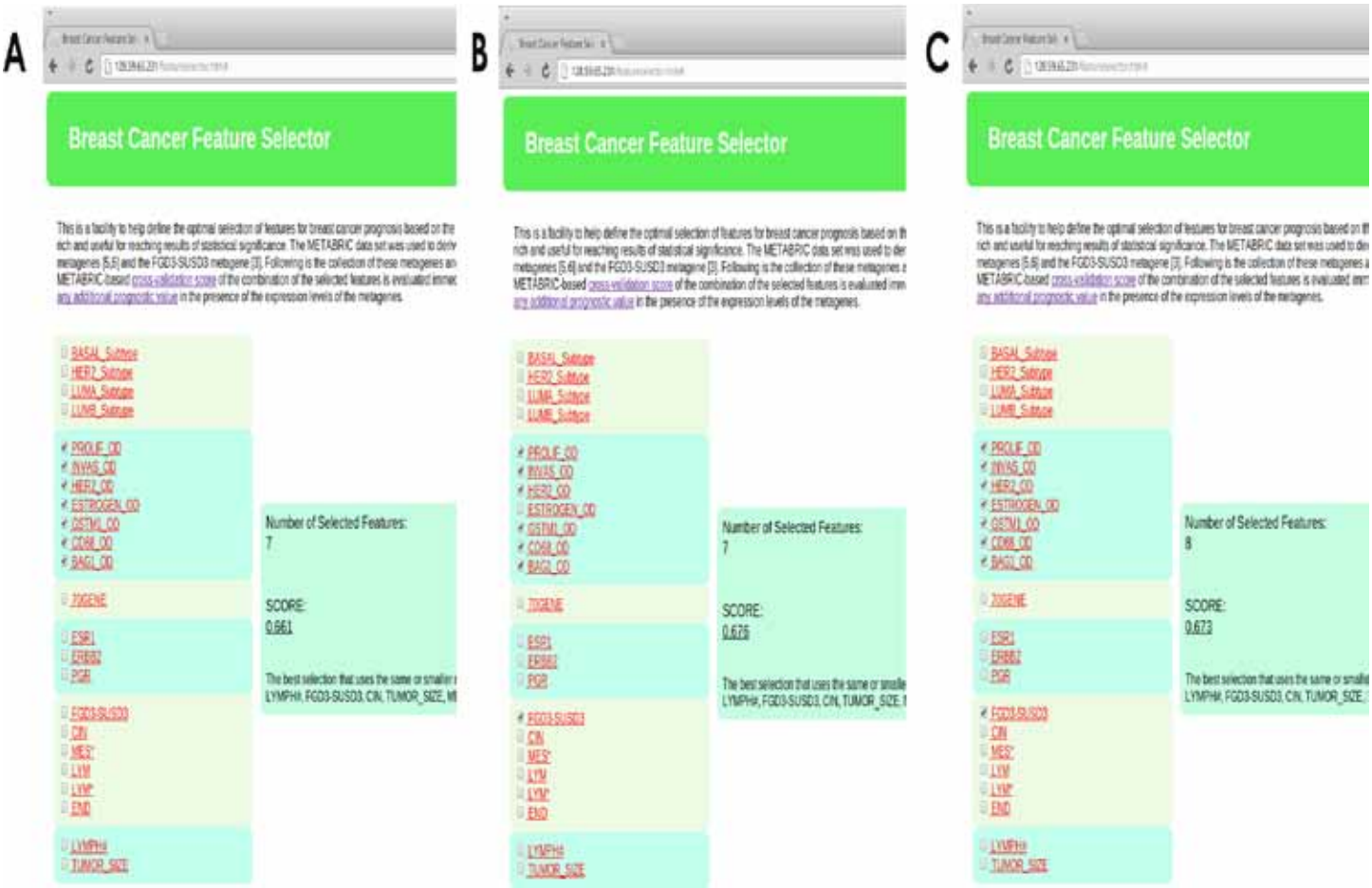


Figure 1: The screenshot selects **(A)** all Oncotype DX features, **(B)** Oncotype DX features replacing the Estrogen group with the FGD3-SUSD3 metagene, **(C)** Oncotype DX features with both the Estrogen group and the *FGD3-SUSD3* metagene. The scores suggest that replacing the Estrogen group with *FGD3-SUSD3* improves performance, and furthermore that the Estrogen group does not provide any additional information.

REFERENCES

1. Cheng, W.Y., T.H. Ou Yang, and D. Anastassiou, Development of a prognostic model for breast cancer survival in an open challenge environment. Sci Transl Med, 2013. 5(181): p. 181ra50.

2. McCarthy, N., Prognostic models: rising to the challenge. Nat Rev Cancer, 2013. 13(6): p. 378.

MAPPING TUMORIGENESIS MECHANISMS USING INSERTIONAL MUTAGENESIS

HARMEN BUSSEMAKER LAB

Building on their innovative methodology for mapping genetic loci that genetically modulate transcription factor activity [1], we recently developed a new method for analyzing parallel genomewide insertion and expression data for a panel of mouse tumors generated by mouse insertional mutagenesis. This research was performed in close collaboration with the Netherlands Cancer Institute. Each individual tumor harbored a unique combination of genetic lesions, which together were responsible for the aberrant behavior of its cells. Graduate student Eunjee Lee developed a novel computational methodology named locus expression signature analysis (or LESA). It integrates information at the genetic and molecular level to construct a genomewide signature that models the effect of an individual genetic lesion on the gene regulatory network of the cell. The signatures were exploited to gain insight into the regulatory pathways perturbed by each lesion, and predict therapeutic locus-drug combinations [2].

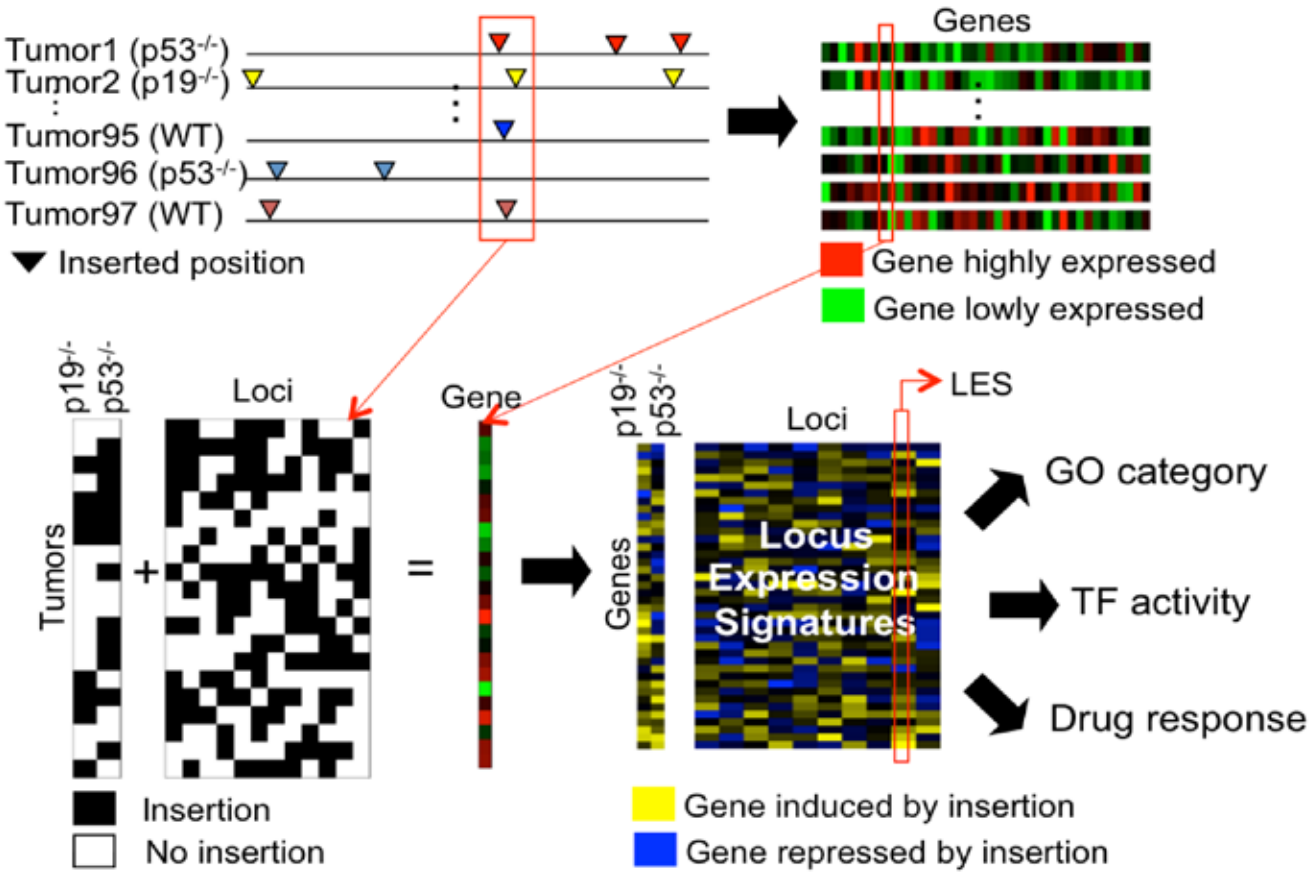


Figure 2: Locus Expression Signature Analysis. Proviral insertions into the mouse genome contribute to tumorigenesis. Recurring mutations each have a characteristic gene expression response associated with them, which is inferred by the locus expression signature analysis (LESA) algorithm. These signatures are further analyzed to identify regulatory mechanisms underlying tumorigenesis.

REFERENCES

1. Lee, E. and H.J. Bussemaker, Identifying the genetic determinants of transcription factor activity. Mol Syst Biol, 2010. 6: p. 412.

2. Lee, E., J. de Ridder, J. Kool, L.F. Wessels, and H.J. Bussemaker, Identifying regulatory mechanisms underlying tumorigenesis using locus expression signature analysis. Proc Natl Acad Sci U S A, 2014. 111(15): p. 5747-52.

A NEW KEY REGULATOR OF AGING AND LONGEVITY

HARMEN BUSSEMAKER LAB

Aging is a fundamental feature of all life, yet remains largely an unsolved problem of biology. Using the nematode C. *elegans* as a model, and using a combination of computational and experimental approaches, Ronald Tepper in the Bussemaker Lab, in close collaboration with the laboratory of Coleen Murphy at Princeton University, discovered that the little-studied transcription factor *PQM-1* is a key regulator of development and longevity[1]. *PQM-1* complements the well-known and conserved aging transcription factor *DAF-16/FOXO* in many important respects. Both are transcriptional activators controlling distinct sets of target genes (stress response vs. growth). Nuclear localization of *DAF-16* or *PQM-1* is controlled by the insulin/*IGF-1* signaling pathway, but in opposite ways. Thus only one of the factors is active at any given time, depending on the conditions (e.g. low nutrients) and genetic background (e.g., loss of the *DAF-2* receptor). The two factors interact in an essential way, each opposing the nuclear localization of the other, and both become cytoplasmic with advancing age.

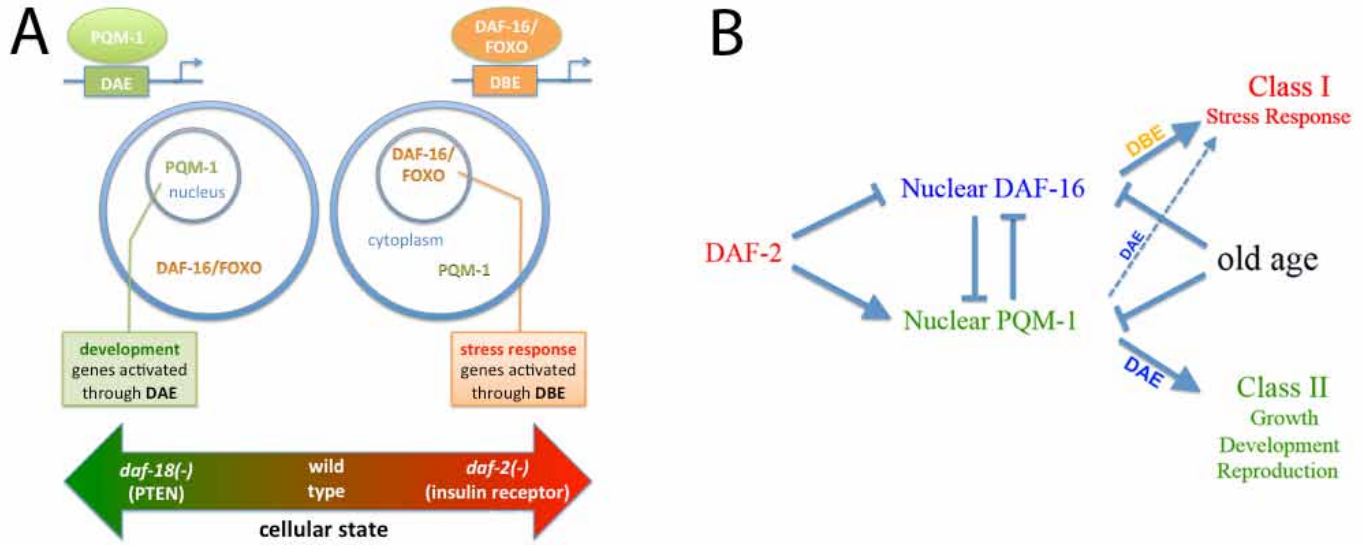


Figure 3: **(A)** Cellular states of short-lived wild-type vs. long-lived mutant animals. On the left, *PQM-1* is nuclear and transcriptionally activates genes for growth and development. On the right, loss of *DAF-2*/insulin signaling causes *PQM-1* to translocate to the cytoplasm, while *DAF-16/FOXO* enters the nucleus where it activates stress response genes. The animals in this state age more slowly and live twice as long as the wild-type ones. **(B)** Schematic diagram summarizing the findings of the paper, and the interplay between *DAF-16* and *PQM-1*. (Figure adapted from [1], with permission.)

REFERENCES

1. Tepper, R.G., J. Ashraf, R. Kaletsky, G. Kleemann, C.T. Murphy, and H.J. Bussemaker, PQM-1 complements DAF-16 as a key transcriptional regulator of DAF-2-mediated development and longevity. Cell, 2013. 154(3): p. 676-90.

SINGLE-CELL APPROACH PROVIDES MAP OF HUMAN B CELL DEVELOPMENT

DANA PE'ER LAB

Many human diseases result from malfunctions in the molecular programs that control development. However, identifying these aberrations has been difficult due to limitations in our basic understanding of the molecular programs that drive development in normal, healthy cells. This problem is compounded by the fact that key regulatory steps in development occur within rare precursor populations that have yet to be characterized. Developmental disease often results from malfunction in these uncharacterized, rare (~1/10,000 cells), hard-to-isolate variants.

We recently undertook a collaboration with Garry Nolan at Stanford University aimed at developing a high-resolution approach to mapping development. Our approach combined an emerging single-cell technology called mass cytometry with a new algorithm called Wanderlust, which utilizes concepts from graph theory to analyze single-cell data. Using mass cytometry, we simultaneously recorded the expression of 44 molecular markers in a population of approximately 200,000 individual B cells taken from a single bone marrow sample. Using Wanderlust, we then converted the 44-dimensional measurement for each cell into a single value that corresponded to its place

within the chronology of development. By using nearest neighbor analysis to plot these values on a graph, Wanderlust correctly ordered all 200,000 cells, including all of the primary molecular landmarks known to be present in human B cell development. By comparing the changes in marker expression across the developmental trajectory, Wanderlust also pinpointed key regulatory signaling checkpoints that are required for B cell development, and identified novel subtypes of B cell progenitor cells that correspond to important developmental stages. The resulting map constitutes the most comprehensive analysis of human B cell development ever conducted.

Having an accurate map of normal cell development provides a critical framework for understanding the origins of developmental diseases, and should provide insights that could be used for the identification of new diagnostics and therapeutics. And because this approach can be applied not just to B cells, but also to any type of cell, it should provide a compass capable of guiding research in regenerative medicine.

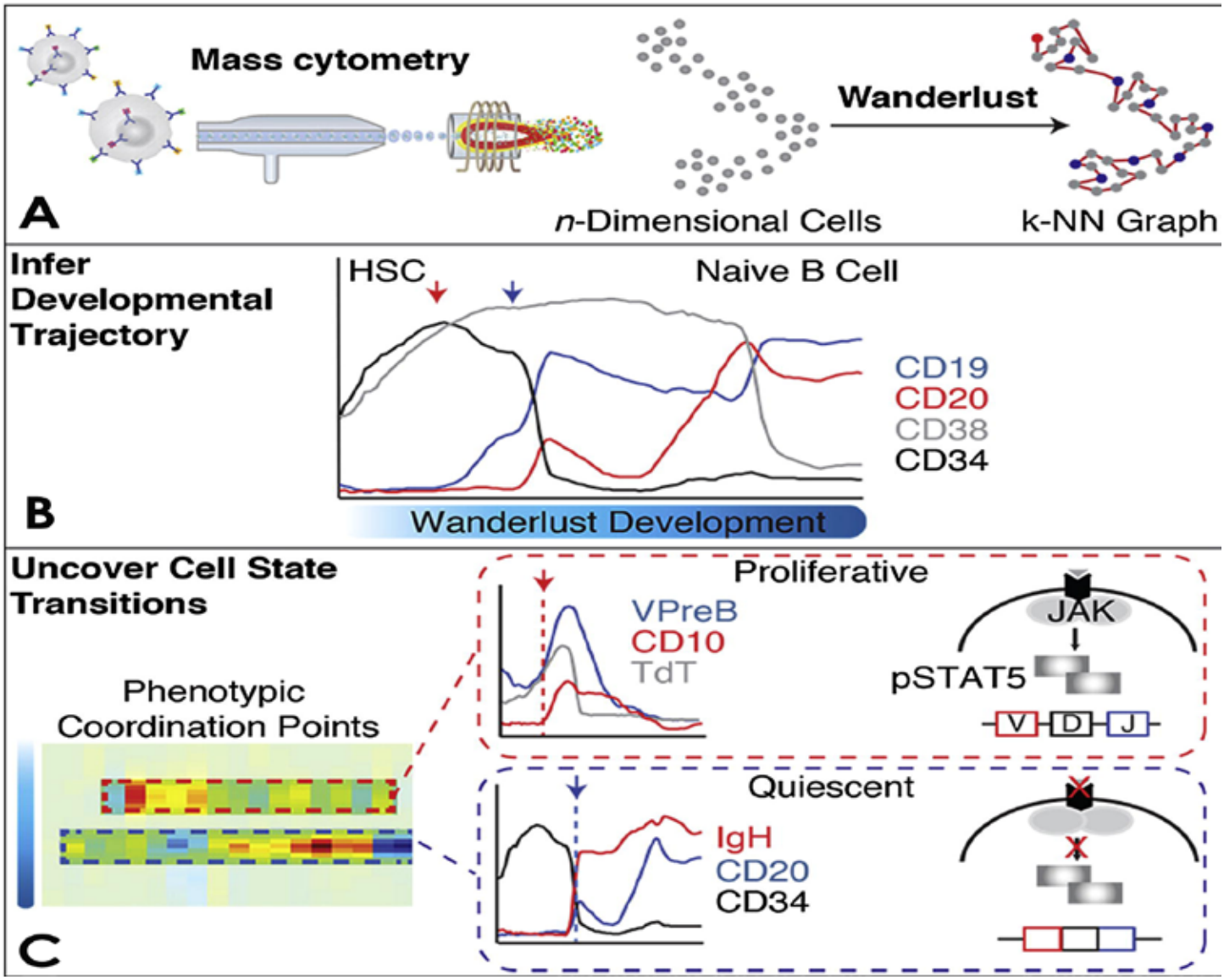


Figure 4: Investigating development in single cells. In **(A)** individual cells are profiled using mass cytometry, producing a high-dimensional data set in which *n* properties are measured in each cell. Wanderlust converts this data into a graph in which cells are represented as points connected to their most similar cells. Using this graph structure, Wanderlust is able to place each cell within its chronology of development. Once the cells have been aligned in the chronology, the plot in **(B)** shows how four key markers gradually change as a cell develops from a hematopoietic stem cell into a naïve B cell. In **(C)** the newly inferred trajectory reveals previously unrecognized rewiring of regulatory signaling and coordinated changes that can now be recognized as hallmarks of developmental progression.

IMAGE-GUIDED RNA-SEQ REVEALS SUBTYPE-SPECIFIC ALTERATIONS IN MOLECULAR AND CELLULAR COMPOSITION AT THE MARGINS OF GLIOBLASTOMA

PETER A. SIMS LAB

Glioblastoma (GBM), the most deadly type of malignant brain tumor, has been the subject of multiple large-scale genomic and expression analysis efforts. These studies revealed a discrete set of expression signatures or subtypes that stratify the majority of patients and co-occur with specific genetic alterations. This information, while extremely valuable, was obtained mainly from specimens isolated during surgical resection. Because GBMs are diffusely infiltrative, populations of transformed cells intermingle with brain tissue surrounding the tumor and are inevitably left behind after surgery. Despite the clinical importance of these cells, which are the targets of pharmacological intervention and the basis of inevitable recurrence, they have not been characterized systematically.

A team of researchers at the Columbia University Medical Center (CUMC) from the Departments of Neurological Surgery, Pathology and Cell Biology, and Systems Biology has developed and implemented methods for large-scale transcriptomic analysis of brain tumor tissue harvested from the infiltrative margins of GBMs. We obtained multiple biopsies from radiographically distinct regions of GBMs using the Brainlab Neuronavigation system from over 70 adult patients who received surgical resection for high-grade glioma at CUMC, and have conducted RNA-Seq and histological analysis on a subset of these specimens at the JP Sulzberger Columbia Genome Center. In addition, we have obtained non-neoplastic brain tissue from shunt placement procedures for comparison.

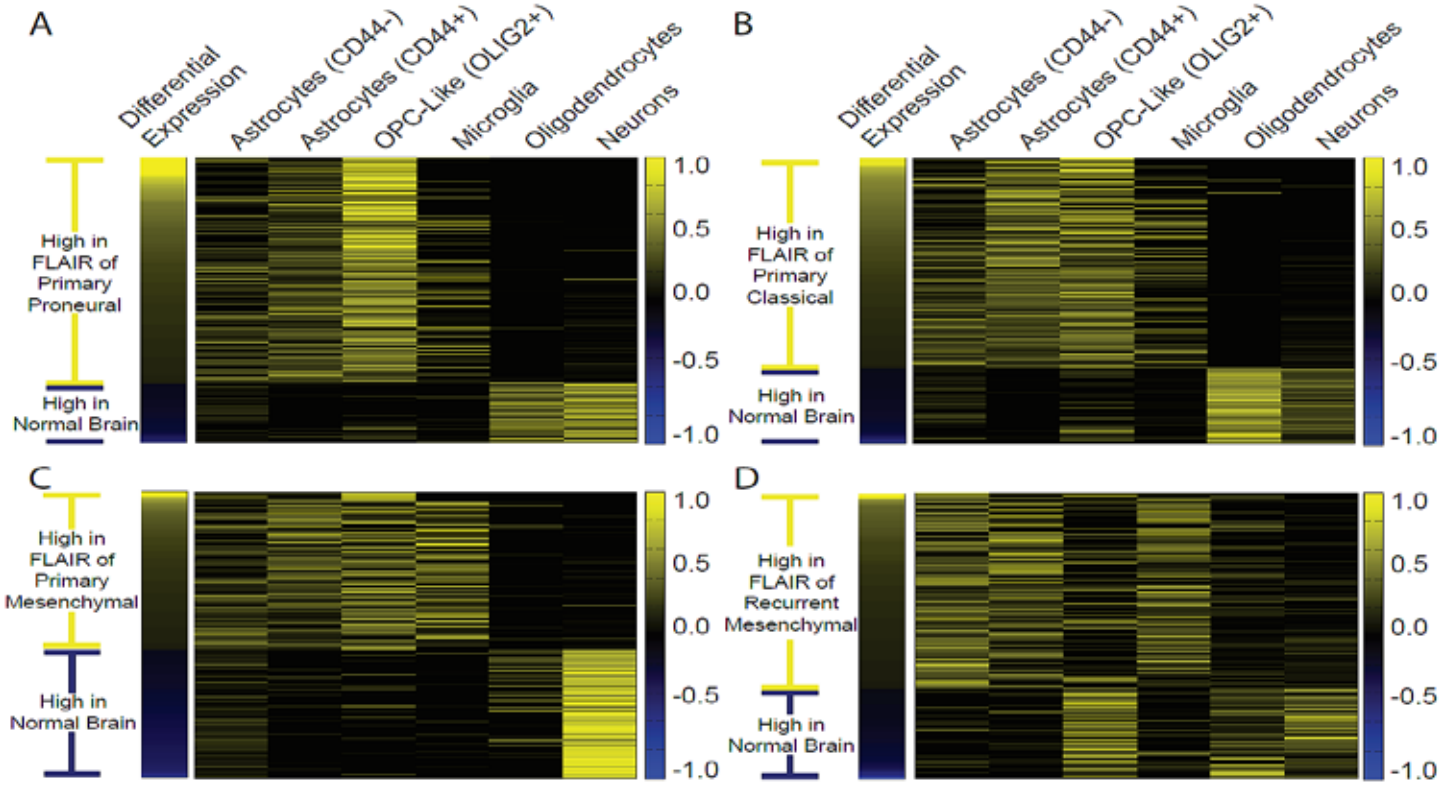


Figure 5. Heatmaps showing the deconvolved cellular distribution of differentially expressed genes compared to non-neoplastic brain tissue for the diffuse margins of **(A)** primary proneural, **(B)** primary classical, **(C)** primary mescenhymal, and **(D)** recurrent mesenchymal GBMs.

A central challenge in analyzing this data set has been the presence of multiple cell types in addition to the transformed populations, particularly in biopsies obtained from the diffuse tumor margins. By computationally deconvolving expression profiles using cell type-specific marker genes, we can generate estimates of the cellular distribution of each gene's expression level. The resulting deconvolved data set provides insights into which neural lineages (including the tumor cells) contribute to differential expression compared to non-neoplastic tissue. While we have only scratched the surface of the information that can be gleaned from our analysis, one feature was immediately obvious. The differentially expressed genes associated with each subtype of GBM exhibit distinct cellular distributions. Differentially expressed genes from the diffuse margins of proneural and classical tumors are largely expressed in Olig2+ cells, resembling glial progenitors, and in Cd44+ cells, resembling reactive astrocytes, respectively. However, the expression signature of primary mesenchymal tumors is distributed substantially in the microglial or monocytic lineage, which is a non-neoplastic cell type. Surprisingly, repeating this analysis on recurrent mesenchymal

tumors revealed a selective loss of the contribution from Olig2+ progenitor-like cells. These patients were treated with radiation and temozolomide, which appear to have a disproportionate effect on this specific lineage. Taken together, these results show that we can associate cell type-specific expression patterns in the diffuse margins of the tumor with distinct subtypes of GBM which are determined from the resected tumor core as well as with response to therapy. Going forward, we hope to use this growing resource to assess the cellular distribution of putative drug targets for post-surgical treatment.

IDENTIFICATION OF 3'-UTR REGULATORY ELEMENTS IN HUMAN TRANSCRIPTS

SAEED TAVAZOIE LAB

The post-transcriptional fate of messenger RNAs contributes substantially to the requisite repertoire of protein products across cellular phenotypes. The underlying post-transcriptional regulatory networks control RNA levels and spatio-temporal patterns through *cis*-regulatory sequences recognized by regulatory factors. As such, RNA binding proteins and mi-RNAs have recently received much attention and studies have documented their critical role in many biological processes. However, the discovery of functional RNA *cis*-regulatory sequences that act as recognition sites remains a challenge.

Our recent work[1], provides an integrated experimental and computational methodology

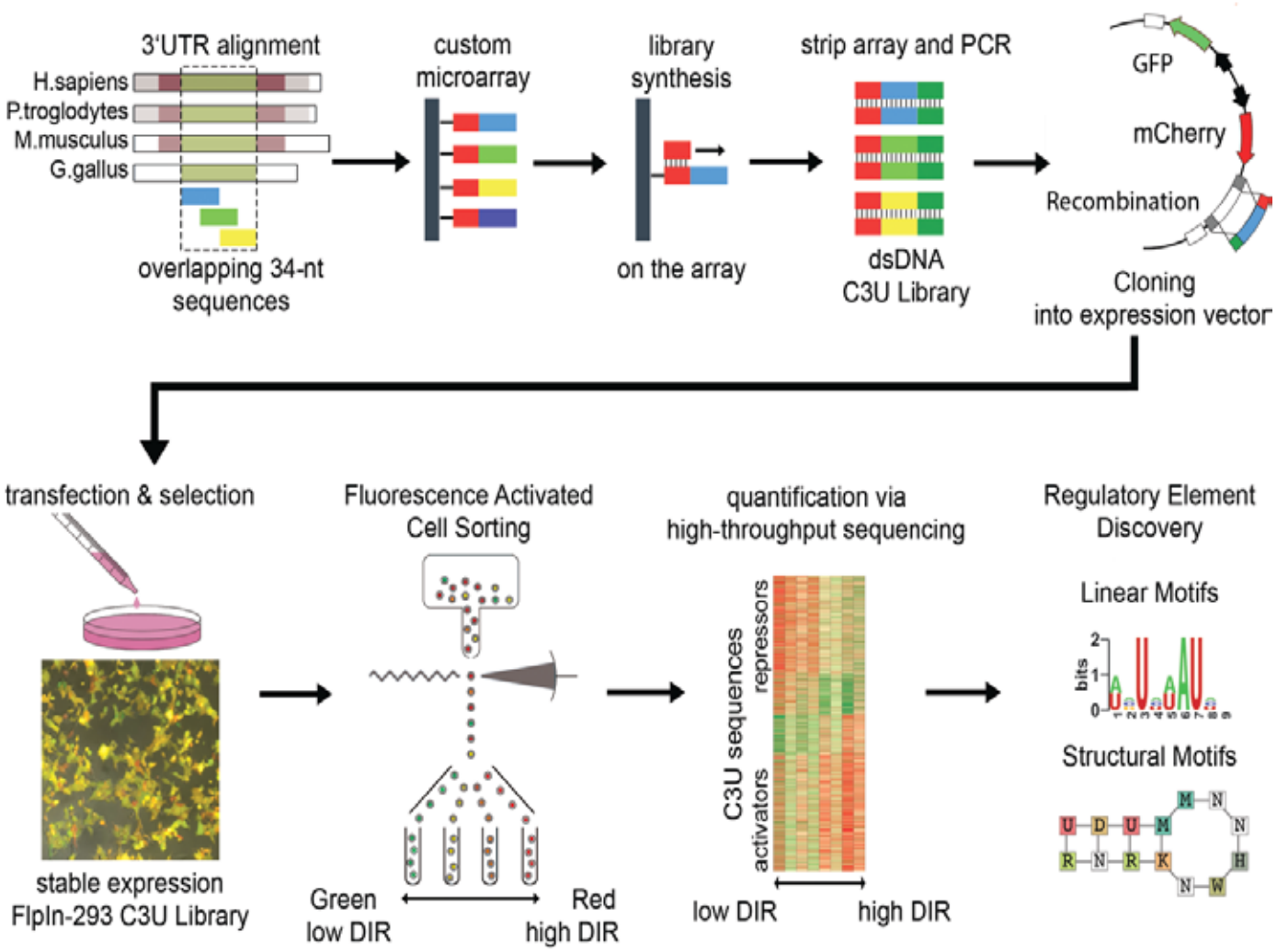


Figure 6: Vertebrate-conserved 3'-UTR sequences were identified and synthesized on a custom microarray. Single stranded DNA sequences were stripped and PCR amplified using universal adapters. Library sequences were cloned downstream of a fluorescent mCherry reporter in a bidirectional construct via recombination. Transfection of the vector into human 293 cells produced a library of cells which stably expressed the bidirectional construct controlled by the 34-nt conserved 3'-UTR sequences. Cells from this library were FACS-sorted into expression bins and analyzed via high-throughput sequencing. Based on over- and under- representation patterns for each sequence in each expression bin, sequences were predicted to be either gene expression repressors or activators. Results were further analyzed to identify new linear and structural RNA regulatory motifs.

that characterizes the functional consequences of thousands of RNA *cis*-regulatory sequences embedded in 3'-untranslated regions of human genes. Starting from a library of 16,332 short, vertebrate-conserved 3'-UTR sequences, we used a bidirectional reporter system coupled with flow cytometry sorting and high-throughput sequencing to quantify the effect of each sequence on expression. As a result, we identified a catalogue of over 2,000 sequences with significant positive or negative contributions to gene expression.

The functional characterization of thousands of sequences in parallel enables *de novo* discovery of regulatory motifs that are informative of the observed post-transcriptional effects across our library. We used computational methods previously developed by our group to reveal a list of 14 linear[2] and 8 structural[3] RNA cis-regulatory elements that are involved in mRNA stability or translation modulation. These motifs, which act as potential targets of trans-factors, included a mix of both repressive and activating elements. Many of the newly identified motifs were informative of mRNA stability measurements in mouse as well as in several human cells, highlighting their functional conservation in different species and cell types.

The short, functional RNA elements revealed in this study can be viewed as potential "building blocks" for modulating expression in the combinatorial context of multiple cis-regulatory elements in the native full length 3'UTR or in synthetic biology applications. Knowing these initial building blocks enables us to map the post-transcriptional regulatory networks and reveal modules of co-regulated transcripts, thus enhancing our systems level understanding of post-transcriptional gene regulation.

REFERENCES

1. Oikonomou, P., H. Goodarzi, and S. Tavazoie, Systematic Identification of Regulatory Elements in Conserved 32 UTRs of Human Transcripts. Cell Reports, 2014.

2. Elemento, O., N. Slonim, and S. Tavazoie, A universal framework for regulatory element discovery across all genomes and data types. Molecular cell, 2007. 28(2): p. 337-350.

3. Goodarzi, H., et al., Systematic discovery of structural elements governing stability of mammalian messenger RNAs. Nature, 2012. 485(7397): p. 264-268.

BIO-ECONOMIC APPROACHES TO MODEL MICROBIAL TRADE USING SYNTHETIC SYNTROPHIC BACTERIAL COMMUNITIES

HARRIS WANG LAB

Metabolic crossfeeding is an important process that governs natural microbial communities, affecting population composition, stability and diversity under many environments. However, little is known about specific crossfeeding principles that drive the formation and maintenance of individuals within a mixed population. Based on comparative genomic analysis of >6000 sequenced bacteria from diverse environments, we have found evidence suggesting that amino acid biosynthesis has been broadly optimized to reduce individual metabolic burden in favor of enhanced crossfeeding to support synergistic growth across the biosphere.

Borrowing key principles from economics, we have built a framework based on general equilibrium theory to predict the population dynamics of crossfeeding biotic communities. Our model for two crossfeeding microbes yields important insights including the impact of comparative advantage on trade-based mutualism between species, and growth-dominance trade-offs in a mixed community. We find that reliance of trade-partners can be a selective advantage that stabilizes microbial ecosystems.

We devised a series of synthetic syntrophic communities to probe the complex interactions underlying metabolic exchange of amino acids using multi-member, multi-dimensional communities of auxotrophic Escherichia coli. We find that biosynthetically costly amino acids tend to promote stronger cooperative interactions and that cells that share common intermediates along branching pathways yielded more synergistic growth. In more complex communities, we find certain members exhibiting keystone species-like behavior that drastically impact the community dynamics. These results improve our basic understanding of the systems biology of microbial communities.

REFERENCES

1. Mee, M.T., J.J. Collins, G.M. Church, and H.H. Wang, Syntrophic exchange in synthetic microbial communities. Proc Natl Acad Sci U S A, 2014.

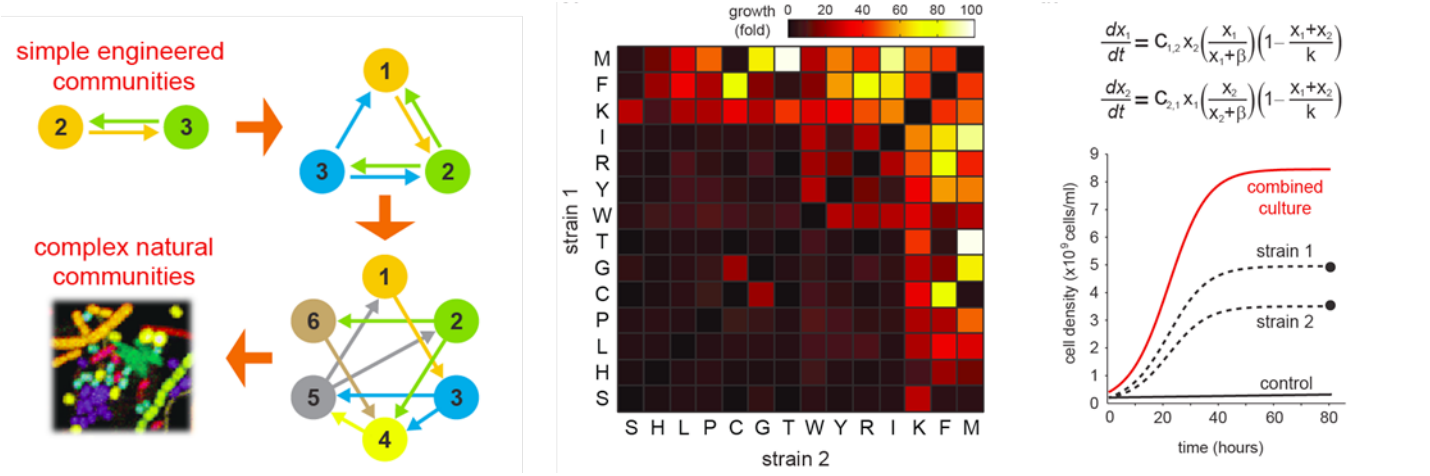


Figure 7. Metabolic crossfeeding in syntrophic communities. Engineering syntrophic interactions between microbial communities can be scaled to network hierarchies matching those of natural systems. Pairwise syntrophic growth between 14 single-KO auxotrophs (strain 1) and all pairwise combinations (strain 2) are shown with color intensity denoting fold growth. A simple two-equation dynamic model can be used to captures essential features of the pairwise consortium, but more sophisticated models using economics-based principles reveal important features of the syntrophic exchange.

DIRECT MUTAGENESIS OF THOUSANDS OF GENOMIC TARGETS USING MICROARRAY-DERIVED OLIGONUCLEOTIDES AND POOLED DEGENERATE LIBRARIES

HARRIS WANG LAB

Recent advances in high-throughput biology have enabled large-scale interrogation of genome function. We previously developed a genome perturbation tool, Multiplex Automated Genome Engineering (MAGE), to allow simultaneous mutagenesis of multiple target sites in bacteria genomes using short oligonucleotides. However, large-scale mutagenesis requires hundreds to thousands of unique oligos, which are costly to synthesize and impossible to scale-up by traditional phosphoramidite column-based approaches. In order to address this issue, we developed a novel method, Microarray-Oligonucleotide (MO)-MAGE, to amplify oligos from microarray chips for direct use to perturb thousands of genomic sites simultaneously.

We designed and synthesized 13,000 oligos that mutated specific regions of the Escherichia coli genome including most ribosomal binding sites, promoters, and protein coding regions possible. A sub-population of the oligo pool was designed to insert a T7 promoter upstream of each of 2587 E. coli genes in the untranslated regions (UTR) to introduce new transcriptional regulation. We developed a protocol to convert oligo pool harvested from an Agilent Oligo Library Synthesis microarray to create 90 bp oligos compatible with MAGE (Figure 1). We applied the pool of 2587 microarray-amplified T7-promoter insertion oligos to mutagenize the E. coli genome. The resulting cell library was characterized by deep-sequencing (Illumina MiSeq), leading to identification of 150 unique insertion sites. Amplicon sequencing of 12 randomly selected targets showed T7 promoter insertion, supporting the notion that the cell library contains most of the designed target insertions.

We further extended our approach to combine MAGE and deep-sequencing into MAGE-seq using large-scale pooled oligo libraries to interrogate the function of essential genes in Escherichia coli. Comprehensive single-codon mutagenesis of the gene *infA*, which is essential for the initiation of protein translation, revealed important constraints on the evolvability of a gene sequence with respect to factors contributed by amino acid choice and RNA structure. MO-MAGE and MAGE-seq enable rapid mutagenesis of bacterial genomes using oligos derived from microarrays and library pools without intermediate cloning or cassette selection steps to generate combinatorial genomic diversity of targeted cell populations. These advances will further foster the broad adoption of genome engineering technologies for manipulating and understanding microbial and eukaryotic systems.

REFERENCES

1. Bonde MT, K.S., Genee HJ, Sarup-Lytzen K, Church GM, Sommer MOA, Wang HH, Direct Mutagenesis of Thousands of Genomic Targets using Microarray-derived Oligonucleotides, ACS Synthetic Biology, 2014.

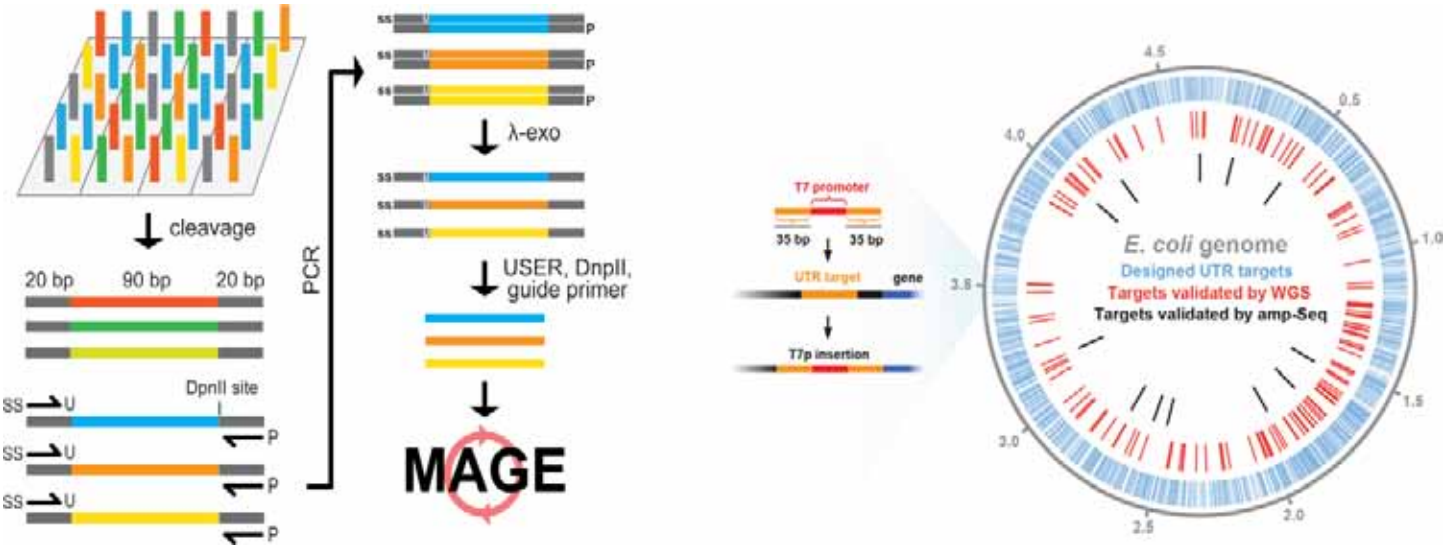


Figure 8. MO-MAGE method for targeted whole genome mutagenesis. 130 base oligonucleotides were designed and synthesized on a DNA microarray. Oligos from different subpools are amplified by PCR and enzymatically treated into 90 bases single-stranded form for MO-MAGE. On the left, MO-MAGE of 2587 genomic targets corresponding to untranslated regions (UTR) upstream of genes for insertion of 20-bp T7 synthetic promoter. Designed targets are shown in blue. Mutated targets verified by whole-genome sequencing are shown in red. Mutated targets verified by amplicon sequencing are shown in black.

THE ALTERNATIVE SPLICING REGULATION NETWORK OF RBFOX AND ITS IMPLICATIONS IN BRAIN DEVELOPMENT AND AUTISM

CHAOLIN ZHANG LAB

The Rbfox family of RNA-binding proteins are developmentally regulated, brain- and muscle-specific splicing factors that are highly conserved through evolution. They have also been implicated in complex neurological diseases such as autism. Our goal was to define the target network of the Rbfox proteins in the mammalian brain and characterize its functional significance.

We identified Rbfox-dependent changes in alternative splicing using both *in vitro* and *in vivo* perturbation experiments. To distinguish direct from indirect targets, we used crosslinking and immunoprecipitation followed by high-throughput sequencing (HITS-CLIP) to map the *in vivo* binding sites of each of the three Rbfox family members in the mouse brain at a single-nucleotide resolution on a genome-wide scale. These datasets were combined with bioinformatically-predicted Rbfox binding sites and evolutionary signatures using a Bayesian network framework to identify the direct and functional targets of Rbfox. This resulted in an extensive molecular network composed of over 1,000 alternative exons whose splicing is predicted to be under direct regulation by Rbfox.

We investigated the functional significance of these targets in developing brains. Over half of the Rbfox targets potentially undergo dynamic changes during development, consistent with the dynamic expression of the regulators. In this process, Rbfox proteins mostly promote the adult splicing pattern. The impact of Rbfox on the developmental molecular program and the previous identification of Rbfox1 mutations in autism prompted us to investigate its link with this devastating neurodevelopmental disorder which affects 1% of children worldwide. Indeed, Rbfox targets are enriched in candidate autism susceptibility genes, including three genes considered causal for syndromic autism spectrum disorders. Our study suggests that disruption of Rbfox or its target genes could potentially contribute to the disease.

REFERENCES

1. SM Weyn-Vanhentenryck, A Mele, Q Yan, S Sun, N Farny, Z Zhang, C Xue, PA Silver, MQ Zhang, AR Krainer, RB Darnell, C Zhang. HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. Cell Rep. 2014 Mar 6.

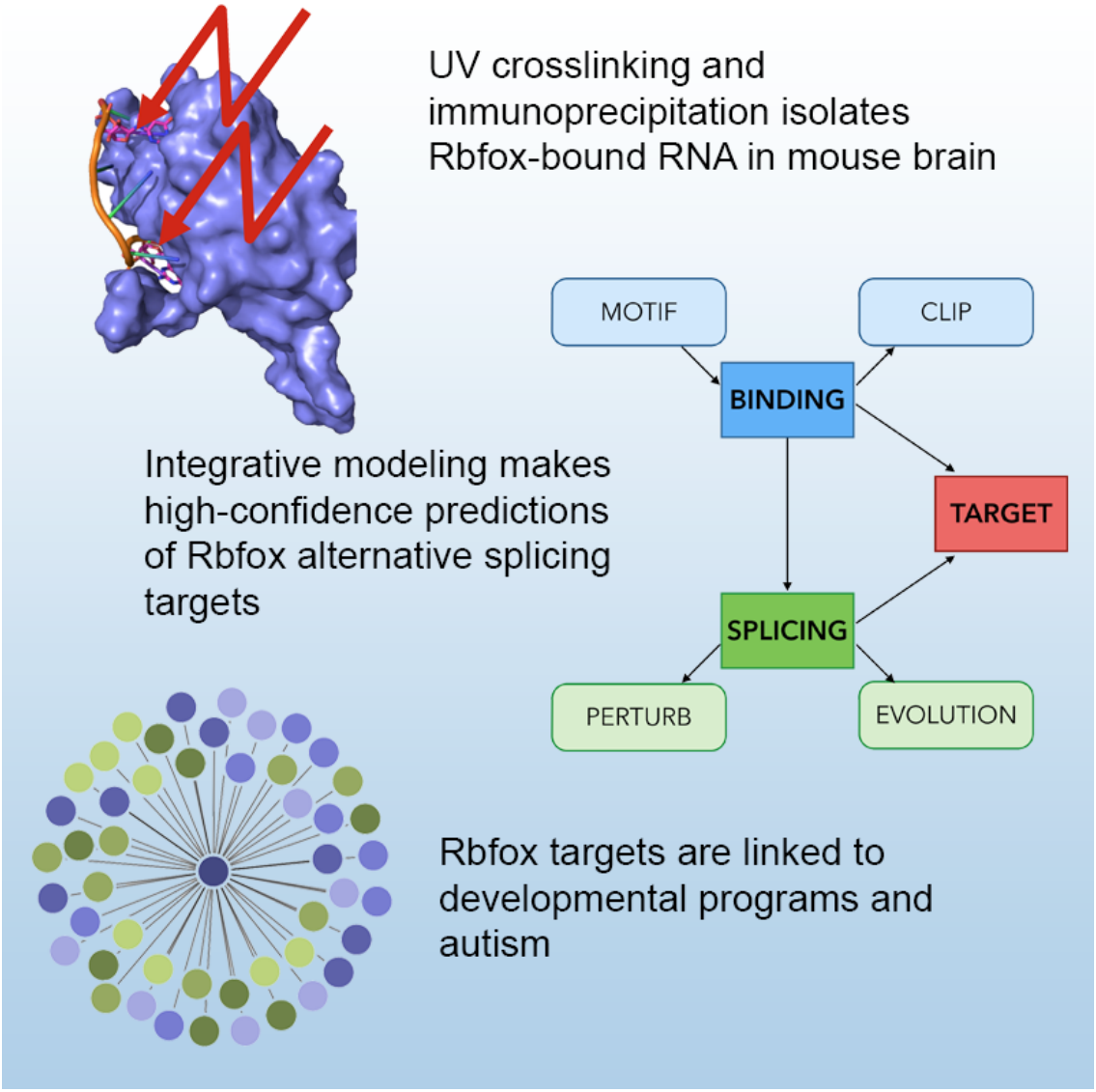
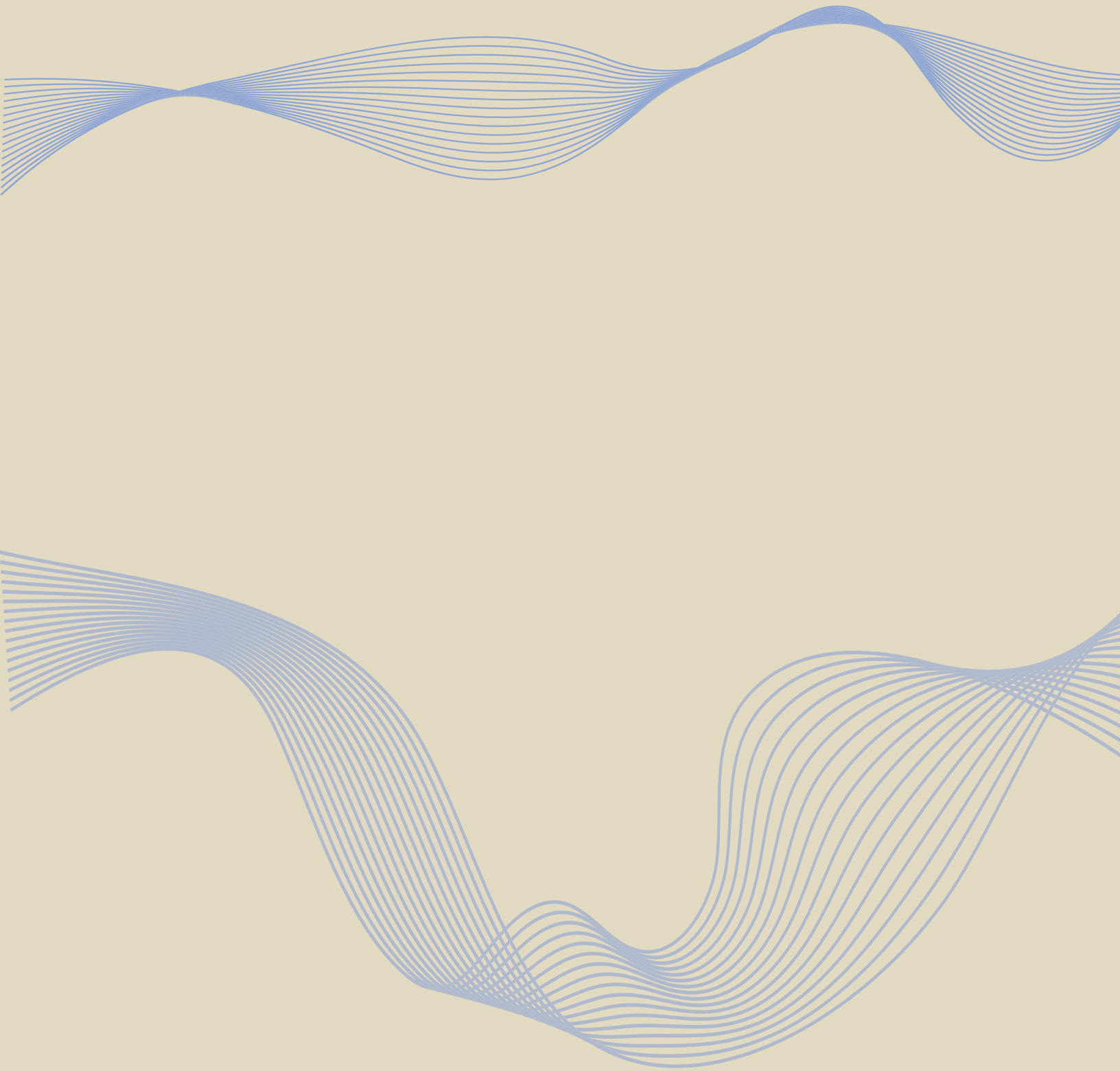
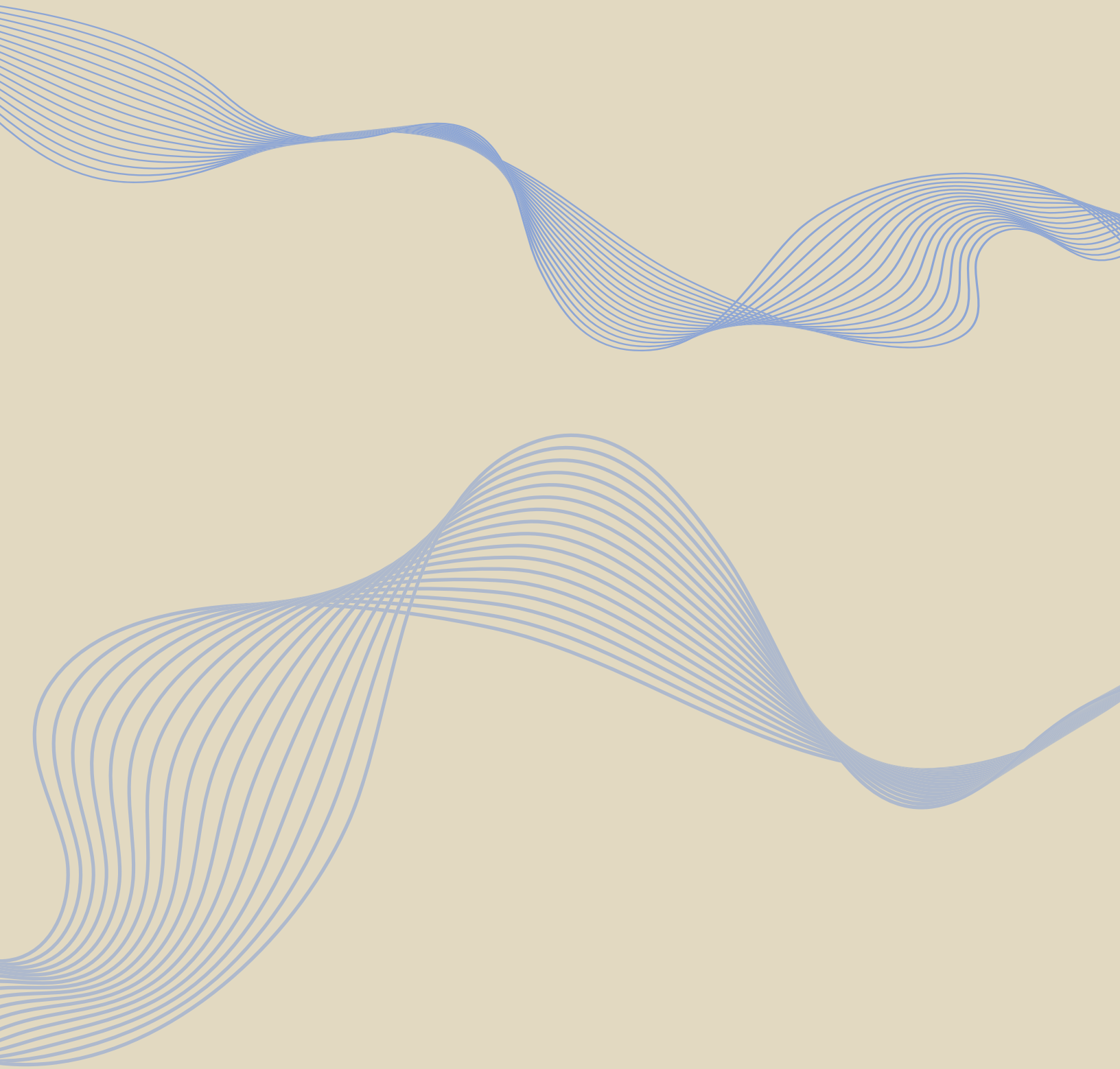


Figure 9: We used HITS-CLIP to identify genome-wide in vivo Rbfox binding sites at a single-nucleotide resolution (top). We used integrative modeling to combine these data with other evidence of Rbfox regulation to obtain over 1,000 high-confidence Rbfox splicing targets (middle). We found that the predicted splicing targets are developmentally regulated and are enriched in candidate autism genes (bottom)





Columbia University
Department of Systems Biology
1130 St. Nicholas Avenue
New York, NY, 10032

SPRING 2014
Issue No. 7