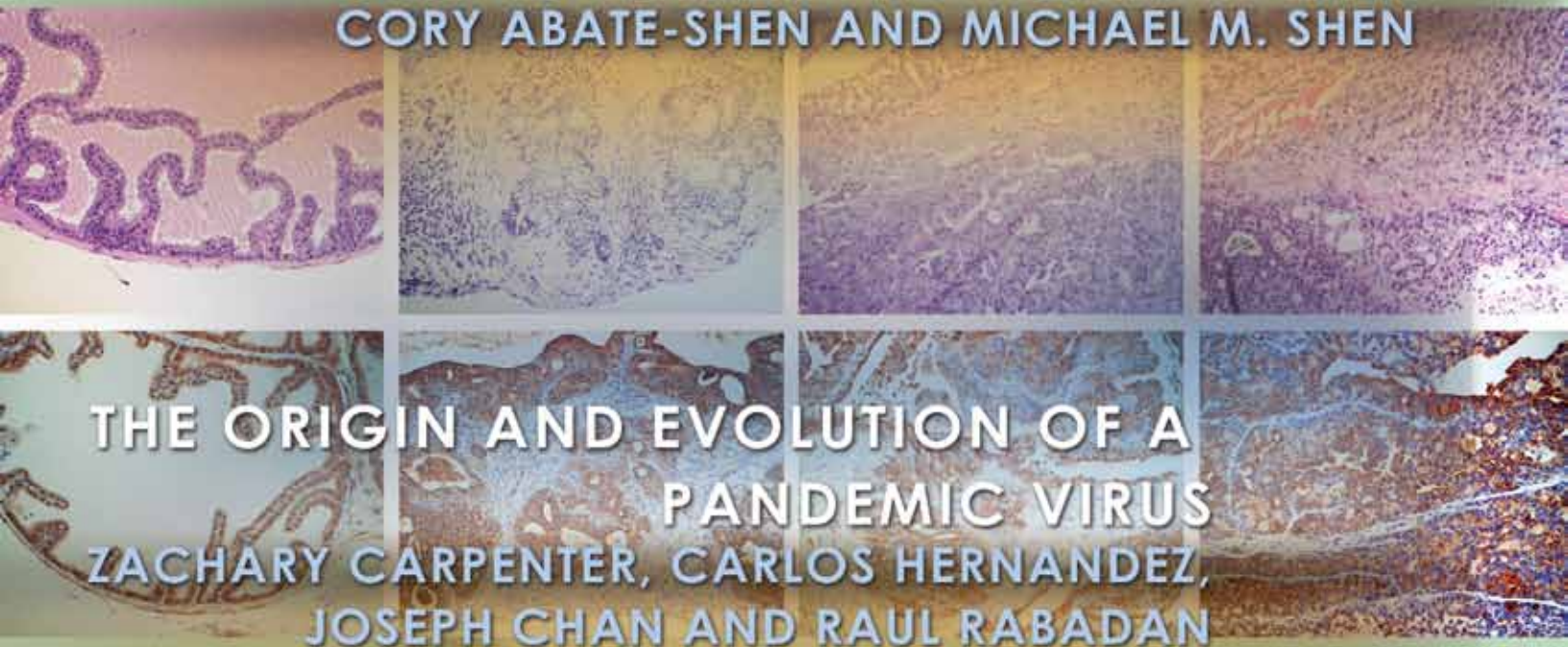
NATIONAL CENTER FOR THE MULTISCALE ANALYSIS OF GENOMIC AND CELLULAR NETWORKS

SPRING 2011 ISSUE NO.4

MAGNet NEWSLETTER

INTERROGATING MOLECULAR PATHWAYS OF PROSTATE TUMORIGENESIS IN MICE AND MEN

CORY ABATE-SHEN AND MICHAEL M. SHEN



THE ORIGIN AND EVOLUTION OF A PANDEMIC VIRUS

ZACHARY CARPENTER, CARLOS HERNANDEZ, JOSEPH CHAN AND RAUL RABADAN

THE TRANSCRIPTIONAL NETWORK FOR MESENCHYMAL TRANSFORMATION OF BRAIN TUMOURS

ANDREA CALIFANO

SPRING 2011
ISSUE NO.4

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

MAGNet NEWSLETTER

FEATURES

03 **FEATURE ARTICLE:**
Interrogating Molecular Pathways of Prostate Tumorigenesis in Mice and Man
CORY ABATE-SHEN, MICHAEL M. SHEN

15 **FEATURE ARTICLE:**
The origin and evolution of a pandemic virus
ZACHARY CARPENTER, CARLOS HERNANDEZ, JOSEPH CHAN, RAUL RABADAN

19 **FEATURE ARTICLE:**
The transcriptional network for mesenchymal transformation of brain tumours
ANDREA CALIFANO

SECTIONS

02 **INTRODUCTION**
ANDREA CALIFANO, BARRY HONIG

23 **FEATURED NEWS**

MICRORNA MODULATORS REGULATE ONCOGENES AND TUMOR SUPPRESSORS IN GLIOBLASTOMA

PREDICTING DISEASE PHENOTYPE FROM GENOTYPE

IDENTIFYING VIRUS HOSTS FROM SEQUENCE DATA

CHASING DRIVERS OF CANCER WITH DATA INTEGRATION

RECONSTRUCTION AND ANALYSIS OF THE PLASMODIUM FALCIPARUM METABOLIC NETWORK

GENSPACE: COMMUNITY-DRIVEN KNOWLEDGE SHARING IN GEWORKBENCH

IMPROVING THE PERFORMANCE OF THE GENSPACE RECOMMENDER SYSTEM

FINDING "aQTLS" – THE GENETIC LOCI THAT MODULATE TRANSCRIPTION FACTOR ACTIVITY

DEMOGRAPHIC INFERENCE FROM HIDDEN RELATEDNESS

CRACKING THE HOX SPECIFICITY CODE

PROBING THE ELECTOSTATIC POTENTIAL OF DNA USING HYDROXYL RADICAL CLEAVAGE

INTRODUCTION

Ready, Set, Go!

The first five years of MAGNet have been characterized by the creation of a community at Columbia University with common interests, common methodologies, and common taste for science. Objectively, based on the number of publications, collaborative grants emerging from this initiative, and traction We have had a very successful beginning. What we started as colleagues we are now continuing as collaborators and friends, enriched by recent recruits that have further enriched our community. So the natural question to ask is what will happen next. Will it be more of the same or will this enterprise be again transformational for our community and hopefully for the field? There are clear signs that the latter is more than a dream and actually quite achievable from the platform that has been assembled by the MAGNet center.

To start, MAGNet has created significant resonance, both within the confines of our institution and across institutions. For instance, thanks to the community that has emerged around MAGNet, Columbia has created a new Initiative in Systems Biology that combines both the Center for Computational Biology and Bioinformatics (C2B2) and the Columbia Genome Center (CGC). Thanks to the generous endowment of the Sulzberger family, the initiative has been able to restructure the genomics cores of the Columbia University Medical School and create the opportunity for a number of scientific recruitments in computational and experimental systems biology, starting with Prof. Saeed Tavazoie, from Princeton University, who will join our faculty in June as a Full Professor. Saeed will be the first of nine new recruits who will participate in the creation of a new Department of Systems Biology in 2012.

Outside of the confines of our institutions, recent results on the role of DNA structure and, in particular, of minor groove width in determining binding specificity of Hox transcription factors and histones has had profound impact on the way we think of the role of DNA in determining protein binding specificity, no longer as a string of nucleic acids but as a true molecular structure. Similarly, in cancer, pioneering MAGNet pioneering work on the interrogation of regulatory networks and molecular signatures to identify key determinants of tumorigenesis, progression, and drug sensitivity, has had significant impact in the study of melanoma, leukemia, lymphoma, and glioblastoma. MAGNet investigators have organized and co-organized a number of cancer systems biology meetings, including the recent AACR Cancer Systems Biology, and the annual meeting of the AACR, where systems biology was prominently featured. Finally, the RECOMB Systems BIology, Regulatory Genomics, and the DREAM challenges are truly helping to create an invisible bond between experimentalists and computational biologists in these fields by establishing rigorous blind tests, based on experimental evidence, to determine the quality of computational predictions. This year, the meeting will take place in Barcelona on Oct 14-19, in an attempt to further bridge our ties with European and Israeli colleagues.

Finally, as the NCBCs are transitioning from a pure roadmap to an NIH institute-based activity, MAGNet is establishing closer ties with the NCI and is further focusing on the study of cancer systems biology, as also illustrated by its dual membership in the Integrated Cancer Biology Program (ICBP) as one of the 11 Centers for Cancer Systems Biology (CCSB). This is reflected in some of the new research directions and Driving Biological Projects. The feature articles in this issue discuss three projects. The first one introduces some exciting results on mouse models of prostate cancer, starting from the analysis of lesions that produce the first fully penetrant model of metastatic prostate cancer in the mouse and continuing with and attempt, for the first time, to assemble a full interactome (including both transcriptional and post-translational interactions) from in vivo data. Specifically, a set of distinct transgenic mouse models and human xenografts are treated with a panel of 14 drugs + vehicle and then their tumors are analyzed to infer transcriptional and post-translational interactions both in the human and in the mouse tumors, using the ARACNe [1] and MINDy [2] algorithms among others. The resulting interactomes are then interrogated using signatures of tumor initiation, progression to metastasis, and drug sensitivity, to elucidate key master regulators of these phenotypes. The second article discusses the evolutionary origins of the first pandemic virus of the 21st century, the H1N1 flu virus of 2009. By comparing the new virus to the 10,000 different genomes that have been collected and sequenced since 1918, the authors trace back its ancestors to swine viruses isolated in two continents (North America and Eurasia) and attempt to reconstruct its reassortment profile. Finally, the third article illustrates some of the continuing work in High-Grade Glioma, where collaborations between MAGNet and Cancer Center investigators have helped elucidate mechanisms of subtype emergence, as well as associated genetic lesion and therapeutic targets.

In summary, with the second 5-year cycle of MAGNet starting, our community is again coming together to address a number of exciting scientific project as a collaborative group.

Andrea Califano

Barry Honig

REFERENCES

1. Margolin, A.A., et al., ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics, 2006. 7 Suppl 1: p. S7.

2. Wang, K., et al., Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nat Biotechnol, 2009. 27(9): p. 829-39.

INTERROGATING MOLECULAR PATHWAYS OF PROSTATE TUMORIGENESIS IN MICE AND MAN

CORY ABATE-SHEN AND MICHAEL M. SHEN
HERBERT IRVING COMPREHENSIVE CANCER CENTER, COLUMBIA UNIVERSITY

Why study prostate cancer?

Prostate cancer has been recognized as a clinical entity since antiquity, when it was first described by the ancient Egyptians, and surgical procedures to remove the prostate were developed more than 100 years ago [1]. However, in the past three decades the availability of a highly-accessible blood test for Prostate-Specific Antigen (PSA) has revolutionized the diagnosis of prostate cancer. Men with elevated PSA levels typically undergo biopsy to assess the potential presence of prostate cancer. If diagnosed, conventional treatment regimens include surgical excision of the prostate or irradiation. In the case of advanced cancer, these regimens are usually followed by androgen-deprivation therapy, which initially reduces tumor burden to low or undetectable levels but ultimately the disease will recur in most cases.

Recent changes in recommendations that have suggested later and less frequent PSA screening highlight a major clinical challenge for prostate cancer diagnosis and treatment [2]. These new recommendations were proposed because the wide-spread use of PSA testing has led to a vast increase in the diagnosis of patients with clinically-localized low Gleason grade carcinomas that may not require treatment, since their tumors are relatively indolent. Consequently, a major clinical challenge is posed by the current inability to readily distinguish indolent from aggressive tumors in prostate cancer patients who present with low Gleason grade tumors [3]. The absence of this prognostic information has led to a significant "overtreatment" of patients who would otherwise only require conservative management. This prognostic challenge could be addressed by better understanding of the molecular basis of cancer progression, which should lead to identification of biomarkers that distinguish indolent and aggressive forms of prostate cancer.

Furthermore, circulating androgens are essential for normal prostate development as well as the onset of prostate cancer through their interactions with the androgen receptor (AR), a nuclear hormone receptor whose signaling plays a key role in normal prostate development as well as in prostate cancer [4]. As shown by Huggins and colleagues in the 1940s, removal of testicular androgens by surgical or chemical castration will lead to regression of prostate tumors [5]. However, androgen depletion is usually associated with the recurrence of prostate cancer, as monitored by rising PSA levels, and this recurrent disease is termed "castration-resistant" [6]. Unfortunately, castration-resistant prostate cancer has been essentially untreatable, with the most effective standard chemotherapeutic regimens resulting in a mean increase in survival time of 2 months [7, 8]. Therefore, a second major clinical challenge is the elucidation of pathways of castration-resistance that work in conjunction with AR, which could lead to the identification of new therapeutic approaches.

A third major clinical challenge corresponds to the propensity for advanced prostate cancer to metastasize to bone, which is primarily responsible for its effect on patient morbidity as well as mortality [9]. Despite the clinical relevance of bone metastasis, the molecular mechanisms that underlie the bone tropism of prostate cancer are not well understood. This gap in knowledge is due in part to difficulties in obtaining metastatic tissue from patients, as well as difficulties in generating mouse models that display bone metastasis.

What do we know about the pathways that are relevant for prostate tumorigenesis

More than 10 years ago, we reviewed the molecular mechanisms of prostate cancer

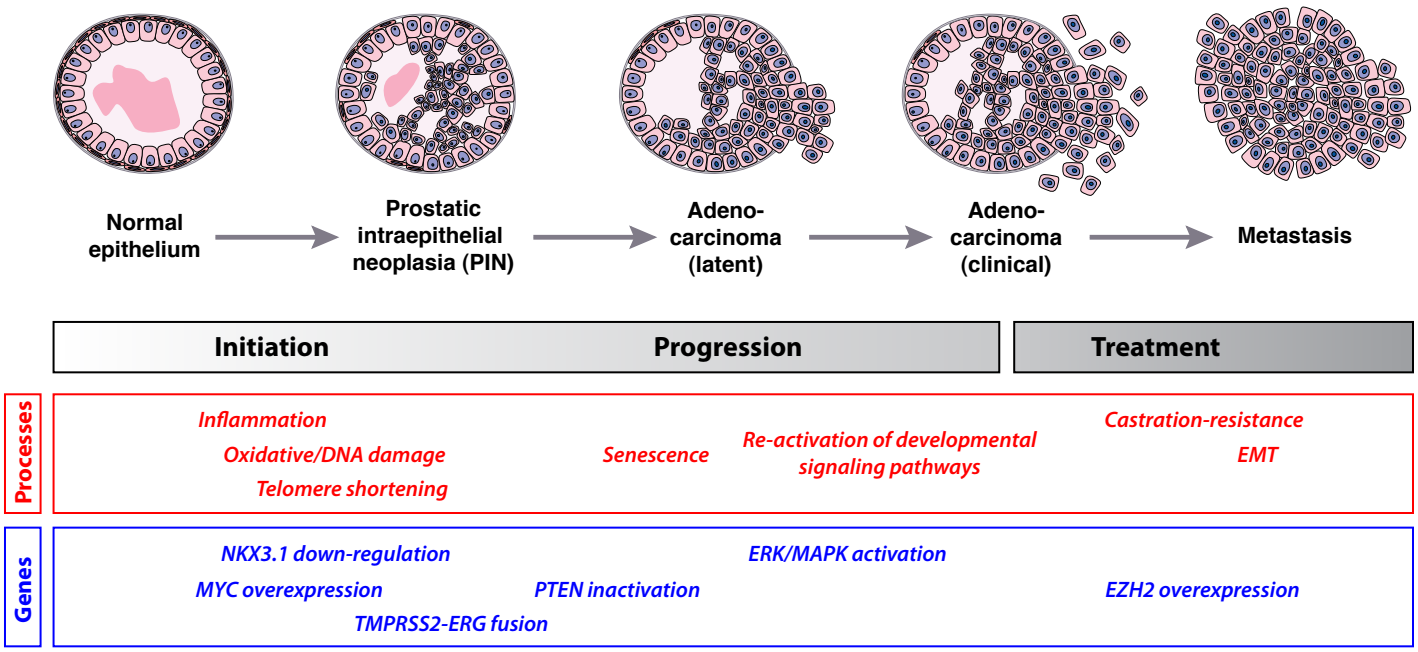


Figure 1. Prostate Cancer Progression. See text; reprinted from [3]

pathogenesis [10], which we recently updated [11] (Fig. 1). Like other epithelial cancers, prostate cancer arises from pre-neoplastic precursors termed prostatic intraepithelial neoplasia (PIN), and progresses to locally-invasive cancer and ultimately metastases. Unlike most other cancers, prostate pathogenesis is critically dependent on androgen signaling and, as discussed above, depletion of androgens ultimately results in castration-resistant disease, which is ultimately fatal.

Studies from our groups led to the identification of NKX3.1 as a prostate-specific homeobox gene located on chromosome 8p21, a hotspot for prostate cancer [12-14]. Analyses of Nkx3.1 function have provided insights into its potential roles in cancer initiation, which have led to a model in which NKX3.1 has been proposed to act as a "gatekeeper" for prostate cancer initiation [15]. Among the tumor suppressor genes that play key roles in prostate cancer, PTEN is particularly significant because of its considerable importance for androgen receptor signaling [13, 16-18]. Thus, loss-of-function of Pten in mouse models collaborates with loss-of-function of Nkx3.1, and is sufficient for castration-resistance [16, 19]. Considerable evidence indicates that Pten loss in prostate cancer results in up-regulation of Akt/mTOR signaling [17, 18, 20], the functional consequences of which are particularly relevant for castration-resistance [16, 21-23].

In addition to Akt/mTOR signaling, Erk (p42/44) MAPK signaling is also frequently activated in prostate cancer, particularly in advanced disease, where it is often coordinately deregulated with Akt/mTOR signaling [24-29]. Thus, simultaneous activation of these signaling pathways promotes tumor progression and castration-resistance [21, 30], while their combinatorial inactivation inhibits castration-resistance prostate cancer [26]. However, in contrast to the Akt/mTOR pathway, which is known to be deregulated in prostate cancer primarily by PTEN loss of function, the upstream events that lead to activation of Erk MAPK signaling have not been elucidated, although they are thought to be linked to aberrant growth factor signaling [31].

Although neither RAS nor RAF are frequently mutated in human prostate cancer [32-34], their pathways are deregulated in a majority of advanced prostate cancers [35]. Interestingly, a small percentage of aggressive prostate tumors contain a translocation of B-RAF or C-RAF that results in its activation [36], suggesting that perturbations of Ras or Raf signaling in prostate cancer may occur through mechanisms other than

activating mutations.

Modeling prostate cancer mice

For nearly a decade, the foundation of our research has been a series of genetically engineered mouse (GEM) models based on germ-line loss-of-function of Nkx3.1 and Pten, which recapitulate the spectrum of prostate cancer phenotypes; these models have provided significant novel insights that are of clinical relevance [16, 37-45]. Nonetheless, these mutant mice have important limitations, not uncommon with other "first-generation" GEM models. Consequently, we have now developed "next generation" GEM models that address these shortcomings, and which form the basis for our ongoing research. Our "next-generation" models are based on a unique Nkx3.1CreERT2 allele, which expresses a fusion protein of Cre with a mutated estrogen receptor (ERT2) under the control of the Nkx3.1 promoter; the resulting CreERT2 fusion protein is completely inactive in vivo, but rapidly activated by administration of tamoxifen [46].

Because we wanted to generate GEM models based on activation of Akt/mTOR and MAP kinase signaling pathways, we used loss of function of Pten to achieve activation of Akt/mTOR signaling and, since the actual means by which MAP kinase signaling is activated in human prostate cancer is not known, we used oncogenic Braf and/or Kras alleles as a surrogate means to activate this

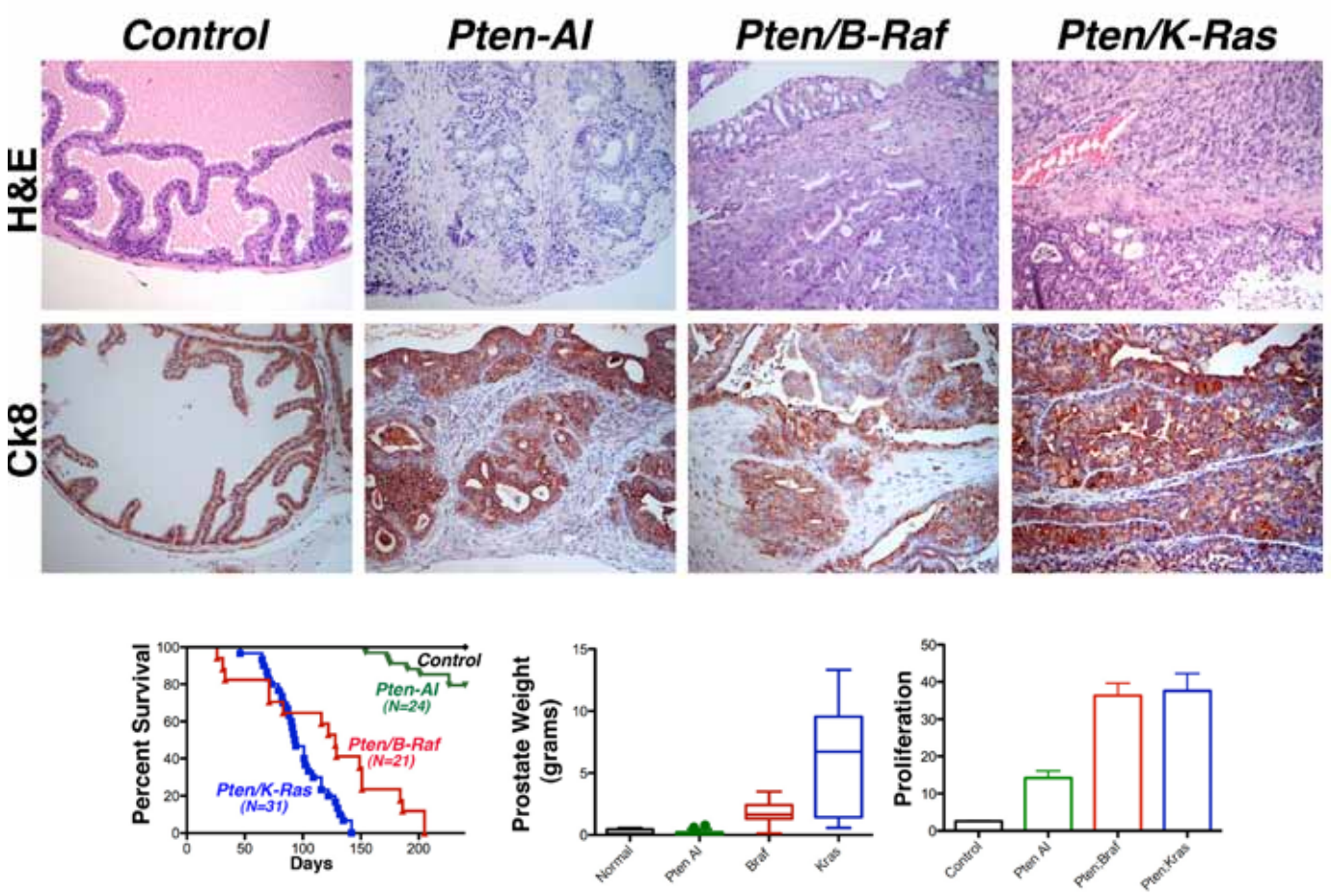


Figure 2. Prostate cancer phenotypes of "next generation" mouse models. **Top:** Mice were injected with tamoxifen or not injected (Control) at 2 months and analyzed at 9 months (Control and Pten-AI) or 4 months (Pten/K-Ras and Pten/B-Raf). Sections of anterior prostate show H&E or immuno staining for cytokeratin 8. **Bottom:** Mice were monitored for survival for the indicated period. Graphs show the number of mice monitored in each group (P<0.0001). Mice were also evaluated for tumor weight and proliferation, the latter by immunostaining with Ki67.

pathway. Therefore, we generated a new series of GEM models using the Nkx3.1CreERT2 driver with a conditional Pten allele [47], a Cre-activatable mutant K-ras allele [48], and a Cre-activatable mutant B-Raf allele [49].

Mice lacking Nkx3.1 and Pten function in the prostate develop high-grade PIN with invasion by 6 months of age, and poorly differentiated adenocarcinoma in mice older than 12 months of age, while androgen deprivation results in the emergence of castration-resistant prostate tumors (Fig. 2). (Hereafter the castration-resistant Nkx3.1CreERT2/+; Ptenflox/flox mice are termed Pten-AI). The Pten-AI mice develop large tumors (~2-7 mm) that display features of highly aggressive prostate cancer, yet there is no adverse effect on their survival, as most of the Pten-AI mice live for up to two years (Fig. 2).

Mice lacking Nkx3.1 and Pten together with activation of either the B-Raf or K-Ras alleles (hereafter termed Pten/B-Raf and Pten/K-Ras, respectively) display highly aggressive prostate tumors that result in ~100% lethality by 6 or 4 months of age, respectively (Fig. 2). Compared to the Pten-AI mice, the prostate tumors in Pten/B-Raf and Pten/K-Ras mice have more poorly differentiated histology, but are primarily luminal, highly proliferative and express androgen receptor (AR) as well as phosphorylated forms of Akt and MAP kinase (Fig. 2).

Although all three models display metastases to lymph nodes, particularly the lumbar node, which is nearest to the prostate, the Pten-AI, Pten/B-Raf and Pten/K-Ras models display differing degrees of metastases to distant organs, including lung and liver. In particular, Pten-AI mice do not (or at least rarely) develop distant metastases, while Pten/B-Raf mice display distant metastases in ~ 20% of cases, and Pten/K-Ras mice with 100% penetrance (Fig. 3). Furthermore, the occurrence of distant metastases was well-correlated with the number of cases having disseminated tumor cells in the bone marrow. This was evident using a PCR-based approach to quantify the targeted allele in captured disseminated tumor cells, as well as by direct visualization of disseminated cells by immunofluorescence staining of bone marrow (Fig. 3). Thus, these GEM mice provide a progressive series of prostate cancer phenotypes with an increasing tendency to metastasize.

Finally, these GEM provides an ideal resource to elucidate molecular pathways that distinguish

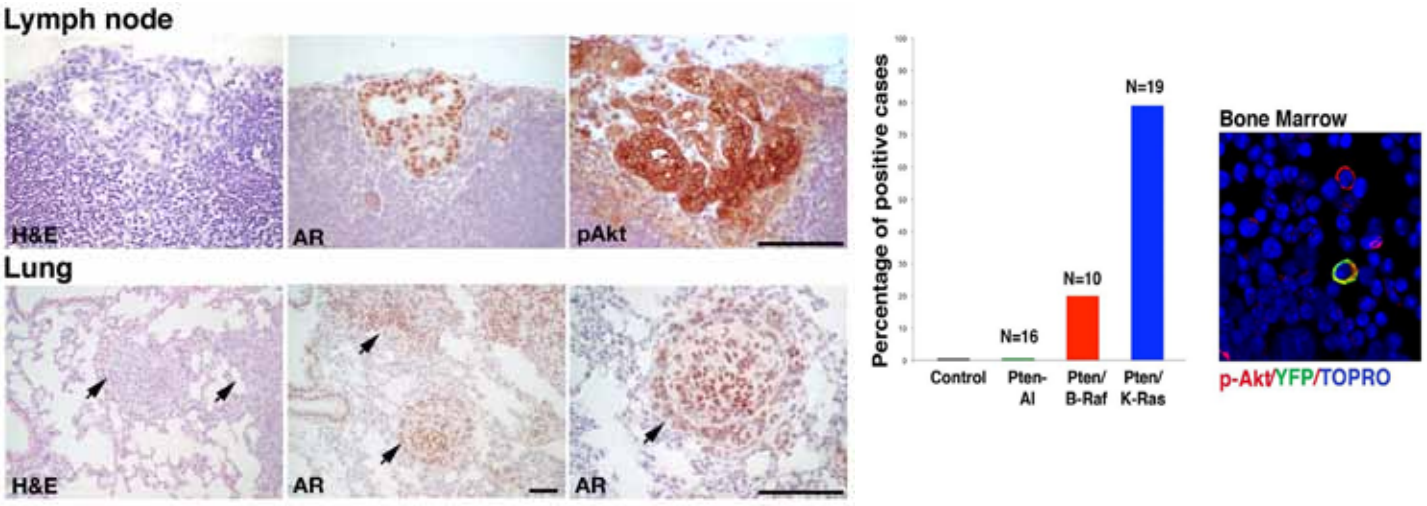


Figure 3. Metastasis in GEM models: **Right:** Disseminated tumor cells in bone marrow: The percentage of positive cases is scored graphically. Image of disseminated cell in bone marrow from Pten/K-Ras mice, expressing p-Akt. **Left:** Metastases phenotype: H&E and IHC images (Akt and AR) of lymph node and lungs from Pten/K-Ras mice.

non-metastatic from metastatic tumors, as well as to identify genes/pathways involved in aggressive prostate cancer. Thus, expression profiling to compare the two most extreme cases — namely the Pten-AI mice (i.e., tumors with no distant metastases) versus the Pten/K-Ras mice (i.e., tumors with frequent distant metastases) — has identified 198 genes whose expression was significantly (P value <1X10-3) different between these two models (Fig. 4), including many genes conserved with human prostate cancer. Furthermore, use of gene set enrichment analyses (GSEA) identified biologically relevant signaling pathways as well as allowed evaluation of the conservation of these gene expression changes in the mouse models for human prostate cancer (Fig. 4). The identification of known genes conserved with human prostate cancer is an important starting point as it provides confidence that the molecular pathways of cancer progression in these GEM models are at least partially conserved in humans and, therefore, that the identification of novel genes will be likely to include those relevant for human prostate cancer.

Curing cancer in mice

We have been pursuing preclinical studies in our "next generation" mouse models, which allow us to address clinically-relevant issues. In particular, we have performed preclinical studies on the

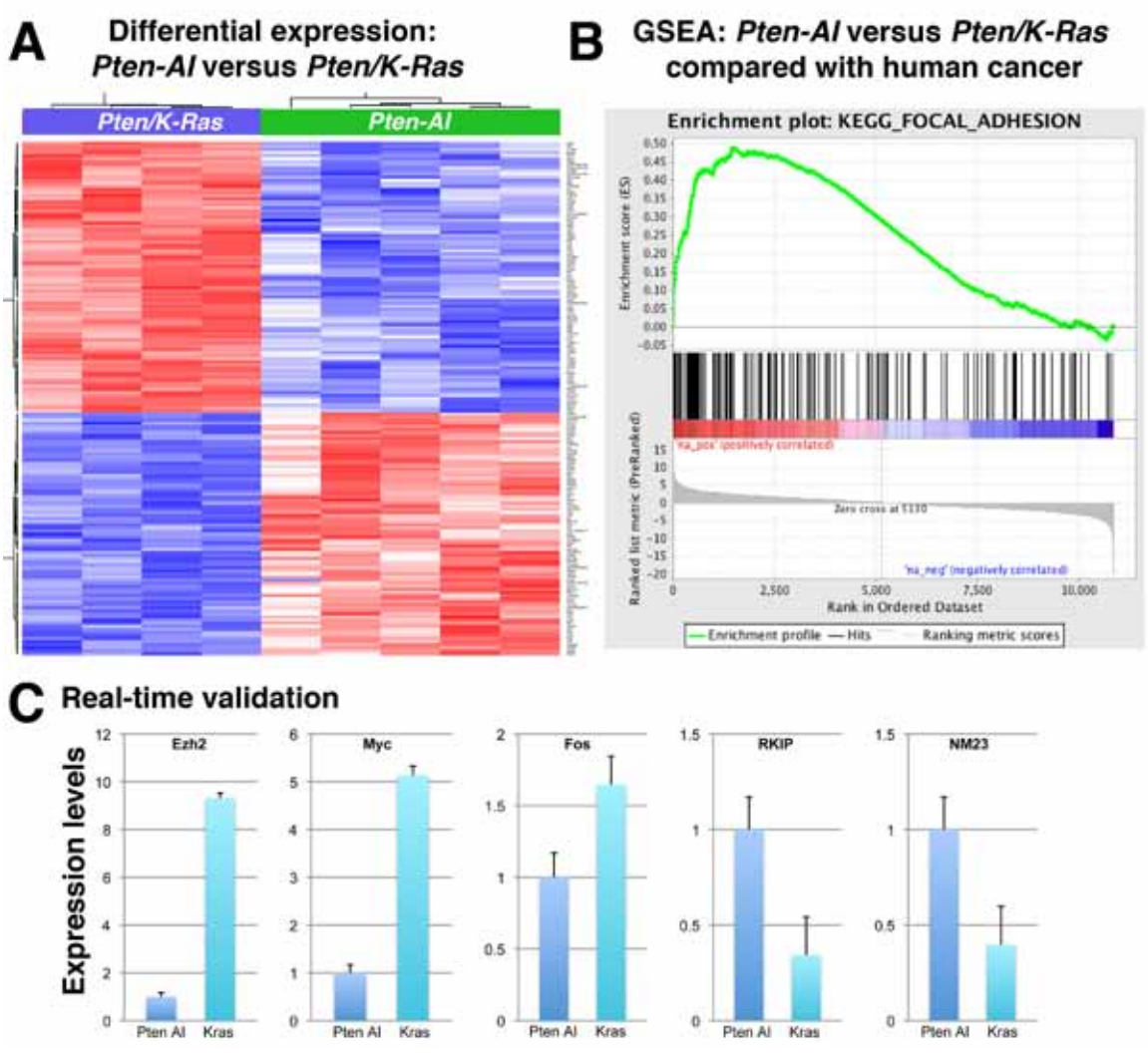


Figure 4: Analyses of differentially expressed genes. **A**. Heat map showing differentially expressed genes in Pten/K-Ras versus Pten-AI mice. **B**. Gene Set Enrichment Analyses comparing differentially expressed genes from Pten/K-Ras versus Pten-AI with the human the C2 MSigDB database(C2 Molecular Signature DataBase, Broad Institute). **C**. Real-time PCR validation of selected genes. **Methods**: RNA from macrodissected prostate tumors was used to interrogate Illumina expression arrays. Gene expression profiles were normalized with the lumi package from R bioconductor. Data were transformed with variance stabilization method and normalized by robust spline normalization. The human to mouse homology mapping was obtained from the Mouse Genome Informatics website (MGI http://www.informatics.jax.org/

Pten; Kras model to compare the consequence of combinatorial targeting of Akt/mTOR and Erk MAP kinase signaling with standard of care chemotherapy (Fig. 5). For these studies, we have used rapamycin to target the Akt/mTOR signaling and a Pfizer MEK inhibitor (PD035901) to target MEK/MAP kinase signaling and Docetaxel as chemotherapy agent (Fig. 5A). Treatment with Rapamycin and PD035901 (Rap + PD) resulted in suppression of tumor growth as evident both by inspection of histology and measurement of prostate tumor weight (p=0.0067) (Fig. 5B). Moreover, the Rap + PD combination also led to a significant (p=0.0006) improvement in survival as well as sustained effects on tumor burden (Fig. 5C), as well as a significant benefit for metastatic burden as evidenced by the reduction of disseminated tumor cells and distant metastases (Fig. 5D). In contrast, treatment of these Pten; Kras mice with chemotherapy (Docetaxel) had a modest effect on suppression of the tumor phenotype (Fig. 5B), similar to its limited efficacy in human prostate cancer [50]. These findings emphasize the benefit of combinatorial targeting key signaling pathways in advanced prostate cancer, and underscore the value of these new mouse models as a preclinical resource for evaluating clinically-relevant end-points.

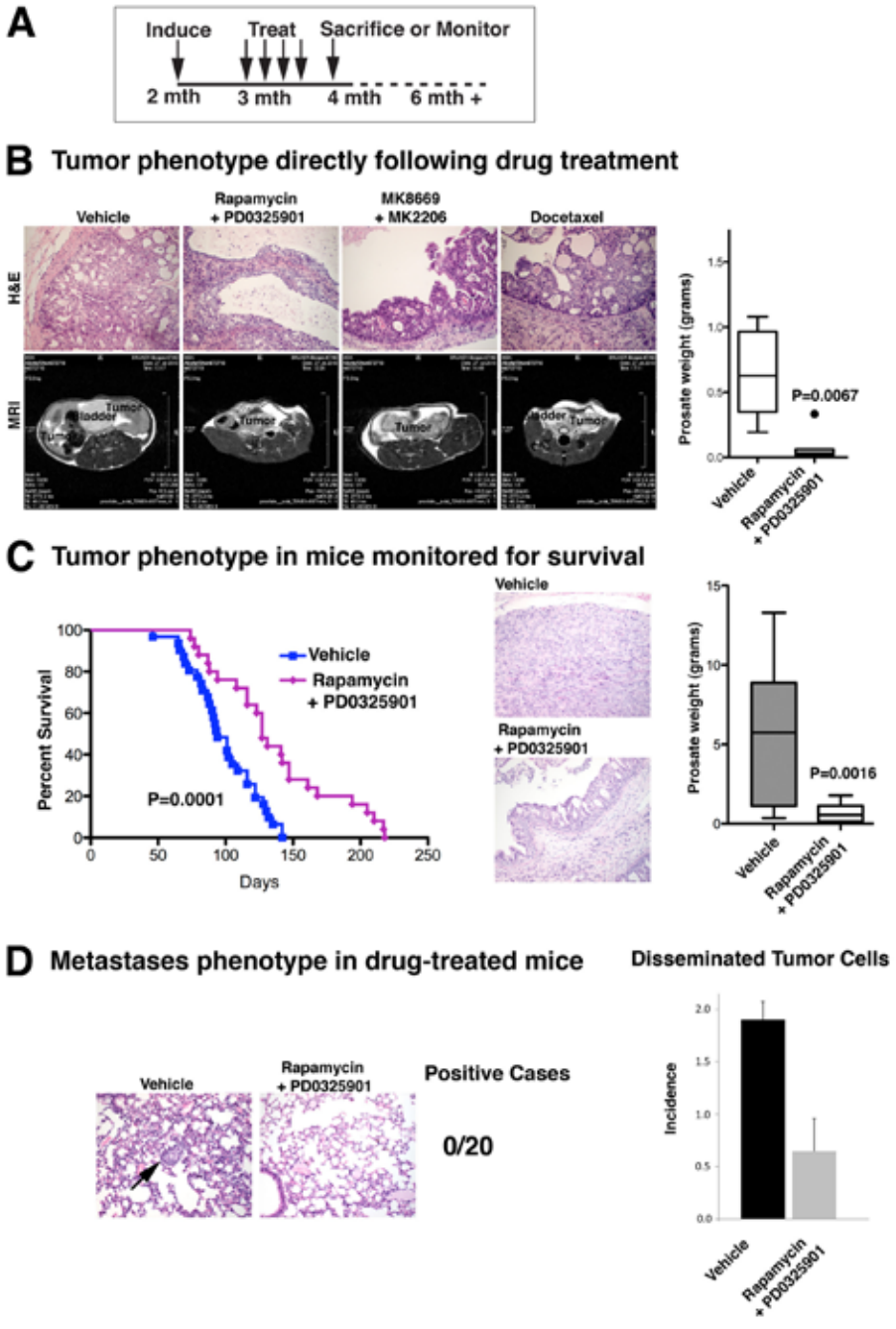


Figure 5. Preclinical analyses.
A. Strategy: Tumors were induced in Pten/K-Ras mice by transient delivery of tamoxifen (at 2 months). Mice were then treated with vehicle or the indicated agents starting at 3 months for a period of 1 month followed by sacrifice (at 4 months) or monitoring for survival (up to ~7 months).
B. Prostate phenotype following treatment. Representative sections showing histology (H&E), MRI images, and prostate weights.
C. Prostate phenotype in the survival cohort. Mice treated with RAP + PD show improved survival relative to mice treated with Vehicle. Prostate histology and weights show persistent tumor remission.
D. Metastatic phenotype: Following RAP + PD treatment, we observed no (0/20) cases of lung metastases compared to the Vehicle-treated mice. The incidence of disseminated tumor cells was also reduced following treatment.
Methods: Dosage and treatment schedule was as follows: Rapamycin (10 mg/kg) delivered IP in combination with PD0325901 (10 mg/kg) delivered PO 3 days a week. MK-8669 (10 mg/kg) delivered by IP injection in combination with MK-2206 (120 mg/kg) delivered PO 3 days a week. Docetaxel (10 mg/kg) was delivered by IP injection twice a week. During the period of the experiment, mice were imaged to evaluate tumor volume using MRI imaging.

Molecular determinants of therapeutic response from mouse to man

We have undertaken an ambitious project to assemble a network of interactions — called interactomes — for both mouse and human prostate cancer to enable the identification of "driver" genes for the cancer phenotype and the therapeutic response. Thus, Andrea Califano's group has pioneered the generation of algorithms for reverse engineering of transcriptional and post-translational networks in normal and neoplastic cells, which produced the first cell-context specific, genome-wide maps of transcriptional and post-translational interactions in human cells (i.e., termed interactomes) [51-53]. This approach is distinct from discovery methods that identify differentially expressed genes, since it identifies "drivers" (i.e., key regulators) that causally affect tumor progression rather than "passenger" genes whose expression is correlated with a particular phenotype.

The production of accurate interactomes is predicated on the availability of large Gene Expression Profiling (GEP) datasets (~300), representative of natural phenotypic variability, which can be achieved by genetic diversity and/or pharmacological perturbations. We have been generating both human and mouse prostate cancer interactomes to facilitate cross-species comparative analyses. The human interactome has been assembled from published data sets available in the public domain (e.g. [35]).

However, the generation of the mouse interactome has required an ambitious undertaking. In particular, to generate both genetic and pharmacologic diversity, we obtained 14 different GEM models that cumulatively represent the spectrum of prostate cancer phenotypes and encompass a range of differing cancer initiating/progression events (Fig. 6).

For pharmacological diversity, each models was treated with various pharmacological agents (Fig. 6). Using this approach, we obtained mouse prostate tissues having 196 independent perturbations (i.e., 14 GEM models each treated with different 14 agents), not including experimental duplicates.

Once the gene expression arrays are collected (i.e., from the published for the human and experimentally for the mouse), the interactomes are then constructed using the ARACNe algorithm developed by Andrea [54], which uses Mutual Information analysis and the Data Processing Inequality, a theoretical property of the mutual information, to infer direct physical interactions

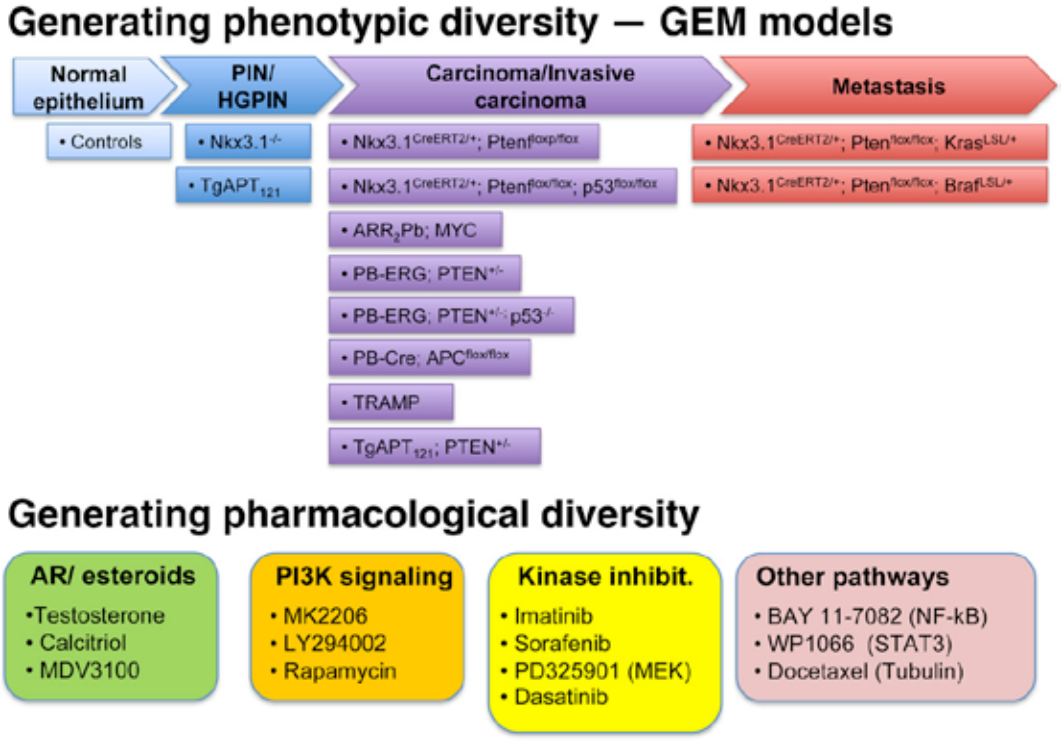


Figure 6: Generating a mouse prostate cancer interactome. GEM models representing the spectrum of prostate cancer phenotypes were treated with each agent listed to generate a large GEP dataset comprised of 196 (14 X 14) independent perturbations.

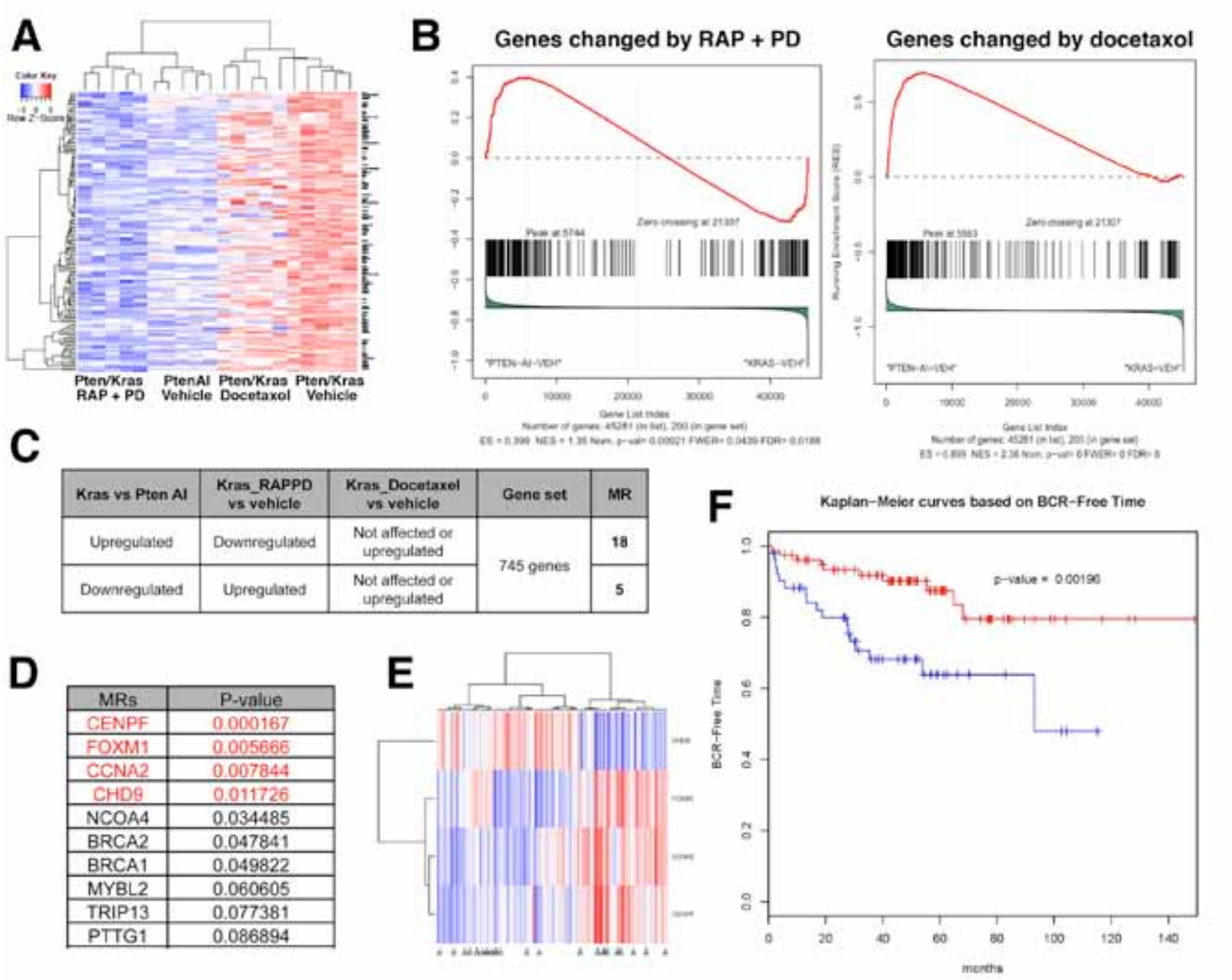


Figure 7: Molecular determinants of therapeutic response. **A.** Heat map showing differentially expressed genes in Pten-AI and Pten/K-Ras mice (given vehicle), compared with Pten/K-Ras mice treated with RAP + PD or Docetaxel for 5 days. **B.** Gene Set Enrichment Analyses comparing the gene expression changes following treatment with RAP + PD or Docetaxel compared with genes differentially expressed between Pten/AI and Pten/Kras. **C.** Strategy for identifying master regulators (MR) of the therapeutic response. **D.** Prioritizing regulators based on disease outcome (biochemical recurrence). **E.** Heat map showing data stratification. **F.** Kaplan-Meier based on biochemical recurrence.

between transcription factors (TFs) and their target genes, while eliminating indirect interactions. The next step uses the MINDy algorithm [55], also developed by Andrea, to dissect multivariate interactions (i.e., three-way associations). Inferences from these algorithms are then integrated with existing empirical evidence to produce a complete, context-specific network, the same procedure used to assemble the first human interactome in a human B cell [52]. Once the interactomes are generated, they are then interrogated using Master Regulator Analysis (MRA) to infer Master Regulators (MR) (the "driver" genes) for specific cancer phenotypes, drug response or a range of experimental questions. We have focused on molecular determinants of therapeutic response (see below).

Once assembled, the interactomes are then validated using biochemical and functional approaches. For biochemical validation, we are using the Nanostring nCounter technology to report on the expression levels of selected candidates after siRNA-mediated knockdown of individual as well as synergistic pairs of candidate master regulators [56]. This approach (now in progress) aims to uncover the hierarchical relationship between the candidate master regulators and hence, their prioritization for subsequent functional assays. Functional validation will be done using gain- or loss-of-function studies in mouse and human prostate cells/tissues using the appropriate lentiviral vectors, and their consequences evaluated for prostate tumor growth using tissue recombination

and orthotopic grafting assays [30, 57, 58].

Although we are still in the process of validating our draft interactomes for mouse and human prostate cancer, we have already benefited by interrogating these to identify molecular indicators that respond to combination targeted therapy in the GEM mice and are predictive of human prostate cancer (Fig. 7). In particular, we devised a strategy to identify genes that are critical for progression to advanced prostate cancer and that respond to combination therapy with RAP + PD, but not to chemotherapy (Docetaxel) (see Fig. 5). Specifically, we compared differentially-expressed genes between the Pten-AI versus Pten-Kras mice (i.e., progression genes; see Fig. 4) with the Pten-Kras mice treated with RAP + PD OR Docetaxel (or vehicle) for 5 days, to focus on genes that immediately respond to the treatment (e.g., therapeutic rather than phenotypic response).

Supervised clustering of the differentially-expressed genes revealed that the Pten-Kras mice treated with Docetaxel tend to cluster with the more aggressive cancer phenotype, namely the Pten-Kras mice, whereas the Pten-Kras mice treated with RAP + PD tend to cluster with the less aggressive phenotype, namely the Pten-AI mice (Fig. 7A). Furthermore, GSEA analyses comparing genes affected by treatment with the RAP + PD combination treatment showed a strong and significant enrichment for genes responsible for progression, which was not evident 10.2relevant for progression to aggressive prostate cancer .

Next, we used these data from the mouse model to interrogate our draft human prostate cancer interactome, with the goal of identifying molecular determinants of therapeutic response for human prostate cancer. For this purpose, we interrogated the human interactome with a gene set (745 genes) comprised of genes differentially expressed in the Pten-AI versus Pten-Kras mice (i.e., the "progression" genes) and then oppositely changed in the Pten-Kras mice following combination therapy but not changed following chemotherapy (Fig. 7C). This led to the identification of 23 candidate master regulators (18 up-regulated and 5 down-regulated) that are predicted determinants of the response to combination treatment.

These master regulators were prioritized by evaluating their relevance for biochemical recurrence in human prostate cancer [35, 59], which led to the identification of a 4-gene signature that has strong predictive value (p=0.001) for both disease recurrence and therapeutic response (Fig. 7D-F). We are now validating these findings using human tissue microarrays (TMAs). These findings highlight the value of integrating preclinical studies in mice with human clinical data using sophisticated bioinformatic resources for predicting drug response.

Summary

In summary, we have undertaken a comprehensive strategy that incorporates preclinical analyses of state-of-the-art mouse models with molecular interrogation of our mouse and human prostate cancer interactomes to evaluate novel therapeutic approaches for advanced prostate cancer and to define molecular determinants that predict drug response. We are now capitalizing on these models, resources and approaches to systematically evaluate drug response in GEM models with the ultimate goal of translating this information to predict patient response to particular agents and in specific phenotypic contexts. Cumulatively, our approach will provide a wealth of clinically-relevant insights that will aid in the vetting of effective drug combinations as well as predicting patients likely to benefit from such treatments.

REFERENCES

1) Capasso, L. L. (2005). Antiquity of cancer. Int J Cancer 113, 2-13.

2) Wolf, A. M., Wender, R. C., Etzioni, R. B., Thompson, I. M., D'Amico, A. V., Volk, R. J., Brooks, D. D., Dash, C., Guessous, I., Andrews, K., DeSantis, C. and Smith, R. A. American Cancer Society guideline for the early detection of prostate cancer: update 2010. CA Cancer J Clin 60, 70-98.

3) Sartor, A. O., Hricak, H., Wheeler, T. M., Coleman, J., Penson, D. F., Carroll, P. R., Rubin, M. A. and Scardino, P. T. (2008). Evaluating localized prostate cancer and identifying candidates for focal therapy. Urology 72, S12-24.

4) Gelmann, E. P. (2002). Molecular biology of the androgen receptor. J Clin Oncol 20, 3001-3015.

5) Huggins, C. and Hodges, C. V. (1941). The effect of castration, of estrogens, and of androgen injection on serum phosphatase in metastatic carcinoma of prostate. Cancer Res 1, 293-297.

6) Scher, H. I. and Sawyers, C. L. (2005). Biology of progressive, castration-resistant prostate cancer: directed therapies targeting the androgen-receptor signaling axis. J Clin Oncol 23, 8253-8261.

7) Petrylak, D. P., Tangen, C. M., Hussain, M. H., Lara, P. N., Jr., Jones, J. A., Taplin, M. E., Burch, P. A., Berry, D., Moinpour, C., Kohli, M., Benson, M. C., Small, E. J., Raghavan, D. and Crawford, E. D. (2004). Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer. N Engl J Med 351, 1513-1520.

8) Tannock, I. F., de Wit, R., Berry, W. R., Horti, J., Pluzanska, A., Chi, K. N., Oudard, S., Theodore, C., James, N. D., Turesson, I., Rosenthal, M. A. and Eisenberger, M. A. (2004). Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer. N Engl J Med 351, 1502-1512.

9) Logothetis, C. J. and Lin, S. H. (2005). Osteoblasts in prostate cancer metastasis to bone. Nat Rev Cancer 5, 21-28.

10) Abate-Shen, C. and Shen, M. M. (2000). Molecular genetics of prostate cancer. Genes Dev 14, 2410-2434.

11) Shen, M. M. and Abate-Shen, C. (2010). Molecular genetics of prostate cancer: new prospects for old challenges. Genes Dev 24, 1967-2000.

12) Abate-Shen, C., Shen, M. M. and Gelmann, E. (2008). Integrating differentiation and cancer: the Nkx3.1 homeobox gene in prostate organogenesis and carcinogenesis. Differentiation 76, 717-727.

13) Dong, J. T. (2006). Prevalent mutations in prostate cancer. J Cell Biochem 97, 433-447.

14) Shen, M. M. and Abate-Shen, C. (2003). Roles of the Nkx3.1 homeobox gene in prostate organogenesis and carcinogenesis. Dev Dyn 228, 767-778.

15) Gelmann, E. P. (2003). Searching for the gatekeeper oncogene of prostate cancer. Crit Rev Oncol Hematol 46 Suppl, S11-20.

16) Gao, H., Ouyang, X., Banach-Petrosky, W. A., Shen, M. M. and Abate-Shen, C. (2006). Emergence of androgen independence at early stages of prostate cancer progression in Nkx3.1; Pten mice. Cancer Res 66, 7929-7933.

17) Mulholland, D. J., Dedhar, S., Wu, H. and Nelson, C. C. (2006). PTEN and GSK3beta: key regulators of progression to androgen-independent prostate cancer. Oncogene 25, 329-337.

18) Shen, M. M. and Abate-Shen, C. (2007). Pten inactivation and the emergence of androgen-independent prostate cancer. Cancer Res 67, 6535-6538.

19) Kim, M. J., Bhatia-Gaur, R., Banach-Petrosky, W. A., Desai, N., Wang, Y., Hayward, S. W., Cunha, G. R., Cardiff, R. D., Shen, M. M. and Abate-Shen, C. (2002). Nkx3.1 mutant mice recapitulate early stages of prostate carcinogenesis. Cancer Res 62, 2999-3004.

20) Paez, J. and Sellers, W. R. (2003). PI3K/PTEN/AKT pathway. A critical mediator of oncogenic signaling. Cancer Treat Res 115, 145-167.

21) Uzgare, A. R. and Isaacs, J. T. (2004). Enhanced redundancy in Akt and mitogen-activated protein kinase-induced survival of malignant versus normal prostate epithelial cells. Cancer Res 64, 6190-6199.

22) Xin, L., Teitell, M. A., Lawson, D. A., Kwon, A., Mellinghoff, I. K. and Witte, O. N. (2006). Progression of prostate cancer by synergy of AKT with genotropic and nongenotropic actions of the androgen receptor. Proc Natl Acad Sci U S A 103, 7789-7794.

23) Majumder, P. K., Yeh, J. J., George, D. J., Febbo, P. G., Kum, J., Xue, Q., Bikoff, R., Ma, H., Kantoff, P. W., Golub, T. R., Loda, M. and Sellers, W. R. (2003). Prostate intraepithelial neoplasia induced by prostate restricted Akt activation: the MPAKT model. Proc Natl Acad Sci U S A 100, 7841-7846.

24) Abreu-Martin, M. T., Chari, A., Palladino, A. A., Craft, N. A. and Sawyers, C. L. (1999). Mitogen-activated protein kinase kinase kinase 1 activates androgen receptor-dependent transcription and apoptosis in prostate cancer. Mol Cell Biol 19, 5143-5154.

25) Gioeli, D., Mandell, J. W., Petroni, G. R., Frierson, H. F., Jr. and Weber, M. J. (1999). Activation of mitogen-activated protein kinase associated with prostate cancer progression. Cancer Res 59, 279-284.

26) Kinkade, C. W., Castillo-Martin, M., Puzio-Kuter, A., Yan, J., Foster, T. H., Gao, H., Sun, Y., Ouyang, X., Gerald, W. L., Cordon-Cardo, C. and Abate-Shen, C. (2008). Targeting AKT/mTOR and ERK MAPK signaling inhibits hormone-refractory prostate cancer in a preclinical mouse model. J Clin Invest 118, 3051-3064.

27) Malik, S. N., Brattain, M., Ghosh, P. M., Troyer, D. A., Prihoda, T., Bedolla, R. and Kreisberg, J. I. (2002). Immunohistochemical demonstration of phospho-Akt in high Gleason grade prostate cancer. Clin Cancer Res 8, 1168-1171.

28) Paweletz, C. P., Charboneau, L., Bichsel, V. E., Simone, N. L., Chen, T., Gillespie, J. W., Emmert-Buck, M. R., Roth, M. J., Petricoin, I. E. and Liotta, L. A. (2001). Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. Oncogene 20, 1981-1989.

29) Thomas, G. V., Horvath, S., Smith, B. L., Crosby, K., Lebel, L. A., Schrage, M., Said, J., De Kernion, J., Reiter, R. E. and Sawyers, C. L. (2004). Antibody-based profiling of the phosphoinositide 3-kinase pathway in clinical prostate cancer. Clin Cancer Res 10, 8351-8356.

30) Gao, H., Ouyang, X., Banach-Petrosky, W. A., Gerald, W. L., Shen, M. M. and Abate-Shen, C. (2006). Combinatorial activities of Akt and B-Raf/Erk signaling in a mouse model of androgen-independent prostate cancer. Proc Natl Acad Sci U S A 103, 14477-14482.

31) Gioeli, D. (2005). Signal transduction in prostate cancer progression. Clin Sci (Lond) 108, 293-308.

32) Cho, N. Y., Choi, M., Kim, B. H., Cho, Y. M., Moon, K. C. and Kang, G. H. (2006). BRAF and KRAS mutations in prostatic adenocarcinoma. Int J Cancer 119, 1858-1862.

33) Carter, B. S., Epstein, J. I. and Isaacs, W. B. (1990). Ras gene mutations in human prostate cancer. Cancer Res 50, 6830-6832.

34) Konishi, N., Hiasa, Y., Tsuzuki, T., Tao, M., Enomoto, T. and Miller, G. J. (1997). Comparison of ras activation in prostate carcinoma in Japanese and American men. Prostate 30, 53-57.

35) Taylor, B. S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B. S., Arora, V. K., Kaushik, P., Cerami, E., Reva, B., Antipin, Y., Mitsiades, N., Landers, T., Dolgalev, I., Major, J. E., Wilson, M., Socci, N. D., Lash, A. E., Heguy, A., Eastham, J. A., Scher, H. I., Reuter, V. E., Scardino, P. T., Sander, C., Sawyers, C. L. and Gerald, W. L. (2010). Integrative genomic profiling of human prostate cancer. Cancer Cell 18, 11-22.

36) Palanisamy, N., Ateeq, B., Kalyana-Sundaram, S., Pflueger, D., Ramnarayanan, K., Shankar, S., Han, B., Cao, Q., Cao, X., Suleman, K., Kumar-Sinha, C., Dhanasekaran, S. M., Chen, Y. B., Esgueva, R., Banerjee, S., LaFargue, C. J., Siddiqui, J., Demichelis, F., Moeller, P., Bismar, T. A., Kuefer, R., Fullen, D. R., Johnson, T. M., Greenson, J. K., Giordano, T. J., Tan, P., Tomlins, S. A., Varambally, S., Rubin, M. A., Maher, C. A. and Chinnaiyan, A. M. (2010). Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat Med 16, 793-798.

37) Abate-Shen, C., Banach-Petrosky, W. A., Sun, X., Economides, K. D., Desai, N., Gregg, J. P., Borowsky, A. D., Cardiff, R. D. and Shen, M. M. (2003). Nkx3.1; Pten mutant mice develop invasive prostate adenocarcinoma and lymph node metastases. Cancer Res. 63, 3886-3890.

38) Banach-Petrosky, W., Jessen, W. J., Ouyang, X., Gao, H., Rao, J., Quinn, J., Aronow, B. J. and Abate-Shen, C. (2007). Prolonged exposure to reduced levels of androgen accelerates prostate cancer progression in Nkx3.1; Pten mutant mice. Cancer Res 67, 9089-9096.

39) Banach-Petrosky, W., Ouyang, X., Gao, H., Nader, K., Ji, Y., Suh, N., DiPaola, R. S. and Abate-Shen, C. (2006). Vitamin D inhibits the formation of prostatic intraepithelial neoplasia in Nkx3.1;Pten mutant mice. Clin Cancer Res 12, 5895-5901.

40) Gao, H., Ouyang, X., Banach-Petrosky, W., Borowsky, A. D., Lin, Y., Kim, M., Lee, H., Shih, W. J., Cardiff, R. D., Shen, M. M. and Abate-Shen, C. (2004). A critical role for p27kip1 gene dosage in a mouse model of prostate carcinogenesis. Proc Natl Acad Sci USA 101, 17204-17209.

41) Gao, H., Ouyang, X., Banach-Petrosky, W. A., Gerald, W. L., Shen, M. M. and Abate-Shen, C. (2006). Combinatorial activities of Akt and B-Raf/Erk signaling in a mouse model of androgen-independent prostate cancer. Proc Natl Acad Sci USA 103, 14477-14482.

42) Kim, M. J., Bhatia-Gaur, R., Banach-Petrosky, W. A., Desai, N., Wang, Y., Hayward, S. W., Cunha, G. R., Cardiff, R. D., Shen, M. M. and Abate-Shen, C. (2002). Nkx3.1 mutant mice recapitulate early stages of prostate carcinogenesis. Cancer Res. 62, 2999-3004.

43) Kim, M. J., Cardiff, R. D., Desai, N., Banach-Petrosky, W. A., Parsons, R., Shen, M. M. and Abate-Shen, C. (2002). Cooperativity of Nkx3.1 and Pten loss of function in a mouse model of prostate carcinogenesis. Proc. Natl. Acad. Sci. USA 99, 2884-2889.

44) Ouyang, X., DeWeese, T. L., Nelson, W. G. and Abate-Shen, C. (2005). Loss-of-function of Nkx3.1 promotes increased oxidative damage in prostate carcinogenesis. Cancer Res 65, 6773-6779.

45) Ouyang, X., Jessen, W. J., Al-Ahmadie, H., Serio, A. M., Lin, Y., Shih, W. J., Reuter, V. E., Scardino, P. T., Shen, M. M., Aronow, B. J., Vickers, A. J., Gerald, W. L. and Abate-Shen, C. (2008). Activator protein-1 transcription factors are associated with progression and recurrence of prostate cancer. Cancer Res 68, 2132-2144.

46) Wang, X., Kruithof-de Julio, M., Economides, K. D., Walker, D., Yu, H., Halili, M. V., Hu, Y.-P., Price, S. M., Abate-Shen, C. and Shen, M. M. (2009). A luminal epithelial stem cell that is a cell of origin for prostate cancer. Nature 461, 495-500.

47) Lesche, R., Groszer, M., Gao, J., Wang, Y., Messing, A., Sun, H., Liu, X. and Wu, H. (2002). Cre/loxP-mediated inactivation of the murine Pten tumor suppressor gene. Genesis 32, 148-149.

48) Jackson, E. L., Willis, N., Mercer, K., Bronson, R. T., Crowley, D., Montoya, R., Jacks, T. and Tuveson, D. A. (2001). Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras. Genes Dev 15, 3243-3248.

49) Dankort, D., Filenova, E., Collado, M., Serrano, M., Jones, K. and McMahon, M. (2007). A new mouse model to explore the initiation, progression, and therapy of BRAFV600E-induced lung tumors. Genes Dev 21, 379-384.

50) Petrylak, D. P. (2007). New paradigms for advanced prostate cancer. Rev Urol 9 Suppl 2, S3-S12.

51) Basso, K., Liso, A., Tiacci, E., Benedetti, R., Pulsoni, A., Foa, R., Di Raimondo, F., Ambrosetti, A., Califano, A., Klein, U., Dalla Favera, R. and Falini, B. (2004). Gene expression profiling of hairy cell leukemia reveals a phenotype related to memory B cells with altered expression of chemokine and adhesion receptors. J Exp Med 199, 59-68.

52) Mani, K. M., Lefebvre, C., Wang, K., Lim, W. K., Basso, K., Dalla-Favera, R. and Califano, A. (2008). A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. Mol Syst Biol 4, 169.

53) Carro, M. S., Lim, W. K., Alvarez, M. J., Bollo, R. J., Zhao, X., Snyder, E. Y., Sulman, E. P., Anne, S. L., Doetsch, F., Colman, H., Lasorella, A., Aldape, K., Califano, A. and Iavarone, A. (2010). The transcriptional network for mesenchymal transformation of brain tumours. Nature 463, 318-325.

54) Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I. and Califano, A. (2006). Reverse engineering cellular networks. Nat Protoc 1, 662-671.

55) Wang, K., Alvarez, M. L., Bisikirska, R. L., Basso, K., Dalla Favera, R. and Califano, A. (2008). Dissecting the relationship between signaling and transcriptional regulation in human B cells. Nature Medicine In press,

56) Geiss, G. K., Bumgarner, R. E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D. L., Fell, H. P., Ferree, S., George, R. D., Grogan, T., James, J. J., Maysuria, M., Mitton, J. D., Oliveri, P., Osborn, J. L., Peng, T., Ratcliffe, A. L., Webster, P. J., Davidson, E. H., Hood, L. and Dimitrov, K. (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol 26, 317-325.

57) Gao, H., Ouyang, X., Banach-Petrosky, W., Borowsky, A. D., Lin, Y., Kim, M., Lee, H., Shih, W. J., Cardiff, R. D., Shen, M. M. and Abate-Shen, C. (2004). A critical role for p27kip1 gene dosage in a mouse model of prostate carcinogenesis. Proc Natl Acad Sci U S A 101, 17204-17209.

58) Kim, M. J., Cardiff, R. D., Desai, N., Banach-Petrosky, W. A., Parsons, R., Shen, M. M. and Abate-Shen, C. (2002). Cooperativity of Nkx3.1 and Pten loss of function in a mouse model of prostate carcinogenesis. Proc Natl Acad Sci U S A 99, 2884-2889.

59) Glinsky, G. V., Glinskii, A. B., Stephenson, A. J., Hoffman, R. M. and Gerald, W. L. (2004). Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest 113, 913-923.

THE ORIGIN AND EVOLUTION OF A PANDEMIC VIRUS

ZACHARY CARPENTER, CARLOS HERNANDEZ, JOSEPH CHAN AND RAUL RABADAN

CENTER FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
COLUMBIA UNIVERSITY

April 9th 2009, a 54 year old man and his 16 year old daughter visiting a clinic in San Diego presented with acute respiratory illness. Several similar cases of febrile respiratory illness followed within the next two weeks in California and Texas. Concurrently, Mexican public health authorities were reporting an increasing number of respiratory diseases, some with serious complications, severe pneumonia and death. In the capital, the number of patients presenting severe respiratory illness increased dramatically with 854 cases of pneumonia and 59 reported deaths by the 24th of April. None of the initial cases in the United States had traveled to Mexico before the onset of the respiratory illness.

The analysis of the first cases released on the 21st of April by the Centers for Diseases Control (CDC) indicated that the respiratory illness was caused by an influenza virus, similar to the H1N1 viruses that were circulating in pigs in North America. Genetically and antigenically, swine H1N1 viruses are sufficiently different to previous human H3N2 or H1N1 strains that vaccination or previous exposure to seasonal viruses probably would not provide protection. In addition, none of the new cases seemed to have been in contact or proximity to pigs, suggesting that the emergent virus was able to transmit from human to human. The same swine origin H1N1 influenza virus was isolated in specimens from patients in Mexico. In contrast to the seasonal influenza, the new virus preferentially infected young healthy adults. By the end of April, it was clear that a novel influenza virus was spreading in North America from human to human transmission. Mexican authorities reacted immediately by cancelling classes and public events. In April, the World Health Organization (WHO) alerted that the new strain had the potential to become pandemic.

By the 28th of April, mild cases were reported in Canada, Spain, Israel, United Kingdom and New Zealand. On the 11th of June, WHO officially declared the first pandemic of the 21st century.

Influenza A viruses, like the seasonal or the pandemic H1N1 viruses, are RNA viruses that possess an enormous capacity to evolve and diversify. The size of a human genome is 3 billion bases that, collectively, code for more than 20,000 genes. Replication in humans occurs every 20 to 30 years and maintains remarkable genetic similarity across generations. viruses produce an enormous quantity of highly divergent copies.

This similarity is due to our intrinsic mutation rate:

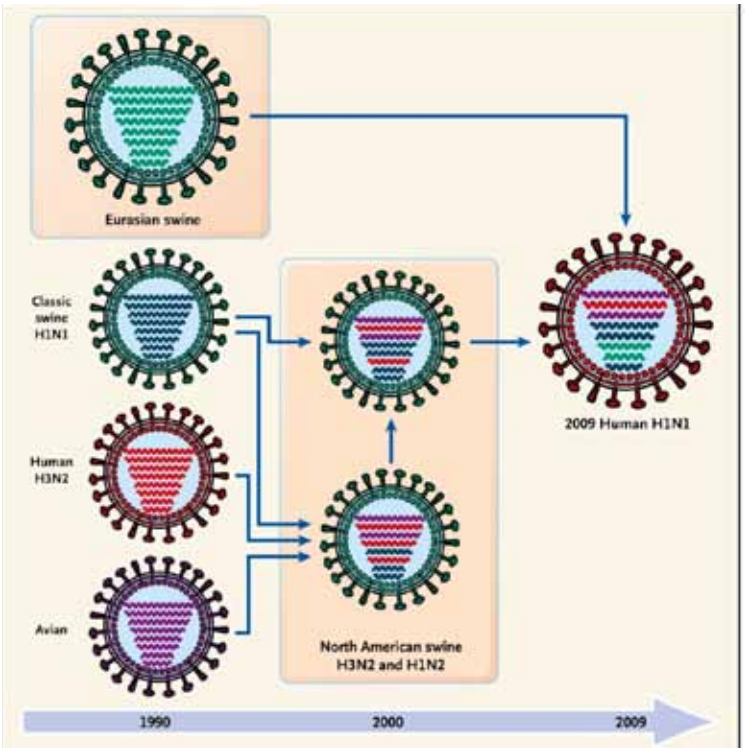


Figure 1: History of Reassortment Events in the Evolution of the 2009 Influenza A (H1N1) Virus (Trifonov, Khiabanian, Rabadan, New England Journal of Medicine).

10^{-8} per replication per base, a relatively low rate in biology. In contrast, genomes of RNA viruses, like flu or HIV, are very small that is 4 orders of magnitude higher than that of humans. Humans and RNA viruses, thus, employ two opposite evolutionary strategies: humans produce a small number of high fidelity copies, while RNA (10,000 bases and only 10 or so genes), and produce millions of copies every day. This high replication rate is compounded by a mutation rate number of high fidelity copies, while RNA viruses produce an enormous quantity of highly divergent copies.

Flu virions contain a segmented negative-sense RNA genome comprised of eight regions responsible for encoding 10 or 11 functionally distinct proteins. In spite of this apparent proteomic simplicity, these viruses are able to hijack and utilize the complex machinery of the cell. The possession of a segmented genome allows for independent assortment of segments during viral assembly. When two different viruses co-infect the same cell, they can produce progeny containing the genetic material of both parental strains, and this process allows for remarkable acceleration of influenza's already rapid rate of evolution. It is held that the emergence of pandemics tightly associates with reassortment of genome segments from divergent influenza strains and subsequent zoonotic transitions. This series of events furnishes influenza viruses unrecognized by existing immune surveillance in human populations. The epidemiological severity of these emerging strains depends on the collective fitness and transmissibility of the virus, constituted by its particular assortment of genome segments. The recent H1N1 arose through a complicated reassortment process whose story is still not fully elucidated. By comparing the new virus to the 10,000 different genomes that have been collected and sequenced since 1918, we are able to trace back its ancestors. The new virus is related to swine viruses isolated on two continents, North America and Eurasia. The North American ancestors were isolated in pigs between 1998 and 2002, and were the result of a reassortment between swine, avian and human viruses, the so-called triple reassortment, at the end of the 90s. The origin of the Eurasian ancestors is even more mysterious. These isolates were found in Europe at the beginning of the 90s; yet, since this acquisition the data has been sparse. How, when, and where these two ancestors recombined remains an open question; one that probably will only be solved through implementation of better worldwide surveillance systems.

In this direction, more than a thousand influenza H1N1 viruses have been isolated and sequenced around the globe. Viruses from New York, Tokyo, Paris, Mexico, Argentina, etc. show a very similar genome with subtle differences. By taking all this information into account, we can estimate that the most recent

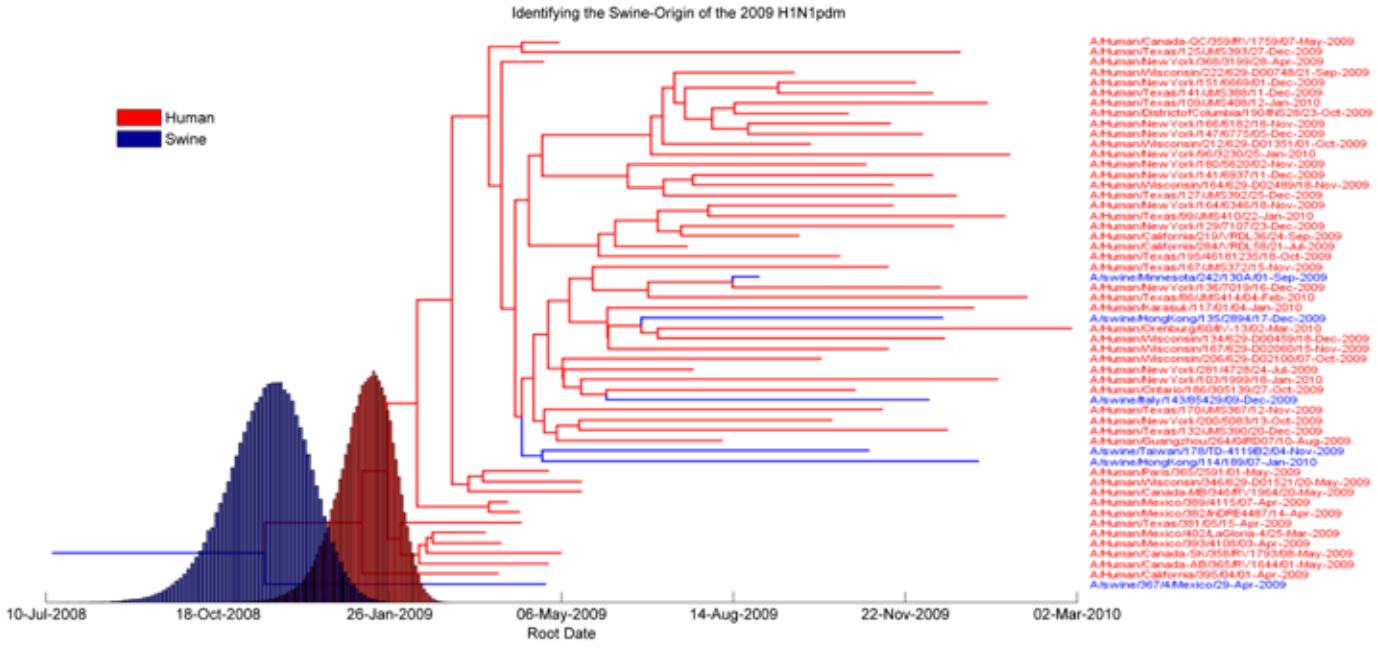


Figure 2: Bayesian Markov chain Monte Carlo analysis of 54 human (red) and swine (blue) influenza sequences. The most recent common ancestor of all the strains dates back to the last weeks of 2008, prior to the date of the most recent common ancestor of the human pandemic strains.

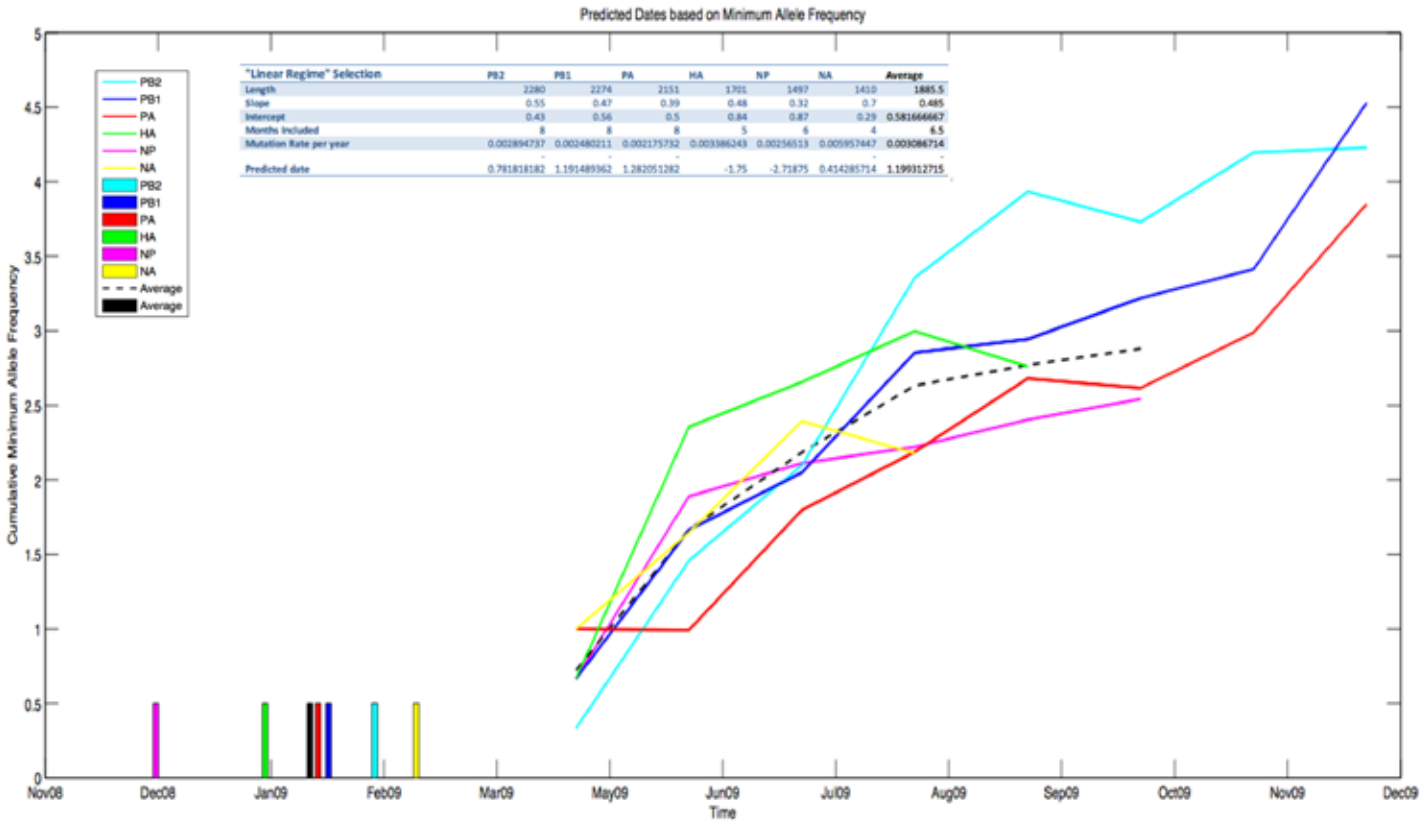


Figure 3: Segregating site analysis of available pandemic strain influenza sequences. Figure taken from submitted manuscript.

common ancestor to the viruses currently spreading in the world is very recent, having most likely arisen during the middle of January of 2009, compatible with the first reported cases in Mexico and the United States. This analysis has been done by several methods that independently suggest similar dates. One method employs classical phylogenetic techniques such as maximum likelihood and neighbor joining, and is complemented by more sophisticated algorithms such as the Bayesian Markov chain Monte Carlo method employed by BEAST (Figure 2). A second novel method currently being developed in the lab utilizes segregating sites to estimate a date of pathogenic emergence and results in a similar prediction. Additionally, because this analysis is independent of the aforementioned classical techniques, its findings serve as a strong control (Figure 3).

Although most of influenza proteins are represented in viral particles, they are differentially recognized by the immune system. One of the main factors that determine immunogenicity of a particular protein is its relative proximity to the viral surface, and resultant accessibility to antibody binding. These antibodies provide selective pressures that result in selection for what are termed "antibody escape mutants". A portion of the derived mutations become fixed within particular strains and represent a major subset of the mutations that define antigenic drift in influenza viruses.

At the moment, the most effective strategy to combat influenza viruses is through vaccination, but antigenic drift adds complexity to vaccine development. The first step in the process of creating an influenza vaccine is to choose a virus that we think will represent the circulating virus in the near future. A high evolutionary speed, attained through mutations and reassortment, is a very successful strategy that makes RNA viruses, like flu, very hard to predict. Tracking this antigenic drift is critical in the development of effective vaccines.

It has been proposed that detecting viral movement by sequence analysis is further complicated by skewed geographic and seasonal distributions in viral isolates. Implementation of spatiotemporal clustering through binomial modeling of regional and seasonal

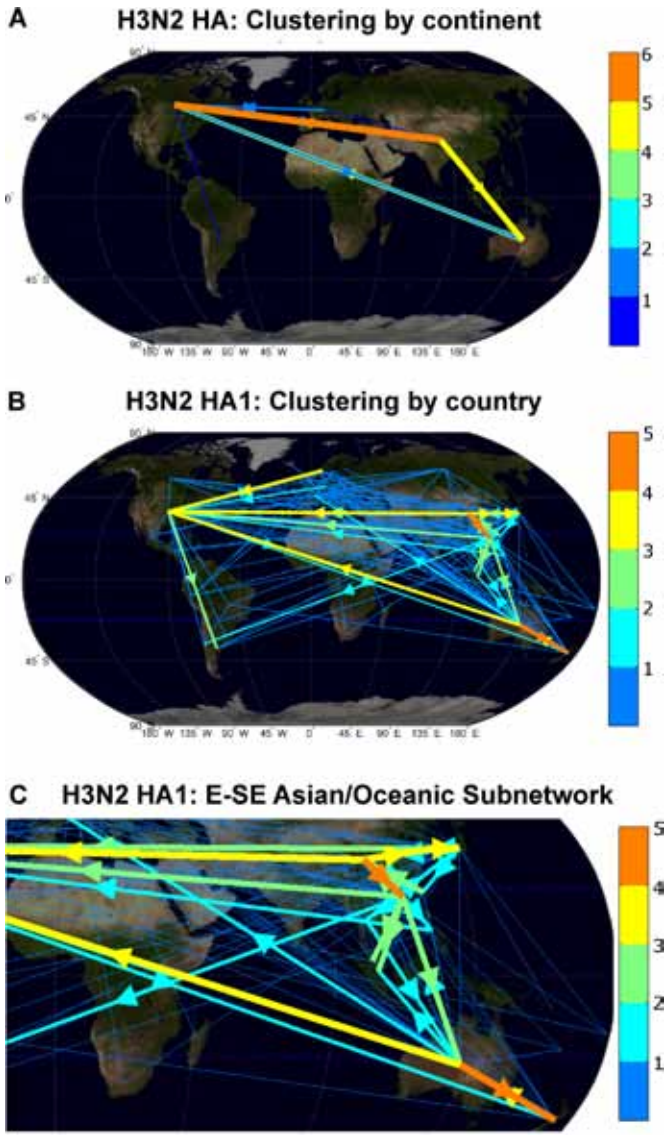


Figure 4: Global network of statistically significant seeding seasons for H3N2 after clustering by (A) continent and (B) country. **A.** Seasonal variants emerge from Asia and make their way to North America. A smaller connection from North America to South America is consistent with the finding that South American isolates are antigenically delayed [8]. **B.** Clustering by country showed tropic-centric movement patterns. **C.** H3N2 seeding events in the South-East Asian/Oceanic subnetwork showing China, Hong Kong, and Australia as major hubs. Arrows signify the direction of the seeding event. Each edge is color-coded according to weight (the number of seeding events represented). For visual simplicity, arrowheads were omitted for edges of unit weight. Edges connect between the centroids of two continents or countries. World map image taken from: onearth.jpl.nasa.gov. doi:10.1371/journal.pcbi.1001005.g004

transmission has been used to counter this data bias, producing networks of influenza spread. Graph theory techniques can then be used to predict the source of new H3N2 seasonal variants as well as to suggest areas in the world where increased vaccination could act to maximally disrupt widespread geographical progression (Figure 4).

Influenza A represents one of mankind's greatest and most interesting pathogens. To bolster this claim, one need only review the repercussions of the 1918 Spanish Flu, which by some estimates was responsible for culling approximately six percent of the human population. Although Influenza A pandemics occur discretely and do not always attain such high levels of virulence, they remain a significant global health threat and economic burden, particularly when compounded with those imposed by seasonal epidemics. Further understanding of Influenza biology, and that of other pandemic associated pathogens, is imperative to the mitigation of their associated risk; the recently emerged swine-origin H1N1 virus pandemic in 2009 provides a very lucid justification of this point. Fundamental to understanding a pandemic is elucidating its point of origin both in space and time. Resolving this information will allow for better prevention strategies to be implemented in the future.

Although the virulence of the 2009 swine flu was relatively quite low, the ease of its transmission and rapid global spread highlights the imperative need for an effective worldwide system of surveillance for emerging viruses. Efforts are needed to map and to identify mutations that could confer resistance to current drugs or vaccines, to learn how to assess the transmissibility of a pathogen in humans, and to identify the molecular factors that determine the virulence of a pathogen. Basic science provides the best rationale for implementing effective public health measures.

THE TRANSCRIPTIONAL NETWORK FOR MESENCHYMAL TRANSFORMATION OF BRAIN TUMOURS

ANDREA CALIFANO

CENTER FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
COLUMBIA UNIVERSITY

Introduction

The enormous amount of data that we are able to obtain through high-throughput genomic technologies in the post-genomic era represents an advance as well as a challenge in our ability to dissect the molecular underpinning of disease. We no longer have a paucity of data from which to draw conclusions, but we are presented with the charge of condensing the data into rational and testable hypotheses in order to efficiently focus experimental efforts to identify the genes responsible for a given pathophysiological state. Human regulatory network reconstruction, although still in its infancy, is proving to be a highly useful interrogative tool which can tease apart some of the key cellular regulatory networks and the master regulatory genes that reside within those networks. Moreover, several in silico tools introduced by MAGNet investigators are being used to interrogate regulatory networks using signatures of differentially expressed genes. Instead of asking the question "which genes are differentially expressed in a specific neoplasm", they are asking "which transcriptional, post-transcriptional and post-translational regulators control the genes that are differentially expressed in a neoplastic phenotype", with the goal of identifying so called Master Regulator (MR) genes causally involved in a pathologic transformation or a physiologic differentiation event (1).

Recently MAGNet investigators have used information theoretic methods to build and interrogate the transcriptional regulatory network that is responsible for malignant gliomagenesis (2). High-grade gliomas, which include anaplastic astrocytoma (AA) and glioblastoma multiform (GBM), are the most common intrinsic brain tumors in adults and are almost invariably lethal, largely as a result of their lack of responsiveness to current therapy (3). Physician's efforts to treat these aggressive tumors have been stymied by their incredibly destructive nature and capacity to invade nearby brain tissue and form new blood vessels that support and fuel growth. Over the last several years numerous studies have shown that the genomic profiles of these tumors offers clues in classifying them into subclasses based on aggressiveness, and may help classify patients as to their response to therapies. Recently, glioma samples have been segregated into three groups with distinctive gene expression (GEP) signatures, displaying preferential expression of genes characteristic of neural tissues (proneural), proliferating cells (proliferative) or mesenchymal tissues (mesenchymal) (4). Malignant gliomas in the mesenchymal group express genes linked with the most aggressive properties of GBM tumors (migration, invasion and angiogenesis) and invariably coincide with disease recurrence and the shortest post-diagnosis survival of GBM patients.

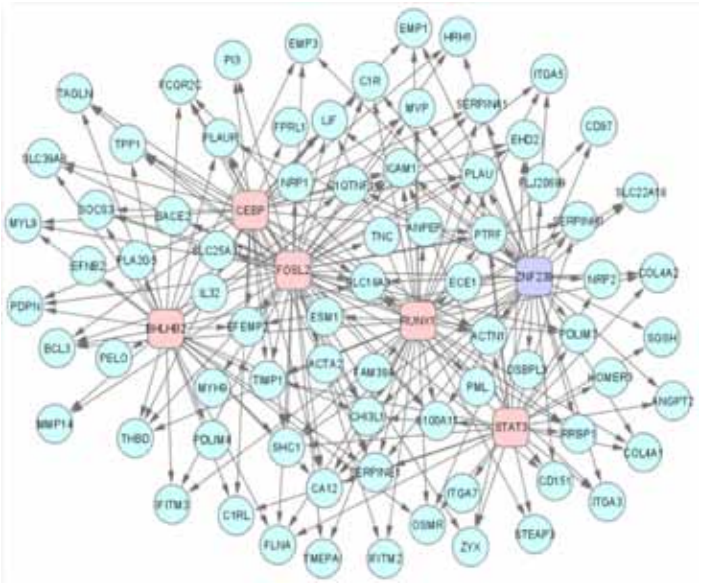


Figure 1: Mesenchymal subnetwork with TFs involved in activation of MGES in pink, and those involved in repression in purple. MGES targets are in cyan. These 6 TFs control over 74% of the genes in the mesenchymal signature.

Defining and Validating an in silico Predicted Regulatory Module

Using gene expression profiles from a large number of high-grade gliomas (grade III and IV), the ARACNe (5) algorithm was employed to assemble a genome-wide repertoire of transcriptional interactions for these tumors. This allowed the definition of a regulon for each transcription factor (TF) within the network (i.e., the full complement of the TF targets). This network was then interrogated using the master regulator analysis algorithm (MARINa) in order to compute the statistical significance of the overlap between the regulon of each TF and the mesenchymal gene expression signature (MGES) genes. This analysis yielded 53 MGES-specific TFs. Following ranking of these MRs on the basis of the total number of MGES targets they regulated, a mesenchymal regulation subnetwork of 6 TFs could be identified, which collectively controlled over 74% of the MGES (Figure 1). Remarkably, network reconstruction and interrogation using 3 completely independent datasets (4, 6, 7) yielded virtually the same master regulators, indicating that there are common regulatory mechanisms associated with the emergence of specific expression signatures (proneural, proliferative, and mesenchymal) (Figure 3).

Experimental validation in vitro and in vivo confirmed the role of the candidate MRs. Specifically, chromatin immunoprecipitation (ChiP) experiments confirmed that on average 80% of the tested targets were indeed bound by their respective TFs in the module and gene expression profiling following both silencing and ectopic expression of the 6 TFs confirmed control of the inferred programs. Computational analysis, ChiP experiments, and shRNA knockdown assays predicted a small, tightly connected, self-regulating, highly modular hierarchy comprising the 6 transcription factors that appear to regulate the mesenchymal signature, with the transcription factors CCAAT/enhancer-binding protein beta and delta (C/

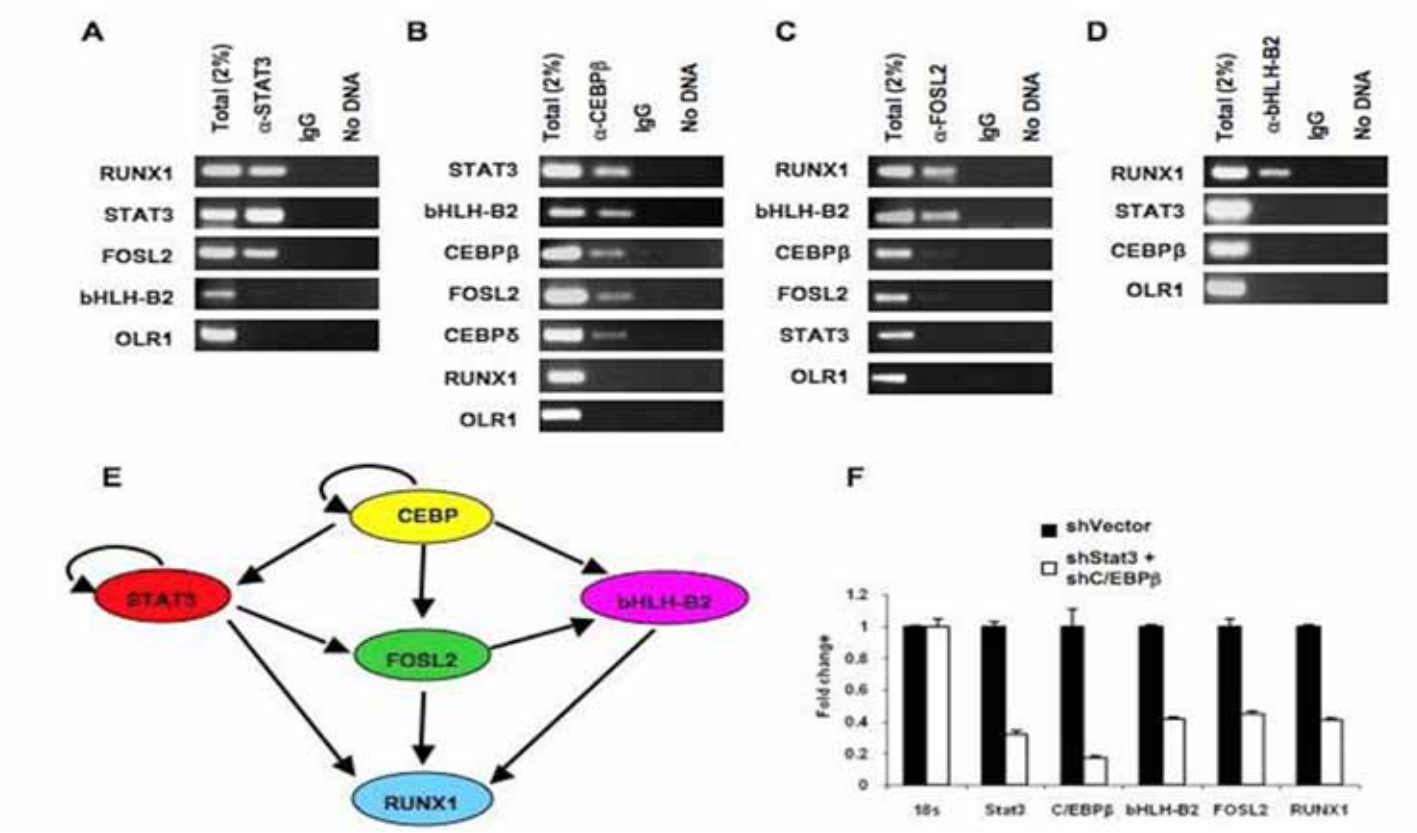


Figure 2 : A hierarchical transcription module regulates the interactions between the master TFs in the mesenchymal module. ChiP experiments with antibodies specific for **(A)** STAT3, **(B)** C/EBP β, **(C)** FOSL2, **(D)** bHLH-B2 showing genomic regions of genes containing binding sites for specific TFs within the module. **(E)** A graphical representation of the module after promoter occupancy analysis showing autoregulatory and feed-forward loops among the TFs and showing C/EBPβ/δ and STAT3 on top of the hierarchy. **(F)** Quantitative RT-PCR of TFs after lentivirus mediated shRNA knockdown of C/EBPβ and STAT3.

EBP/β/δ and signal transducer and activator of transcription 3 (STAT3) on top of the hierarchy (2) (Figure 2). Strikingly, C/EBPβ/δ and STAT3 are not differentially expressed in the glioma tumors and would never have been identified by traditional methods focused solely on their expression profile variance.

Biological Assays Have the Final Word

The value of a computational model lies only in as far as it is predictive and, therefore, biologically testable. The introduction of both C/EBPβ/δ and STAT3 (but not either one alone) was sufficient to reprogram mouse neural stem cells into cells exhibiting a fibroblast-like morphology and the expression of mesenchymal markers, and confers an invasive growth potential in wound healing assays. Conversely, silencing of both genes, either in human GBM-derived tumor initiating cells or in a stable human glioma cell line, resulted in loss of mesenchymal signature and abrogation of invasive growth in vitro and in mouse xenografts. Notably, immunohistological analysis of an independent glioblastoma cohort, with associated patient survival outcome data, showed that tumors that were double positive for C/EBPβ and activated STAT3 had the worst clinical outcome, compared to either single positive or double negative staining tumors. Taken together these experiments showed that C/EBPβ/δ and STAT3 are synergistic master regulators of the mesenchymal subtype of glioblastoma and are necessary and sufficient to confer mesenchymal transformation.

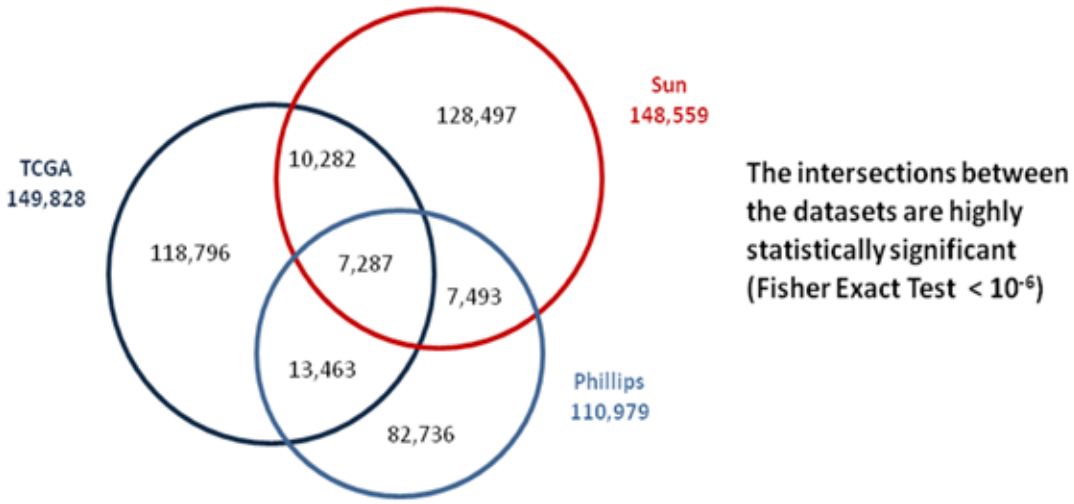


Figure 3: Venn Diagram showing the intersection between the 3 ARACNe ouptputs obtained with the 3 high-grade glioma gene expression profile datasets.

Conclusions

Regulatory network based approaches may dramatically increase our ability to identify therapeutic targets and disease biomarkers by extending them from a small set of alteration-harboring genes to a much broader set of genes that act as master regulators and integrators of an entire spectrum of aberrant signals originating from upstream genetic alterations. Specifically, we have hypothesized and experimentally confirmed that non-mutated transcriptional regulators that are affected by numerous oncogenes may constitute better therapeutic targets and biomarkers than any of the upstream oncogene alterations whose signals they integrate. For instance, C/EBPβ and STAT3 constitute optimal biomarkers and valuable genetic targets when inhibited in combination, as they abrogate tumorigenesis in vivo and can effectively discriminate between worst and best prognosis in GBM patients. This result could not have been produced by genetic analysis precisely because these genes are not mutated in this tumor subtype, nor are they differentially expressed. Additionally these genes provide an entry point to investigate which genes may in turn regulate them, which could lead to additional targets for potential therapeutic intervention.

REFERENCES

1. Lim W, Lyashenko E, Califano A. Master Regulators Used As Breast Cancer Metastasis Classifier. Pac Symp Biocomp. 2009;14:492-503, PMID: 19209726

2. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature. 2010;463(7279):318-25.

3. Ohgaki H, Kleihues P. Population-based studies on incidence, survival rates, and genetic alterations in astrocytic and oligodendroglial gliomas. J Neuropathol Exp Neurol. 2005;64(6):479-89.

4. Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell. 2006;9(3):157-73.

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

6. Network CGAR. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061-8.

7. Sun L, Hui AM, Su Q, Vortmeyer A, Kotliarov Y, Pastorino S, et al. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. Cancer Cell. 2006;9(4):287-300.

MICRORNA MODULATORS REGULATE ONCOGENES AND TUMOR SUPPRESSORS IN GLIOBLASTOMA

ANDREA CALIFANO LAB

A key interest in the Califano lab is the dissection of molecular interaction networks that are dysregulated in various neoplasms, including high-grade human gliomas. Although microRNAs have emerged as key regulators of both normal and pathologic phenotypes, including cancer, an understanding of their regulation in glioblastoma, the most common and the most aggressive type of primary human adult brain tumor, is still in its infancy. Surprisingly, there appear to be a number of previously unsuspected post-transcriptional regulation processes that may account for a significant portion of the unexplained genetic variability associated with tumor initiation and progression. Indeed we identified and experimentally validated a large number of genes capable of modulating both microRNA biogenesis and their on-target activity, including that of microRNA previously characterized as oncogenic, such as miR-21, MiR-26a and miR-221/2. Furthermore, we have identified several gene clusters that regulate each other by competing for large, common microRNAs programs.

Specifically, we extended the MINDy algorithm to systematically identify candidate modulators of microRNA biogenesis and on-target activity in glioblastoma. The algorithm identified hundreds of candidate modulators, including activators and suppressors of miRISC-mediated regulation, and mRNAs that compete for one or more microRNAs. Experimental validation was performed in SNB19 cell line by measuring differential expression of mature microRNAs and their targets after ectopic expression or inhibition of predicted regulator proteins. 3' UTR luciferase assays were further used to establish the post-transcriptional nature of the regulation. Focusing on two known drivers of gliomagenesis and progression, we showed that PTEN is post-transcriptionally up-regulated by WNT7A through miR-21 and down-regulated by PALB2 through miR-106a, and that RUNX1 is post-transcriptionally down-regulated by WIPF2 through miR-17-5p. Further, we showed that PTEN and RUNX1, which are widely considered functionally unrelated, are part of a cluster of genes that regulate each other by competing for a common microRNA program. In addition, other modulators with prognosis-predictive genetic alterations act through common microRNA-regulatory programs to regulate PTEN and RUNX1 expression. For example, prognosis-predictive genetic alterations at the CTBP2 locus significantly alter both CTBP2 and PTEN expression in glioblastoma; siRNA-mediated silencing of CTBP2 resulted in a 30% reduction in the expression of PTEN 3' UTR luciferase reporters. Taken together, genetic alterations of these modulators account for virtually all the cases (~25%) where PTEN is not deleted but is functionally down-regulated in glioblastoma. These results suggest that microRNA modulation may play a significant role in tumorigenesis and progression of glioblastoma.

PREDICTING DISEASE PHENOTYPE FROM GENOTYPE

CHRIS WIGGINS LAB

Spurred by technological advances in high-throughput sequencing and genotyping of Single Nucleotide Polymorphisms (SNP), Genome-Wide Association Studies have become a promising tool to answer fundamental questions on the genetic basis of complex diseases. The Wiggins lab, in collaboration with Dr. Trey Ideker and Dr. Yoav Freund (UC San Diego), has been applying powerful, large-margin classification algorithms like Adaboost to genotype data obtained from large-scale studies conducted by the Wellcome Trust Case Control Consortium. The algorithm infers predictive models on a class of highly expressive, tree-structured models called Alternating Decision Trees, complex models built from simple single-SNP genotype-based decision rules.

Armed with large amounts of high-dimensional data, the Wiggins lab has been able to learn models that are highly predictive of disease state for complex diseases like Type-1 diabetes, Type-2 diabetes and Bipolar Disorder. Additionally, the inferred models are highly sparse, facilitating functional interpretation, encoding predictive epistatic interactions between loci and robustly identifying regions of the genome

that contain putative, disease-relevant causal variants.

IDENTIFYING VIRUS HOSTS FROM SEQUENCE DATA

CHRIS WIGGINS LAB

Emerging pathogens constitute a continuous threat to our society, with recent outbreaks including the West Nile virus in New York (1999), LUJO virus in Lusaka (2008) and H1N1 influenza pandemic virus in Mexico and the US (2009). An integral part of rapid and effective public health measures during viral pandemics is the accurate identification and characterization of the pathogen; a notoriously difficult task in the initial stages of the pandemic when, often, very little reliable, biological information about the virus is known.

Using the mismatch feature space representation of sequence data and Adaboost, a popular classification algorithm, our labs have been able to learn models from protein sequence data that are strongly predictive of hosts of viruses within a family. Furthermore, the sparse models learned by Adaboost have helped identify sequence motifs or genomic regions that are strongly conserved among viruses sharing a particular host type. Such strong motif-conservation suggests putative host-specific functional adaptation, allowing us to tease out the specific mutations necessary for a virus to infect a new host.

CHASING DRIVERS OF CANCER WITH DATA INTEGRATION

DANA PE'ER LAB

The Pe'er lab has combined alterations in copy number with gene expression to develop a method for identifying key genes driving cancer. The method, called CONEXIC (COpy Number and EXpression In Cancer) was applied to studying melanoma. It focuses on genes which have alterations of copy number in some, but not necessarily all the tumor samples. CONEXIC finds gene expression signatures (groups of co-expressed genes) that serve as "genomic footprint", and, unique to our approach, the single strongest combination of a small number of master regulators controlling these signatures. This analysis correctly identified known drivers of melanoma such as MITF and connected them to many of their targets and biological functions. In addition, it predicted novel melanoma tumor dependencies, two of which, TBC1D16 and RAB27A, were confirmed experimentally. This work was published in Cell, December 2010.

RECONSTRUCTION AND ANALYSIS OF THE PLASMODIUM FALCIPARUM METABOLIC NETWORK

DENNIS VITKUP LAB

Malaria remains one of the most severe public health challenges worldwide (W.H.O., 2008). In response to the urgent need for new drugs and treatments on the face of rapid emergence of acquired drug resistance to existing drugs by the most lethal malaria causative agent, Plasmodium falciparum (Mackinnon & Marsh, 2010), we built a genome-scale flux-balance model of the P. falciparum metabolism to identify new therapeutic drug targets (Plata et al, 2010). The model was constructed using metabolic reconstruction tools develop in the Vitkup lab; it includes 366 genes, 1001 reactions, 616 metabolic species, and 4 cellular compartments. We applied flux-balance analysis (Orth et al, 2010) to identify the genes and reactions that are required to produce a set of biomass components necessary for growth of the parasite. When compared to the yeast metabolic network (Duarte et al, 2004), the Plasmodium network has a significantly higher proportion of essential genes. We confirmed this result with 90% accuracy using a comparative analysis of known gene knockouts in the two microbes. This low level of genetic robustness, which is likely due to the parasitic lifestyle, suggests that many metabolic genes of the parasite can be used as effective drug targets. We further verified experimentally one of the enzymes identified as essential: nicotinate mononucleotide adenylyltransferase (NMNAT, Figure 1A) using compound 1_03 (Sorci et al, 2009), a candidate for new anti-malarials that completely blocked host cell escape and reinvasion by arresting parasites in the trophozoite growth stage (Figure 1B).

The metabolic model of the parasite can be also used to integrate various genomic data, such as gene expression (Oberhardt et al, 2009). To illustrate these possibilities, we applied gene-expression data as constraints for the flux-balance model (Colijn et al, 2009) in order to predict changes in metabolic exchange fluxes. We found that the model was able to correctly predict the changes in external metabolite concentrations (Olszewski et al, 2009) with about 70% accuracy. The availability of a human metabolic network reconstruction (Duarte et al, 2007) would allow, in the future, to analyze the combined parasite-host network, which would deepen understanding of the P. falciparum metabolic vulnerabilities.

REFERENCES

Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, Cheng TY, Moody DB, Murray M, Galagan JE (2009) Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. PLoS Comput Biol 5: e1000489

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A 104: 1777-1782

Duarte NC, Herrgard MJ, Palsson BO (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Res 14: 1298-1309

Mackinnon MJ, Marsh K (2010) The selection landscape of malaria parasites. Science 328: 866-871

Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5: 320

Olszewski KL, Morrisey JM, Wilinski D, Burns JM, Vaidya AB, Rabinowitz JD, Llinas M (2009) Host-parasite interactions revealed by Plasmodium falciparum metabolomics. Cell Host Microbe 5: 191-199

Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nat Biotechnol 28: 245-248

Plata G, Hsiao TL, Olszewski KL, Llinás M, Vitkup D (2010) Reconstruction and flux-balance analysis of the Plasmodium falciparum metabolic network Mol Syst Biol. 2010 Sep 7; 6:408

Sorci L, Pan Y, Eyobo Y, Rodionova I, Huang N, Kurnasov O, Zhong S, MacKerell AD, Jr., Zhang H, Osterman AL (2009) Targeting NAD biosynthesis in bacterial pathogens: Structure-based development of inhibitors of nicotinate mononucleotide adenylyltransferase NadD. Chem Biol 16: 849-861

W.H.O. (2008) World malaria report 2008, Geneva: World Health Organization.

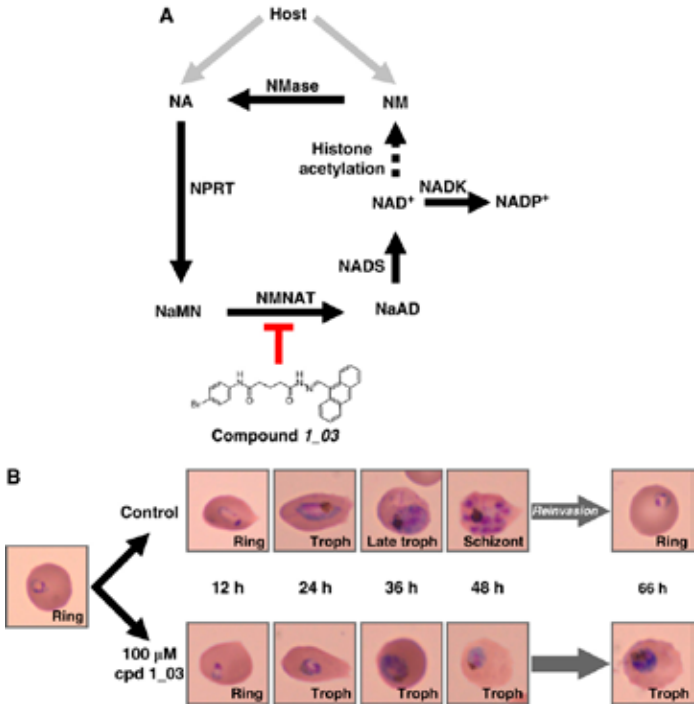


Figure 1: Small-molecule inhibition of the parasite nicotinate mononucleotide adenylyltransferase (NMNAT). **A.** Schematic of the P. falciparum NAD(P) synthesis and recycling pathway determined from the genome sequence. Nicotinamide (NM) and nicotinic acid (NA) can be scavenged from the host. Compound 1_03 is an inhibitor targeting NMNAT. **B.** Compound 1_03 causes growth arrest of intraerythrocytic P. falciparum. Cultures were resuspended in niacin-free medium containing 0 or 100 µM of compound 1_03 at early ring stage and observed for 66 hours (see Methods). Untreated parasites undergo normal development and reinvasion, while drug-treated parasites arrest at the trophozoite ("troph") stage and do not reinvade. Abbreviations: NM, nicotinamide; NA, nicotinic acid; NaMN, nicotinate mononucleotide; NaAD, nicotinate adenine dinucleotide; NAD(P)+, nicotinamide adenine dinucleotide (phosphate), reduced; NMase, nicotinamidase; NPRT, nicotinate phosphoribosyltransferase; NMNAT, nicotinate mononucleotide adenylyltransferase; NADS, NAD synthase; NADK, NAD kinase.

GENSPACE: COMMUNITY-DRIVEN KNOWLEDGE SHARING IN GEWORKBENCH

GAIL KAISER LAB

geWorkbench is frequently expanded through the addition of new analysis and visualization modules. The number of geWorkbench users is also increasing. Most users may be familiar with a few of the tools in geWorkbench, but typically they are not familiar with all of them. geWorkbench includes extensive documentation on the tools, but this static documentation can lag behind the addition of new or updated tools. Choosing the right tool to use and how to chain these tools in sequence (workflows) can be very daunting, especially to new users. Our geWorkbench plug-in called genSpace aims to alleviate this problem. It provides users with recommendations in geWorkbench directly using collaborative filtering, which is a technique for generating recommendations for users through a "people like you" interface and is used by websites such as Amazon.com and Netflix to suggest new products to customers based upon their purchasing history.

One substantial way that we diverge from and expand upon the collaborative filtering provided by popular websites is that we address the ordering of related activities conducted in sequence, i.e., as a workflow. For example, a common workflow in geWorkbench is to run the ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks) analy sis followed by the MINDy (Modulator Inference by Network Dynamics) analysis. Through genSpace, users can search for recommendations such as most common workflows including a given tool, and most common next steps given the user's current activity.

genSpace logs, aggregates, and analyzes geWorkbench users' activities to provide these recommendations. All of the recommendations are derived using data mining and collaborative filtering techniques, which use geWorkbench users'

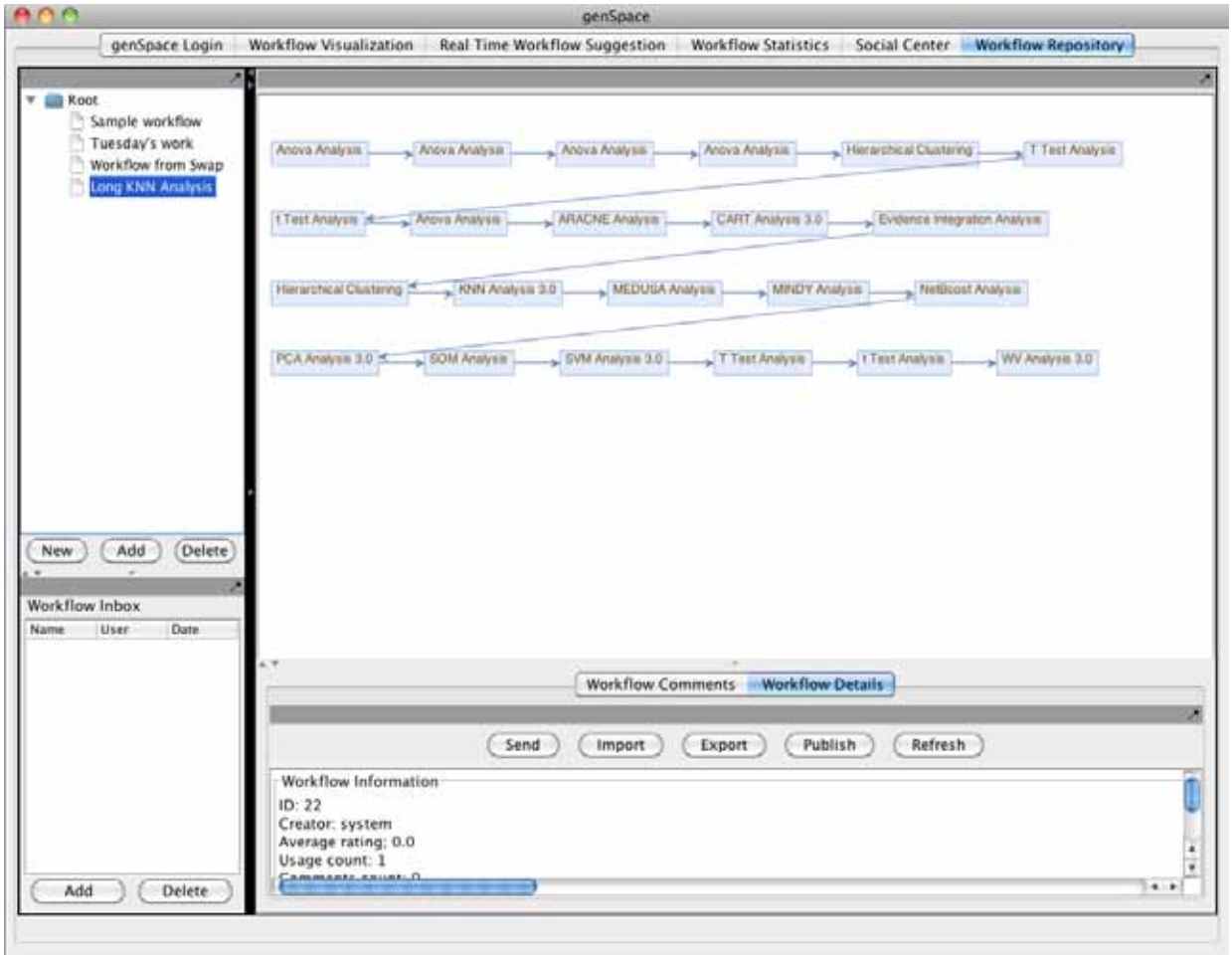


Figure 2: genSpace's Workflow Repository, displaying a long sample workflow

past usage history to provide recommendations. As of 12 April 2011, we have collected over 19,000 user logs (a 70% increase from last year) from roughly 360 distinct users, including users from Germany, Italy, France, New Zealand, and Brazil.

Our genSpace plug-in was first included in the January 2009 release of geWorkbench. For the May 2011 release, we have developed new social networking features for genSpace to further enhance knowledge sharing and collaboration amongst researchers. genSpace now allows users to select "friends" (similar to Facebook) and to create and join "networks" (for example, an "Investigator Lab" network, or a "MAGNet" network). Users can chat and share workflows with their friends, directly within geWorkbench. We also have a new "workflow repository" feature, which allows users to save and share workflows for future use. The new data collected offer many opportunities for future research. For example – if we know that a user is affiliated with other users in a lab (i.e., they are in the same network), how can we improve workflow suggestions?

Further information is available at http://www.psl.cs.columbia.edu/genspace

IMPROVING THE PERFORMANCE OF THE GENSPACE RECOMMENDER

GAIL KAISER LAB

geWorkbench and genSpace have over 360 distinct users and we have collected over 19,000 user logs. A user log is generated every time a user runs an analysis tool in geWorkbench. The number of users and user logs have been growing steadily over the past few years. Since we anticipate a significant increase in usage when geWorkbench soon introduces a Web-based client, we wanted to study how our genSpace recommender system would respond to and/or if it could cope with a large increase in the number of users and user data.

We considered two approaches for implementing our recommender system. One was to directly access a database and generate recommendations on the fly as user requests come in. The alternative approach, which we implemented in genSpace, makes use of server-side caching of recommendations. Our genSpace cache is a "prefetch cache" that prefetches all types of recommendations supported by the system. Every recommendation that we need will be present in the cache, removing the need to access the database for any information. This helps us answer user queries for recommendations with low response time and high throughput.

We carried out an empirical study to evaluate the performance of our caching system. Figure 3.a shows

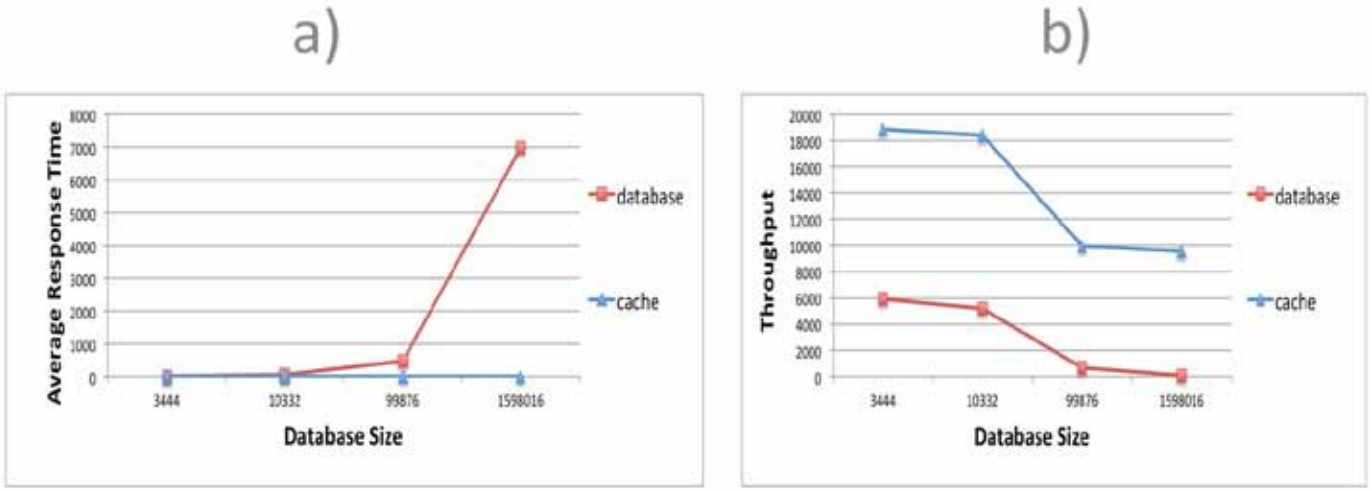


Figure 3: a) Database Size vs. Average Response Time, for "Get Most Popular Workflow Heads", b) Database Size vs. Throughput, for "Get Most Popular Tools"

the plot of the database size (in number of rows) versus the Average Response Time for a recommendation that gets the most popular workflow heads, i.e., tools at the start of a workflow. Figure 3.b shows the plot of the database size (in number of rows) versus the Throughput for a recommendation that gets the most popular tools in the system. The red lines with squares as data points shows the performance when using direct database access and the blue lines with triangles as data points shows the performance when using our cache. From the graphs, we see that the cache consistently outperforms the direct database approach by a factor of at least 3 to as much as 200 as database size increases.

More details can be found in our paper titled "Towards using Cached Data Mining for Large Scale Recommender Systems", which was published at the 2011 International Conference on Data Engineering and Internet Technology (DEIT 2011). The paper can be viewed at http://www.psl.cs.columbia.edu/genspace/genspace-DEIT2011.pdf

FINDING "aQTLS" – THE GENETIC LOCI THAT MODULATE TRANSCRIPTION FACTOR ACTIVITY

HARMEN BUSSEMAKER LAB

Analysis of parallel SNP genotyping and messenger RNA expression profiling datasets – also known as expression quantitative trait locus or "eQTL" data – has shown that mRNA levels are highly heritable in organisms ranging from yeast to human. Only a small fraction of this genetic variance can currently be accounted for in terms of specific molecular network mechanisms. This leaves a major challenge for the field of Systems Biology.

The influence of trans-acting polymorphisms on gene expression traits is often mediated by transcription factors (TFs). This motivated the laboratory of Dr. Harmen Bussemaker to develop a computational method that can use prior knowledge about the in vitro DNA binding specificity of any given TF to map the genomic loci whose genetic inheritance modulates its protein-level regulatory activity. Genome-wide regression of differential mRNA expression on predicted promoter affinity is used to estimate segregant-specific TF activity, which is subsequently mapped as a quantitative phenotype.

In budding yeast, the novel algorithm identifies six times as many locus-TF associations and more than twice as many trans-acting loci as all existing approaches combined. Application to mouse data from an F2 intercross identified an aQTL on chromosome VII modulating the activity of Zscan4 in liver cells. The method is mechanism-based, strictly causal, computationally efficient, and generally applicable. The research was published in Molecular Systems Biology and highlighted in Nature Biotechnology.

REFERENCES

E. Lee and H.J. Bussemaker. (2010) Identifying the genetic determinants of transcription factor activity. Molecular Systems Biology 6:412.

DEMOGRAPHIC INFERENCE FROM HIDDEN RELATEDNESS

ITSIK PE'ER LAB

Distantly related individuals may share long segments co-inherited and therefore Identical-by-Descent (IBD) from their common ancestor. We have previously shown that IBD is ubiquitous in human populations, and provides strong evidence for the recent shared ancestry of individuals and populations. Over the last year we have published application of this methodology to multiple Jewish communities, resolving common genetic origins across Jewish groups (Atzmon et al., AJHG 2010).

We have since developed the mathematical theory behind IBD segments, deriving a closed form equation for the distribution of their lengths for every practical population history. This facilitates reconstruction of recent demographic events such as bottlenecks and expansions. We infer histories for multiple populations and show that East African groups have a unique recent demography: it is best explained by a model of multiple cross-migrating small "villages" (Figure 4).

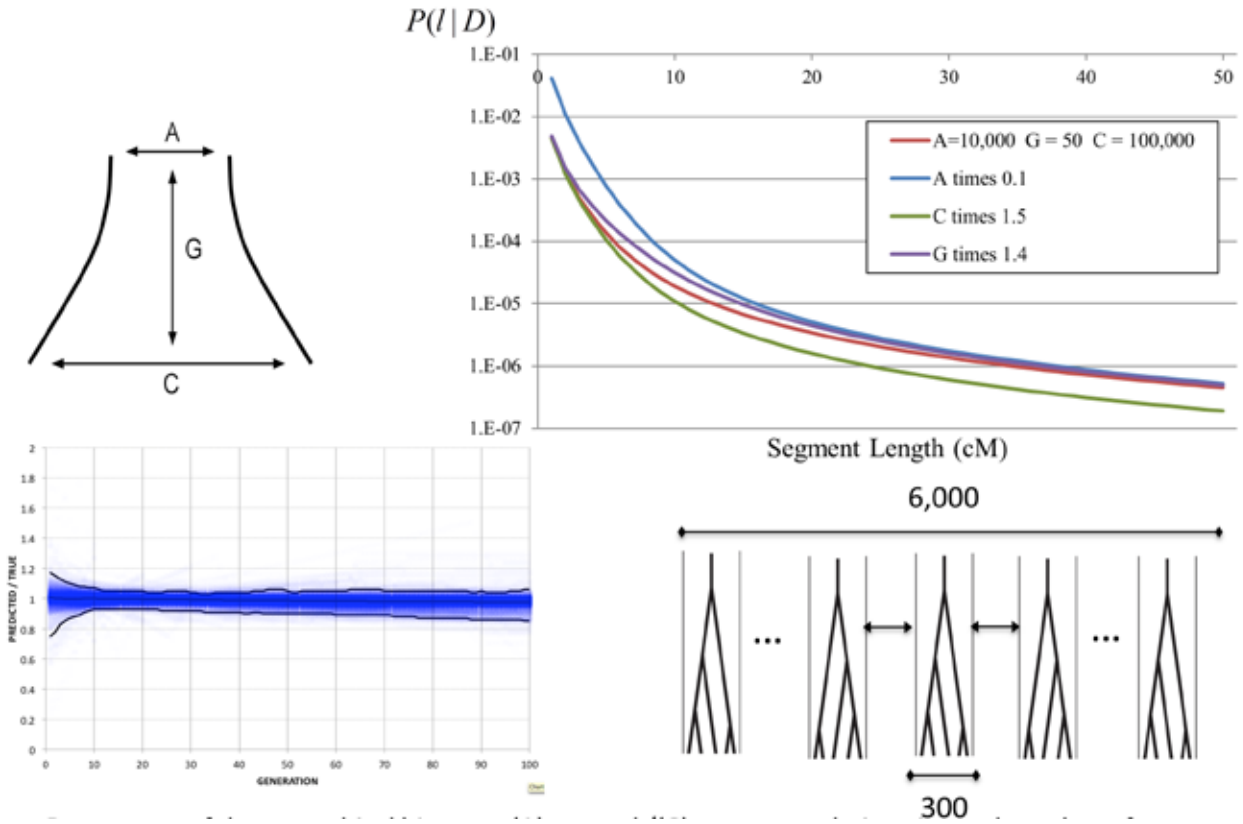


Figure 4: Parameters of demographical history – (A)ncestral /(C)urrent population size and number of (G) enerations (top left) affect distribution of IBD segment lengths (top right), and enables accurate reconstruction of population size (bottom left). The best model for East African groups is of multiple, cross-migrating "villages" (bottom right).

CRACKING THE HOX SPECIFICITY CODE

RICHARD MANN LAB

Transcription factors, proteins that play key roles in deciphering genomic information, are often members of large families that share highly similar DNA binding properties. Yet different members of the same transcription factor family often carry out distinct functions in vivo, suggesting that they somehow achieve specificity in vivo that is not seen in their in vitro DNA binding properties. In collaborative work with the Bussemaker, Honig, and Tullius laboratories, we have been analyzing the degree to which cofactors can resolve this paradox. The protein family we are focusing on is the Hox family of homeodomain transcription factors, which carry out many essential functions during animal development. To address this question, we developed a novel experimental and computational procedure called SELEX-seq, which combines the traditional method of SELEX (Systematic Evolution of Ligands by Exponential Enrichment) with massively parallel sequencing. Using this approach, we have been able to distinguish the DNA binding specificities of all of the Drosophila Hox proteins when they bind with the same cofactor, Extradenticle. In general terms, these studies go a long way towards explaining how members of the same transcription factor family can achieve specificity in vivo.

PROBING THE ELECTOSTATIC POTENTIAL OF DNA USING HYDROXYL RADICAL CLEAVAGE

TOM TULLIUS LAB

A collaboration involving the labs of MAGNet investigators Tom Tullius, Barry Honig, and Richard Mann has shown that the hydroxyl radical cleavage

pattern of DNA yields a map of minor groove electrostatic potential. Why is this important? The new work extends earlier findings from the Honig and Mann laboratories which demonstrated that recognition of minor groove electrostatic potential is an important but previously unappreciated mechanism for specificity in protein binding to DNA. The earlier work depended on analysis of high-resolution X-ray crystal structures of DNA and DNA-protein complexes, which aren't available for genome-scale investigations. The ability to use the results of chemical probe experiments (hydroxyl radical cleavage) to map electrostatic potential will allow this new recognition principle to be applied to entire genomes.

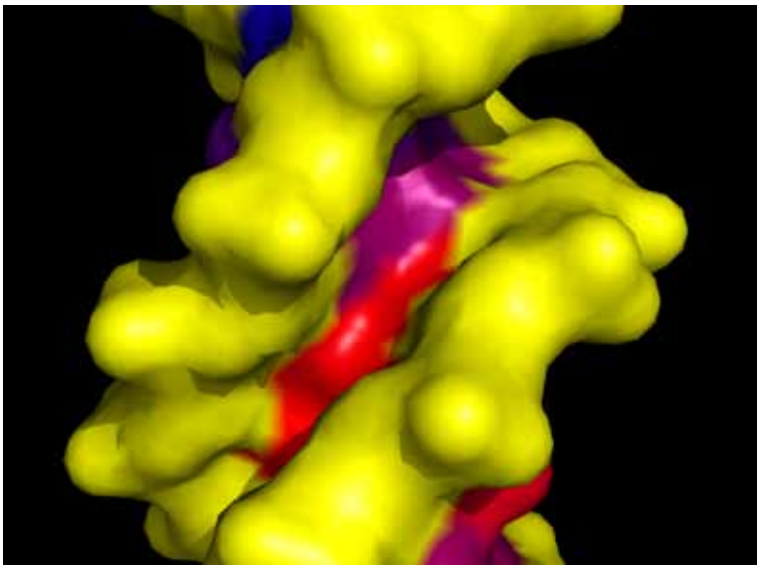
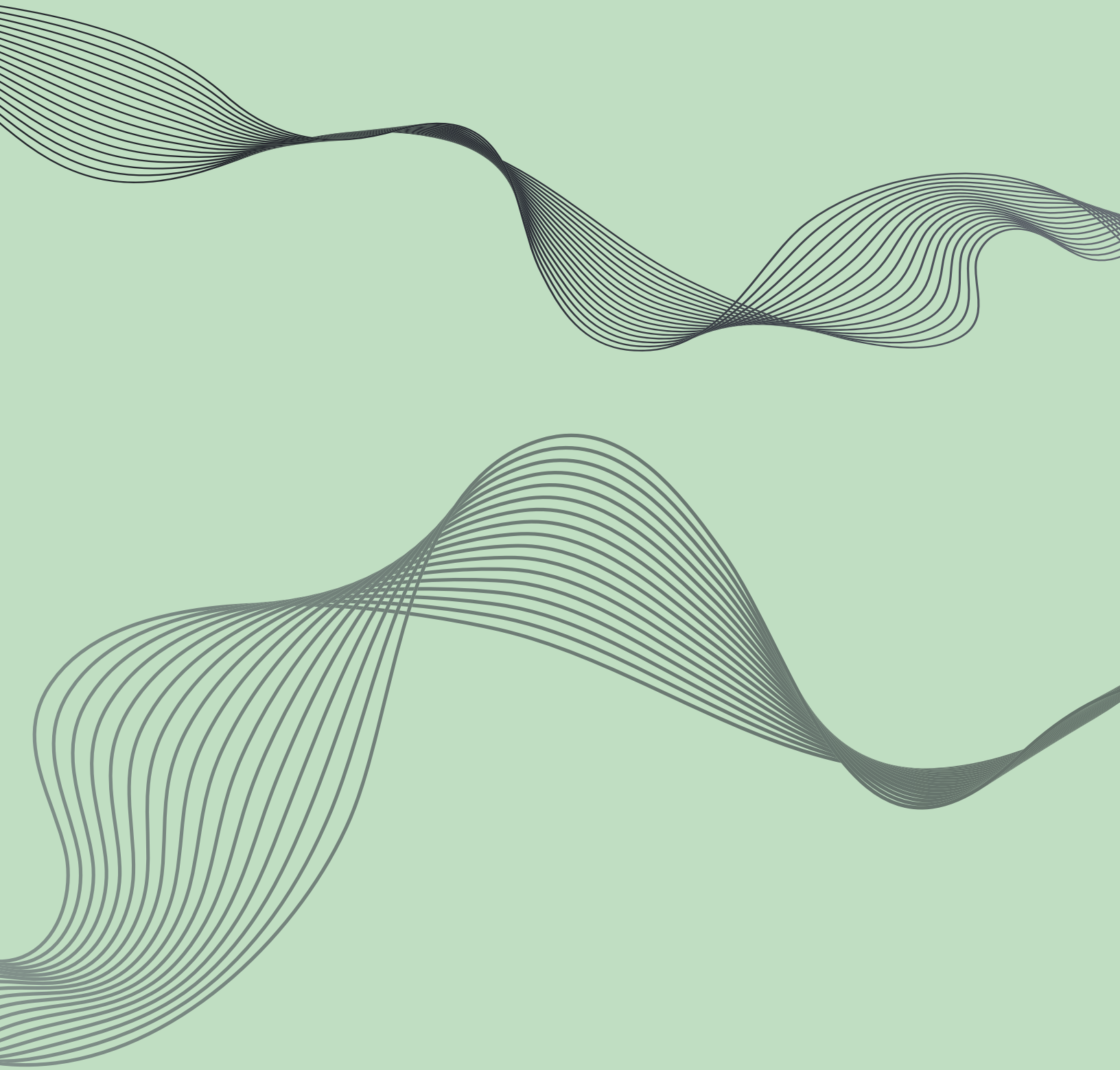


Figure 5: Electrostatic potential varies along the DNA minor groove. Where the groove is wide, the potential is less negative (purple). Where the groove narrows, the potential is more negative (red).



Columbia University
Center for Computational Biology and Bioinformatics
1130 St. Nicholas Avenue
New York, NY, 10032

SPRING 2011
Issue No. 4

COPYRIGHT © 2011 MAGNET - MULTISCALE ANALYSIS OF GENOMIC AND CELLULAR NETWORKS