NATIONAL CENTER FOR THE MULTISCALE ANALYSIS OF GENOMIC AND CELLULAR NETWORKS

SPRING 2013

ISSUE NO.6

MAGNet NEWSLETTER

PrePPI: A NOVEL STRUCTURE-BASED METHOD FOR PREDICTING PROTEIN-PROTEIN INTERACTIONS

BARRY HONIG

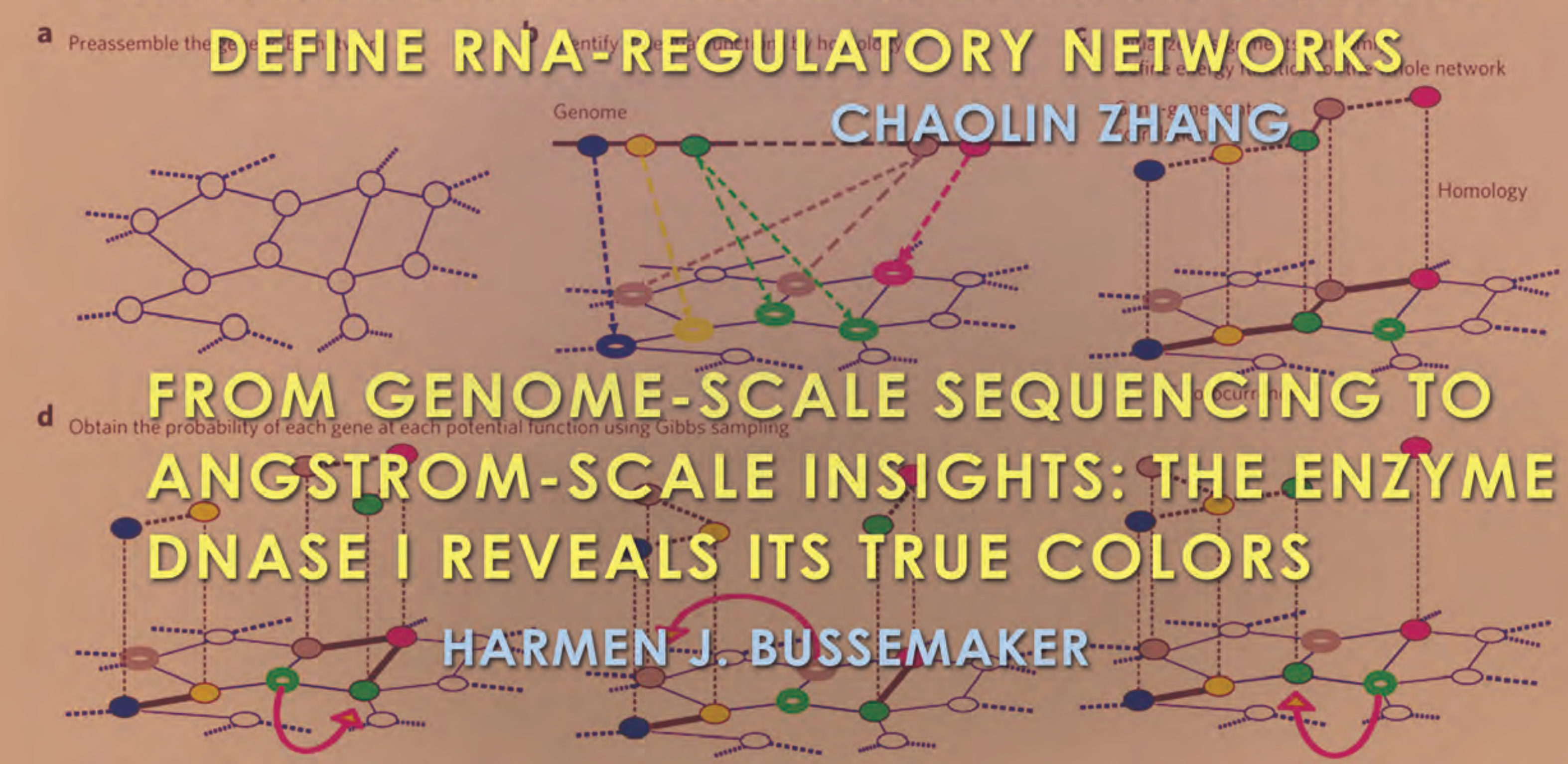
PDB structures/
homology models
Structural neighbours
Template complexes from PDB
Interaction models
Structural-based score
QA
QB
?
Sequence similarity
Structural similarity
MA
MB
NA1
NA2
NB1
NB2
NB3
Structural superposition
Model evaluation
SM
Bayesian classification
PrePPI score
Non-structural evidence
E1
E2
En

AN INTEGRATIVE GENOMICS APPROACH TO DEFINE RNA-REGULATORY NETWORKS

CHAOLIN ZHANG

FROM GENOME-SCALE SEQUENCING TO ANGSTROM-SCALE INSIGHTS: THE ENZYME DNASE I REVEALS ITS TRUE COLORS

HARMEN J. BUSSEMAKER


a
Preassemble the
b
c
whole network
Genome
Homology
d
Obtain the probability of each gene at each potential function using Gibbs sampling

SPRING 2013
ISSUE NO.6

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

MAGNet NEWSLETTER

FEATURES

03 FEATURE ARTICLE:
PrePPI: A NOVEL STRUCTURE-BASED METHOD FOR PREDICTING PROTEIN-PROTEIN INTERACTIONS
BARRY HONIG

09 FEATURE ARTICLE:
AN INTEGRATIVE GENOMICS APPROACH TO DEFINE RNA REGULATORY NETWORKS
CHAOLIN ZHANG

15 FEATURE ARTICLE:
FROM GENOME-SCALE SEQUENCING TO ANGSTROM-SCALE INSIGHTS: THE ENZYME DNASE I REVEALS ITS TRUE COLORS
HARMEN J. BUSSEMAKER

SECTIONS

02 INTRODUCTION

21 FEATURED NEWS
- FIRST METHOD FOR GLOBAL PROBABILISTIC ANNOTATIONS OF METABOLIC NETWORKS
- PAN-CANCER GENE SIGNATURES ARE PROGNOSTIC IN BREAST CANCER
- CERNA REGULATORY INTERACTIONS ARE HIGHLY CONSERVED AND MODULATE DRIVERS OF TUMORIGENESIS
- USING WEB SEARCH LOGS TO DETECT ADVERSE DRUG INTERACTIONS
- THE GENOME SEQUENCE OF THE ASHKENAZI JEWISH POPULATION
- A LARGE-SCALE STUDY OF TUMOR-INDUCED GENE EXPRESSION CHANGES IN HUMAN METABOLISM
- VIRTUAL PROTEOMICS
- STATISTICAL INFERENCE IN NANOPORE SEQUENCING
- A DNA LANDING PAD FOR PROTEINS
- SUPPORT FOR CELL-OF-ORIGIN MODEL FOR PROSTATE CANCER HETEROGENEITY

In this sixth issue of the MAGNet center newsletter, we report on several recent results by investigators in the Center for the Multiscale Analysis of Genomics and Cellular Networks (MAGNet) at Columbia University. We also report on the creation of the Department of Systems Biology (DSB) in June 2013, an event that is a direct result of the MAGNet center scientific and educational success over the past eight years. The DSB at Columbia University ultimately embodies the vision of a fully integrative approach, combining computational and experimental high throughput approaches, to elucidate biological processes, both physiological and disease related. Not surprisingly, this is the same model that originally led to the creation and successful operation of the MAGNet Center. To implement these goals, the DSB integrates the Columbia Center for Computational Biology and Bioinformatics (C2B2), the original home of MAGNet, and the Sulzberger Columbia Genome Center. This will provide a unique collaborative and multidisciplinary environment that uniquely benefits MAGNet center investigators by providing access to key resources, infrastructure, and programs: from high-performance computing to large scale next generation sequencing and high throughput screening. Furthermore, it will help catalyze translation of basic science discoveries by MAGNet investigators to address diagnosis and treatment of human disease on a more quantitative and predictive basis.

Achieving a more “systems level” understanding of biological processes, to seamlessly combine multi-scale data from the atomic to the functional level, is emerging as an increasingly relevant theme in all areas of biology and medicine. The MAGNet center plays a central role in providing innovative algorithms, computational tools, methodologies, and information databases to the scientific community, enabling the integration and interrogation of structural and functional information towards elucidating the internal logic of biological systems. In addition, MAGNet investigators continue their tradition of both performing and enabling high impact science by supporting research collaborations and stimulating scientific exchange, both within their labs and in collaboration with labs at many other institutions. Among the many scientific accomplishments by MAGNet investigators, which have resulted in over 50 publications in 2012-2013, over a third of which in high impact journals, this issue features studies by three MAGNet labs that have contributed significant advances at the interface of systems and structural biology. Each of these studies reports on the development of novel computational algorithms that can dissect specific classes of molecular interactions that are crucial for the overall understanding of biological processes. The highlighted studies describe how analysis of the molecular interaction network models produces biologically relevant hypotheses that can be experimentally validated.

The first feature article describes a new algorithm called PrePPI (Predicting Protein-Protein Interactions), published in Nature in 2012, which resulted from a collaboration of the Honig and Califano labs. PrePPI is truly the culmination of the original plan by MAGNet investigators to combine structural and functional data to achieve a molecular level understanding of regulatory and complex interactions in the cell. Specifically, the algorithm is used to systematically map protein-protein interactions by using remote homology matching to complex templates in structural databases, such as the PDB. Candidate interactions are then scored and prioritized based on a variety of clues coming from the putative contact interface and from other functional data, such as co-expression within specific cellular contexts. PrePPI’s performance, which was rigorously experimentally validated, was shown to significantly outperform previous methods, including high-throughput experimental methods. Overall, it contributes about 300,000 protein-protein interactions, most of which are novel, to the human repertoire.

In the second article, Dr. Chaolin Zhang, a recent DSB recruit, presents a unique integrative genomics approach to infer RNA regulatory networks by integrating computational and experimental methods. His work elucidates RNA regulatory networks by identifying an extensive number of RNA-binding proteins (RBPs) and their interactions with target RNA transcripts. Most RBPs recognize short sequence motifs with limited information, posing a challenge to infer RNA-regulatory networks. Dr. Zhang’s research utilizes biochemical and high-throughput assays, such as HITS-CLIP (or CLIP-Seq) and RNA-Seq to profile transcriptomes and protein-RNA interactomes. By integrating data directly relevant to RNA regulation, including those from exon-sensitive microarrays and RNA-Seq, with genetic perturbation data, they are able to identify RBP-dependent changes in target transcripts and to predict protein-RNA binding. Predictions are made using statistical algorithms, whose results complement and enhance the information obtained from experimental data.

Finally, in a third study by MAGNet investigator Harmen Bussemaker, published in the Proceedings of the National Academy of Sciences, insights derived from the activity of a small DNase enzyme shed light on how variations in methylation patterns can regulate mRNA expression changes. Understanding the association between methylation patterns and mRNA regulation is an unsolved problem in regulatory genomics. The Bussemaker lab developed a robust new methodology to elucidate the dependency of DNase I activity on the nucleotide sequence. This information suggests that DNA shape modulates the cleavage rate of the enzyme. Their results show DNase I to be a highly sensitive probe of the geometry of DNA’s minor groove, which can serve as an important recognition site within the DNA-binding interface of many regulatory proteins. High throughput sequencing provides a map of all the methylation events in the genome and the methylation pattern is highly cell type specific. This constitutes a method to determine the relationship between DNase I cleavage rate and cytosine methylation. This MAGNet funded collaborative effort reveals cytosine methylation as a possible mechanism for the modulation of DNA shape at the epigenetic level. This information is increasingly important for transcription factors and their complexes because DNA shape mediated binding by cytosine methylation applies there as well.

In summary, MAGNet investigators continue in their tradition of highly innovative scientific work at the boundary of computational and experimental biology, to achieve a more systems level understanding of biological processes. The studies highlighted in this issue of the MAGNet newsletter demonstrate exciting new methodologies by which genomic, proteomic, biophysical and biochemical data can be integrated to improve our understanding of basic cellular processes that affect both cell physiology and cell pathology.

-Andrea Califano

PrePPI: A NOVEL STRUCTURE-BASED METHOD FOR PREDICTING PROTEIN-PROTEIN INTERACTIONS

BARRY HONIG
DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOPHYSIC
INITIATIVE IN SYSTEMS BIOLOGY
COLUMBIA UNIVERSITY

INTRODUCTION

A basic tenet of research at the Center for Multiscale Analysis of Genomic and Cellular Networks (MAGNet) has been that developing models of molecular-level interactions across multiple levels of granularity will dramatically improve our understanding of cellular systems and phenotypes. In recent years, systems biologists have made remarkable progress in developing methods for integrating a variety of ‘omics data types and generating predictive, genome-wide models of molecular networks. However, protein structure has not been a significant factor in the generation of these large-scale cell regulatory networks.

The limited role of protein structure is due in large part to the fact that the number of proteins with experimentally known structures is still relatively small. As of early 2010, for example, the Protein Data Bank (PDB) provided structures for ~600 of the total complement of ~6,500 yeast proteins (~10%). The structural coverage of protein–protein complexes is even sparser, with only ~300 structures available out of the approximately 75,000 PPIs (<0.5%) recorded in publicly available databases. The effort necessary to experimentally solve the structures for the vast number of remaining proteins and protein complexes would be enormous, and so if structure is to be useful on the genome-wide scale that systems biology requires, new computational approaches for modeling PPIs are critically needed.

One of MAGNet’s long-range goals has been to develop methods for integrating computational structure analysis into systems biology research. In a 2012 paper in Nature, we reported on what we believe to be a significant advance toward achieving this goal [1]. For the first time, a novel algorithm, which we have named PrePPI (Predicting Protein-Protein Interactions), uses homology modeling, Bayesian statistics, and information from the PDB to make high-confidence predictions of protein-protein interactions on a genome-wide scale. In what follows we explain our approach and discuss how it enables predictions with accuracy comparable to that of high-throughput experimental methods.

DESIGN OF THE ALGORITHM

To predict an interaction for a given a pair of query proteins (QA and QB, see Figure 1), we first use sequence alignment to identify structural representatives (MA and MB) that correspond to either experimentally determined structures or to computationally derived models. These structural representatives are taken directly from the PDB, where available, or from the ModBase [2] and SkyBase [3] homology model databases. PDB structures are identified by sequence homology, using PSI-BLAST [4]; matching structures in the PDB are required to have >90% sequence identity and cover >80% of the query target. Homology models are selected based on a combination of sequence homology and structure-based scores (pG score for SkyBase models; MPQ score for ModBase models). When multiple structures are available for a target/domain we chose only one representative using (1) the PDB structure with the best resolution, if available; (2) the ModBase model with the highest MPQ score; or (3) the SkyBase model with the highest pG score.

After assignment of the structural representatives we use structural alignment to find both close and remote structural neighbors (NAi and NBj) of MA and MB; (an average of ~1,500 neighbors are found for each structure). Structural alignment compares a protein of interest to proteins with similar folds. In contrast to other methods, we do not just compare global structures but look for local structural similarity as well. Whenever a pair of neighbors of MA and MB (for example, NA1 and NB3) form a complex reported in the PDB, we use this complex as a template for modeling the interaction of QA and QB. The model is constructed by superimposing the representative structures on their corresponding structural neighbors in the template (that is, MA on NA1 and MB on NB3). This

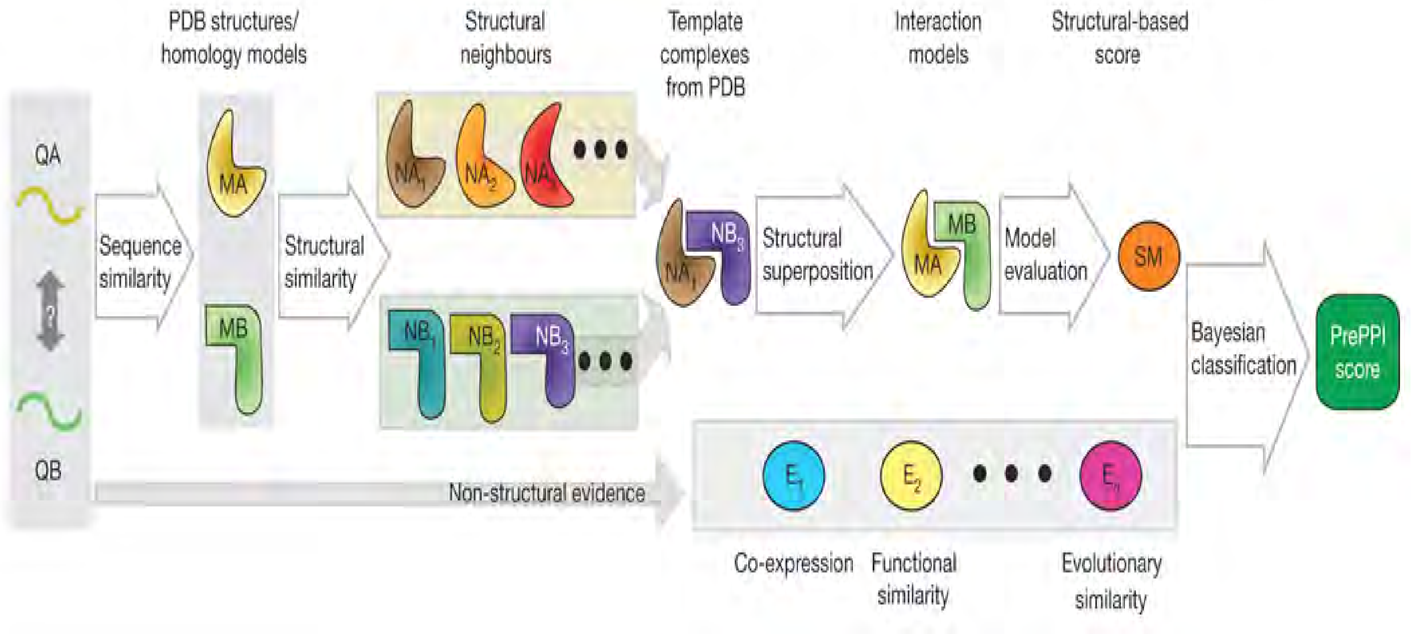


Figure 1: Predicting protein–protein interactions using PrePPI.

procedure produces about 550 million "interaction models" for about 2.4 million PPIs involving about 3,900 yeast proteins, and about 12 billion models for about 36 million PPIs involving about 13,000 human proteins.

Once an interaction model has been created, it is evaluated using a combination of five empirical scores that measure properties derived from alignments of the individual monomers to their templates. The first score depends on the structural similarity between models of the two query proteins (that is, MA and MB) and those in the template complex (that is, NA1 and NB3). The next two scores determine whether the interface in the template complex actually exists in the model. The final two scores reflect whether the residues that appear in the model interface have properties consistent with those that mediate known PPIs (for example, residue type, evolutionary conservation, or statistical propensity to be in protein–protein interfaces). This information is obtained from three publically available servers that predict interfacial residues based on the sequence and structure of the individual subunits of the model [5-7].

The five empirical scores are combined using a Bayesian network to yield a likelihood ratio (LR) that a candidate protein–protein complex represents a true interaction. The network is trained on positive and negative "gold standard" reference data sets comprising interactions from multiple databases, to ensure a broad coverage of true interactions. We divide these sets into high-confidence (HC) and low-confidence (LC) subsets; the HC sets contain 11,851 yeast interactions and 7,409 human interactions that have more than one publication supporting their existence; interactions with only one supporting publication compose the LC set. All potential PPIs in a given genome not in the HC plus LC set form the negative (N) reference set. Using the Bayesian network classifier trained on the yeast HC set, we select the best interaction model with the highest LR for each PPI.

COMPARISON TO OTHER METHODS

To assess the performance of structural modeling (SM), we compared it with a number of non-structural clues previously used to infer PPIs [8-10]: (1) essentiality of the proteins in the interacting pair; (2) co-expression level; (3) gene ontology (GO) functional similarity; (4) Munich Information Centre for Protein Sequences (MIPS) functional similarity; and (5) phylogenetic profile similarity. We found that SM yields comparable performance over the entire range of false positive rate (FPR) values but is considerably more effective at low FPRs (FPR≤ 0.1%). This difference is critical because, owing to the huge number of negative interactions, only very low FPR rates can produce a small enough number of false positives to be used effectively in practice. At low FPRs, SM by itself outperforms even the naive Bayesian classifiers that combine all non-structure-based clues. Because it combines structural and non-structural clues, PrePPI also yields many more high-confidence predictions than either source of information alone.

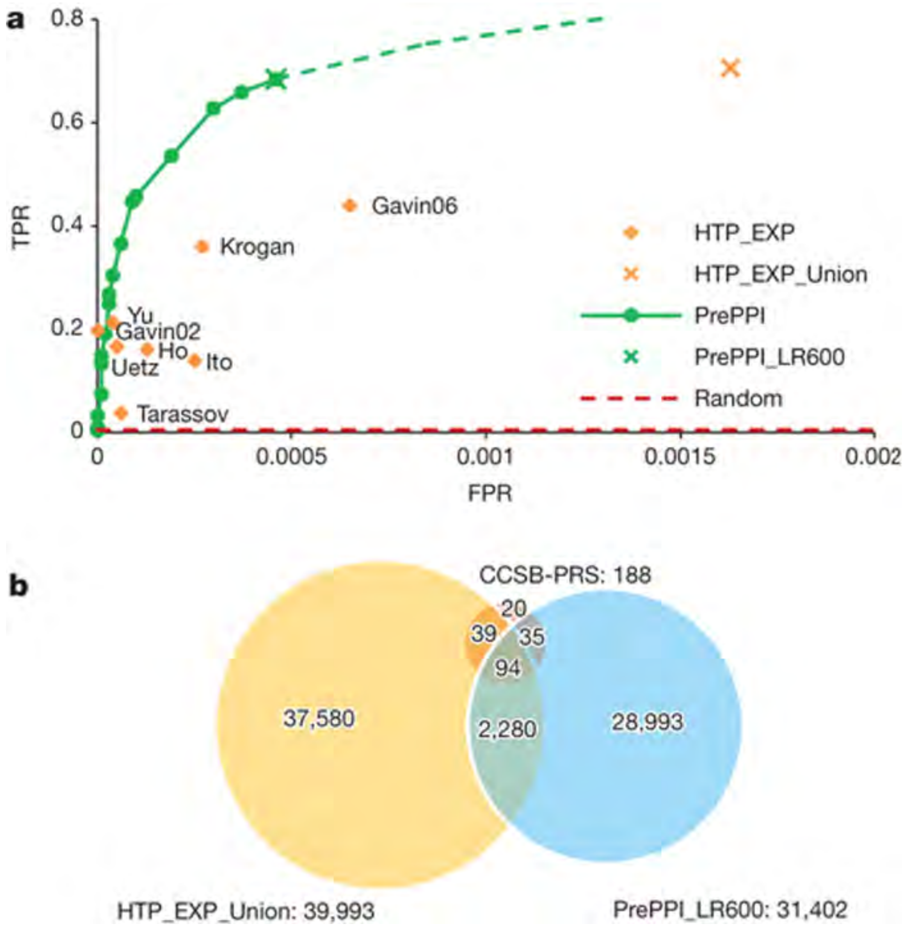


Figure 2: Comparison of PrePPI predictions and high-throughput experiments in yeast. In graph (a), the ROC curve compares true positive rates for PrePPI and several high-throughput experiments, labeled with the first author of the relevant publication. The Venn diagram (b) shows the number of interactions using PrePPI and in reference sets, as well as intersections among the sets.

As an independent test, we evaluated PrePPI against a PPI prediction challenge from the 2009 Dialogue for Reverse Engineering Assessments and Methods (DREAM) competition [11]. PrePPI outperformed all other methods for cases where structural information was available. We also compared PrePPI's performance to data in a previously reported comparison of high-throughput experimental techniques [12], finding PrePPI to be somewhat better overall than high-throughput methods for most data sets that were tested. As can be seen in Figure 2, many of the interactions inferred by PrePPI were different from those identified by high-throughput assays. This suggests that methods which combine both approaches may prove to be highly effective in expanding the coverage of PPIs.

RESULTS AND VALIDATION

At an LR cutoff of 600, PrePPI predicts 31,402 high-confidence interactions for yeast and 317,813 interactions for human. These, as well as predictions with lower LR scores, are available in a database from the PrePPI website (http://bhapp.c2b2.columbia.edu/PrePPI/). As a further validation of PrePPI we tested its performance on approximately 24,000 new interactions involving human proteins that were added to public databases after August 2010. Among these interactions, 1,644 are predicted by PrePPI to have an LR > 600.

Four separate laboratories used co-immunoprecipitation assays to validate 19 individual PrePPI predictions experimentally, leading to confirmation of 15 interactions. The investigators in each laboratory queried the PrePPI database for previously uncharacterized interactions involving proteins of interest that had relatively high SM and PrePPI scores. Findings from the validation assays include:

- PrePPI predicts high-confidence interactions between the nuclear receptor peroxisome proliferator-activated receptor γ (PPAR-γ) and the transcription factors LXR-β (also known as NR1H2), PAX7, PDX1, NKX2-2 and HHEX. Except for HHEX, all of the interactions were validated. PPAR-γ has a pivotal role in regulating glucose and lipid metabolism, the inflammatory response and tumorigenesis, and is known to heterodimerize with retinoid X receptors (RXRs) and to recruit cofactors to regulate target gene transcription. The predicted interaction with nuclear receptor LXR-β had not previously been characterized and suggests an unrecognized convergence of signaling and metabolic pathways regulated by these two nuclear receptors. The interactions between the ligand-binding domain of PPAR-γ and the homeodomains of PAX7, PDX1 and NKX2-2 are new observations that require further studies, suggesting that PPAR-γ may have a role in endocrine progenitor and pancreatic β-cell development.

- PrePPI predicts that suppressor of cytokine signaling (SOCS3) forms complexes with GRB2 and RAF1, two key components in the RAS/MAPK pathway. The algorithm also predicts the formation of a complex between SOCS3 and BTK, a cytoplasmic tyrosine kinase important in B-lymphocyte development, differentiation, and signaling. These interactions were validated.

- Experiments also validated predictions that several kinases interact with the clustered protocadherin proteins (PCDH-a, -β and -γ). The PCDHs have six cadherin-like extracellular domains, and unique cytoplasmic domains. They assemble into large complexes at the cell surface, and associate with a variety of proteins, including signaling adaptors, kinases, and phosphatases. Analysis of potential PCDH-kinase PPIs in mice confirmed published interactions between PCDH-a and -γ with the tyrosine kinase RET, and predicted interactions with ROR2, VEGFR2 and ABL1. PrePPI predicts that these PPIs are mediated by the extracellular cadherin domains and immunoglobulin (Ig) domains, a result that was confirmed experimentally. This and other results suggest that, in addition to predicting binary interactions, PrePPI has the potential to reveal novel and unsuspected interfaces.

- A fourth group of experiments, aiming to identify new components of large protein–protein complexes, validated previously uncharacterized interactions between the special AT-rich sequence-binding protein SATB2 and the Emerin 'proteome' complex 32; and between the pre-mRNA-processing factor PRPF19 and the centromere chromatin complex.

It is important to emphasize that any PPIs detected using PrePPI must be confirmed through appropriate in vivo experiments. Taken together, however, these findings suggest that PrePPI has sufficient accuracy and sensitivity to provide a wealth of novel hypotheses that can drive biological discovery.

DISCUSSION

The accuracy and range of applicability of PrePPI, and the crucial role of structural modeling, were unanticipated but should not come as a complete surprise. Most protein complexes in the PDB have structural neighbors that share binding properties [13], and protein interface space may well be close

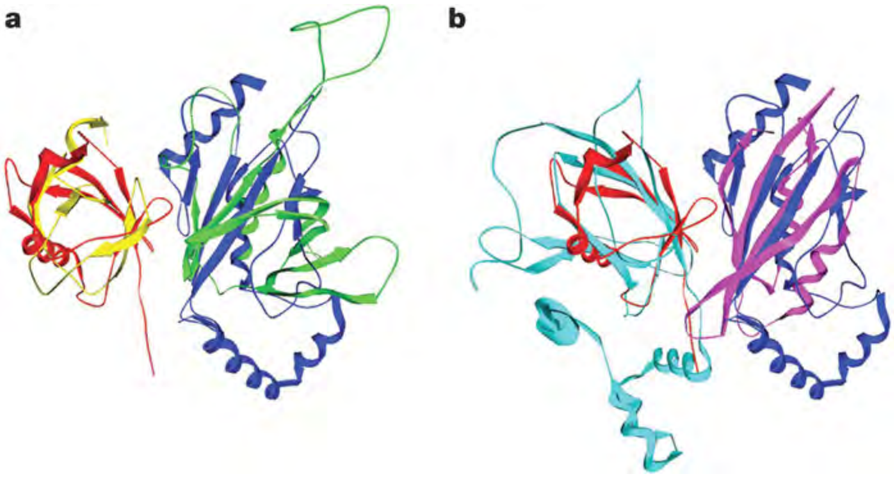


Figure 3: Models for the PPI formed between PRKD1 and PRKCE, and EEF1D and VHL using homology models and remote structural relationships. The structures of the PH domain of PRKD1 and the GNE domain of EEF1D (shown in green and purple) are homology models from ModBase; the structure of a C1 domain of PRKCE (yellow) is a homology model from SkyBase; the structure of VHL (cyan) is from PDB (accession 1LM8; V chain). In each case, the relevant homology models are structurally superimposed on one of the two templates in the UBE2D3–ubiquitin complex.

to ‘complete’ [14]. Moreover, these elements can be identified with geometric alignment methods [13, 15], a fact that has been exploited in the approach introduced here.

Thus, although the information required to predict whether two proteins interact appears to be present in the PDB, the question has been how to mine the data. There are three key elements that are responsible for the success of structural modeling and PrePPI:

- A marked expansion in the number of interactions that can be modeled. When compared to using only experimentally determined structures that are available in the PDB, the utilization of homology models and remote structural relationships increases the number of human PPIs that can be modeled from ~2.5 million to ~36 million. Not only is the number of PrePPI predictions an order of magnitude larger but it also comprises interactions of a more diverse nature. As might be expected, predictions based on the use of PDB structures alone are more likely to recover mainly known interactions (defined by their presence in databases). By contrast, the combined use of homology models and remote structural relationships yields a marked expansion in the total number of interaction models with many more high-confidence predictions. Further, the huge number of low-confidence structural interaction models leads to an even greater expansion when combined with functional, evolutionary, and other sources of evidence.

- The efficiency of our scoring scheme for interaction models which allows us to evaluate an extremely large number of models while still discriminating among closely related family members. Discrimination among complexes involving members of the same protein family—that is, specificity—is obtained from the properties of the predicted interface, for example, the statistical propensity of certain amino acids to appear in interfaces [6, 7] (and, additionally, from non-structural clues such as whether two proteins are co-expressed).

- The use of Bayesian evidence integration methods which allow independent and possibly weak interaction clues to be combined to make reliable predictions and to improve prediction specificity.

The exploitation of homology models and of remote structural relationships imply that each new experimentally validated structure can be used to detect large numbers of new functional relationships, even if the protein itself is of limited biological interest. PrePPI also appears to offer a viable alternative to high-throughput experiments; in addition to determining the likelihood of a given interaction, our approach can produce useful (if somewhat crude) models of the domains and residues that form protein–protein interfaces. This should facilitate the generation of experimentally testable hypotheses of true physical interactions. In conclusion, our study suggests that using structural information to the study of PPIs is possible, and that structural biology has an important role to play in molecular systems biology.

REFERENCES

1. Zhang, Q.C., D. Petrey, L. Deng, L. Qiang, Y. Shi, C.A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano, and B. Honig, Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature, 2012. 490(7421): p. 556-60.

2. Pieper, U., N. Eswar, F.P. Davis, H. Braberg, M.S. Madhusudhan, A. Rossi, M. Marti-Renom, R. Karchin, B.M. Webb, D. Eramian, M.Y. Shen, L. Kelly, F. Melo, and A. Sali, MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res, 2006. 34(Database issue): p. D291-5.

3. Mirkovic, N., Z. Li, A. Parnassa, and D. Murray, Strategies for high-throughput comparative modeling: applications to leverage analysis in structural genomics and protein family organization. Proteins, 2007. 66(4): p. 766-77.

4. Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997. 25(17): p. 3389-402.

5. Zhang, Q.C., L. Deng, M. Fisher, J. Guan, B. Honig, and D. Petrey, PredUs: a web server for predicting protein interfaces using structural neighbors. Nucleic Acids Res, 2011. 39(Web Server issue): p. W283-7.

6. Liang, S., C. Zhang, S. Liu, and Y. Zhou, Protein binding site prediction using an empirical scoring function. Nucleic Acids Res, 2006. 34(13): p. 3698-707.

7. Chen, H. and H.X. Zhou, Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. Proteins, 2005. 61(1): p. 21-35.

8. Jansen, R., H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, and M. Gerstein, A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 2003. 302(5644): p. 449-53.

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

10. von Mering, C., L.J. Jensen, B. Snel, S.D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M.A. Huynen, and P. Bork, STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res, 2005. 33(Database issue): p. D433-7.

11. Stolovitzky, G., R.J. Prill, and A. Califano, Lessons from the DREAM2 Challenges. Ann N Y Acad Sci, 2009. 1158: p. 159-95.

12. Yu, H., P. Braun, M.A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.F. Rual, A. Dricot, A. Vazquez, R.R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.S. de Smet, A. Motyl, M.E. Hudson, J. Park, X. Xin, M.E. Cusick, T. Moore, C. Boone, M. Snyder, F.P. Roth, A.L. Barabasi, J. Tavernier, D.E. Hill, and M. Vidal, High-quality binary protein interaction map of the yeast interactome network. Science, 2008. 322(5898): p. 104-10.

13. Zhang, Q.C., D. Petrey, R. Norel, and B.H. Honig, Protein interface conservation across structure space. Proc Natl Acad Sci U S A, 2010. 107(24): p. 10896-901.

14. Gao, M. and J. Skolnick, Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. Proc Natl Acad Sci U S A, 2010. 107(52): p. 22517-22.

15. Keskin, O., R. Nussinov, and A. Gursoy, PRISM: protein-protein interaction prediction by structural matching.

AN INTEGRATIVE GENOMICS APPROACH TO DEFINE RNA REGULATORY NETWORKS

CHAOLIN ZHANG
DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOPHYSICS
CENTER FOR MOTOR NEURON BIOLOGY AND DISEASE
INITIATIVE IN SYSTEMS BIOLOGY
COLUMBIA UNIVERSITY

In the past decade, we have witnessed a rapid expansion in our understanding of how DNA binding proteins regulate transcription, giving a new appreciation for the remarkable complexity of the mammalian transcriptome. Post-transcriptional regulation, which can significantly affect the activity of RNA transcripts, is much less understood, however. Increasingly, researchers now recognize that RNA-binding proteins (RBPs) play critical roles in post-transcriptional regulation, affecting RNA diversification, maturation, and function. RNA regulation contributes to specification of different cell types and developmental stages[1, 2], and disruption of RNA regulation has been implicated in genetic diseases such as neurodegenerative disorders and cancer [3, 4]. Characterizing the networks that define functional interactions between RBPs and their target RNA transcripts could thus improve our understanding of a variety of important biological phenomena.

We currently face two big challenges to our ability to dissect RNA-regulatory networks at the systems level. First, the RNA binding sites that RBPs recognize, known as motifs, tend to be very short (~3-7 nucleotides) and are degenerate, containing limited information. In addition, experimental methods to determine in vivo protein-RNA interactions on a genome-wide scale have been lacking until recently. Second, although emerging high-throughput approaches such as exon- or exon-junction microarrays and next-generation sequencing (RNA-Seq) now make it possible to profile the transcriptome at exon or nucleotide resolution, our ability to quantify the levels of specific RNA variants (e.g., one resulting from the inclusion or skipping of a particular exon) suffers from a low signal-to-noise ratio.

In our lab we study the regulation of RNA processing, particularly alternative splicing, in the nervous system. Initially, we attempted to define global RNA-regulatory networks by focusing on a family of brain-, heart-, and muscle-specific RBPs, the Rbfox proteins, which have been implicated in several neurological diseases. We chose Rbfox proteins because it is known that they recognize a specific hexameric element, UGCAUG; this gives us leverage for using tools from bioinformatics to predict RBP binding targets [5]. Using a phylogenetic tree-based branch length score (BLS)[6] in 28 vertebrate species, we modeled the conservation of this sequence element quantitatively, and were able to predict >1,000 Rbfox target exons, a majority of which are amenable to being validated experimentally. In the list were many transcripts with important neuromuscular functions, a finding that is consistent with the expression pattern of their regulators. This work provided a first glimpse at the extensiveness of tissue-specific RNA-regulatory networks.

However, the binding specificity of Rbfox represents an exception, not a rule. For example, the high-affinity binding sites of the neuron-specific splicing factor Nova, a protein that is important for synaptic function, are characterized by the tetramer YCAY (Y can be either C or U). One would expect YCAY sites to appear approximately once in every 64 nucleotides by chance, so this motif provides very limited predictive power. Excitingly, the Robert Darnell lab at Rockefeller University developed a biochemical assay named crosslinking and immunoprecipitation (CLIP) to isolate RNA fragments that are directly bound by an RBP [7, 8]. Combining CLIP with next-generation high-throughput sequencing, an approach called HITS-CLIP, makes it possible to map in vivo protein-RNA interactions on a genome-wide scale [8, 9]. The analysis of HITS-CLIP data involves mapping short CLIP tags (~50 nucleotides) to the genome and identifying where they cluster. This approach can be used to determine an RBP binding footprint region at a resolution of 30-60 nucleotides. Given the degeneracy of RBP motifs, our next step was to ask whether we could pinpoint the sequences that directly interact with the protein.

One important application of HITS-CLIP data is that it defines a large number of high-quality RBP

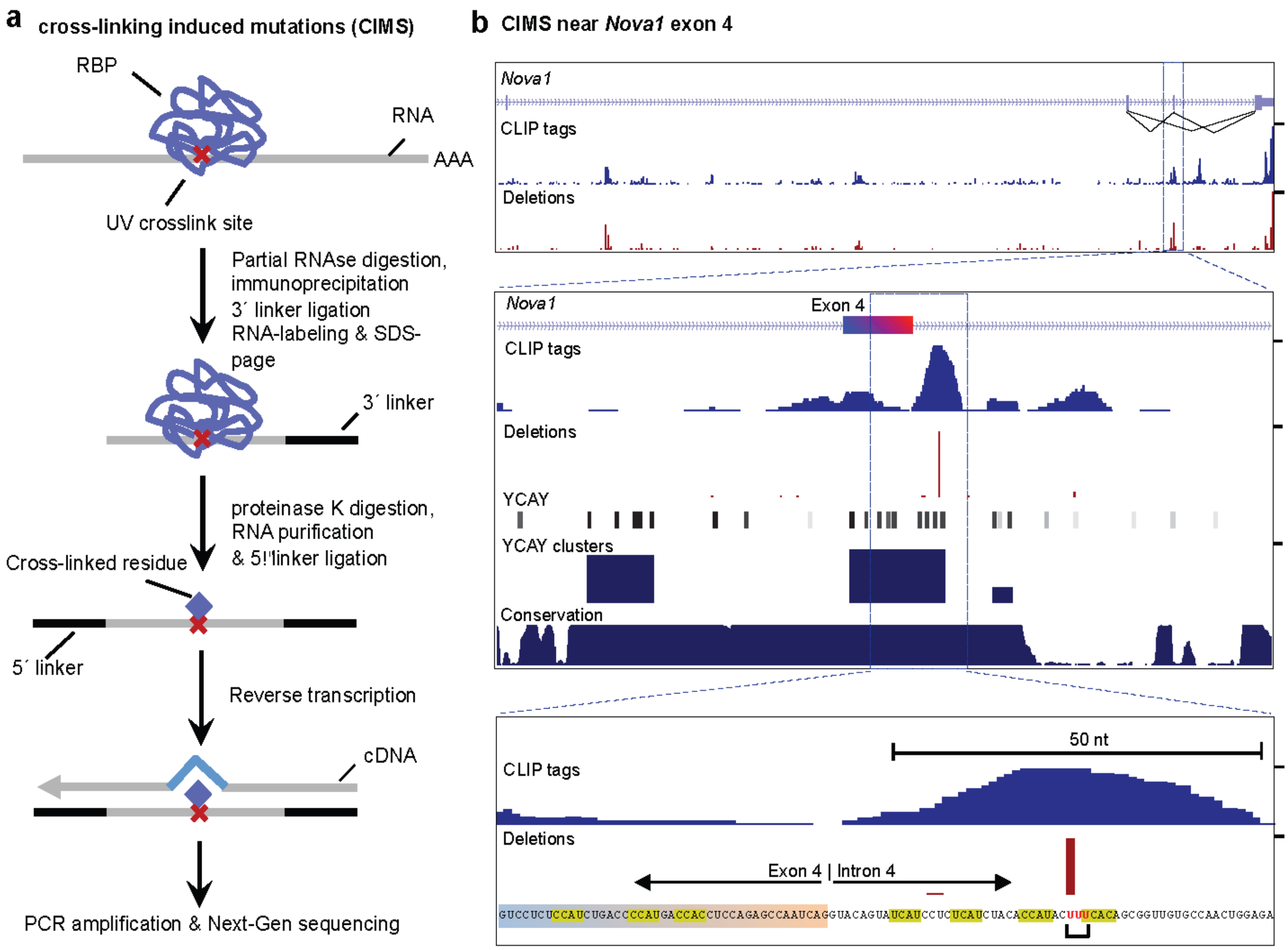


Figure 1: CIMS analysis to determine protein-RNA interactions at a single-nucleotide resolution. (a) Schematic representation of HITS-CLIP and cross-linking induced mutations. Protein-RNA complexes are purified by immunoprecipitation and stringent washing, followed by treatment with proteinase K, a broad-specificity enzyme which cleaves peptide bonds. The remaining cross-linked amino acid(s) attached to RNA, as indicated by the red cross, impose an obstacle for reverse transcription, so that a mutation may be induced during the conversion of RNA to cDNA. CLIP tags are then PCR-amplified and read out by high-throughput sequencing. (b) An example (Nova1 exon 4) of CIMS that precisely maps Nova-RNA interactions. Top panel: the Nova1 gene locus, with the number of CLIP tags and frequency of deletions shown in blue and red, respectively. Inclusion or exclusion of exon 4 is autoregulated by Nova [10]. Middle panel: a zoom-in view of exon4 and flanking intronic sequences. In addition to CLIP tags and deletions, positions of YCAY elements, scores of bioinformatically predicted YCAY clusters (see below), and cross-species sequence conservation in mammals are shown. Bottom panel: A further zoom-in view of sequences around the CIMS. YCAY elements and the nucleotides with deletions are highlighted.

binding sites at a high resolution. This information can be used as a training set for building richer probabilistic models, which can be used to determine the more sophisticated and sometimes subtle rules behind protein-RNA interactions. Despite the small size and degeneracy of individual RBP motifs, we can take advantage of the fact that many RBPs can multimerize or have multiple RNA binding domains (RBDs); thus, several copies of the same sequence motif may be required for high-affinity and functional protein-RNA interactions. For example, three copies of YCAY elements are generally necessary and sufficient for Nova to bind RNA (through three RBDs) [10, 17-19].

We now have multiple types of data that are directly relevant to RNA regulation. This includes exon-sensitive microarray or RNA-Seq data, which in combination with genetic perturbation data can be used to identify RBP-dependent changes in target transcripts. We can also now make bioinformatic predictions and use biochemical methods to detect protein-RNA interactions. However, we now face the problem of maximizing the information we can extract from these data and defining the target regulatory networks of specific RBPs. Each individual dataset is somewhat limited by its signal-to-noise ratio and scope. Profiles of RNA isoforms based on exon-sensitive microarrays or RNA-Seq are subject to a higher level of noise or low coverage at the exon level, and they are unable to distinguish direct

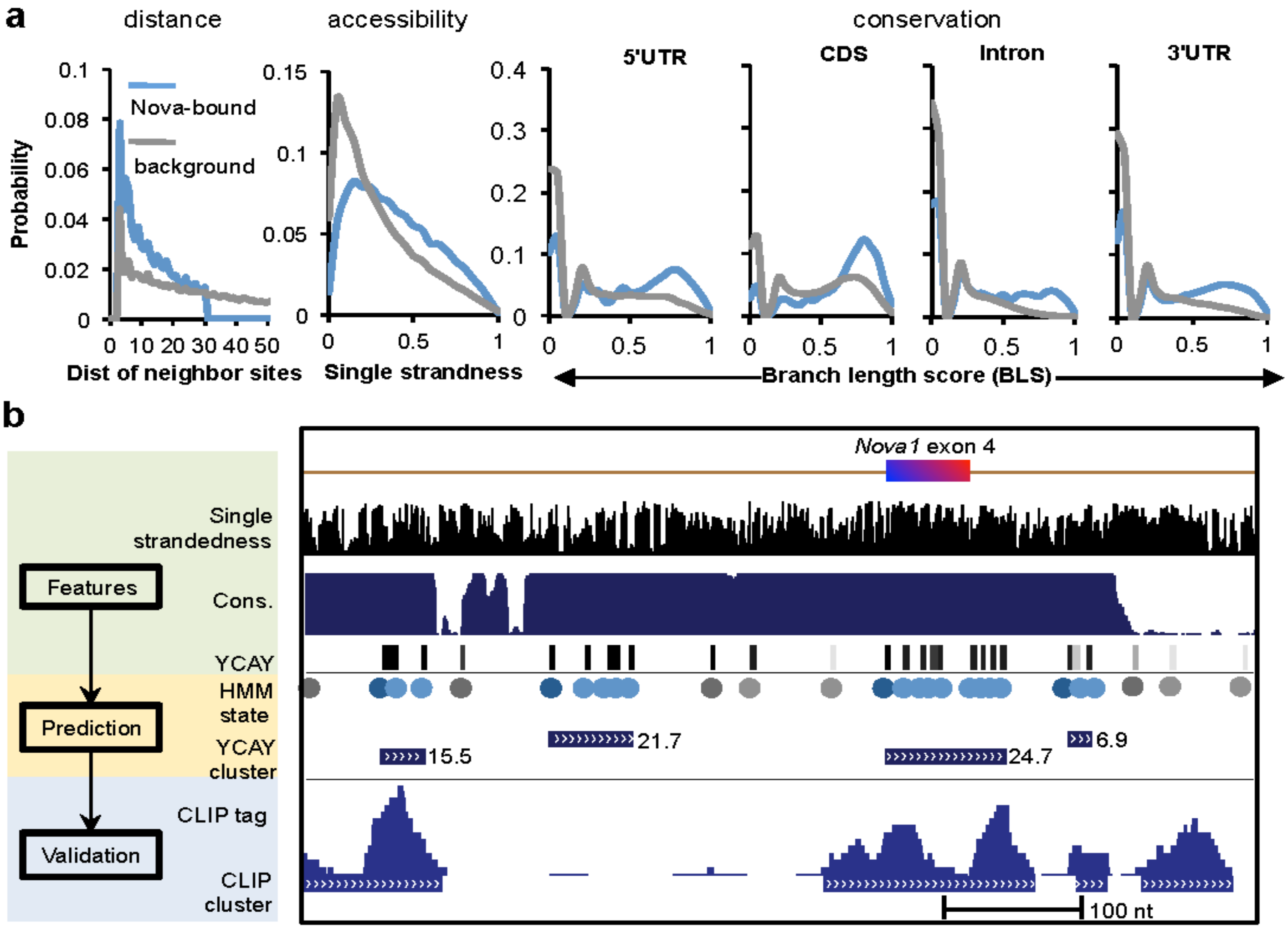


Figure 2: Bioinformatic prediction of clustered RBP motif sites. (a) Features that distinguish in vivo Nova binding sites (blue) and control sequences (gray) learned from Nova CLIP data: spacing between YCAYs, their accessibility and cross-species conservation. (b) YCAY cluster prediction by a hidden Markov model (HMM), which integrates multiple features using sequences flanking Nova1 exon 4 as an example. Predicted high-scoring YCAY cluster (score = 24.7) overlaps with CLIP data, and exactly match detailed mutation analysis. On a genome-wide scale, bioinformatic prediction and CLIP assay identified overlapping, but distinct sets of Nova-binding sites. In the CLIP assay, UV light is used to induce the irreversible crosslinking of protein and RNA, a step that is critical for isolating specific RNAs of interest. The protein is then digested and undergoes the standard cloning procedure of converting RNA into cDNA (i.e., reverse transcription), followed by PCR amplification (Fig. 1a). Interestingly, because one or a few amino acids remain crosslinked with RNA, reverse transcription sometimes introduces errors, such as deletions, at the crosslinked nucleotides, resulting in a mutation in the cDNA. Such crosslinking induced mutation sites (CIMS) provide a signature of the exact protein-RNA interaction and crosslink sites. We developed a statistical method capable of identifying reliable CIMS with reproducible errors, and distinguishing them from random sequencing or alignment errors. CIMS analysis provides a way to determine protein-RNA interactions at single-nucleotide resolution. After we demonstrated the effectiveness of the method using Nova and Argonaut (another RBP) HITS-CLIP data[11], it was validated in a number of additional RBPs [12-16]. In the example shown in Fig. 1b, the Nova1 gene, which encodes Nova proteins, has an alternative exon 4, which is autoregulated by the Nova proteins. The CLIP data defined a Nova-binding footprint of ~50 nucleotides. Using CIMS analysis, we determined that the exact crosslinked nucleotide was a uridine in a Nova-binding YCAY element.

and indirect targets. Predictions of protein-RNA interactions generated by bioinformatics methods or CLIP experiments do not necessarily imply functional regulation. Even when we compared CLIP data and bioinformatics predictions of RBP motif sites, which show substantial overlap, we found binding sites that are distinct to each set.

Therefore, our next step was to design an integrative strategy for properly weighing and combining multiple types of datasets, so that individually weak bits of information can be synthesized to make confident predictions of direct, functional RBP targets. Our solution was to develop a Bayesian network model for such purposes (Fig. 3a). When we applied this method to study Nova, we successfully identified ~700 direct Nova target alternative splicing events; these predictions had a high validation rate (~90%) and sensitivity (~78%) (Fig. 3b) [21]. A majority of these targets cannot be predicted confidently using individual datasets.

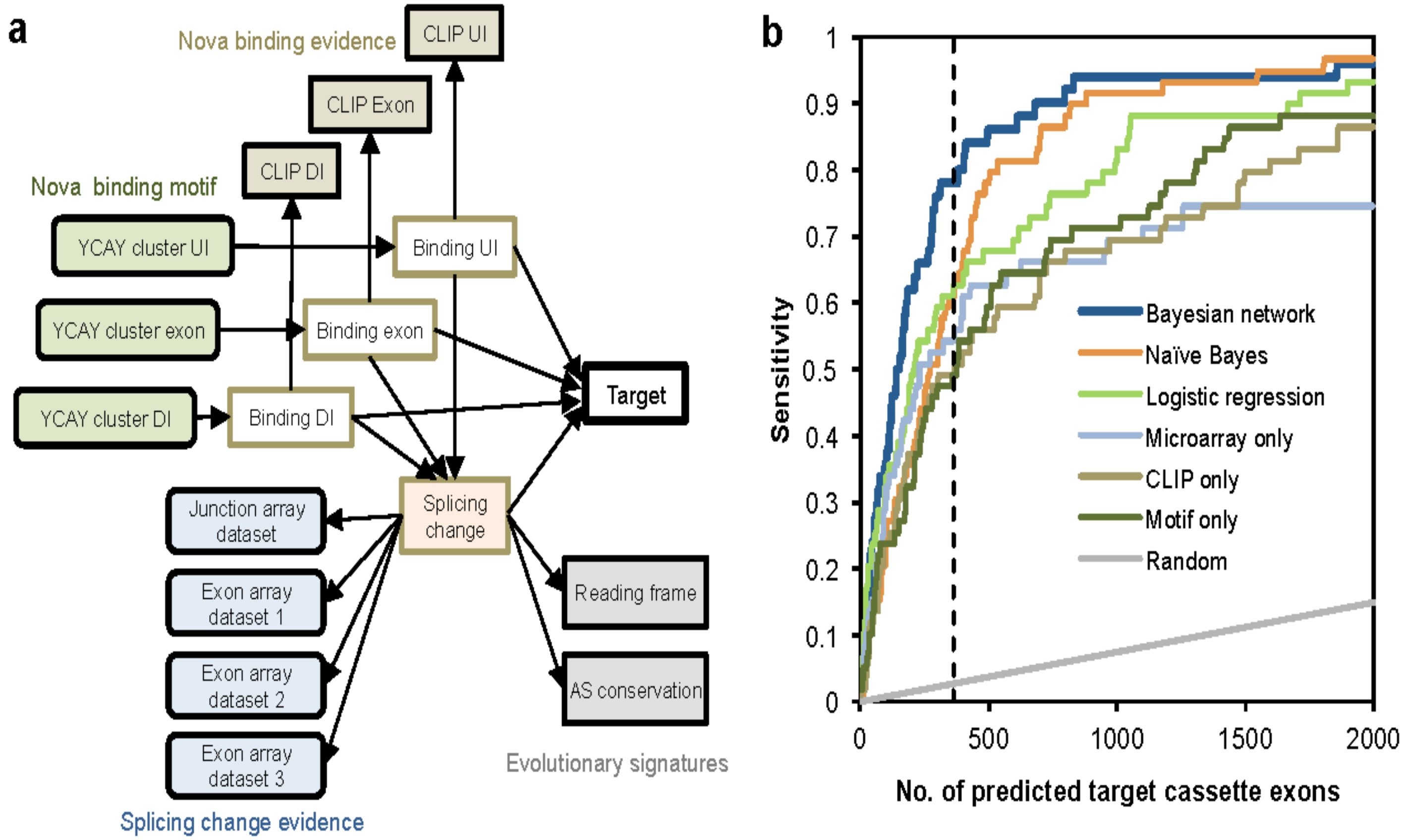


Figure 3: Integrative prediction of Nova targets using a Bayesian network. (a) Design of the Bayesian network. The 17 nodes (variables) model four types of data, including YCAY clusters and CLIP clusters in each cassette exon or flanking upstream (UI) and downstream introns (DI), splicing microarrays comparing wild-type and Nova knockout brains, and evolutionary signatures. Edges reflect causal relationships among variables. (b) Comparison of Nova target prediction by different methods (Bayesian network, naïve Bayes, and logistic regression) or using individual datasets (exon microarrays, CLIP clusters, and YCAY clusters). Each curve represents the prediction sensitivity with varying stringency; the performance of random predictions is also shown. The dotted line indicates the top 363 predictions. Additional analysis revealed several other insights. First, when we compared the YCAY elements that CLIP data revealed to be bound by Nova to elements that Nova does not bind, Nova-bound YCAY elements tended to cluster together more closely (Fig. 2a). Second, RNA also presents complex secondary structures, so that some sequences are more accessible to RBPs than others. Indeed, we observed that Nova-bound YCAYs are more frequently located in single-stranded regions. Finally, evolution selects against mutations in regulatory sequences of functional importance; this leads to a higher level of cross-species conservation that can be observed in Nova binding sites. We can now integrate these features together quantitatively, using a probabilistic model (a hidden Markov model in this case) to predict RBP motif sites. The resulting algorithm, named mCarts, is able to predict clustered RBP motif sites with sufficient specificity for successful experimental validation [20]. For example, among the alternative exons with nearby YCAY clusters that mCarts ranks highest, up to 90% showed Nova-dependent inclusion when we compared wild type and Nova knockout mouse brains. Using the same example of Nova1 exon 4 (Fig. 2b), we are now able to predict the YCAY cluster that is critical for Nova function, which exactly matches previous detailed mutation analysis [10].

The comprehensiveness of the RNA-regulatory networks we have defined provides much-improved statistical power in revealing general principles of these networks. For example, exploratory analysis of the Nova target network unexpectedly revealed a subset of exons that are regulated by both Nova and Rbfox—these exons are more frequently disrupted in neurological disorders such as autism. In addition, we found a direct coupling between splicing regulation and post-translational modification, particularly phosphorylation (Fig. 4); specifically, Nova target transcripts are enriched in RNAs that encode phosphoproteins and, interestingly, the phosphorylation sites are frequently encoded by the Nova-regulated alternative exons. Furthermore, kinases and phosphatases are also over-represented in Nova targets. This observation is important because alternative exons are in general much smaller than constitutive exons. It is therefore unclear how alterations in several amino acids, frequently in disordered regions, can have large effects on the function of the resulting protein products. The two-tiered regulation, first through alternative splicing, and then through post-translational regulation, provides a likely mechanism for fine-tuning protein-protein interactions and cell signaling. Recently, independent studies validated this observation [22].

In conclusion, we have developed a range of tools to dissect RNA-regulatory networks. While we have discussed Nova as an example in order to demonstrate the effectiveness of this integrative

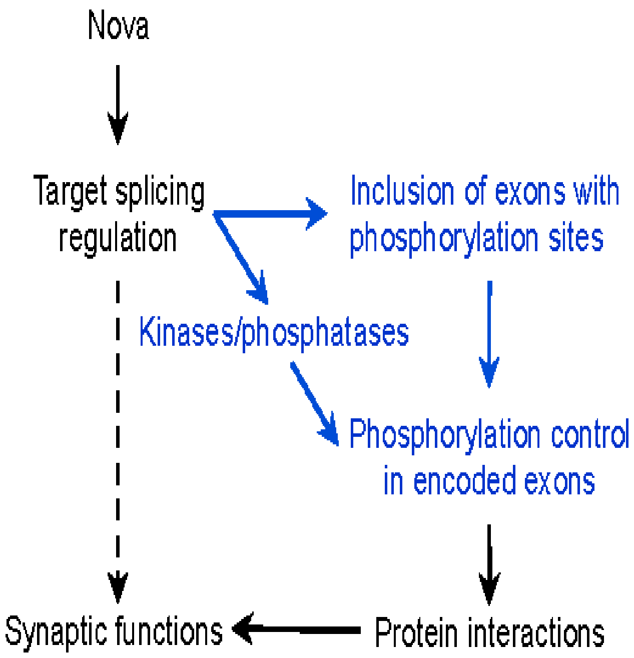

Nova
Target splicing regulation
Inclusion of exons with phosphorylation sites
Kinases/phosphatases
Phosphorylation control in encoded exons
Synaptic functions
Protein interactions

Figure 4: A model of Nova regulated alternative splicing regulation to control protein phosphorylation patterns. Nova can potentially affect the activity of kinases and phosphatases, enzymes to add or remove phosphate groups in substrate proteins, or affect the inclusion or exclusion of alternative exons encoding phosphorylation sites. This two-tiered regulation provides a mechanism to modulate downstream protein-protein interactions and synaptic functions.

strategy, these methods can be readily extended to other RBPs, a direction that we are currently exploring. We would like to understand how the combinatorial and dynamic regulation of RNA contributes to specification of cell types during the differentiation of embryonic stem cells into neurons, and how such regulation is perturbed in neural degenerative disorders and brain tumors. We believe our strategy holds the potential to provide mechanistic insights that will have clinical relevance.

REFERENCES

1. Licatalosi, D.D. and R.B. Darnell, RNA processing and its regulation: global insights into biological networks. Nat Rev Genet, 2010. 11(1): p. 75-87.

2. Kalsotra, A. and T.A. Cooper, Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet, 2011. 12(10): p. 715-729.

3. Cooper, T.A., L. Wan, and G. Dreyfuss, RNA and disease. Cell, 2009. 136(4): p. 777-793.

4. Licatalosi, D.D. and R.B. Darnell, Splicing regulation in neurologic disease. Neuron, 2006. 52(1): p. 93-101.

5. Zhang, C., Z. Zhang, J. Castle, S. Sun, J. Johnson, A.R. Krainer, and M.Q. Zhang, Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. Genes Dev, 2008. 22(18): p. 2550-2563.

6. Stark, A., M.F. Lin, P. Kheradpour, J.S. Pedersen, L. Parts, J.W. Carlson, M.A. Crosby, M.D. Rasmussen, S. Roy, A.N. Deoras, J.G. Ruby, J. Brennecke, E. Hodges, A.S. Hinrichs, A. Caspi, B. Paten, S.-W. Park, M.V. Han, M.L. Maeder, B.J. Polansky, B.E. Robson, S. Aerts, J. van Helden, B. Hassan, D.G. Gilbert, D.A. Eastman, M. Rice, M. Weir, M.W. Hahn, Y. Park, C.N. Dewey, L. Pachter, W.J. Kent, D. Haussler, E.C. Lai, D.P. Bartel, G.J. Hannon, T.C. Kaufman, M.B. Eisen, A.G. Clark, D. Smith, S.E. Celniker, W.M. Gelbart, and M. Kellis, Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature, 2007. 450(7167): p. 219-232.

7. Ule, J., K.B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R.B. Darnell, CLIP identifies Nova-regulated RNA networks in the brain. Science, 2003. 302(5648): p. 1212-1215.

8. Darnell, R.B., HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley Interdiscip Rev RNA, 2010. 1(266-286): p. 266-286.

9. Licatalosi, D.D., A. Mele, J.J. Fak, J. Ule, M. Kayikci, S.W. Chi, T.A. Clark, A.C. Schweitzer, J.E. Blume, X. Wang,

J.C. Darnell, and R.B. Darnell, HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature, 2008. 456(7221): p. 464-469.

10. Dredge, B.K., G. Stefani, C.C. Engelhard, and R.B. Darnell, Nova autoregulation reveals dual functions in neuronal splicing. EMBO J., 2005. 24: p. 1608-1620.

11. Zhang, C. and R.B. Darnell, Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat Biotech, 2011. 29: p. 607-614.

12. Cho, J., H. Chang, S.C. Kwon, B. Kim, Y. Kim, J. Choe, M. Ha, Y.K. Kim, and V.N. Kim, LIN28A is a suppressor of ER-associated translation in embryonic stem cells. Cell, 2012. 151(4): p. 765-777.

13. Wang, E.T., N.A.L. Cody, S. Jog, M. Biancolella, T.T. Wang, D.J. Treacy, S. Luo, G.P. Schroth, D.E. Housman, S. Reddy, E. Lv©cuyer, and C.B. Burge, Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. Cell, 2012. 150(4): p. 710-724.

14. Charizanis, K., K.-Y. Lee, R. Batra, M. Goodwin, C. Zhang, Y. Yuan, L. Shiue, M. Cline, M.M. Scotti, G. Xia, A. Kumar, T. Ashizawa, H.B. Clark, T. Kimura, M.P. Takahashi, H. Fujimura, K. Jinnai, H. Yoshikawa, M.r. Gomes-Pereira, G.v. Gourdon, N. Sakai, S. Nishino, T.C. Foster, M. Ares Jr, R.B. Darnell, and M.S. Swanson, Muscleblind-like 2-mediated alternative splicing in the developing brain and dysregulation in myotonic dystrophy. Neuron, 2012. 75(3): p. 437-450.

15. Licatalosi, D.D., M. Yano, J.J. Fak, A. Mele, S.E. Grabinski, C. Zhang, and R.B. Darnell, Ptbp2 represses adult-specific splicing to regulate the generation of neuronal precursors in the embryonic brain. Genes Dev, 2012. 26: p. 1626-1642.

16. Pandit, S., Y. Zhou, L. Shiue, G. Coutinho-Mansfield, H. Li, J. Qiu, J. Huang, G.W. Yeo, M. Ares Jr, and X.-D. Fu, Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. Mol Cell: p. pii: S1097-2765(13)00204-9. doi: 10.1016/j.molcel.2013.03.001.

17. Jensen, K.B., B.K. Dredge, G. Stefani, R. Zhong, R.J. Buckanovich, H.J. Okano, Y.Y.L. Yang, and R.B. Darnell, Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. Neuron, 2000. 25(2): p. 359-371.

18. Dredge, B.K. and R.B. Darnell, Nova regulates GABAA receptor ?2 alternative splicing via a distal downstream UCAU-rich intronic splicing enhancer. Mol. Cell. Biol., 2003. 23(13): p. 4687-4700.

19. Ule, J., G. Stefani, A. Mele, M. Ruggiu, X. Wang, B. Taneri, T. Gaasterland, B.J. Blencowe, and R.B. Darnell, An RNA map predicting Nova-dependent splicing regulation. Nature, 2006. 444: p. 580-586.

20. Zhang, C., K.-Y. Lee, M.S. Swanson, and R.B. Darnell, Prediction of clustered RNA-binding protein motif sites in the mammalian genome. Nucleic Acids Res, 2013. in press.

21. Zhang, C., M.A. Frias, A. Mele, M. Ruggiu, T. Eom, C.B. Marney, H. Wang, D.D. Licatalosi, J.J. Fak, and R.B. Darnell, Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. Science, 2010. 329: p. 439-443.

22. Merkin, J., C. Russell, P. Chen, and C.B. Burge, Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science, 2012. 338(6114): p. 1593-1599.

FROM GENOME-SCALE SEQUENCING DATA TO ANGSTROM-SCALE INSIGHTS: THE ENZYME DNASE I REVEALS ITS TRUE COLORS

HARMEN J. BUSSEMAKER
DEPARTMENT OF BIOLOGICAL SCIENCES
INITIATIVE IN SYSTEMS BIOLOGY
COLUMBIA UNIVERSITY

Bovine deoxyribonuclease I — better known as DNase I — is a DNA-nicking enzyme that has been a major workhorse of molecular biology for decades. An ideal reagent for mapping the "footprints" of proteins bound to DNA, it has been used on a genome-wide scale in conjunction with high-throughput sequencing to probe the detailed structure of chromatin, which is often very different between cell types. This approach was used most notably in the context of the NIH-funded ENCODE project.

Surprisingly for such a widely used enzyme, it has until recently remained unclear whether DNase I exhibits intrinsic sequence preferences when cleaving "naked" DNA. Such biases would obviously have to be taken into account when interpreting in vivo DNase footprinting patterns. One reason for the lack of clarity has been that two different studies, each published well over a decade ago and based on the quantification of traditional gels, arrived at incompatible conclusions.

The current availability of high-throughput sequencing technology presented an opportunity to revisit this long-standing question. In a recent paper published in the Proceedings of the National Academy Sciences [1], MAGNet investigator Dr. Harmen Bussemaker and his graduate student Allan Lazarovici, in close collaboration with the laboratory of Dr. John Stamatoyannopoulos at the University of Washington and the group of Dr. Remo Rohs at the University of Southern California, an unexpectedly rich spectrum of insights was obtained from an in-depth study of the properties of DNase I.

The Bussemaker group has long been interested in building accurate biophysical models of the sequence preferences of DNA-binding transcription factors from high-throughput genomics data. Such knowledge is highly valuable in the quest to understand and predict the biological function of these regulatory proteins within the regulatory network of the cell. Being able to quantify the often subtle but functionally important differences in DNA preference makes it possible to attribute genome-wide changes in mRNA expression to changes in protein-level activity of the correct transcription factors, and thus helps with the identification of the "master regulators" of specific biological processes [2, 3]. Recently, the FeatureREDUCE algorithm, developed by postdoc Todd Riley in the Bussemaker group, emerged as the top-performing algorithm in a benchmark challenge to infer accurate sequence-to-affinity models from high-throughput protein binding microarray (PBM) data [4]. This challenge was partly based on data from the DREAM5 competition held under the auspices of MAGNet.

From the in-depth study of DNase I led by the Bussemaker group it rapidly became clear that in addition to its value for chromatin footprinting, the enzyme is an ideal vehicle for investigating fundamental aspects of protein-DNA interaction. Whenever a DNase I protein binds to double-stranded DNA, there is a small probability that one of the strands of the sugar-phosphate backbone gets cleaved. This leaves a permanent mark on the DNA, which records the binding event at single-base-pair resolution. When DNase I is allowed to digest a sample of purified genomic DNA, such marks will accumulate across the genome. After the reaction is stopped, the genomic location of each "cut" can be mapped by sequencing the ends of millions of the resulting single-stranded fragments of genomic DNA in parallel using high-throughput sequencing technology.

As part of the collaboration, the Stamatoyannopoulos laboratory collected deep DNase-seq data for genomic DNA purified from the fibroblast cell line IMR90, which has been widely used in the ENCODE project. Using these data, the Bussemaker group built a model of the intrinsic DNA sequence preferences of DNase I. What emerged is that the activity of the enzyme is mainly affected by the base identity at the first three nucleotides on either side of the location of cleavage. With this motivation, they constructed a lookup table of the relative cleavage rate at the center of each of the 46 possible hexamers by tabulating all cleavage events, while normalizing by the frequency of

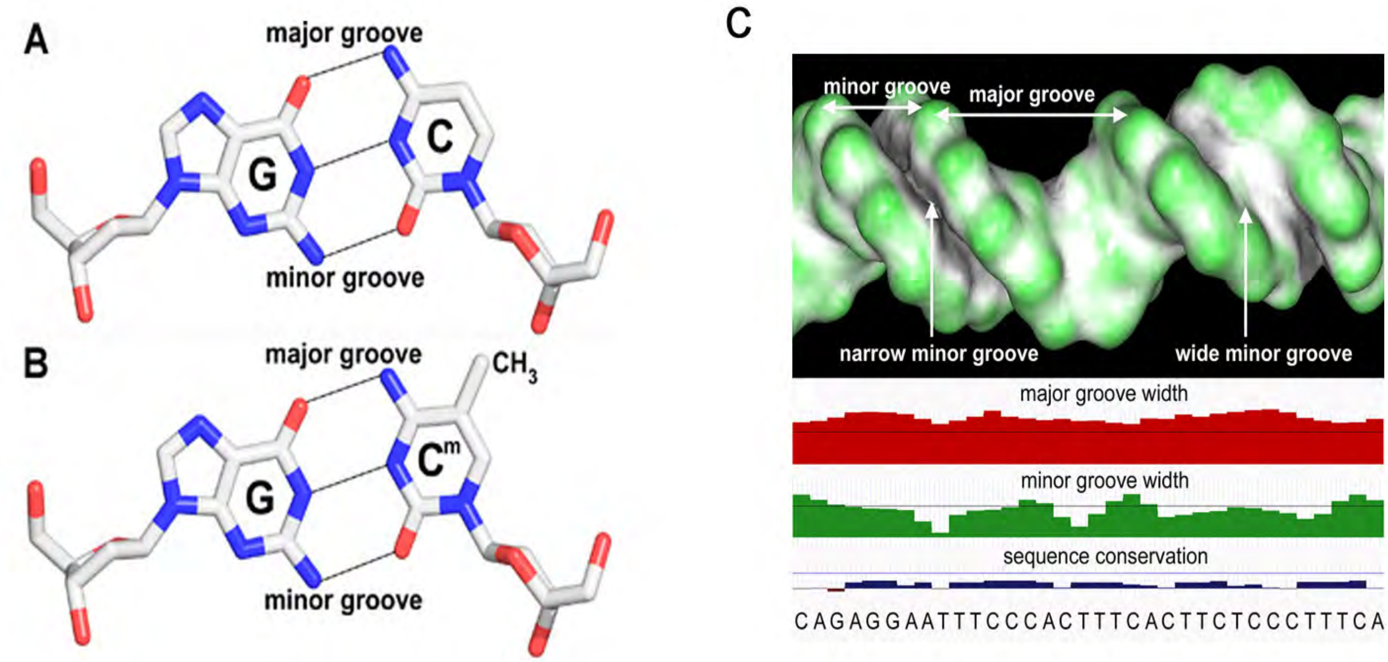


Figure 1: Background information about DNA structure. (A,B) The methyl group that gets added to cytosine is located in the major groove. (C) Variation in DNA sequence drives variation in minor groove width, even in the absence of DNA-binding factors

each hexamer in the genome. Strikingly, the difference in cleavage rate between the most and the least cleavable hexamer was found to be roughly a thousand-fold!

When low-throughput data are used, the modeling of DNA binding preferences of transcription factors is almost invariably done in the form of weight matrices and sequence logos. Examples are the widely used TRANSFAC and JASPAR databases. The underlying assumption for weight matrices is that the effects of mutations at different nucleotide positions within the protein-DNA binding interface on binding affinity are independent. The Bussemaker lab explicitly tested the validity of this assumption using the accurate hexamer tables constructed from the DNase-seq data. Extreme dependencies were found to exist, especially between adjacent nucleotide positions (Figure 2). In other words, traditional weight matrices are inadequate for capturing the DNA sequence preferences of DNase I. Since oligomer based lookup tables are not always a feasible option for other DNA-binding proteins, especially large transcription factor complexes whose binding interface over a larger number of base pairs, the Bussemaker lab recently developed robust new methodology to capture nucleotide dependencies in a biophysical model as part of the FeatureREDUCE software suite (Riley et al., submitted).

For DNase I, the dependencies that were observed between neighboring nucleotides suggested a possible role for DNA shape in the observed strong variation of cleavage rate with base sequence. Previous work by Dr. Rohs, performed while he was a postdoc in the laboratory of MAGNet investigator Dr. Barry Honig, had uncovered the broad importance of DNA shape readout for protein-DNA interactions whenever positively charged amino-acid side-chains such as arginine contact the DNA via the minor grove. Another particularly successful collaboration, between the Honig and Bussemaker groups and the laboratory Dr. Richard Mann, had focused on the shape readout by Hox proteins in the fruit fly Drosophila [5, 6]. Because x-ray crystal structures indicate that DNase I interacts with DNA only via a single location in the minor groove, it therefore made a lot of sense to think of the enzyme as a pure sensor of minor grove shape. Consequently, any insights obtained from the DNase-seq data regarding DNA shape readout might apply more generally to protein-DNA interactions.

To address the relationship between DNase I cleavage rate and minor groove geometry for DNase I, the Rohs group used detailed all-atom computer simulations of free DNA molecules to predict DNA shape parameters for a variety of hexanucleotides, covering the entire range from highly to poorly cleavable. As hoped, the variation in these shape parameters turned out to be highly predictive of the variation in cleavage rate (Figure 3). The shape of a DNA molecule is of course determined by its base sequence, but in a way that is not at all easy to predict. As it turns out, the most cleavable hexamer sequences have a narrower minor groove. This causes the electrostatic interaction with arginine side-chains of DNase I to be enhanced. This structural mechanism had been previously described for transcription factors such as the Hox proteins mentioned above. However, the depth

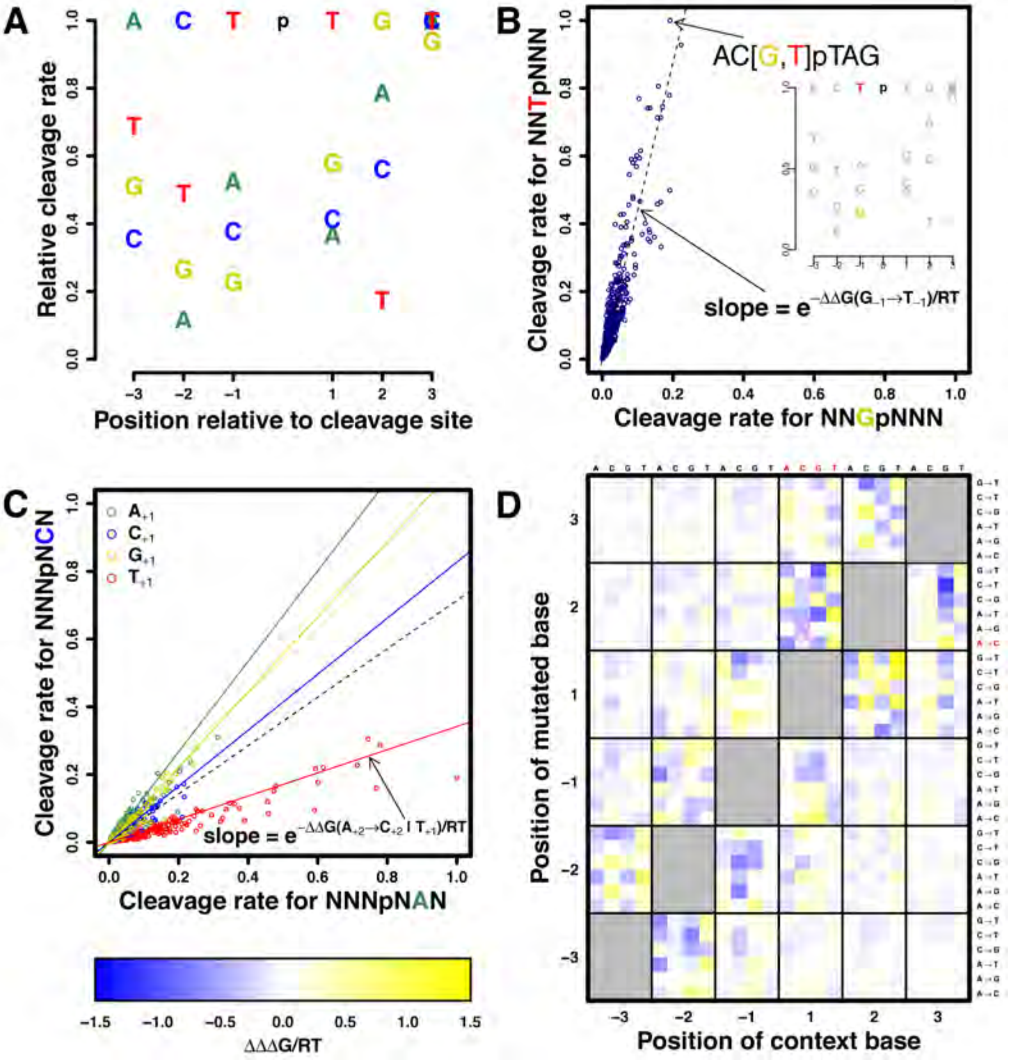


Figure 2: Deep sequencing reveals striking positional dependencies between nucleotide positions within the DNase I recognition site. (A) Position-specific relative cleavage rate parameters as derived from DNase I digestion of human genomic DNA from IMR90 fibroblasts under the assumption of independence between nucleotides. Dependence on local sequence context is largely limited to a hexamer centered at the cleaved backbone bond. (B) Comparison between cleavage rates for pairs of hexamers that are related by a single-nucleotide substitution. The slope of the dashed line corresponds to the position-specific cleavage rate in panel (A), and is directly related to the "unconditional" ΔΔG, the change in binding free energy associated with the point mutation. The fold-change in cleavage rate due to a mutation from G to T at position -1 is largely independent of the base identity of the five neighboring nucleotides. (C) Breakdown of the independence assumption (dashed line). The effect on cleavage rate of a point mutation from A to C at position +2 is highly dependent on the base identity at the "modulating" position +1. Using a "conditional" ΔΔG for each possible base at position +1 (colored lines) provides a far more accurate description. (D) The strength of the positional dependencies can be quantified in terms of a new quantity "ΔΔΔG", defined as the difference between the conditional and unconditional ΔΔG. The values in the highlighted row and columns correspond to the ratio in slope between each of the colored solid lines and the dashed line in (C). Far away from the diagonal ΔΔΔG becomes numerically small (white in heat map), indicating an increasing degree of independence.

of the DNase-seq data allowed it to be analyzed at an unprecedented level of quantification across thousands of different sequences.

The last and most unexpected insight obtained from the DNase I project was related to DNA methylation. In each human cell, methyltransferase enzymes convert cytosine to 5-methylcytosine by adding a methyl group in the DNA major groove, usually in the context of CpG dinucleotides. Only a specific subset of the genomic cytosine bases gets methylated, and this methylation pattern is highly cell-type dependent. High-throughput sequencing had been used to map all methylation events in the genome [7]. Notably, this was done for the same IMR90 cell line that the Stamatoyannopoulos lab had used for their DNase-seq experiment. This provided a unique opportunity to analyze whether DNase I cleavage rate was sensitive to cytosine methylation. This was not necessarily an obvious question, because DNase I interacts with the minor groove and the methyl group attached to cytosine sits in the major groove. However, analysis of DNase-seq data from a digest of purified genomic DNA from a different organism (the yeast Saccharomyces cerevisae) had revealed that certain CpG-containing hexamers were cleaved quite differently when the two organisms were compared. Since the computational analysis controlled for the genome sequence, this represented a puzzle that needed to be resolved.

To make a long story short, it turns out that even though cytosine methylation happens in the major groove, one of its key effects is to narrow the minor groove (Figure 4). Thus, varying the base sequence of genomic DNA is not the only way in which the cell can modulate the landscape of minor groove shape along its genome. Cytosine methylation provides another avenue for this at the epigenetic level. Because contacts with the minor groove, long under-studied, are increasingly found to be important for a variety of transcription factors and their complexes, it is likely that the same mechanism of minor-groove-shape-mediated modulation of binding affinity by cytosine methylation applies there as well. Thus, the insights derived from the small DNase I enzyme could go a long way toward understanding how variation in methylation patterns can drive changes in mRNA expression, which is currently one of the key unsolved problems in regulatory genomics.

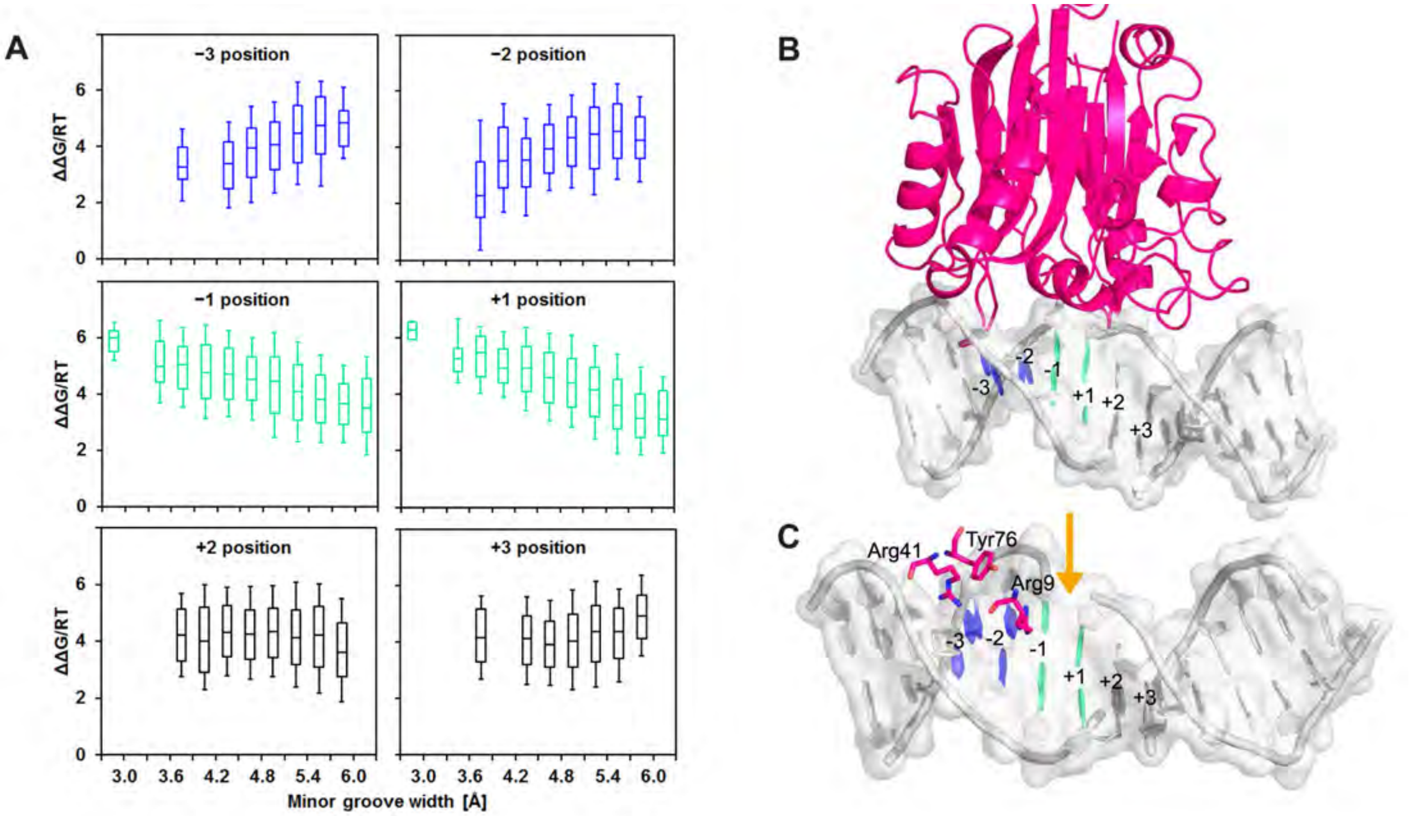


Figure 3: Minor groove width is predictive of DNase I cleavage rate. (A) ΔΔG derived from the negative logarithm of cleavage rate as a function of minor groove width (MGW) at the six positions of all 4,096 unique hexamers. MGW of this region was predicted for naked binding sites based on a pentamer-based high-throughput (HT) shape prediction approach used in Slattery et al. (2011). HT predictions for all possible 16 dinucleotide flanks were averaged and values of MGW that fall within intervals of 0.3 Å assigned to groups of sequences for which cleavage rates are shown as box plots. (B) DNase I-DNA complex based on crystal structure (PDB ID 2DNJ). Base pairs at positions -3 and -2, where DNase I cleavage anticorrelates with MGW, are highlighted in blue. Base pairs at positions -1 and +1, where DNase I cleavage correlates positively with MGW, are highlighted in green. Regions where no correlation could be detected are shown in gray. The color code of the base pairs in the crystal structure is equivalent to the one used for the box plots. (C) DNase I-minor groove contacts within a distance of 5 Å from any base atom are shown for the same crystal structure. Arg41 and Arg9 bind upstream of the cleavage site, where MGW anticorrelates with DNase I cleavage (blue base pairs). This anticorrelation likely arises from the attraction between the positively charged arginine residues and the locally enhanced negative electrostatic potential. The cleavage site (indicated by the orange arrow), by contrast, is located in a region where MGW correlates positively with DNAse I cleavage (green base pairs).

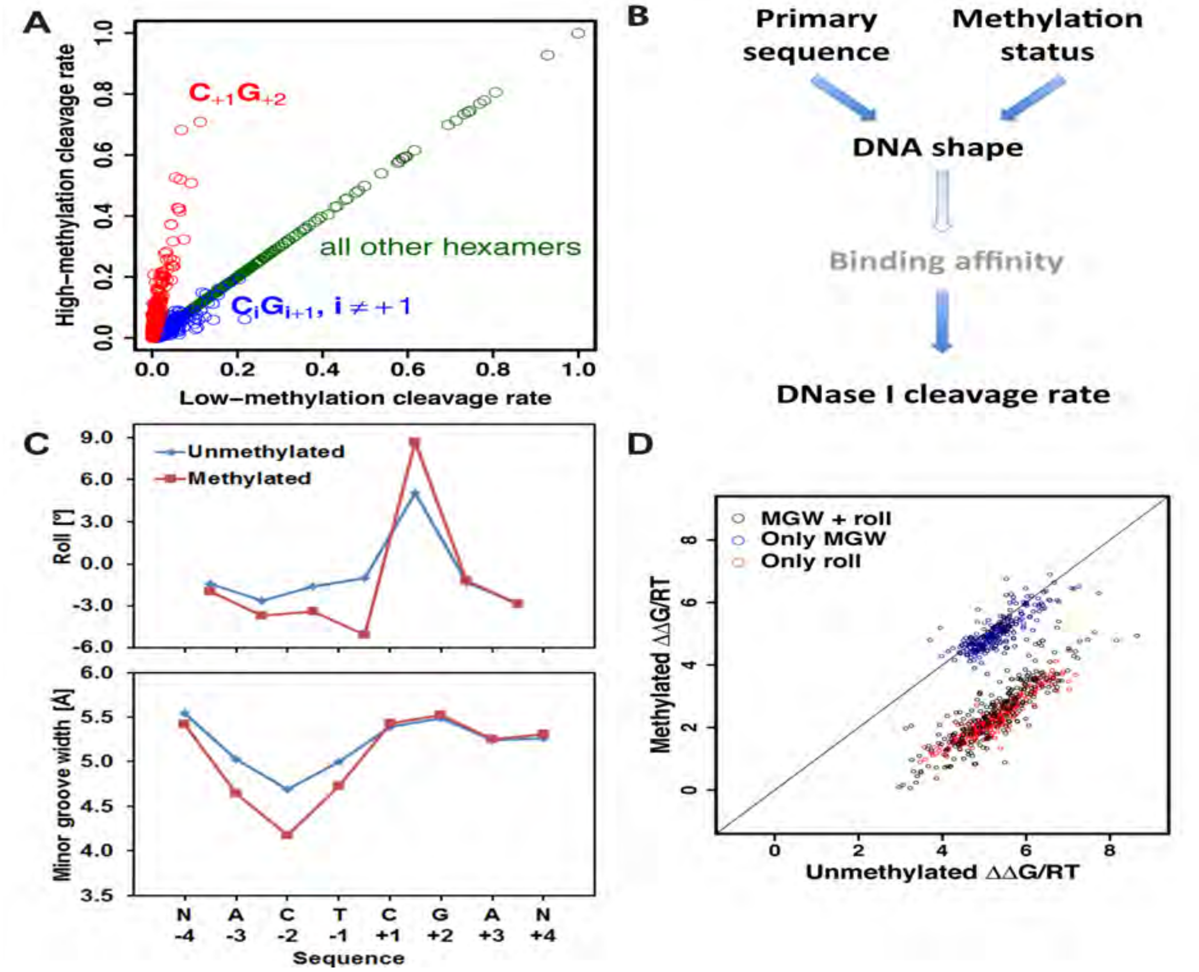


Figure 4: Observation and analysis of the effect of methylation on DNase I cleavage rate. (A) The rate of cleavage depends strongly on the DNA methylation status. We used a positional map of DNA methylation in IMR90 (Lister et al.) to delineate subsets of genomic positions with low/high degrees of CpG methylation, respectively. Comparison between the hexamer cleavage rates derived from these respective subsets shows an ~8-fold increase in cleavage rate for hexamers with a methylated CpG immediately downstream of the cleaved phosphate (red points). (B) Interplay between DNA sequence and methylation status, DNA geometry, and DNase I cleavage suggested by our analysis. (C) Roll and MGW of methylated and unmethylated versions of the same hexamer based on MC predictions. Methylation leads to an increase in the positive roll angle at the CpG dinucleotide and a narrowing of the MGW at position -2 by roughly 0.5 Å. (D) The effect of methylation on DNAse I cleavage can be predicted in silico by training a model to predict the cleavage rates of unmethylated DNA sequences of type NNNCGN using information on DNA minor groove width and roll angle along these same unmethylated sequences. An increase in cleavage rate (i.e., data points shifting downward) is predicted when minor groove widths and roll angles for the methylated versions of the sequences are supplied as input to the model.

It is gratifying to see how the MAGNet-funded collaboration on Hox protein function between the Mann, Honig, and Bussemaker labs helped spawn a novel and highly interdisciplinary project, which, starting from genomewide sequencing data, was able to yield structural insights at Angstrom resolution.

REFERENCES

1. Lazarovici, A., T. Zhou, A. Shafer, A.C. Dantas Machado, T.R. Riley, R. Sandstrom, P.J. Sabo, Y. Lu, R. Rohs, J.A. Stamatoyannopoulos, and H.J. Bussemaker, Probing DNA shape and methylation state on a genomic scale with DNase I. Proc Natl Acad Sci U S A, 2013. 110(16): p. 6376-81.

2. Bussemaker, H.J., B.C. Foat, and L.D. Ward, Predictive modeling of genome-wide mRNA expression: from modules to molecules. Annu Rev Biophys Biomol Struct, 2007. 36: p. 329-47.

3. Lee, E. and H.J. Bussemaker, Identifying the genetic determinants of transcription factor activity. Mol Syst Biol,

2010. 6: p. 412.

4. Weirauch, M.T., A. Cote, R. Norel, M. Annala, Y. Zhao, T.R. Riley, J. Saez-Rodriguez, T. Cokelaer, A. Vedenko, S. Talukder, H.J. Bussemaker, Q.D. Morris, M.L. Bulyk, G. Stolovitzky, and T.R. Hughes, Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol, 2013. 31(2): p. 126-34.

5. Joshi, R., J.M. Passner, R. Rohs, R. Jain, A. Sosinsky, M.A. Crickmore, V. Jacob, A.K. Aggarwal, B. Honig, and R.S. Mann, Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell, 2007. 131(3): p. 530-43.

6. Slattery, M., T. Riley, P. Liu, N. Abe, P. Gomez-Alcala, I. Dror, T. Zhou, R. Rohs, B. Honig, Harmen J. Bussemaker, and Richard S. Mann, Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. Cell, 2011. 147(6): p. 1270-1282.

7. Lister, R., M. Pelizzola, R.H. Dowen, R.D. Hawkins, G. Hon, J. Tonti-Filippini, J.R. Nery, L. Lee, Z. Ye, Q.M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A.H. Millar, J.A. Thomson, B. Ren, and J.R. Ecker, Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 2009. 462(7271): p. 315-22.

FIRST METHOD FOR GLOBAL PROBABILISTIC ANNOTATIONS OF METABOLIC NETWORKS

DENNIS VITKUP LAB

One challenge facing systems biology is to develop accurate maps of the complex networks of genes involved in metabolism. Although a number of databases can be used to predict metabolic gene functions by comparing their genetic sequences to those of enzymes for which function is better understood (a process called homology modeling), these approaches have been known to generate imprecise predictions and can not explain how metabolic activity emerges from complex systems of molecular interactions. Moreover, a probabilistic approach is necessary for predicting biochemical function because of the inherent uncertainty contained within existing data.

We have developed the first genome-wide framework for making probabilistic annotations of metabolic networks. Our algorithm, called Global Biochemical Reconstruction Using Sampling (GLOBUS), combines information about sequence homology with context-specific information including phylogeny, gene clustering, and mRNA co-expression to predict the probability of biochemical interactions between specific metabolic genes. By integrating these different categories of information using a principled probabilistic framework, this approach overcomes limitations of considering only one functional category or one gene at a time, providing an accurate, global prediction of metabolic networks.

To test the effectiveness of GLOBUS, we generated genome-wide predictions about the metabolic networks in the bacteria Bacillus subtilis and Staphylococcus aureus, and tested three predictions experimentally to validate their accuracy. Experiments by collaborators at ETH Zurich confirmed the predicted functions, including an important pathway of genes responsible for spore formation in B. subtilis.

Data from GLOBUS can be combined easily with metabolomic, proteomic, and fluxomic data being generated using experimental methods. By identifying complementary patterns that appear across these different data types, researchers will be able to make increasingly confident predictions

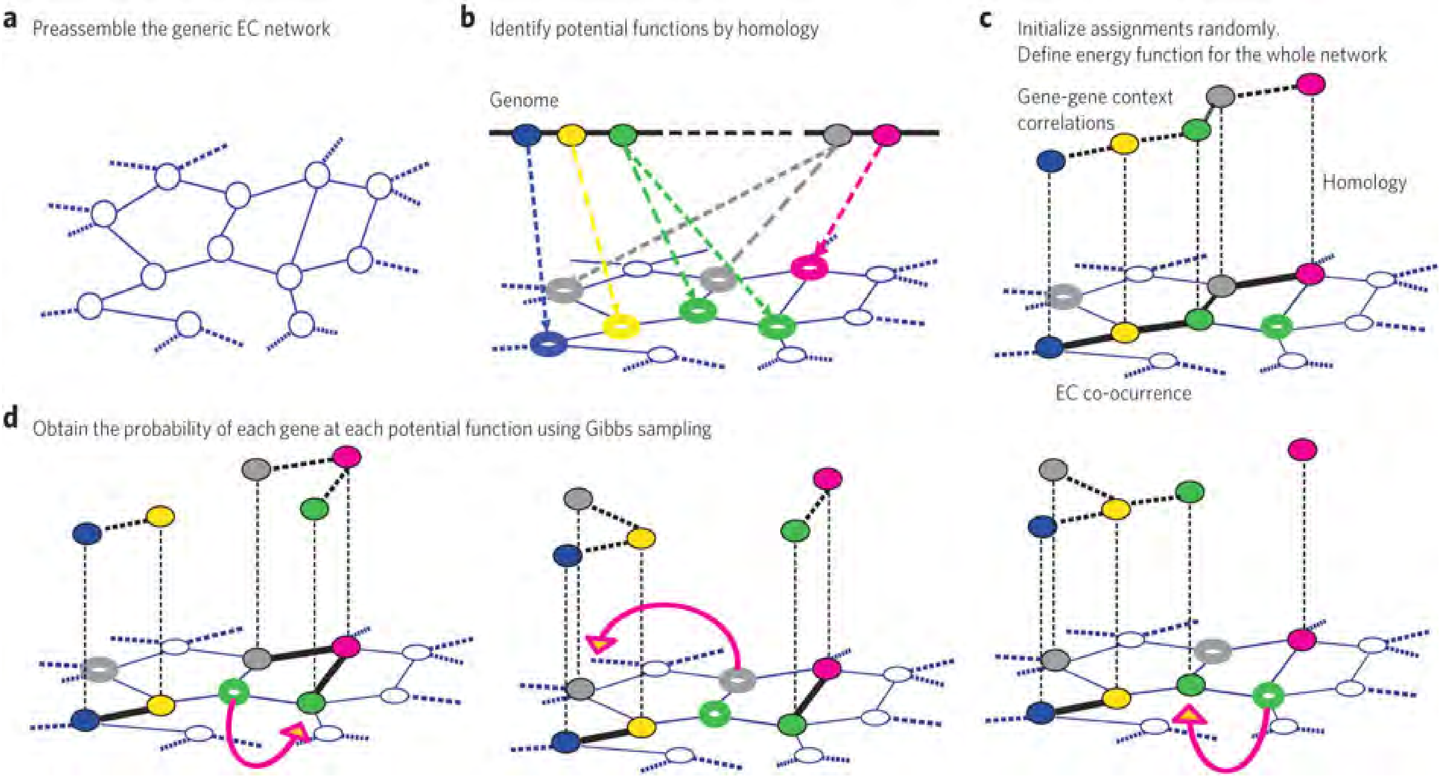

a
Preassemble the generic EC network
b
Identify potential functions by homology
Genome
c
Initialize assignments randomly.
Define energy function for the whole network
Gene-gene context correlations
Homology
EC co-ocurrence
d
Obtain the probability of each gene at each potential function using Gibbs sampling

Figure 1: Overview of the GLOBUS method. (a) A generic ec network, where nodes represent all known biochemical activities and edges indicate metabolites shared between activities. (b) For a genome of interest, the potential network locations of each gene are assigned on the basis of sequence homology to known enzymes. (c) each gene is initially assigned randomly to one of its possible locations. A fitness function is defined such that assignments to locations with high sequence identity and good context correlations with neighboring genes correspond to higher values of the fitness function (higher probability). (d) Gibbs sampling is used to sample all possible assignments of genes to their candidate network locations. At each step of a Gibbs chain, a random gene is selected and reassigned to one of its possible locations (arrows). The marginal probabilities for assigning every gene to each candidate network location are derived from converged Gibbs chains.

about the function of the genes involved in metabolism. We also believe that GLOBUS could be particularly useful for understanding metabolic networks in less-studied organisms as their genomes are sequenced.

PAN-CANCER GENE SIGNATURES ARE PROGNOSTIC IN BREAST CANCER

DIMITRIS ANASTASSIOU LAB

Using a novel data mining methodology designed to point to the genes at the "heart" (core) of co-expression signatures, we found a set of such signatures, which we call attractor metagenes, present in nearly identical form in multiple cancer types. Three of these pan-cancer signatures are related to mitotic chromosomal instability, mesenchymal transition, and a lymphocyte-specific immune response. We hypothesized that these signatures represent important attributes of cancer in general and therefore that they would prove to be prognostic features for breast cancer in particular. Using these signatures as features, we developed a computational model that won the Sage Bionetworks/ DREAM Breast Cancer Prognosis Challenge. We hope that these results will improve biomarker products applicable to multiple cancers and spur the development of treatments that block the molecular mechanisms responsible for making cancers aggressive or invasive.

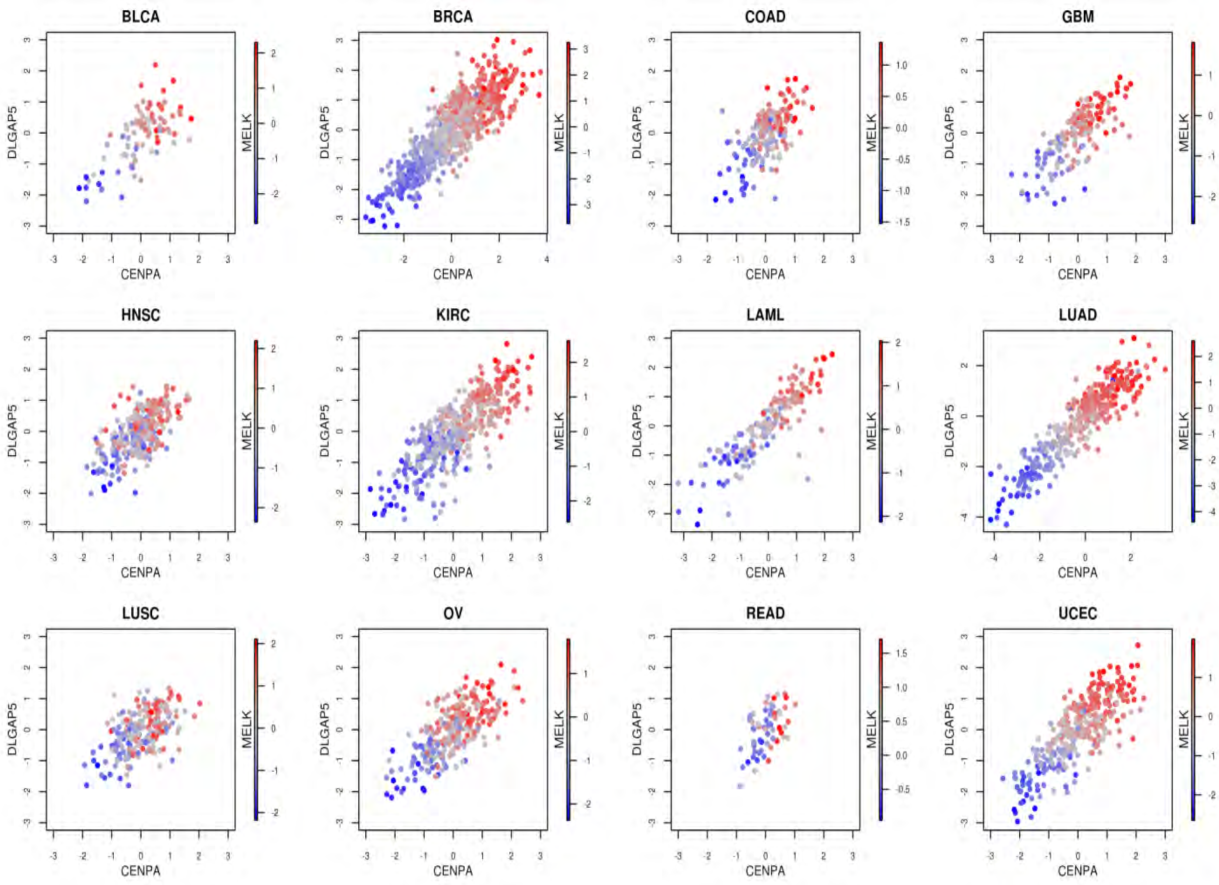


Figure 2: Scatter plots from twelve different cancer types (using the TCGA "pancan12" datasets) indicate that the three top-ranked genes, CENPA, DLGAP5 and MELK, of the "mitotic chromosomal instability attractor metagenes" are indeed co-expressed in all twelve types. In each scatter plot, patients represented by the blue dots at the lower left corner have tumors with low levels of the signature, while those represented by the red dots at the upper right corner have tumors with high levels of the signature.

REFERENCES

1. W.Y. Cheng, T.H. Ou Yang and D. Anastassiou, "Biomolecular events in cancer revealed by attractor metagenes," PLoS Computational Biology, Vol. 9, Issue 2, February 2013.

2. W.Y. Cheng, T.H. Ou Yang and D. Anastassiou, "Development of a prognostic model for breast cancer survival in

an open challenge environment," Science Translational Medicine, Vol. 5, Issue 181, p. 181ra50, April 2013. The Sage Bionetworks/DREAM Breast Cancer Prognosis Challenge is the cover story of the journal.

CERNA REGULATORY INTERACTIONS ARE HIGHLY CONSERVED AND MODULATE DRIVERS OF TUMORIGENESIS

ANDREA CALIFANO LAB

Recent research has discovered that mRNA and non-coding RNAs compete for binding and regulation by microRNAs through a mechanism called competing endogenous RNA (ceRNA). Although initial reports have begun to connect this novel layer of gene regulation with development and disease, these findings have not been detailed quantitatively, and it has not been clear whether ceDNA dysregulation drives tumorigenesis, or whether observations to this effect are the result of experimental artifacts.

Our recent work provides a mechanistic link between dysregulated oncogenes and tumor suppressors and specific loci that drive these events through ceRNA networks. Using a kinetic model that considered the full complement of molecular components, we demonstrated that ceRNA interaction networks, in contrast with transcriptional networks, are highly conserved across cell lineages. We dissected ceRNA networks with more than 400,000 interactions and identified a sub-network containing more than 160,000 interactions that is hyper-conserved across four tumor types. In a collaboration with Todd Golub at the Broad Institute, we validated this network using shRNA-mediated silencing assays targeting predicted ceRNAs in the Library of Integrated Network-based Cellular Signatures (LINCS) database, reporting on 1,171 genes using a multiplexed Luminex approach. These data constitute the most extensive experimental validation of a regulatory network ever published, and will be made publicly available as a resource that will benefit researchers who study cancer and other diseases.

Our studies found that ceRNA interactions can account for missing genetic variability — that is, genes whose function is aberrantly regulated despite the fact that their loci are genetically and epigenetically normal — in the majority of oncogenes and tumor suppressors across eight cancer types. Our investigation also indicates that ceRNA interactions mediated by a single miRNA are rare, but become stronger and independent of individual miRNA variability when a ceRNA network involves more components. Interactions involving multiple components within ceRNA networks occur frequently, are highly conserved, and are often relevant to pathogenesis.

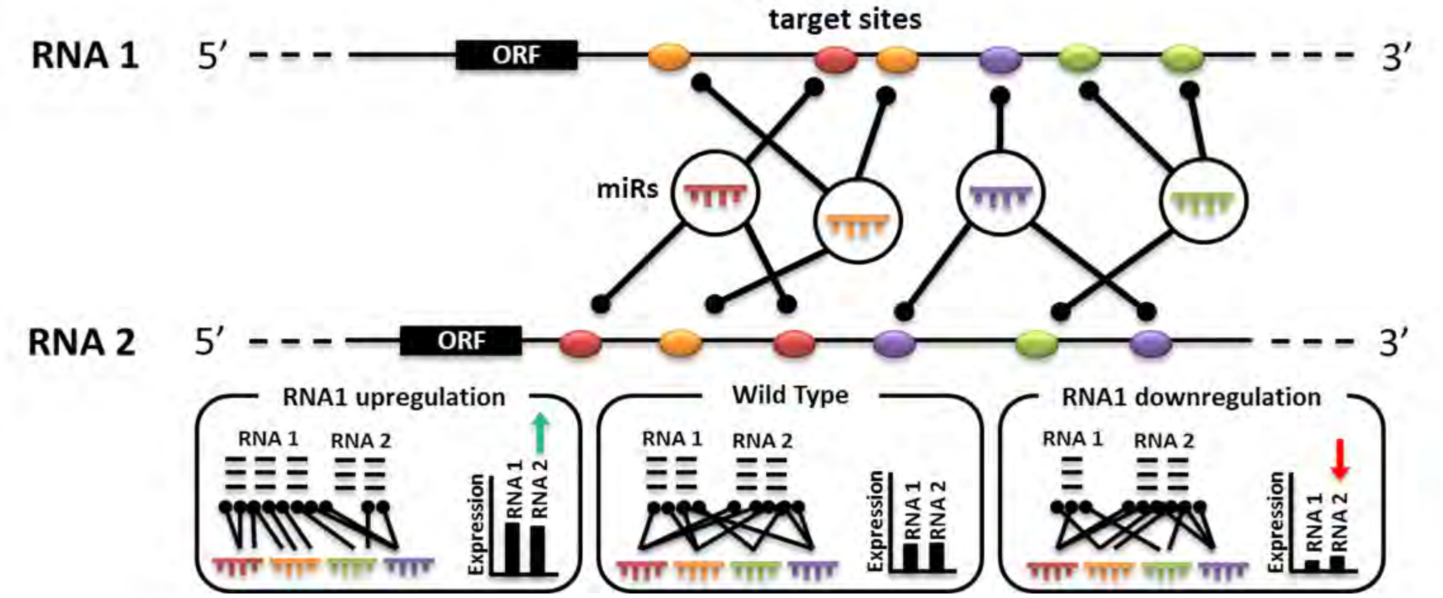


Figure 3: Model for ceRNA regulation. Genes whose transcripts compete for common miRNA regulators up and down regulate one another through a miRNA titration mechanism (ceRNA). Up regulation of RNA 1 sequesters common miRNA regulators, leading to weaker miRNA-mediated repression of RNA 2 transcripts. Analogously, down regulation of RNA 1 leads to an increase in the abundance of common miRNAs that target RNA 2 transcripts, and consequently increases RNA 2 repression.

USING WEB SEARCH LOGS TO DETECT ADVERSE DRUG INTERACTIONS

NICHOLAS TATONETTI LAB

Although the US Food and Drug Organization and other agencies collect and analyze reports on adverse drug effects, alerts for single drugs and drug-drug interactions are often delayed due to the time it takes to accumulate evidence. We hypothesized that Internet users can provide early clues of adverse drug events as they seek information on the web concerning symptoms they are experiencing.

As a test, we asked whether we could detect evidence of an interaction between the antidepressant paroxetine and the anti-cholesterol drug pravastatin by analyzing web search logs from 2010. In a previous study, a data mining algorithm was used to analyze FDA adverse event reporting records, and retroactively found this combination to be associated with hyperglycemia (high blood sugar) in some patients. In this project, we analyzed the search logs of millions of internet users from a period before the above association was identified to see how often users entered search terms related to hyperglycemia and to one or both medications under investigation. (Participants in this study opted in by voluntarily installing a web browser extension that tracked their activity anonymously.)

Using disproportionality analysis, we looked for searches of drug pairs that occurred at higher than expected frequencies by participants who also searched for hyperglycemia. The study found that people who searched for both paroxetine and pravastatin over the 12-month period were more likely to perform searches on terms related to hyperglycemia than those who searched on only one of the drugs. Because the interaction between pravastatin and paroxetine was not known during the period from which the search records were gathered, our analysis was similar to a prediction task, and thus demonstrates the potential for using web logs to identify adverse drug-drug interactions more quickly.

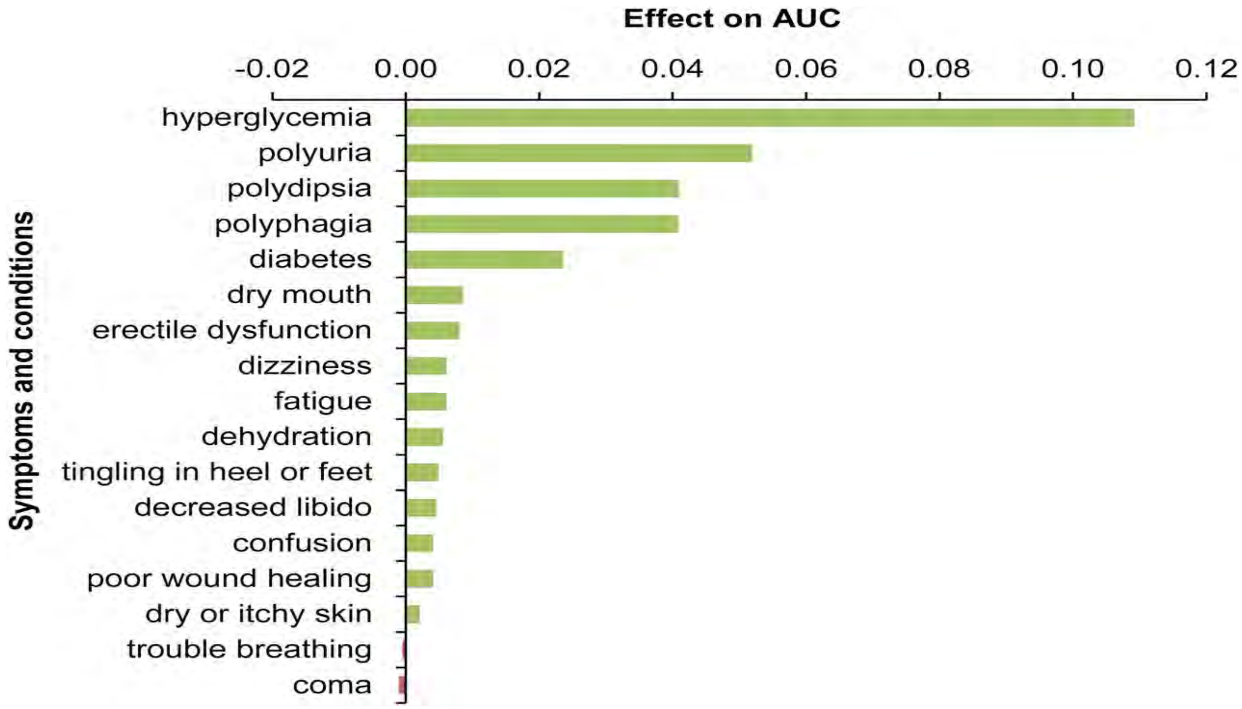


Figure 4: Hyperglycemia (and its synonyms such as 'high blood sugar') has the largest correlation with searches for paroxetine and pravastatin, followed by the three core hyperglycemic symptoms polyuria, polydipsia, and polyphagia.

THE GENOME SEQUENCE OF THE ASHKENAZI JEWISH POPULATION

ITSIK PE'ER LAB

The full DNA sequence of an individual patient can now be sequenced affordably and rapidly. Yet, interpreting such 3-billion-bases-long sequences relies heavily on previously obtained catalogs of deleterious vs. neutral genetic variants present in the population. We focused on the Ashkenazi Jewish (AJ) population, whose genetic history makes it feasible to represent their founding ancestors with a small number of contemporary samples. Following Jewish migration to Eastern Europe, AJ appeared as a relatively genetically isolated population. Ensuing AJ-specific mutations added to our understanding of many genetic diseases. We previously developed methodology to detect long genomic segments that are identical by descent (IBD) from a recent common ancestor to pairs of purported unrelateds, which is an evidence for their remote family ties. We modeled the distributions of lengths of such segments as a function of the population's demographic history. An excess of long IBD segments in AJ made it possible to infer a model for this population. The inferred history includes a narrow (N=300) population bottleneck late-medievals, followed by surprisingly rapid expansion. This means a small number of genomes can be sequenced and be informative with respect to millions alive today.

Through a multi-investigator effort we sequenced a set of elderly, healthy AJ controls (n=128). High quality sequence (depth >62X; total > 25 Terabasepairs) was produced by Complete Genomics. Each newly sequenced genome included 3.4M non-reference variants. Filtering such variants against pre-exiting databases, obtained by sequencing in the general human population, still leaves 130,000 variants of unknown significance per sample, hindering any possibility of genetic diagnosis. Using our catalog, we reduced this to 30,000 variants, and <200 likely-functional changes – a manageable number. Compared to Europeans (Flemish samples), allele frequencies are distinct to indicate distinct history, but with similarities that indicate ancient gene flow (Figure 5, left/middle). Optimizing a population model for both populations shows AJ founders to be an admixed group (Figure 5, right) and reveals Europeans to be descendants of the Neolithic expansion, rather than earlier human migration to Europe.

Ashkenazi sequencing thus confirms the strong and recent founding bottleneck of this population, and sheds new light on its ancestors: an admixed group with roots in Europe as well as the Levant. By sequencing only a manageable number of individuals we have enabled analysis of any personal genome of a patient from the largest isolated population in the US. These insights and lessons will no doubt facilitate extension of our methods to larger and more diverse populations.

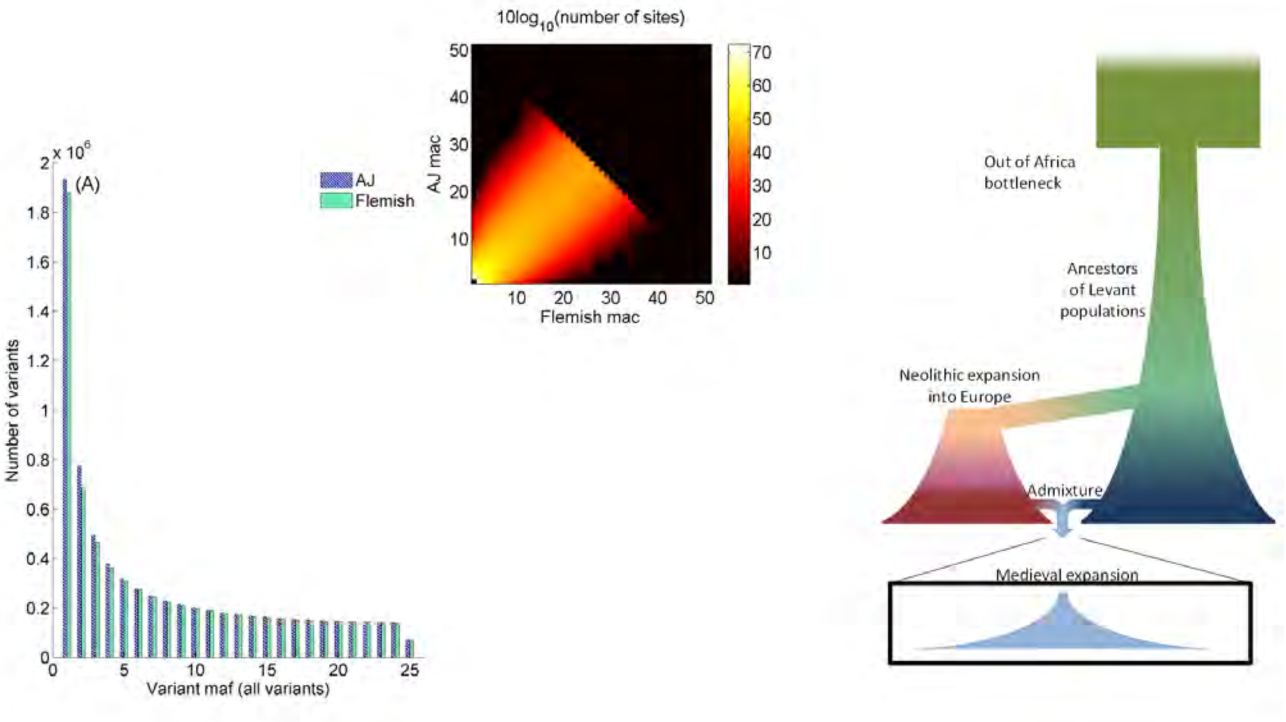


Figure 5: Left, middle: AJ/EU joint spectra of allele frequency counts; Right: Inferred 2-population model shows AJs to be recent admixture of Europeans and Levantine ancestors.

A LARGE-SCALE STUDY OF TUMOR-INDUCED GENE EXPRESSION CHANGES IN HUMAN METABOLISM

DENNIS VITKUP LAB

Using a comprehensive collection of more than 2,500 microarray measurements from 22 cancer types, we systematically analyzed all tumor-induced changes in mRNA expression across metabolic genes, comparing global expression patterns between tumors and normal tissues. These studies indicated that the similarity between metabolic expression patterns of tumors and those of their corresponding tissues is high. However, cancer-induced changes in metabolic gene expression diverged greatly among different cancer types, suggesting that there is no uniform metabolic transformation that is characteristic of all tumors. Moreover, many of these metabolic changes in individual tumor types are not random, but are highly conserved when comparing independent samples of the same tumor type.

Our study generated a wide range of insights into specific expression changes associated with cancer cells. For example, we found that expression of the enzyme isocitrate dehydrogenase increases in glioblastoma and acute myeloid leukemia, creating an efficient mechanism for overproduction of 2-hydroxyglutarate, which is known to promote tumor growth. Interestingly, our analysis also identified hundreds of isoenzymes (enzymes with different amino acid sequences that catalyze the same biochemical process) that showed significant, tumor-specific changes in expression. Many of these isoenzymes are functionally important to the growth of tumors, and in some cases mimic or possibly enhance the effects of recurrent tumor-promoting genetic mutations.

In addition to providing insights into gene expression within metabolic networks, these discoveries hold potential for identifying novel drug targets against cancer. In particular, being able to target isoenzymes and interfere with cancer cell metabolism could offer new strategies for starving tumors of their energy supply and their ability to synthesize compounds that are critical for their survival.

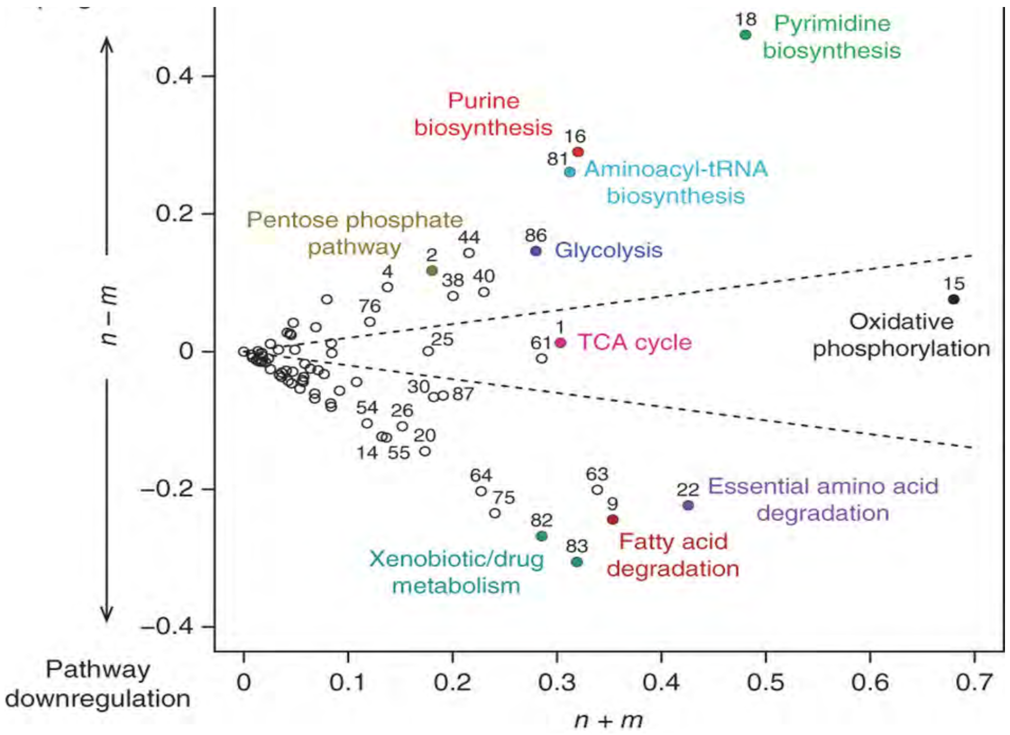


Figure 6: The biochemical pathways defined in the KEGG database are shown in the coordinates of *n+m* and *n-m*, where *n* is the average fraction of tumor samples in which a pathway is significantly up-regulated, and *m* is the average fraction in which a pathway is significantly down-regulated. The averages *n* and *m* were calculated across 22 tumors. The significance of up- and down-regulation was determined using Wilcoxon signed-rank test. The dashed lines demarcate the region where *(n-m) < 0.2*(n+m)* and are shown for visualization purposes only. Metabolic pathways without significant expression changes are primarily clustered on the left of the figure. Pathways that are often significantly up-regulated (high *n* values) occupy positions in the upper right corner, whereas pathways that are primarily down-regulated (high *m* values) occupy positions in the lower right corner. Highly heterogeneous pathways that show, in different tumors, both significant up- and down-regulation are clustered on the right near zero on the vertical axis.

VIRTUAL PROTEOMICS

ANDREA CALIFANO LAB

Because of the technical challenges still facing the field of proteomics, measurement of transcription factor (TF) RNA expression in the cell has been widely used as a proxy for their activity. This approach is not adequate, however, because even if the TF protein level associates with its transcription, the mere presence of TFs in the cell does not necessarily mean that they are actively regulating transcription.

We have developed a new approach called virtual proteomics that can generate a more reliable picture of protein activity. Our algorithm, called Virtual Inference of Protein Activity by Regulon Readout (VIPER), posits that the abundance of mRNA transcripts that are most directly regulated by a specific transcription factor constitute a better proxy for that transcription factor's activity than TF mRNA abundance. Using models of regulatory networks generated by the ARACNe (Basso, Margolin et al. 2005) and MARINA (Wang, Saito et al. 2009) algorithms, we can define the suite of activated and repressed targets for a transcription factor (i.e., its regulon). By then measuring which of these regulon targets are activated or repressed in single samples, we can predict which of their upstream regulatory proteins are active. Because of its ability to infer protein activity within single samples, VIPER constitutes an advance on the capabilities of the MARINA algorithm. Our approach transforms a typical gene expression matrix (multiple mRNAs profiled across multiple samples) into a matrix that represents the relative activity of individual proteins.

We have validated VIPER's accuracy in a series of experiments to predict proteins whose activity was abrogated by RNAi silencing. This approach was used to test five TFs and one protein kinase in different cell contexts, including glioma and human B cells. VIPER was also able to infer 1) the activity of BCL6 and other relevant proteins following B cell receptor mediated modulation of BCL6 turnover, and 2) the activity of NF-κB subunits following ligand mediated CD40 receptor activation. Finally, we used VIPER to estimate inactivation of estrogen receptor (ESR1) activity following inhibition by small-molecule antagonists, suggesting a potential role for the algorithm in elucidating drug mechanism of action.

REFERENCES

1. Basso, K., A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera and A. Califano (2005). "Reverse engineering of regulatory networks in human B cells." Nat Genet 37(4): 382-390.

2. Wang, K., M. Saito, B. C. Bisikirska, M. J. Alvarez, W. K. Lim, P. Rajbhandari, Q. Shen, I. Nemenman, K. Basso, A. A. Margolin, U. Klein, R. Dalla-Favera and A. Califano (2009). "Genome-wide identification of post-translational modulators of transcription factor activity in human B cells." Nat Biotechnol 27(9): 829-839.

STATISTICAL INFERENCE IN NANOPORE SEQUENCING

CHRIS WIGGINS LAB

Rapid advances in biological research that leverage next generation sequencing platforms have been possible only with the development of new bioinformatics algorithms that can make sense of raw sequence data. Likewise, new sequencing technologies will require new algorithms to fully utilize the data for downstream analysis. One promising technology is nanopore sequencing, in which single DNA molecules are threaded through a small pore, and nucleotide-specific current signals are recorded. Advantages include extremely long reads (~10,000 bases) and reduced sample preparation. A challenge intrinsic to this approach is that stochastic motion inside the pore allows the DNA molecule to move both backward and forward. As a result of this diffusive motion, the position of a read on the sequence is not known, and must be inferred from the data.

In order to address this problem, we developed a statistical method of combining a set of *N* output reads generated from an input DNA sequence subject to diffusive motion. Translocation is modeled as a one-dimensional biased random walk with forward bias *p* and noise term *e* (Figure

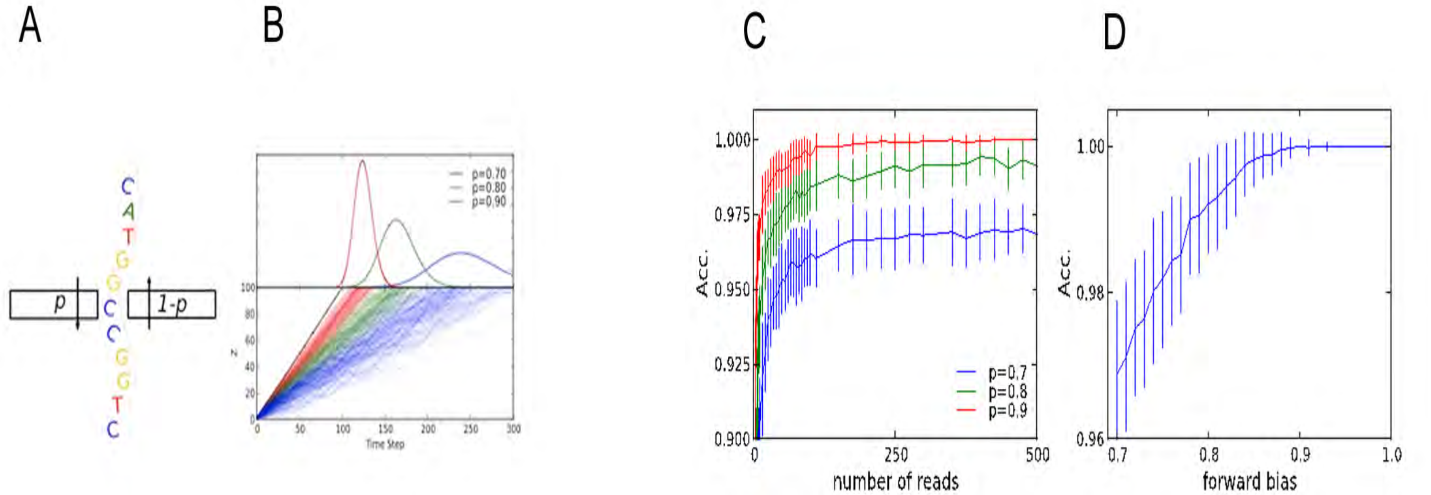


Figure 7. A: cartoon of the nanopore sequencing physical model. B: distribution of read lengths across biased random walks. C: Inference accuracy as a function of coverage. D: asymptotic inference accuracy across forward biases.

7A, 7B). A Hidden Markov Model (HMM) is used to obtain an estimate of the sequence that is most likely to generate the observed data. Each output read is modeled as a set of observed states x, corresponding to the bases read off the current trace, and a discrete set of hidden states z, the unknown position along the true sequence. We use Expectation Maximization (EM) to learn the emission distribution, *S=p(x|z),* which represents our inferred sequence. Accurate inference is possible because the likelihood factorizes over the independent output reads; at each iteration of EM, we aggregate our estimate for *S* across all reads. We find an *N*-asymptotic accuracy exceeding 99% feasible with *p*=0.8, and large *N* compensating entirely for high *e* (Figure 7C, 7D).

Finally, we observed that inference errors take the form of discrete insertion and deletion events. The entropy signature of S allows us to locate these erroneous positions. We implemented a heuristic to systematically remove these positions using restarts embedded in the EM algorithm.

A DNA LANDING PAD FOR PROTEINS

TOM TULLIUS LAB

DNA shape is now recognized as a key determinant of protein-DNA recognition, thanks to the work of MAGNet investigators Barry Honig, Richard Mann, Harmen Bussemaker, and Tom Tullius. In an influential paper published in Nature in 2009, Honig, Mann, and coworkers mined the Protein Data Bank of X-ray crystal structures of DNA-protein complexes and found that many proteins recognize their DNA binding sites by inserting an arginine residue into a narrow DNA minor groove.

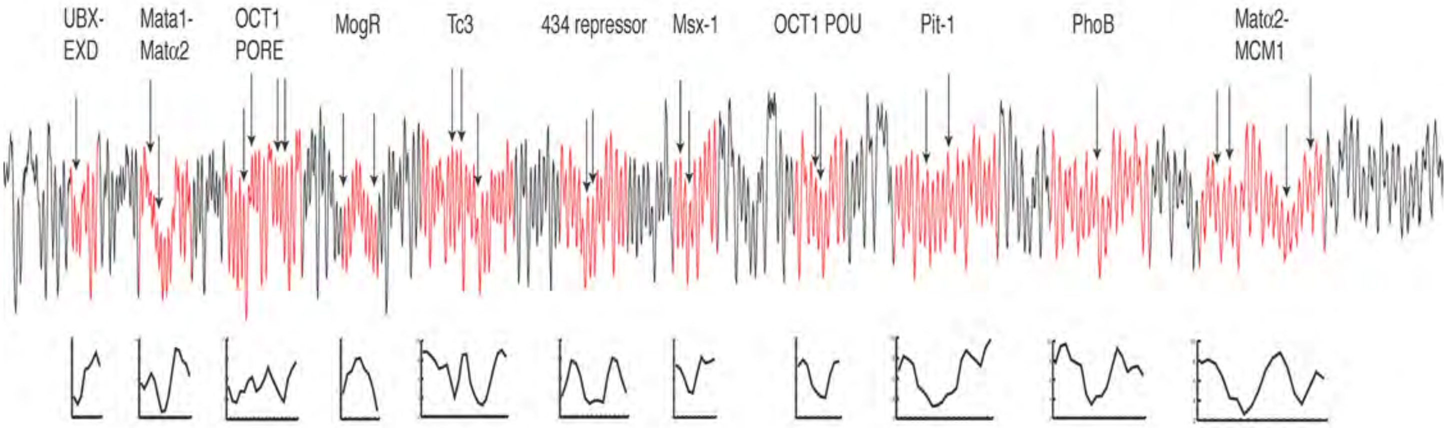


Figure 8:Experimental DNA structure map of eleven protein binding sites. The hydroxyl radical cleavage pattern of a 400 base pair DNA molecule designed to contain eleven protein binding sites is shown in the center of the figure. Protein binding sites are colored red. Minima in cleavage indicate a narrow minor groove. Arrows show where arginine side chains from the protein bind in the DNA minor groove. Below the cleavage pattern are shown plots of minor groove width variation, from X-ray structures of the protein-DNA complexes. For several binding sites, there is a close correspondence in cleavage pattern and X-ray structure-derived minor groove width (Mata1-Mata2, MogR, 434 repressor, Msx-1, for example). Other binding sites do not correspond as well with the DNA-protein X-ray structure (Pit-1 and PhoB, for example), suggesting that protein binding changes the structure of the DNA in those sites.

Because this analysis was based on X-ray structures of DNA-protein complexes, the question arises as to whether the special shape of the DNA binding site exists in naked DNA, or whether it is induced by protein binding. Previous MAGNet-supported work by the Tullius, Honig, and Mann laboratories showed that the hydroxyl radical cleavage experiment yields an accurate map of minor groove width for naked DNA molecules in solution. In work supported by the MAGNet Center, the Tullius laboratory used hydroxyl radical cleavage to experimentally map the shapes of the DNA binding sites of eleven of the proteins studied in the Nature paper. This is the first time that such a large number of binding sites has been directly compared in the same DNA molecule. They found that most of the protein binding sites in naked DNA have a minor groove shape that closely resembles the shape seen in the X-ray structure of the protein-DNA complex. The binding sites that differ in shape from the X-ray structure of the DNA-protein complex may be molded by the protein into a different conformation.

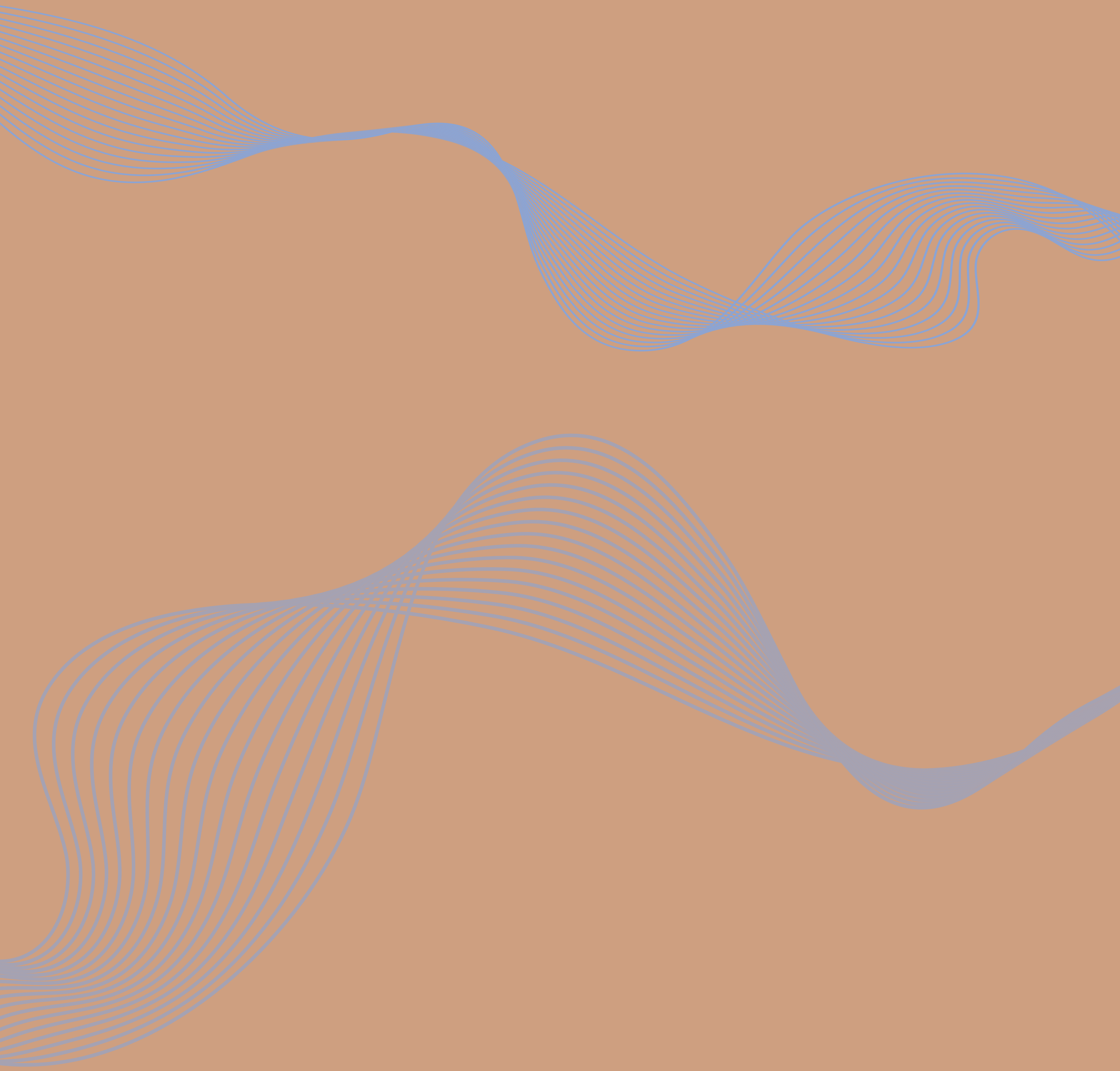
SUPPORT FOR CELL-OF-ORIGIN MODEL FOR PROSTATE CANCER HETEROGENEITY

MICHAEL SHEN, CORY ABATE-SHEN, AND ANDREA CALIFANO LABS

The cell-of-origin model in cancer biology suggests that some tumors are more aggressive than others because of differences in the cell lineages from which they arise. In the prostate gland, there are three types of epithelial stem cell — luminal cells, basal cells, and rare neuroendocrine cells. There has been some discrepancy in the scientific literature, however, about whether luminal cells, basal cells, or both can act as a cell of tumor origin.

We undertook a comprehensive analysis of prostate basal cell properties in mouse models, performing genetic linkage marking to study an identical cell population in multiple assays of stem cell function. Our studies showed that discrepancies in the published literature arise because basal stem cell properties can change when studied outside their endogenous tissue microenvironment; that is, in ex vivo cell culture and tissue grafting assays. To avoid this problem, genetic lineage tracing in vivo should be considered the gold standard for identifying physiologically relevant stem cells. In addition, our molecular and bioinformatics analysis showed that tumors that originate from luminal cells are more aggressive than those that develop from other cell types. Finally, bioinformatics studies identified a cross-species molecular signature within mouse luminal cells that correlates with human patient outcomes.

Identifying cells of tumor origin and determining their specific molecular signatures holds potential for identifying biomarkers of specific prostate cancer subtypes. Further research in this area could provide valuable information that could be used to distinguish high- from low-risk patients and to predict treatment response.



Columbia University
Initiative in Systems Biology
1130 St. Nicholas Avenue
New York, NY, 10032

SPRING 2013
Issue No. 6