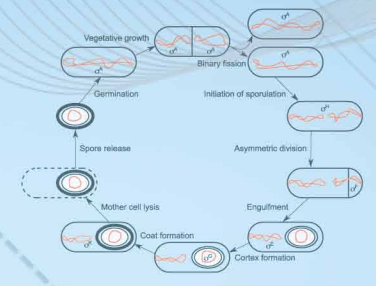
NATIONAL CENTER FOR THE MULTISCALE ANALYSIS OF GENOMIC AND CELLULAR NETWORKS

SPRING 2010

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

ISSUE NO.3

MAGNet NEWSLETTER



STRUCTURAL ORGANIZATION, EVOLUTION,AND HETEROGENEOUS ACTIVATION OF A BACTERIAL SPORULATION NETWORK

DENNIS VITKUP, PHD

TOM TULLIUS, PHD
BREAKING THE CODE OF NON-CODING DNA IN THE HUMAN GENOME

ITSIK PE'ER, PHD
THE FAMILY OF MAN: METHODS AND APPLICATIONS FOR DISCOVERING SHARED ANCESTRY AMONG PURPORTED UNRELATEDS

SPRING 2010
ISSUE NO.3

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

MAGNet NEWSLETTER

FEATURES

SECTIONS

INTRODUCTION

Welcome to the third issue of the MAGNet Center Newsletter and to MAGNet's fifth year of successful operations. The Center was established in 2005 with the mission of creating novel methodologies and tools for the dissection and interrogation of cell-context specific molecular interaction networks, using a combination of structural and systems biology approaches. This year marks an important landmark in the Center's brief history as it is the last year of funding under the original NCBC application and our proposal for the 5-year competitive renewal was just submitted.

The end of this first cycle thus seems like an appropriate moment to reflect upon our accomplishments and challenges. Let us first take a moment to consider our mission. A distinguishing feature of our scientific and technical effort has been the application and integration of diverse principles and analytical "lenses" to the study of molecular interaction networks. These have ranged from the structural and biophysical origins of molecular interactions, to the inference of regulatory networks from molecular profile data, and to the network-mediated effects of genetic and epigenetic variability within populations. Within each layer, our computation-based predictions have undergone rigorous experimental validation and have produced both valuable tools for the community and exciting new discoveries. We are excited about continuing our work by going increasingly between layers, both in terms of integrating distinct modes of regulation as well as dissecting processes that occur at the boundary of distinct cellular populations. In this issue of the magazine we present some exciting new topics that will hopefully provide a valuable, albeit narrow, preview of the research topics and approaches that will be investigated over the next five years.

Our Guest Article is by Dr. Tom Tullius, from Boston University, who is collaborating with MAGNet investigators Drs. Honig, Bussemaker, and Mann to dissect the role of DNA shape in protein binding specificity. Dr. Tullius, in conjunction with MAGNet's own work on the shape of the DNA minor groove, adds an exciting perspective to the long-standing puzzle of how DNA-binding proteins select their correct binding site amongst the 3 billion base pairs of the human genome. Previously, research into this particular question of molecular recognition has been protein-centric, focusing on the complex shapes that proteins adopt upon DNA binding, and defining the DNA consensus sequences that are recognized. With the aid of various high-throughput experimental methods, and in combination with computational approaches, we are now beginning to appreciate how looking at DNA as a molecule, rather than as a string of nucleic acids, is key to elucidating the DNA-protein recognition problem.

Using hydroxyl radical cleavage patterns as a quantitative read-out of the shape of the DNA, Dr. Tullius's group has been able to define how the surface structure of DNA varies throughout the genome. By generating a database of such experimentally defined topographical maps that include crystallographic or NMR structural "gold standard" DNA segments, the group was able to develop an algorithm that predicts a cleavage pattern of any DNA segment of interest. Amazingly, a comparison of regions of DNA sequence conservation with sites experimentally defined to be functionally conserved showed that only half of the areas of sequence conservation were also functionally conserved. Moreover, areas of divergent sequence can take on similar shapes, suggesting that not only DNA sequence but also DNA shape may be subject to evolutionary selection. A computational approach seems to imply just that: comparison of genomic DNA from 36 species against human showed that nearly twice as much of the human genome is conserved in shape as in sequence.

In our first Feature Article, Dr. Dennis Vitkup reflects on the structural and functional organization of bacterial sporulation networks, both through evolutionary analysis and stochastic simulations. By looking at the evolutionary conservation of genes from 24 endospore-forming bacteria and grouping the genes into functional modules for the sporulation program, the Vitkup lab was able to draw several conclusions. The evolution of the bacterial sporulation network is not random, but follows the logic of the functional and structural hierarchy of the network itself. Specifically, the most conserved components are those highest in the hierarchy, and include the regulatory modules regulated by the sigma factors, along with the sequential activation of these modules and the signaling interactions between them. It is also noteworthy that the conservation of regulatory interactions is even greater than gene content conservation. Stochastic simulations gave insight into the asynchronous entry into sporulation that is characteristic of endospore forming species. Model predictions required key protein components of the signaling pathway to be up-regulated post-transcriptionally in order for sporulation to proceed, and additionally predicted that this was rate limited by the phosphate flux in the environment. Both of these predictions were substantiated by experimental work in other laboratories. Lastly, the group has begun to define the interactions between the sporulation regulatory network and computationally predicted metabolic pathways. The *yngJIH-GFE* gene cluster in *B. subtilis* was computationally inferred to be involved in the leucine or fatty acid degradation and labeling experiments showed that this pathway was indeed utilized by

sporulating cells and not vegetatively growing cells, nor *yng* mutant cells.

Finally, in our second Feature Article, Dr. Itsik Pe'er's group is developing new methods to characterize the presence and length of shared genetic sequences, called "identical by descent" (IBD) segments, to study population history and search for disease-related genes. Using the program GERMLINE to detect common IBD segments, the group discovered a high degree of relatedness within a number of separate Jewish communities, and to a lesser extent between these communities. The technique also detected the signature of a known late-medieval bottleneck in the Ashkenazi Jewish population in Europe. The physical distribution of IBD regions on chromosomes was examined, and it was found that certain regions tend to display frequent IBD, while most of the genome does not. Such IBD regions were found to include loci such as HLA and long polymorphic inversions. The group went on to show how IBD SNP data can be used to impute the genotypes of hidden (untyped), causative alleles, using a form of triangulation. A second method to the same end detected clusters within the haplotype structure to pinpoint causative loci. Finally, the use of imputation in the context of personal genomes was examined, a technique that will become of great interest as growing numbers of individuals obtain full genome sequences. In particular, the combination of IBD analysis with whole genome sequencing was shown to be able to pinpoint likely causal mutations.

- Andrea Califano, Ph.D.

BREAKING THE CODE OF NON-CODING DNA IN THE HUMAN GENOME

TOM TULLIUS, PHD
Department of Chemistry and Program in Bioinformatics, Boston University

When you think about DNA, what image comes to mind? These days, you would probably first imagine a (very!) long string of letters. After the sequence of the human genome was determined a decade ago, it seemed like the best way to understand how it works would be to search its 3 billion letters for meaning. The abstraction of representing the DNA molecule by a string of letters is effective when we want to find the parts of the genome that code for proteins, because of the strict, universal three-letter genetic code that specifies each of the 22 amino acids that make up proteins. But when the human genome was sequenced, we found that the easy-to-crack genetic code only makes up around 2% of the genome. What is the other 98% doing? And how does it do it?

One thing that the rest of the genome does is to encode binding sites for the proteins that regulate how the genome works. A major question is how these proteins find the right place to bind among the 3 billion possible places. What these proteins don't do is "read" the letters of the genome. Because proteins and DNA are molecules, the principles of molecular recognition must be the key to directing a protein to its proper binding site.

How does molecular recognition work for proteins searching for their DNA binding sites? Proteins fold into complex three-dimensional shapes that present a surface that can be complementary in some way to the surface of DNA. But while we appreciate the complex forms that proteins can adopt, because of the many examples that are available from structural biology investigations, the other half of the recognition problem, the shape of DNA, is usually neglected, or at best simplified.

If we do go beyond thinking of DNA as a string of letters, our next image is likely that of the iconic double helix. It is unfortunate, though, that most stop at the Watson-Crick depiction of a perfectly uniform helix as always representing the shape of the DNA molecule. Over the past 30 years, structural biologists (crystallographers and NMR spectroscopists) have found that the DNA double helix varies in structure depending on its nucleotide sequence. These subtle structural variations are clear in high-resolution X-ray or NMR structures, and offer insight into how a protein might distinguish one nucleotide sequence from another. So, if we could just determine how DNA structure varies throughout the human genome, both halves of the DNA-protein recognition problem would be in hand.

Determining the structure of the whole human genome is not a job for crystallography or NMR, though – it's too big. My lab has worked for the past 25 years (Tullius 1987) on developing and then using a method for DNA structure determination that can scale to the whole genome. We use the chemistry of one of the most reactive free radicals, the hydroxyl radical (•OH), to make a chemical image of the variation in shape of the DNA double helix.

The hydroxyl radical produces this image by making breaks in the DNA strand. We discovered that how often a break is made at a particular nucleotide translates directly into the shape of the DNA backbone and grooves at that place in the DNA molecule. Because this method can be used on the same short segments of DNA that are studied by crystallography and NMR, the structural information that is obtained from the hydroxyl radical cleavage pattern can be calibrated based on "gold standard" high-resolution structures. From these experiments we find that the hydroxyl radical pattern produces a topographical map of the shape of the surface of DNA.

We can perform the hydroxyl radical cleavage experiment on much larger DNA molecules, too, and map their shape

variation. But the human genome is still too big for us to study its structure directly by this experiment (although we are currently developing a new version of the experiment that takes advantage of ultra-rapid DNA sequencing technology, which may well make it possible to experimentally determine the structure of DNA throughout an entire genome). Nonetheless, we recently found a way to combine experiment with computation to extend the reach of the hydroxyl radical cleavage experiment, so that it can be used to make a structural map of the human genome.

We did this by first making a database of experimental hydroxyl radical cleavage patterns from a large number of different DNA sequences, which we called ORChID, the OH Radical Cleavage Intensity Database (Greenbaum et al. 2007). We then devised a computational algorithm that uses these patterns to predict the cleavage pattern of any DNA sequence of interest. The algorithm is fast and efficient enough to compute the cleavage pattern of the human genome in only a few minutes. Our structural map of the human genome is available in the UC Santa Cruz human genome browser (http://genome.ucsc.edu/ENCODE/) as one of the ENCODE Project datasets (ENCODE Consortium 2007). The ORChID map can be used by anyone as a dataset for investigating how DNA shape contributes to the function of the human genome.

One thing that my group has done with ORChID is to test a revolutionary idea concerning how evolution operates on the human genome (Parker et al. 2009). Evolutionary conservation of DNA nucleotide sequence in the human genome has proven to be a powerful way to find the segments of the genome that contribute to biological function. After all, the argument goes, why would a particular segment of the genome be conserved in sequence among a wide variety of organisms unless it was involved in an important biological function? A problem arose when the ENCODE Project compared sites of evolutionary conservation of sequence in the human genome with sites in the genome that were found by experiment to be performing some biological function (ENCODE Consortium 2007). The result of this comparison was surprising: evolutionary conservation of sequence was found in only about half of the functional regions of the human genome.

Because the shape of DNA has been found by MAGNet investigators to be important for recognition by DNA binding proteins (Joshi et al. 2007; Rohs et al. 2009), my group wondered whether DNA shape might be under evolutionary selection. An important result of our initial analysis of ORChID was that different DNA sequences could adopt very similar shapes (Greenbaum et al. 2007). This would mean that a segment of the human genome could have diverged in sequence from that of other organisms, but maintained the same structure. If that had occurred, and if shape but not sequence was the key to biological function of that site, searching only for sites of nucleotide sequence conservation would miss such a functional site. So, we developed a computational method to compare the shape of the DNA throughout the human genome with the shape of the genomic DNA from 36 different species. We found that nearly twice as much of the human genome is conserved in shape as is conserved in sequence. We also discovered that many of the functional regions that were found by the ENCODE Project not to overlap with sequence-conserved regions did overlap with structure-conserved regions.

We think that our investigation of how DNA shape varies throughout the human genome is uncovering a new "grammar" by which the genome works. We are looking forward to combining our efforts with similar efforts by MAGNet investigators to decode the non-coding 98% of the human genome. Since much of the disease-associated variation in DNA sequence occurs in the non-coding parts of the human genome, this research will contribute to uncovering the link between genome and disease.

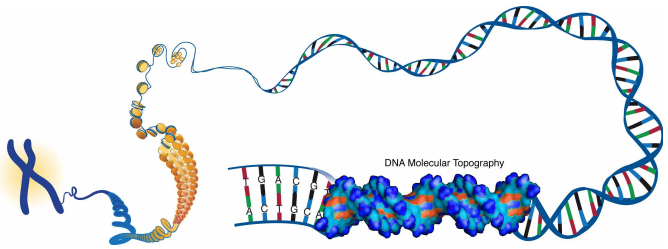


Figure 1. DNA topography and the human genome. DNA is shown streaming out from a chromosome, progressively unfolding as chromatin, through the 30 nm filament, nucleosomes, the DNA double helix, and finally to the letters representing the nucleotide sequence. Work in the Tullius lab aims at finding out how the varying topography of the DNA double helix contributes to the function of the human genome. (Image courtesy of Darryl Leja, NHGRI, NIH).

REFERENCES:

ENCODE Consortium 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.

Greenbaum, J. A., Pang, B., and Tullius, T. 2007. Construction of a genome-scale structural map at single-nucleotide resolution. Genome Res 17: 947-953.

Joshi, R., Passner, J. M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M. A., Jacob, V., Aggarwal, A. K., Honig, B., and Mann, R. S. 2007. Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell 131: 530-543.

Parker, S. C. J., Hansen, L., Abaan, H. O., Tullius, T. D., and Margulies, E. H. 2009. Local DNA topography correlates with functional noncoding regions of the human genome. Science 324: 389-392.

Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B. 2009. The role of DNA shape in protein-DNA recognition. Nature 461: 1248-1253.

Tullius, T. 1987. Chemical "snapshots" of DNA: Using the hydroxyl radical to study the structure of DNA and DNA-protein complexes. Trends Biochem Sci 12: 297–300.

STRUCTURAL ORGANIZATION, EVOLUTION AND HETEROGENEOUS ACTIVATION OF A BACTERIAL SPORULATION NETWORK

DENNIS VITKUP, PHD

Center for Computational Biology and Bioinformatics, Columbia University

Bacterial sporulation provides a fascinating and unique molecular system for studying the evolutionary basis of a complex, yet tractable gene regulatory network. Although the general framework of the bacterial sporulation network has been firmly established, many conceptual questions about the functional organization and evolution of the network have yet to be answered. Here we highlight recent work in our laboratory investigating several important open topics. First, by comparing the conservation of genes versus the conservation of their regulatory interactions, we examine how the evolution of the sporulation network correlates with its functional and hierarchical logic. Next, we use simulation studies to investigate how a population of bacteria relies on stochastic processes to assure that entry into the sporulation state is asynchronous, allowing the population to continually sample and respond to its environment. Finally, we describe methods which identify and use interactions between regulatory and metabolic networks to predict a complete metabolic pathway potentially important in sporulation. In describing each of these endeavors, we emphasize the effective integration of experimental and computational approaches.

Overview of Bacterial Sporulation

Although prokaryotic organisms do not have a complex body plan, they can form multi-cellular structures, such as biofilms and fruiting bodies (Vlamakis et al. 2008). In addition, elaborate developmental processes have been characterized in many bacterial species. Endospore formation (sporulation) is the prime example of such a developmental process. The sporulation process has been characterized in significant detail in the model gram-positive bacterium *Bacillus subtilis*. In rich medium, *B. subtilis* cells divide by binary fission approximately every 30 minutes. By contrast, deterioration of environmental conditions triggers sporulation which takes about 8 to 10 hours to complete. Endospore formation represents a formidable investment of time and energy for the bacterium and is considered to be a survival pathway of last resort (Piggot and Losick 2002, Gonzalez-Pastor et al. 2003).

The successive morphological stages of *B. subtilis* sporulation (Figure 1) have been defined using electron microscopy. (Piggot and Coote 1976). Sporulation begins with an asymmetric cell division and results in the generation of two cell types, a forespore (the smaller compartment) and a mother cell. Shortly after asymmetric division, two parallel programs of gene expression are established in each compartment under the control of transcription factors that are activated in a cell-specific manner. Sporulation commences only after a round of DNA replication is completed to ensure that two chromosome copies are available in the predivisional cell. Following asymmetric cell division, the next morphological stage of sporulation is the engulfment of the forespore by the mother cell. This process is analogous to phagocytosis. As a result of the engulfment the forespore becomes entirely surrounded by the mother cell. Protective coat proteins are synthesized in the mother cell and assembled on the surface of the developing spore. In the final stage of sporulation, the mother cell lyses and the mature spore is released.

Fully formed bacterial spores, recognized as the most resistant form of life on the planet (Nicholson et al. 2000), protect the bacterial genome against heat, desiccation, radiation, and oxidation. In addition, spore formation might be an efficient way to escape predation from higher organisms. As soon as environmental conditions become favorable for vegetative growth, B. subtilis quickly exits from the dormant state. This process is referred to as spore germination (Setlow 2003). Germination is triggered by the presence of nutrients in the environment. The nutrients are sensed by specific spore membrane receptors, and within minutes, the spore core rehydrates and the coat is shed. Ultimately, DNA replication is initiated and the first cell division soon follows.

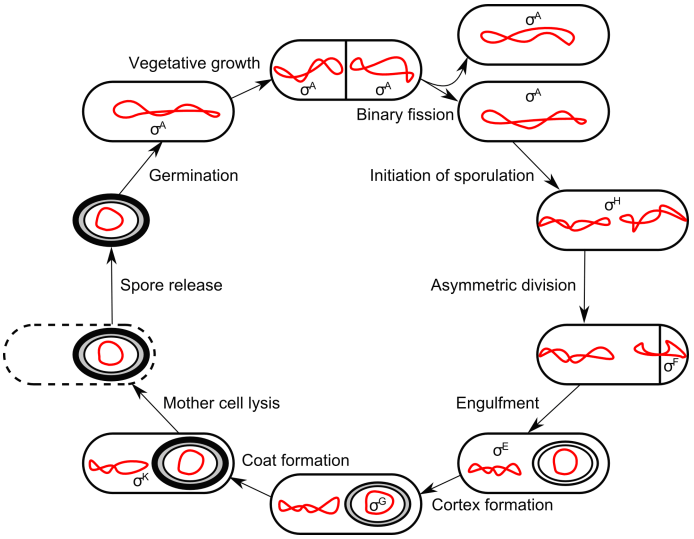


Figure 1. Main morphological stages of the B. subtilis life cycle. The temporal and compartment-specific activity of each sporulation sigma (σ) factor is indicated. Sporulation is initiated in response to starvation. During asymmetric division, two membrane-bounded compartments are generated: a small forespore and a large mother cell. After asymmetric division, the remainder of the forespore chromosome (i.e. the origin-distal region) is pulled into the forespore by translocation. Engulfment of the forespore by the mother cell results in the release of the forespore as a free protoplast in the mother cell. The protective coat (black) is a complex structure that assembles around the forespore surface. Following mother cell lysis, the mature spore is released in the environment. Spores will germinate in response to improvements in environmental conditions and quickly resume vegetative growth.

Sporulation Regulatory Network

Initiation and progression of sporulation is controlled by a complex network of protein-protein and protein-DNA interactions, consisting of regulatory modules, feed-forward network motifs, and post-translational regulation (Errington 2003, Piggot and Losick 2002). The formation of a bacterial spore requires the interrelated and coordinated functioning of protein signaling, regulatory, and metabolic networks. The general structure of the regulatory network controlling *B. subtilis* sporulation is shown in Figure 2. The sporulation program is organized spatially and temporally into several modules, with the sigma factors σ^E, σ^K, σ^F, and σ^G directing gene expression in the early mother cell, the late mother cell, the early forespore, and the late forespore, respectively. In addition, the sigma factor σ^H regulates transcription of genes involved in the initial (predivisional) stage of sporulation. The overall structure of the regulatory network creates two parallel pathways of gene expression (one in the forespore and one in the mother cell) in which one regulatory protein switches the next protein in the cascade. These two parallel pathways are also effectively synchronized through three pathways of intercellular communication.

Importantly, entry into sporulation is controlled by the transcription factor Spo0A. Before sporulation, the gene *spo0A* is transcribed at a low level from a house-keeping sigma factor σ^A promoter. Sporulation is initiated when Spo0A is activated by the phosphorelay system which transfers information about environmental conditions to the sporulation network. The phosphorylated form of Spo0A initiates the transcription of various sporulation genes which respond to different levels of the phosphoprotein (Molle et al. 2003).

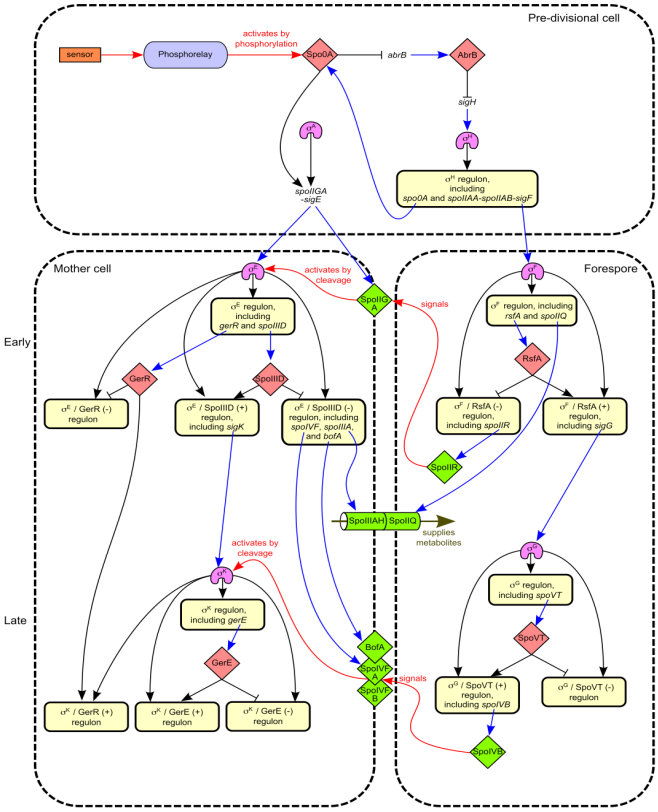


Figure 2. Modular architecture of the sporulation regulatory network in B. subtilis. The temporal progression of sporulation is from the top to bottom. Each cellular compartment (predivisional cell, forespore and mother cell) is surrounded by dotted lines. Sigma factors are shown in purple, transcription factors in green, and regulons in yellow. Signaling proteins are shown in red. Transcriptional regulation is indicated by black arrows, gene expression (protein synthesis) by blue arrows, and signaling pathways by red arrows. Coherent feed-forward loops are indicated by a plus sign, and incoherent feed-forward loops by a minus sign.

A population of the sporulating cells is far from being homogeneous. Often, sporulating bacteria form subpopulations of cells in alternative states with different programs of gene expression. For example, only a subpopulation of nutrient-limited *B. subtilis* cells activates Spo0A during sporulation (Dubnau and Losick 2006)

Endospore-forming bacteria belong to two broad classes: the Bacilli (aerobic Firmicutes, Figure 3a, yellow) and the Clostridia (anaerobic Firmicutes, Figure 3a, cyan). Both classes can be further subdivided into a number of smaller orders. It is generally assumed that the common ancestor of all Bacilli and Clostridia was an endospore-forming organism, even though several genera that presumably evolved from that common ancestor have lost the ability to sporulate. Endospore formation is an ancient process that appeared only once in the course of evolution and likely predated the rise in oxygen in the terrestrial atmosphere about 2.3 billion years ago. Furthermore, given the large number of genes that are essential to endospore formation, it is unlikely that the ability to sporulate could have been gained by other phyla as a result of horizontal gene transfer.

The analyses of evolutionary patterns performed recently in our laboratory allow us to draw several important conclusions about the evolution of the bacterial sporulation network and its regulation. Previously, several eukaryotic developmental systems were considered in detail, notably the sea urchin developmental network (Davidson 2009). In developmental regulatory networks of higher organisms, conservation of regulatory interactions reflects the hierarchy inherent to the formation of a new body part. In the early stage, the domain that will develop into a body part is specified, followed by the middle stage in which the morphology of the body part is determined, and the late stage, specifying the details of the body part. As each stage in this hierarchy builds on the previous stage, gene regulatory interactions in the earlier stages of development have more widespread consequences than those in later stages, and therefore tend to be more evolutionarily conserved (Davidson and Erwin 2006).

Although spore-forming bacteria do not have an elaborate body plan, evolutionary patterns similar to the ones observed in higher organisms are also present in the sporulation network (de Hoon et al. 2010). Importantly, the observed conservation patterns clearly demonstrate that the evolution of the sporulation network is not random, but, to a large extent, follows the functional and hierarchical logic of the sporulation network (Figure 3b). Specifically, the regulatory modules governed by each sigma factor (Figure 3b, red) are conserved in all spore-forming bacteria.

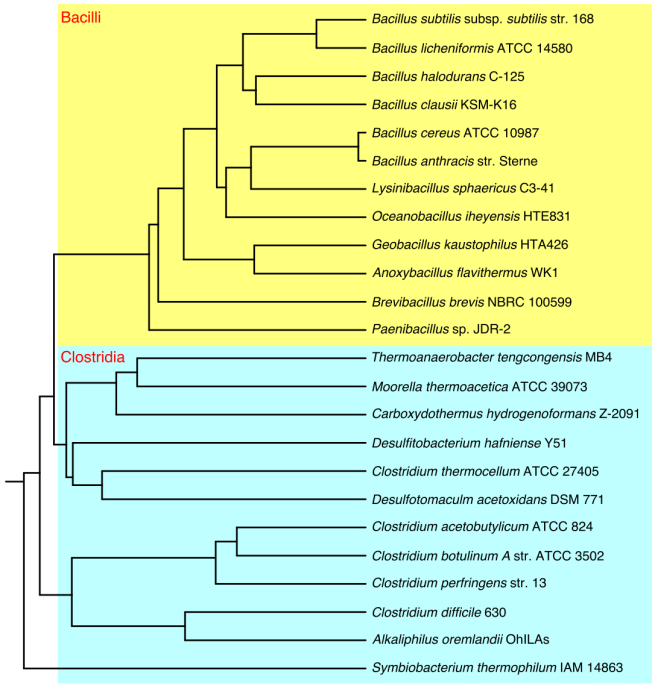


Figure 3a. Phylogenetic tree of 24 representative endospore-forming bacteria. *B. subtilis* belongs to the yellow cluster of 12 aerobic bacteria from the class Bacilli. The other cluster (cyan) includes 12 species from the anaerobic class Clostridia.

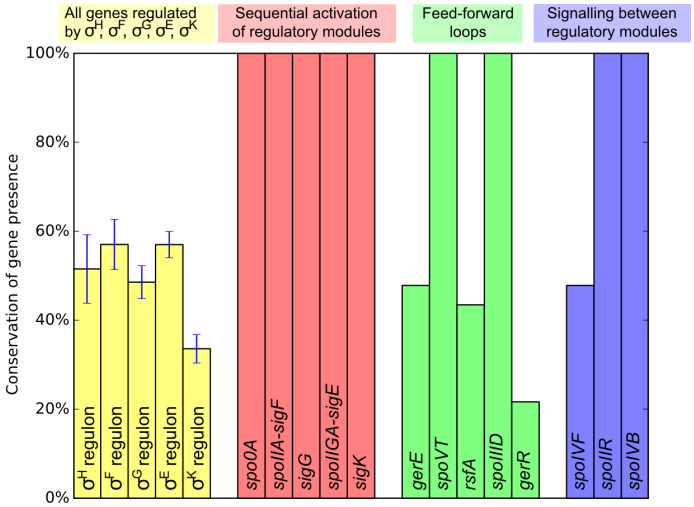


Figure 3b. Conservation of gene presence in the sporulation regulatory network for the 24 representative endospore-formers. Conservation is shown in percentages compared to *B. subtilis*. The figure demonstrates that the observed conservation patterns follow the functional and structural hierarchy of the sporulation network.

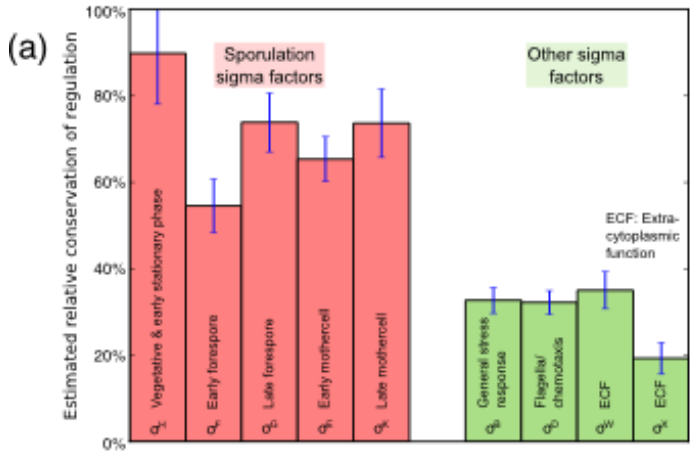


Figure 4a. Estimated conservation of regulation for *B. subtilis* sigma factors. The conservation, given the presence of the regulated gene, is shown for the 305 genes regulated by the sporulation sigma factors, and for the 211 genes regulated by the non-sporulation sigma factors.

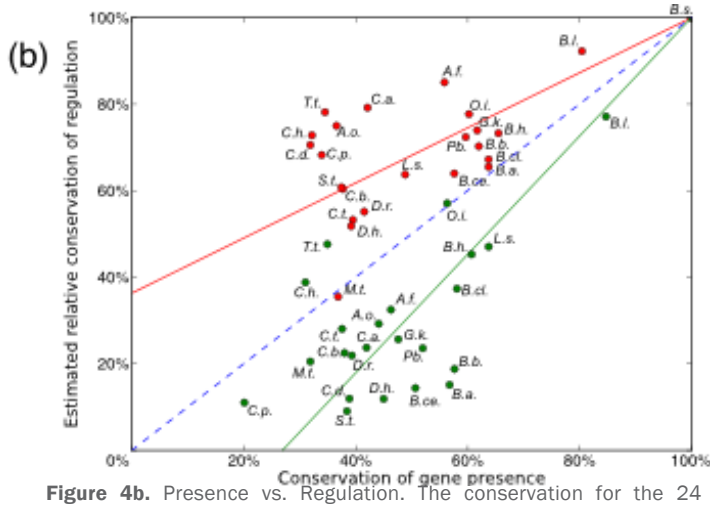


Figure 4b. Presence vs. Regulation. The conservation for the 24 representative species is shown in red for regulation by the sporulation sigma factors and in green for regulation by the non-sporulation sigma factors.

The sequential activation of these global regulatory modules is also strongly conserved, as well as signaling interactions between the modules (Figure 3b, blue). The feed-forward motifs show an intermediate level of conservation (Figure 3b, green), i.e. they are less conserved than the sigma factors but significantly more than the other sporulation genes (Figure 3b, yellow), which occupy the lowest level in the functional and evolutionary hierarchy.

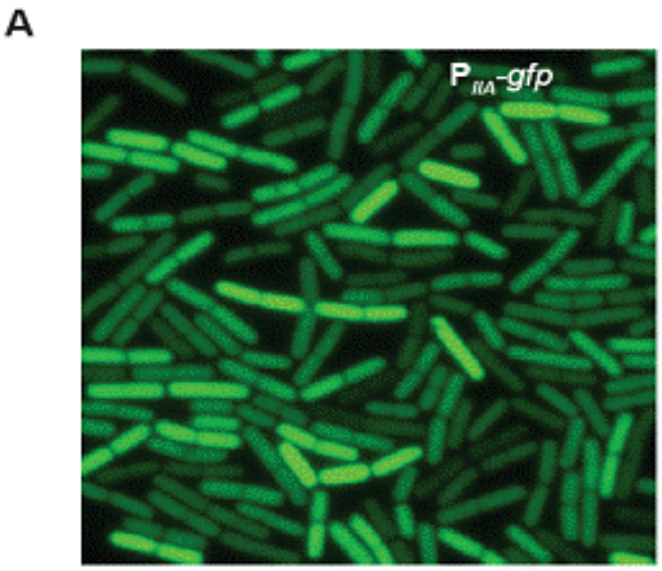
In Figure 4 (a) we contrast the evolutionary conservation of regulatory interactions for sporulation (σ^F, σ^G, σ^E, σ^K, σ^H) and non-sporulation (σ^B, σ^D, σ^W, σ^X) sigma factors. This comparison demonstrates a significantly (2-3 times) higher conservation of regulatory relationships involved in sporulation. It is also interesting to compare the conservation of gene presence with conservation of regulatory interactions (Figure 4b). In eukaryotic organisms, especially at short evolutionary distances, regulatory changes usually serve as the main driving force behind evolution and functional adaptation (Prud'homme et al. 2007, Carroll 2008). In bacteria, gene regulatory networks have also been found to be extremely flexible, with only a small fraction of regulatory interactions conserved across diverged species (Lozada-Chavez et al. 2006). In contrast, while the estimated conservation of regulation by sporulation sigma factors is generally proportional to the conservation in gene content, regulatory relationships involved in sporulation are significantly more conserved (Figure 4b, red). On average, for a bacterium with 50% gene content conservation, about 70% of regulatory interactions involved in sporulation are conserved.

The observed pattern, indicating relatively faster changes in gene content, is likely to be a consequence of the physiological importance of transcriptional regulation in sporulation. In contrast, if we consider the changes in regulatory relationships for sigma factors not involved in sporulation, a more conventional picture emerges (Figure 4b, green) in which regulatory changes are diverging faster compared with changes in gene content.

Stochastic heterogeneous activation of sporulation network

A striking feature of sporulation is its dichotomy: even under optimal conditions only a portion of the population forms spores (Dubnau and Losick 2006). Sporulation is initiated by changes in environmental conditions which are sensed by a group of histidine kinases capable of auto-phosphorylation. The phosphate group is then sequentially transferred from the histidine kinases to the Spo0B and Spo0F proteins of the phosphorelay and finally to Spo0A (Burbulys et al. 1991). The decision of whether to sporulate is dictated by the accumulation of Spo0A~P to high levels. Thus, a tempting hypothesis to explain the bimodality of the sporulation process is that the distribution of the master regulator Spo0A~P is itself broadly heterogeneous.

In order to gain better understanding of the heterogeneous Spo0A activation we took advantage of the available FACS (Fluorescence Activated Cell Sorting) data (Figure 5a,b) to build a computational model for the Spo0A regulatory network (Chastanet et al. 2010). Modeling was performed using, separately, ordinary differential equations and stochastic equations based on Gillespie's algorithm. To build the model we used equations describing transcription, translation, and protein and RNA degradation for important components of the Spo0A phosphorelay. When biochemical parameters were unknown, we relied on previously reported generic parameters for similar systems.



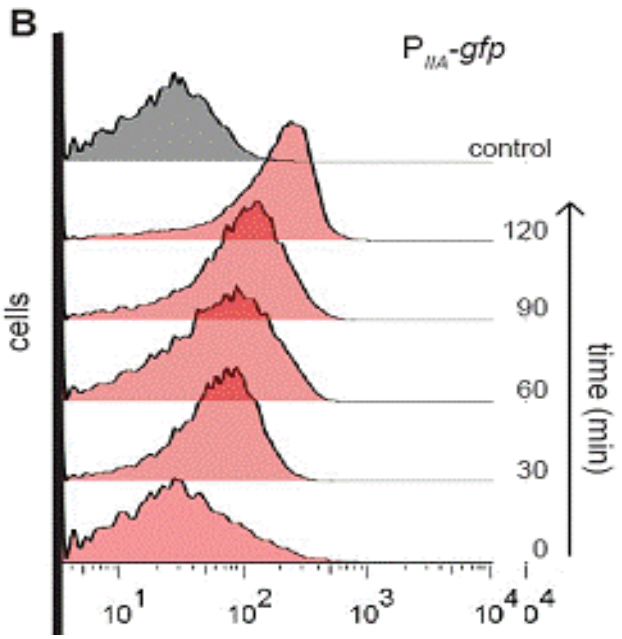


Figure 5a. Visualization of Spo0A~P activation in a population of sporulating cells by fluorescence microscopy. The data was taken 1 hour after induction of sporulation. Variations in color indicate heterogeneity in the Spo0A~P activity in different B. subtilis cells.

Figure 5b. Quantitative FACS (Fluorescence Activated Cell Sorting) analysis of sporulating cells. The data were collected at the indicated times after induction of sporulation. The distributions indicate heterogeneity of Spo0A~P activity across the B. subtilis population.

Two different computational models were able to fit the experimentally measured time-courses of Spo0A~P accumulation reasonably well (Figure 6a,b). The first model (Figure 6a) assumed a linear increase in the concentration of phosphorelay proteins, while the second (Figure 6b) assumed a positive feedback from the Spo0A~P to the phosphorelay proteins.

A striking conclusion from the modeling simulations was a requirement to significantly increase the amount of relay proteins at some time point after the beginning of sporulation. A slightly better fit for the rise in Spo0A~P levels was achieved when we assumed that the rates of accumulation of other relay proteins accelerate due to positive feedback loops. No such feedback loops, operating at the transcriptional level, are known to be active in the network. We therefore speculated that if our model is correct, then an unknown mechanism must exist to stimulate the accumulation of relay proteins at a post-transcriptional level. Gratifyingly, and after the simulations had been finished, we learned that Eswaramoorthy et al., have indeed observed that the levels of all four relay proteins indeed rise as Spo0A is activated (M. Fujita, personal communication). Interestingly, the experiments also demonstrated that the increase in the concentration of relay protein was achieved due to post-transcriptional regulation, exactly as predicted by the computational model.

Overall, our stochastic simulations suggest that one of the major functions of the phosphorelay, in addition to integrating environmental signals, is to create asynchrony in the time of entry into sporulation. Sporulation requires a high threshold concentration of Spo0A~P and only cells attaining this threshold are able to proceed through morphogenesis; not all cells in the population reach this threshold and succeed in forming spores. Indeed, we experimentally confirmed that artificially increasing the phosphate flux in the relay decreases asynchrony and concomitantly increases the proportion of spore-forming cells. Thus, asynchrony helps to ensure that all cells do not commit to spore formation at once, an advantageous strategy given the vicissitudes of the environment.

Interestingly, the experimental data and computational analysis also suggest that, in contrast to other systems that generate cell population heterogeneity, such as genetic competence, the rate-limiting factor in the Spo0A network is not the transcription factor concentration but the phosphate flux. This ensures that en route to committing to sporulation *B. subtilis* constantly monitors environmental conditions through the magnitude of this flux.

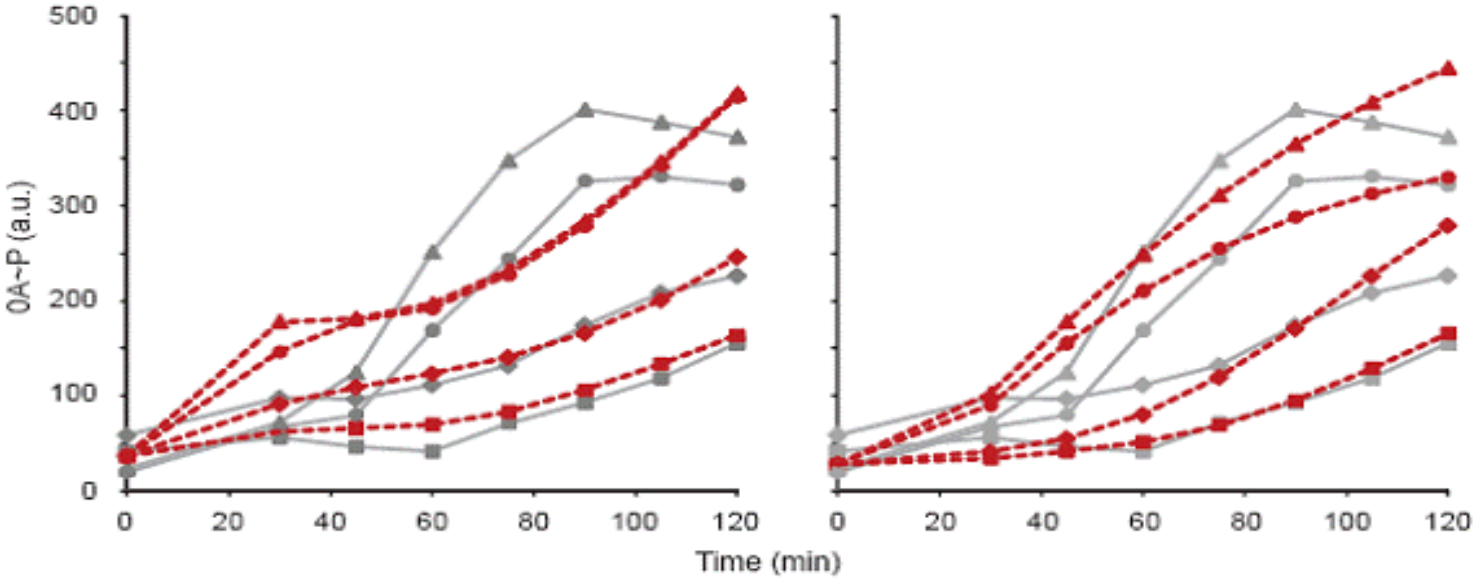


Figure 6a. Spo0A~P accumulation from computational Model 1 fitted to the mean fluorescence experimental data. Grey curves are experimental data and dotted red curves are computational predictions. Model 1 assumes a linear increase of all relay proteins.

Figure 6b. Spo0A~P accumulation from computational Model 2 fitted to the mean fluorescence experimental data. Grey curves are experimental data and dotted red curves are computational predictions. Model 2 assumes that positive feedback loops, driven by Spo0A~P, regulate the increase of relay proteins.

Interaction between the regulatory sporulation network and bacterial metabolism

One of the remaining questions of B. subtilis sporulation concerns the interactions between the regulatory network and bacterial metabolism. The computational methods developed in our laboratory allow us to accurately predict enzymatic activities currently missing from microbial metabolism. Applying these methods to metabolism of B. subtilis allowed us to predict and validate an entire metabolic pathway with a potentially important function in sporulation (Hsiao et al. 2010).

The predicted pathway is formed by the genes from the yngJIHGFE cluster in *B. subtilis* (Figure 7a). Our algorithm suggested that the yng cluster forms a complete degradation pathway from leucine or fatty acids to acetoacetyl-CoA, which can be further catabolized through the bacterial TCA cycle.

What is the biological role of the leucine degradation pathway in B. subtilis? The yng genes are under transcriptional control of the σ^E sigma factor and are primarily expressed early in the mother cell during sporulation, i.e. when extracellular nutrients are limited. Due to the structure of its TCA cycle, B. subtilis cannot grow on leucine as the sole carbon source (Sonenshein et al. 2001).

Nevertheless, the catabolism of leucine and fatty acids through the TCA cycle can provide additional energy during early sporulation stages. The selection of the energy source becomes logical if one considers that leucine is one of the most abundant amino acid in logarithmically growing B. subtilis cells (Sauer et al. 1996), responsible for about 8-10% of all protein residues. In addition, B. subtilis lipids are predominantly (>90%) composed of branched chain fatty acids (Kaneda 1991; Sonenshein et al. 2001); odd-iso fatty acids can be oxidized to 3-methylbutanoyl-CoA. It is likely that during sporulation branched-chain fatty acids and amino acids are present in the extracellular media due to the bacterial cannibalism process (Gonzalez-Pastor et al. 2003), which allows a fraction of B. subtilis cells to kill their non-sporulating siblings and feed on the released nutrients.

To experimentally validate the role of the yng cluster during sporulation we used 13C labeling experiments. First, we analyzed B. subtilis cells in non-sporulating minimal medium supplemented with 13C leucine. Because the degradation pathway leads from leucine to acetyl-CoA (Figure 7a), we measured the fractional labeling of the acetyl-CoA. No 13C labeling above the natural abundance was detected in cells during vegetative growth. This result confirmed that the predicted degradation pathway is not active during favorable environmental conditions.

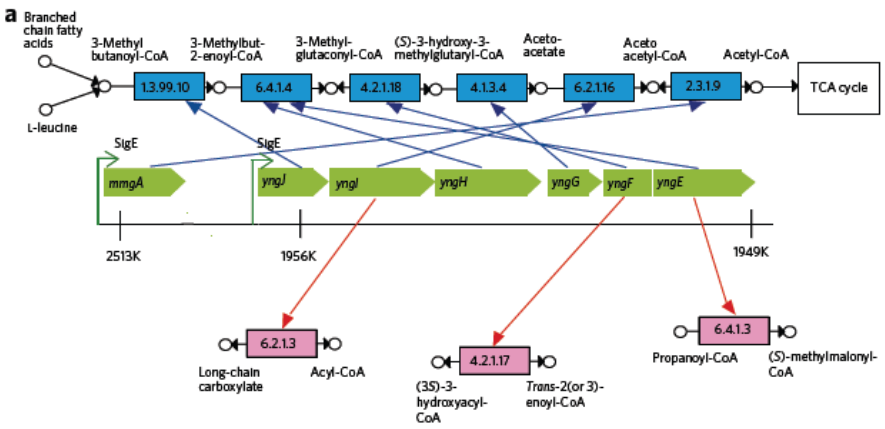


Figure 7a. Predicted function of genes forming the yng cluster in B. subtilis. The genomic positions of the yng genes are shown in green. The detected misannotations are indicated in red. The predicted functions, forming the degradation pathway, are shown in blue.

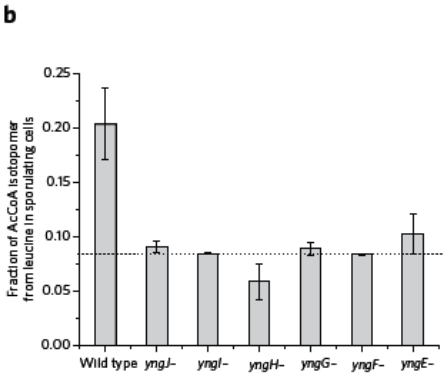


Figure 7b. Experimental validation of the predicted yng functions. To validate the predictions we used the fractional 13C labeling of acetyl-CoA in the wild type and mutant sporulating cells.

In contrast, in sporulating cells the fraction of acetyl-CoA derived from leucine was about 3 times higher than background, while all yng mutants displayed essentially background labeling levels (Figure 7b). Consequently, the yng pathway is indeed active and can supply additional energy to sporulating cells.

Conclusions and Future Directions

The presented results demonstrate how an application of diverse computational tools allows gaining a deeper understanding of the sporulation network at various levels of structural and functional organization. Our study integrates evolutionary analysis with stochastic simulations and with metabolic gene predictions. Because the analyzed sporulation process spans several types of biological networks and timescales, the computational tools provide a natural mechanism to integrate and interpret biological data. In the same way as no single experimental technique is sufficient to fully understand the complexity of sporulation network, no single computational approach is adequate. The complexity of the underlying network needs to be matched by a comprehensive analysis with multiple interconnected computational tools.

We believe that in the future it will be important to complement the simulations on the cellular level with a multicellular analysis. As bacterial cells are capable of forming complex communities, which can rival the complexity of some multicellular organisms, it will be essential to couple simulations of intracellular networks with networks of cell-cell signaling and communications.

Such extensions of our work will ultimately allow a truly multiscale description of the sporulation process.

REFERENCES

Burbulys, D., K.A. Trach, and J.A. Hoch. 1991. Initiation of sporulation in B. subtilis is controlled by a multicomponent phosphorelay. Cell 64: 545-552.

Carroll, S.B. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134: 25-36.

Chastanet, A., D. Vitkup, G.-C. Yuan, T.M. Norman, J. Liu, and R. Losick. 2010. Broadly heterogeneous activation of the master regulator for sporulation in B. subtilis. Proc. Natl. Acad Sci. in press.

Davidson, E.H. 2009. Network design principles from the sea urchin embryo. Curr. Opin. Genet. Dev. 19: 535-540.

Davidson, E.H. and D.H. Erwin. 2006. Gene regulatory networks and the evolution of animal body plans. Science 311: 796-800.

de Hoon, M.J., P. Eichenberger, and D. Vitkup. 2010. Hierarchical evolution of the bacterial sporulation network. Curr. Biol. in press.

Dubnau, D. and R. Losick. 2006. Bistability in bacteria. Mol. Microbiol. 61: 564-572.

Errington, J. 2003. Regulation of endospore formation in Bacillus subtilis. Nat. Rev. Microbiol. 1: 117-126.

Gonzalez-Pastor, J.E., E. Hobbs, and R. Losick. 2003. Cannibalism by sporulating bacteria. Science 301: 510-513.

Hsiao, T.L., O. Revelles, L. Chen, U. Sauer, and D. Vitkup. 2010. Automatic policing of biochemical annotations using genomic correlations. Nat. Chem. Biol. 6: 34-40.

Kaneda, T. 1991. Iso- and anteiso-fatty acids in bacteria: biosynthesis, function, and taxonomic significance. Microbiol Rev 55: 288-302.

Lozada-Chavez, I., S.C. Janga, and J. Collado-Vides. 2006. Bacterial regulatory networks are extremely flexible in evolution. Nucleic Acids Res. 34: 3434-3445.

Molle, V., M. Fujita, S.T. Jensen, P. Eichenberger, J.E. Gonzalez-Pastor, J.S. Liu, and R. Losick. 2003. The Spo0A regulon of Bacillus subtilis. Mol. Microbiol. 50: 1683-1701.

Nicholson, W.L., N. Munakata, G. Horneck, H.J. Melosh, and P. Setlow. 2000. Resistance of Bacillus endospores to extreme terrestrial and extraterrestrial environments. Microbiol. Mol. Bio. Rev. 64: 548-572.

Piggot, P.J. and J.G. Coote. 1976. Genetic aspects of bacterial endospore formation. Bacteriol. Rev. 40: 908-962.

Piggot, P.J. and R. Losick. 2002. Sporulation genes and intercompartmental regulation. In B. subtilis and its closest relatives. ASM Press, Washington, D.C.

Prud'homme, B., N. Gompel, and S.B. Carroll. 2007. Emerging principles of regulatory evolution. Proc. Natl. Acad Sci. 104: 8605-8612.

Sauer, U., V. Hatzimanikatis, H.P. Hohmann, M. Manneberg, A.P. van Loon, and J.E. Bailey. 1996. Physiology and metabolic fluxes of wild-type and riboflavin-producing Bacillus subtilis. Appl. Environ. Microbiol.: 3687-3696.

Setlow, P. 2003. Spore germination. Curr. Opin. Microbiol. 6: 550-556.

Sonenshein, A.L., J. Hoch, and R. Losick. 2001. Bacillus subtilis and its Closest Relatives. ASM PRess. Vlamakis, H., C. Aguilar, R. Losick, and R. Kolter. 2008. Control of cell fate by the formation of an architecturally complex bacterial community. Genes Dev. 22: 945-95.

THE FAMILY OF MAN: METHODS AND APPLICATIONS FOR DISCOVERING SHARED ANCESTRY AMONG PURPORTED UNRELATEDS

ITSIK PE'ER, PHD

DEPARTMENT OF COMPUTER SCIENCE, COLUMBIA UNIVERSITY
CENTER FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, COLUMBIA UNIVERSITY

Introduction

Much of the recent progress in human genetics has been achieved through associating traits to alleles across large collections of unrelated individuals. But are we really unrelated? Obviously individuals share common ancestors that would be detected if one could follow their lineages back into the past. How far back would such lineages meet? Can individuals still be approximated as being unrelated? Can data from remote relatives inform genetic studies? Can it help understand personal genomes that are now becoming available?

Over the last couple of years we have developed new methods for detecting hidden relatedness and using it to make inferences about population structure, and genomic features, and in the analysis of association to observed phenotypes. The key idea is that individuals' genomes are mosaics of the genomes of their ancestors, and relatives that share an ancestor may share genetic segments inherited from that ancestor. Such segments are said to be Identical By Descent (IBD). The length and number of IBD segments shared between two individuals due to a recent common ancestor k generations ago have closed-form distributions. These imply that for k>=6, such remote relatives are quite likely to share no regions due to this common ancestor, but as long as k is not much greater, than, say k≤20, when an IBD region does exist, that region is likely to be long - 3 cM or greater – and therefore detectable from modern genetic polymorphism data in terms of information content.

We have developed GERMLINE, a hashing-based algorithm that overcomes the algorithmic barrier to detecting IBD segments across all pairs of samples from large cohorts (Gusev et al., 2009). This opens a window to improved understanding of recent population genetics, stronger statistical tests of genetic association, and increased information content from high throughput sequencing.

Population Genetics of Identity-By-Descent

We can therefore define the relatedness graph G, with nodes corresponding to individuals and edges to genetic relatedness such that across a cohort of size n, one expects p(n¦2) edges. Under various assumptions, graph theory essentially guarantees that in a random graph, if pn > 1 + C, the majority of nodes will be transitively linked in what is known as a "giant" connected component (Kenny et all., 2009). We set out to test this prediction.

We observed a very different picture. A good example is the IntraGen Database of 917 New York Health Study participants that are self-declared unrelateds of European ancestry, whose whole-genome SNP data are hosted by C2B2 and made publicly available. In this cohort, 87,562 pairs of samples shared a total average of at least 30 cM each, likely intermediate relatives. Analyzing the relatedness graph, only 418 out of the 912 nodes (Figure 1) are spanned by a large connected component. This result represents a consistent deviation from the theoretical expectation for the case of a random graph, where such strong node partitioning is very unlikely to be maintained (p-value << 10-100). The cohort is indeed structured, and the node membership in the connected component is highly correlated with self identification as an Ashkenazi Jewish (AJ). We note that the AJ cohort was not only enriched for likely intermediate relatives, but also included 6,024 out of the 75,466 pairs (8%) of likely remote relatives in IBD, each sharing a single segment.

To investigate whether this increased relatedness amongst AJ is specific to Intragen data we compared sharing to 400 additional AJ samples from the Hebrew University Genetic Research

(HUGR). We observed the two AJ cohorts to have similar levels of sharing within each dataset, as well as across datasets, with the HUGR samples exhibiting the highest amount of sharing internally and to other AJ samples. On average, a pair of Ashkenazi Jewish (AJ) samples from HUGR and Intragen stand out as having larger probability of sharing a single locus (0.018) compared to the non-AJ European populations (average 0.006).

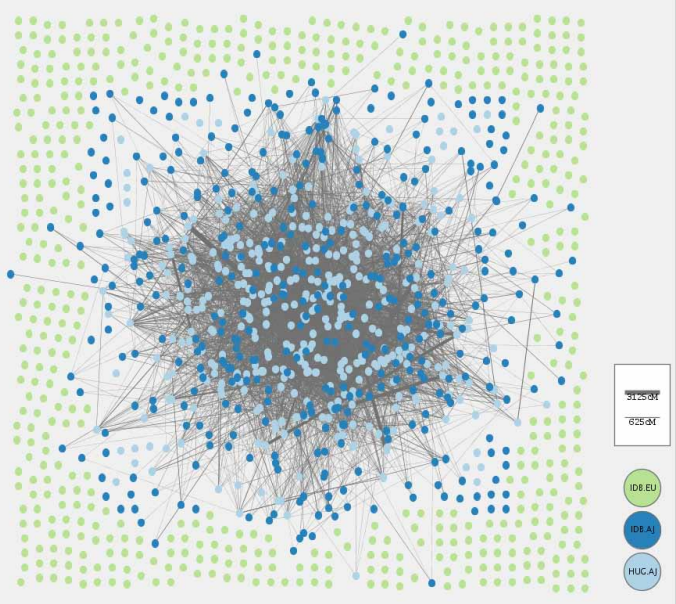


Figure 1. Identity By Descent between Long Islanders. Each node represents an individual with four reported European grandparents. Each edge represents observed multiple IBD segments shared between the pair of individuals, with thickness corresponding to the overall length of shared segments between them. Edges are only visible for pairs sharing > 50cM. Samples include the Intragen Data Base of Long Islanders, of AJ (blue) and non-AJ European (green) ancestry, and AJ samples from the Hebrew University Genetic Resource (cyan) for comparison

We further observed IBD segments to be quantitatively informative regarding population history of AJ. The distribution of their lengths is log-linear, consistent with theoretical predictions for an isolated population with rapid expansion, and enabling determination of late-medieval timing for the AJ population bottleneck, consistent with non-genetic information.

This pattern of increased IBD sharing within populations is not limited to AJ. Studying cohorts from 6 other Jewish communities with genome-wide SNP data as part of the Jewish Haplotype Map we see population-specific sharing. In contrast, the bottleneck of the AJ population, is specific to this Jewish group. In non-Jewish populations from the Human Genome Diversity Panel and the third-generation HapMap we further observe significantly more IBD segments than expected by standard population models.

Genomic regions with Identity-By-Descent

By definition, chances of recombination per centimorgan are constant, and suggest that IBD segments should be uniformly distributed along the genome. In contrast, we observe significant variation between loci when considering the fraction of sample pairs that are IBD at each locus. Some regions are frequently IBD, while most of the genome is rarely so. This observation holds per the basepair or the centimorgan length of regions, therefore is not just reflecting different recombination rates for different loci. High rates of IBD do suggest some of the haplotypes in a region recombine less frequently than others. In the AJ samples, the clearest such region with rarely-recombining haplotypes is the HLA locus (Figure 2). This region is known to have such a structure of long-range haplotypes, with explanations ranging from increased mutation rate to unique selective forces shaping the HLA. In other populations, commonly shared regions are enriched for common, long structural variants, including the long polymorphic inversions on chromosomes 8 and 17, where recombinants are likely to be selected against.

Association Analysis with IBD Segments

SNP genotypes are merely markers for the majority of human variation, which is untyped. We therefore sought to utilize SNP haplotypes that are shared between pairs and subsets of individuals as surrogates for variants carried on the background of such haplotypes. A genetic effect of a causal, untyped variant will be observed as association to a haplotype marker, with the genotype of the former only imputed. This has the potential to improve analysis of genetic association, especially for infrequent alleles, a typical blindspot for genomewide association studies.

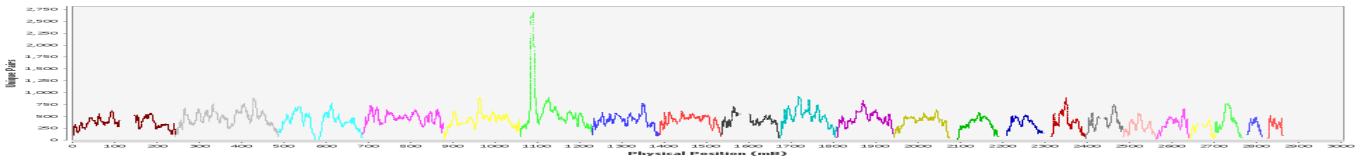


Figure 2. Identity By Descent varies along the genome. The frequency of a pair of individuals sharing a genomic locus IBD (y-axis) is shown along the genome (x-axis; colors represent chromosomes), as evaluated from the Intragen data of AJ samples from Long Island. Although IBD is observed genomewide, frequencies of sharing vary significantly along the genome, with a strong peak on chromosome 6, at the HLA locus.

This idea can come in each of several flavors: with vs. without data to train the imputation process, and on a locus vs. genome scale.

To study imputation of variants at a specific locus with training data, we chose to focus on the HLA. In addition to the unique abundance of long haplotypes, this region offers well-established assays and datasets involving complete typing of coding variants that can train imputation models. We have developed a novel algorithm, based on resolving triangles in a graph of shared haplotypes, that can use such training data for imputing HLA types in large cohorts with only SNP data available (Setty et al., 2010). In our recent work with the Severe Adverse Effects Consortium (www.saeconsortium.org) we have detected association of HLA alleles with drug-induced liver injury, and were able to utilize this imputation methodology for quality-control and backing up of our experimental results (Daly et al., 2009).

When training data is not available, we can still use IBD haplotypes for association. To do this, we need to examine correlation of traits to haplotype carrier status, and therefore to pinpoint the haplotypes that should be tested for such correlation. We have developed a method to locally detect clusters in the graph of IBD between pairs of individuals. We repeat this in a sliding window across a locus, handling complications such as ploidy and inaccurate data.

We have applied this method to genotypes of Pacific Islanders from Kosrae, Micronesia, where multiple metabolic phenotypes have been measured. We reported strong association of plantsterol levels to the ABCG8 gene locus, a functional candidate for this trait. We were indeed able to follow up this discovery by sequencing and confirmed a missense mutation exclusive to carriers of the associated haplotype (Kenny et al., 2009).

We have recently extended the methodology of association to shared haplotypes to handle sliding windows along the entire genomes. This required overcoming algorithmic challenges, e.g. by maintaining information between the clustered graphs in nearby windows. The resulting software tool is called DASH, and it enjoys use by multiple groups pre-publication. As internal proof of concept we chose the Kosrae data, which with multiple phenotypes and genomewide SNP data on 3000 individuals is especially conducive to such analysis. We have previously shown Kosrae to be a uniquely isolated population (Bonnen et al., 2006) with very little admixture (Bonnen at al, 2010). We indeed detect multiple novel variants that are DASH-associated to metabolic traits and blood markers (Kenny et al., 2010), that had not been previously detected (Lowe et al., 2009).

IBD with Personal Genomes

Using IBD for association genome wide with training information holds great potential, but until recently was not feasible due to limited training data. Such data would involve sequences of multiple personal genomes from a genome wide association cohort. The IBD segments shared between sequenced and unsequenced individuals can provide imputed sequence for the latter with near-deterministic accuracy along such segments (Figure 3). This strategy would be most powerful in an isolated population where the genome of any individual is a mosaic of genomes from just a small number of recent founding ancestors; therefore much of the genome of any unsequenced individual will be IBD to a small reference panel of sequenced individuals.

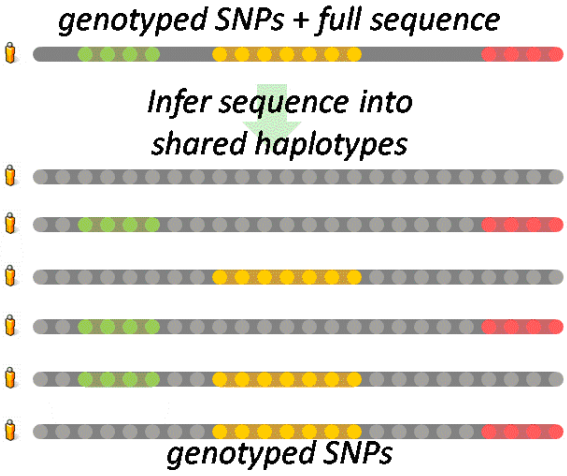


Figure 3. Using IBD to impute personal genomes. In a cohort of multiple individuals (grey chromosome-like stripes), with SNP array data (dots), IBD segments (colorfulcolors) are detectable. Every individual with a sequenced personal genome (top, solid grey) provides imputation information on other individuals with IBD segments (bottom)

In fact, we observed that our Kosrae samples, with a 19th-century bottleneck, rapid subsequent expansion, and an available cohort of >3000 genome wide-typed individuals, comprising the majority of the adult population, is ideal for such an analysis. Specifically, if carefully selected, as few as 30 personal genomes provide imputation information for 80% of the genome of any other individual (Figure 4).

Alas, having multiple personal genome sequences from a single isolated population had, until recently been infeasible. On the positive side, current "next-generation" sequencing technologies are revolutionizing genetics and enabling such endeavors. We were therefore fortunate to collaborate with LifeTechnologies, a leader in this field, as part of their $10k-Genome award. We generated pilot data, with low-pass personal genomes of seven Kosraens, chosen to maximize imputation potential. We observed unique variation patterns, demonstrating the value of resequencing an isolated population. Specifically, these variants are more likely to expose sequence alterations that are rare elsewhere, and would require a much larger sequencing cohort to be discovered in a general population.

Combining the whole genome sequences with DASH-associated IBD haplotypes, we observed sequence-level variants that explain the association signal. Such variants, including a long deletion allele and multiple coding changes, are likely causal changes, provide the first association study using whole-genome sequencing.

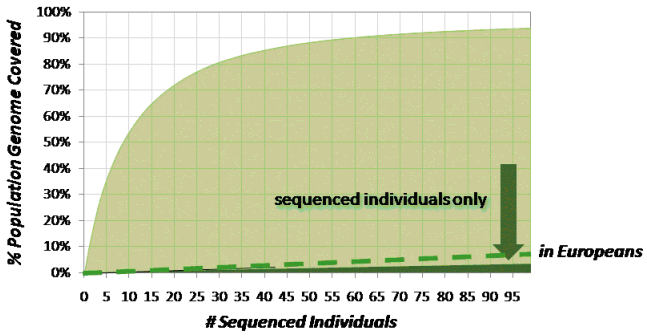


Figure 4. Imputation potential in an isolated population. When sequencing a small reference panel of a few individuals, at each locus imputation information is available on unsequenced individuals based on IBD to sequenced ones. We show the average such fraction along the genome (y-axis) as a function of the sequenced set size (x-axis). In Kosrae (light green), most of the "population genome" is thus imputable, even though only a small fraction of it is directly sequenced (dark green). This is in stark contrast to the potential imputation capacity in a general, non-isolate population (Europeans, dashed).

REFERENCES

1. Bonnen PE, Pe'er, Plenge RM, Salit J, Lowe JK, Shapero MH, Lifton RP, Breslow JL, Daly MJ, Reich DE, Jones KW, Stoffel M, Altshuler D, Friedman JM. Evaluating potential for whole genome studies in Kosrae, an isolated population in Micronesia. Nat Genet. 2006; 38(2):214

2. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I , Whole Population, Genome-Wide Mapping of Hidden Relatedness, Genome Res. 2009. 19(2):318-26

3. Lowe JK, Maller JB, Pe'er I, Neale BM, Salit J, Kenny EE, Shea JL, Burkhardt R, Smith JG, Ji W, Noel M, Foo JN, Blundell ML, Skilling V, Garcia L, Sullivan ML, Lee HE, Labek A, Ferdowsian H, Auerbach SB, Lifton RP, Newton-Cheh C, Breslow JL, Stoffel M, Daly MJ, Altshuler DM, Friedman JM. GWAS in an isolated founder population from the Pacific Island of Kosrae. PLoS Genet. 2009. 5(2):e1000365

4. Daly AK, Donaldson PT, Bhatnagar P, Shen Y, Pe'er I, Floratos A, Daly MJ, Goldstein DB, John S, Nelson MR, Graham J, Park BK, Dillon JF, Bernal W, Cordell HJ, Pirmohamed M, Aithal GP, Day CP for the DILIGEN study and International SAE Consortium, HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin Nat Genet. 2009 Jul;41(7):816-9

5. Kenny EE, Gusev A, Riegel K, Lütjohannd D, Lowe JK, Salit J, Maller JB, Stoffel M, Daly MJ, Altshuler DM, Friedman DM, Breslow JL, Pe'er I, Sehayek E. Systematic haplotype analysis resolves a complex plasma plant sterol locus on the Micronesian Island of Kosrae. Proc Natl Acad Sci U S A. 2009 106(33):13886-91

6. Bonnen PE, Lowe JK, Altshuler DM, Breslow JL, Stoffel M, Friedman JM, Pe'er I. European admixture on the Micronesian island of Kosrae: lessons from complete genetic information. Eur J Hum Genet. 2010 Mar;18(3):309-1.

7. Setty M, Gusev A, Pe'er I, HLA type inference via haplotypes IBD, RECOMB 2010, in press.

8. Kenny EE, Kim M, Gusev A, Lowe JK, Salit J, Smith G, Kovvali S, Kamg HM, Newton-Cheh C. Daly MJ, Stoffel M, Altshuler DM, Friedman DM, Eskin E, Breslow JL, Pe'er I. Increased power of mixed-models facilitates association mapping of 10 loci for metabolic traits in an isolated population. Submitted.

THE RNA BACKBONE REVEALS ITS PERSONALITY

HARMEN BUSSEMAKER LAB

The base sequence of RNA molecules plays an essential role in determining their three-dimensional structure, but it is commonly assumed that the sugar-phosphate backbone provides little specificity. By computationally analyzing a large number of RNA crystal structures, Dr. Xiang-Jun Lu, a Senior Research Associate in the Bussemaker Lab, has been able to show that an "edge" formed by a hydrogen-bonding interaction between two groups on the backbone (highlighted in Figure 1 by red spheres) helps stabilize RNA structures in a highly sequence-specific manner. This insight provides a reason for the thermodynamic stability and evolutionary conservation of two well-known RNA structural motifs. Intriguingly, it might also explain the extreme conservation of the GU dinucleotide observed at the 5′ end of almost all introns. A paper describing this work, to be published in Nucleic Acids Research (Lu *et al.*, in press), and was selected by the Editorial Board as a Featured Article, an honor bestowed on only the top 5% of papers in terms of "originality, significance and scientific excellence".

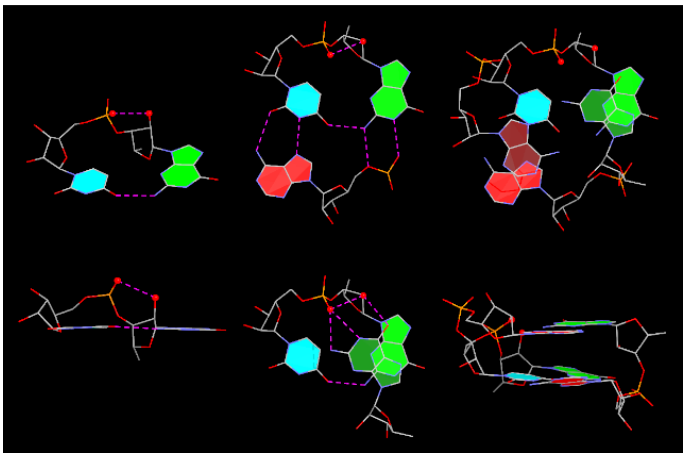


Figure 1. RNA-stabalization via hydrogen-bonding interaction

References

1. SX.-J. Lu, W.K. Olson, and H.J. Bussemaker. The RNA backbone plays a crucial role in mediating the intrinsic stability of the GpU dinucleotide platform and the GpUpA/GpA miniduplex. Nucleic Acids Res., in press..

HIGH-THROUGHPUT SEQUENCING MEETS FLY BODY PLAN SPECIFICATION

HARMEN BUSSEMAKER LAB AND RICHARD MAN LAB

Hox proteins constitute a subclass of the large family of homeodomain transcription factors. They play a crucial role in body plan and tissue specification throughout the animal kingdom, but so far only little is known about the molecular mechanisms that enable them to control different sets of target genes. Recently, the Honig and Mann labs were able to make important progress in this area using a structural approach. Now, the Mann and Bussemaker labs have also teamed up to determine the DNA binding specificity of each of the Hox proteins in the fruit fly Drosophila melanogaster using a genomics approach. They have developed a novel methodology, called "SELEX-seq", which uses in vitro selection of DNA from random pools, followed by high-throughput sequencing and biophysical modeling. The collaboration has already revealed new sequence-based rules explaining why dimers of the co-factor Extradentical (Exd) with different Hox proteins prefer different DNA sequences.

IDENTIFYING THE UNDERLYING NETWORK AND MASTER REGULATORS OF GLIOBLASTOMA MULTIFORME

ANDREA CALIFANO LAB

One of the imperatives of translational research is to decode the relationship between the cancer phenotype and the underlying cancer genotype. Our lab is dissecting this dependency for the highly proliferative and invasive brain tumor, glioblastoma multiforme (GBM). GBM is nearly uniformly fatal with only modest survival benefits with existing therapies.

We recently demonstrated that STAT3 and C/EBP, two synergistic transcriptional master regulators associated with GBM tumor aggressiveness, could be identified by reverse engineering the network underlying the cancer phenotype of high grade gliomas[1]. Neither of these factors had a prior association with brain cancer or a role in determining the most aggressive properties of glioblastoma, including aberrant mesenchymal transformation, invasion of normal surrounding tissue and angiogenesis. In collaboration with Dr. Antonio Iavarone at the Herbert Irving Comprehensive Cancer Research Center, the computational findings were validated by follow-up experiments. Expression of the two genes in neural stem cells caused them to display all the hallmarks of the mesenchymal identity associated with the most aggressive glioblastoma.

Conversely, silencing these genes in cell culture and in GBM-human brain tumor initiating cells reduced their invasiveness and their ability to form tumors in mouse xenografts. Notably, the expression of these two genes together was strongly correlated with increased patient mortality.

We are extending these results by continuing to interrogate the multiple human glioma data sets that are publically available, in part through The Genome Atlas/TCGA consortium, with the following goals:

1. Constructing a fully fully-integrated interactome for high-grade glioma (HGi), including transcriptional, post-transcriptional, and post-translational interaction layers

2. Using the HGi to integrate genetic, epigenetic, and functional assays associated with the molecular mechanisms underlying tumor progression and GBM poor-prognosis

3. Interrogating the HGi to identify candidate therapeutic target genes

4. Integrating experimental evidence from genome-wide RNAi screens in the prioritization of candidate therapeutic target genes

5. Identifying and validating small molecule inhibitors of potential therapeutic target genes

6. Disseminating the algorithms, models, software tools, workflows and integrated datasets to the research community using the geWorkbench platform through the MAGNet Infrastructure.

As we continue to dissect the full complement of molecular interactions that are dysregulated in human high-grade gliomas, we are hopeful that the underlying pathological mechanisms will become evident thus leading to the identification of novel candidate targets for diagnostic and therapeutic intervention.

References

1. Carro, M.S. *et al.*, Nature. 2010 Jan 21;463(7279):318-25.

ZOOMING IN ON THE ROLE OF INDIVIDUAL GENES IN PREDICTING DRUG RESPONSE

DANA PE'ER LAB

The lab of Dana Pe'er has combined genotype, that is, the variations in an individual's DNA, with gene expression (RNA) profiling to develop a new statistical method for predicting resistance or sensitivity to particular drugs. The method is called Camelot (CAusal Modelling with Expression Linkage for cOmplex Traits) and was tested using a diverse set of strains of the yeast Saccharomyces cerevisiae. It makes use of each strain's genotype and a baseline gene expression profile created in a drug-free, unperturbed state. The growth rate of a "training set" of strains in response to a panel of drugs was measured and the three types of data were integrated in Camelot to create a statistical model. A premise of this design is that gene expression integrates information from multiple genetic loci that are individually too weak to detect but which, in combination, contribute significantly to the phenotype (the growth rate). The resulting model was very successful at predicting the response of other, test strains to the drugs, based only on their genotype and their gene expression profile in the unperturbed state. Camelot correctly predicted the responses to 87 out of the 94 drugs tested. The procedure was very successful in identifying genes causally related to the drug response, and in finding novel gene–drug interactions.

An important distinguishing feature of Camelot is that the gene expression of an individual need only be assayed once. This single-gene expression profile can be harnessed to analyze the connection between genotype and phenotype for a large number of traits that manifest under many different conditions. Moreover, the response to a drug can be predicted before treatment, a critical feature for clinical application. The approach is robust, applicable to other phenotypes and species, and has potential for applications in personalized medicine, for example, in predicting how an individual will respond to a previously unseen drug.

References

1. Chen BJ, Causton HC, Mancenido D, Goddard NL, Perlstein EO, Pe'er D (2009). Harnessing gene expression to identify the genetic basis of drug resistance. Mol Syst Biol. 5:310. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2779083

IDENTIFYING GENETIC DETERMINANTS OF SERIOUS ADVERSE EVENTS

ARIS FLORATOS LAB AND ITSIK PE'ER LAB

The International Serious Adverse Event Consortium (iSAEC) is a nonprofit organization whose mission is to identify genetic factors for predicting the risk of drug-related serious adverse events (SAEs). Our labs have been providing infrastructure and expertise to support the data analysis and coordination needs of the iSAEC. In the past three years we have conducted several Genome-Wide Association Studies (GWAS) on a variety of SAEs, including serious skin rash (SSR), drug-induced liver injury (DILI), and drug induced elongated QT and Torsades de pointes. We have successfully identified genetic variants associated with predisposition to some of these adverse reactions. In particular, we found HLA-B*5701 is a major determinant of flucloxacillin-induced liver injury (Daly AK et al 2009), and several Major histocompatibility complex (MHC) variants increase the risk of coamoxiclav-induced livery injury (publication in preparation).

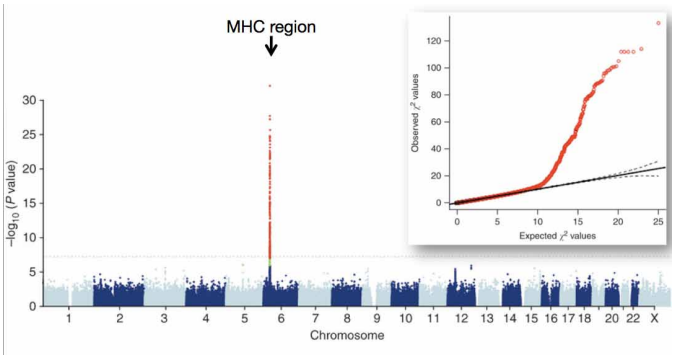


Figure 2. Manhattan and QQ-plots demonstrating the location and strength of genome wide association in a cohort of 51 subjects with flucloxacillin-induced DILI (Daly AK *et al.* 2009).

In general, the statistical power of pharmacogenetically motivated GWAS studies can be compromised by a number of factors. First, given the rarity of most adverse events, it is usually hard to compile case cohorts of significant size. The fact that genetic effects are likely to be drug-specific exacerbates the problem as most clinically compiled subject populations are heterogeneous in their causal drug. Second and most important, the commercial genotyping arrays used in GWAS interrogate almost exclusively variants that are common in the population; it is now understood that such variants have limited power of predicting the risk of SAEs, a situation similar to the missing heritability problem in the study of common diseases (Manolio TA et al 2009). Instead, there is growing evidence that rare genetic variants may be significant contributors to the unexplained heritability. In response to these new insights, the iSAEC has started using novel genomic technologies, including whole genome re-sequencing, to explore the role of rare variants for deepening our understanding of the genetics of SAEs.

References

1. Daly, A. K., P. T. Donaldson, *et al.* (2009). "HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin." Nat Genet 41(7): 816-819.

2. Manolio, T. A., F. S. Collins, *et al.* (2009). "Finding the missing heritability of complex diseases." Nature 461(7265): 747-753.

NEW USES FOR PROTEIN STRUCTURES

BARRY HONIG LAB

Although structural relationships can be used to infer function when sequence information fails to provide adequate clues, structure has not been widely exploited for this purpose. Reasons include the increased probability of incorrect annotation transfer, the difficulty of defining just what it means for proteins to be structurally similar, and the lack of structural information. To more fully exploit the potential of the increasingly large quantities of structural information becoming available, we have been developing strategies to analyze and predict remote functional relationships based on structure, with particular focus on the prediction of protein-protein interactions. We carried out a comprehensive analysis of the degree to which the location of a protein interface is conserved in sets of proteins that share varying degrees of similarity. Our results show that while the interface conservation is most significant among close neighbors, it is still conserved to a surprising extent even for remote structural neighbors. Further, we developed an interface prediction method, PredUs, that outperforms currently available tools.

Our work has opened up new avenues for combining techniques from computational structural biology and systems biology. In a preliminary investigation of the yeast interactome, we found that coarse modeling of interacting proteins based on remote structural neighbors followed by a bioinformatics-based evaluation of the reliability of the interaction can be highly successful. Our approach yields a much larger number of hypotheses for protein-protein interactions than has been previously possible. Based on cross-validation, our results also suggest that the range of applicability of structural information could be expanded to a scale and reliability comparable to that of other approaches, e.g. sequence similarity, expression profiles, phylogenetic analyses and even high-throughput experiments.

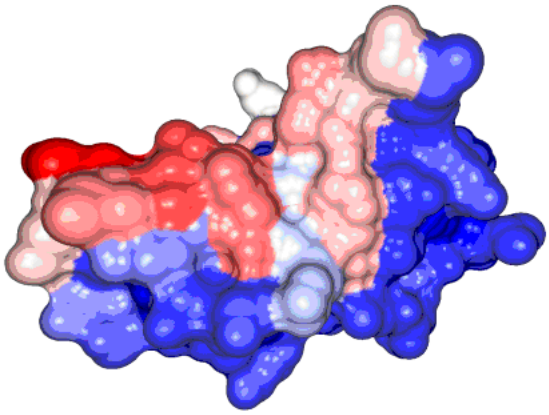


Figure 3. The surface of T-cell receptor protein CD8 frequency with which interactions made by its structural neighbors are "mapped" to individual residues on its surface (red/white/blue = high/intermediate/low frequency). Our structural alignment tool, Ska, found 978 structure neighbors for this protein. The fact that only 188/317/348 of these neighbors come from the same SCOP family/superfamily/fold of 1akj.D suggests that there are many remote structural relationships beyond the fold of 1akj.D suggests that there are many remote structural relationships beyond the fold level. The red high-contact frequency regions show that protein-protein interaction sites can be conserved among both close and remote structural neighbors. level. The red high high-contacting frequency regions show that protein-protein interaction sites can be conserved among both close and remote structural neighbors.

P53 RECOGNIZES THE SHAPE OF DNA BINDING SITES WITH AN EXTENDED GENOMIC ALPHABET

BARRY HONIG LAB

The tumor suppressor protein p53, also known as the guardian of the genome, binds as a tetramer to DNA response elements consisting of two half-sites. Mutations in the DNA binding domain of p53 are responsible for about 50% of human cancers. In particular, mutations of the Arg248 residue, which binds to the DNA minor groove, are the most frequent mutations found in human tumors. New high-resolution crystal structures of p53-DNA complexes with contiguous half-sites published by Barry Honig and coworkers suggest a novel protein-DNA readout mechanism[1]. The two central base pairs in each half-site form Hoogsteen base pairs, which deviate from the common Watson-Crick geometry (see Figure 4). This different geometry leads to a specific variation of the shape of the p53-DNA binding site. The double-helix forms a local waist, which allows for enhanced interactions between p53 protomers, and the minor groove becomes narrow in the regions where the Arg248 side chains bind.

The narrowing of the minor groove results in an enhanced negative electrostatic potential, which stabilizes the interactions of the Arg248 residues with the p53 response element. This observation provides a molecular explanation of for why Arg248 is a cancer hotspot mutant. The recognition of narrow minor groove regions by arginines through enhanced negative potentials was recently identified as a readout mode that contributes to the DNA-binding specificity of Hox proteins[2] and other protein families[3]. Although related, the readout mode employed by p53, which is based on an alternate base pairing geometry, is distinct from the previously described shape readout of sequence-dependent variations in groove width[4]. Moreover, the observation of Hoogsteen base pairs in undistorted B-DNA extends the four letter alphabet of Watson-Crick base pairs, and illustrates that structural information is needed for the complete understanding of the genome.

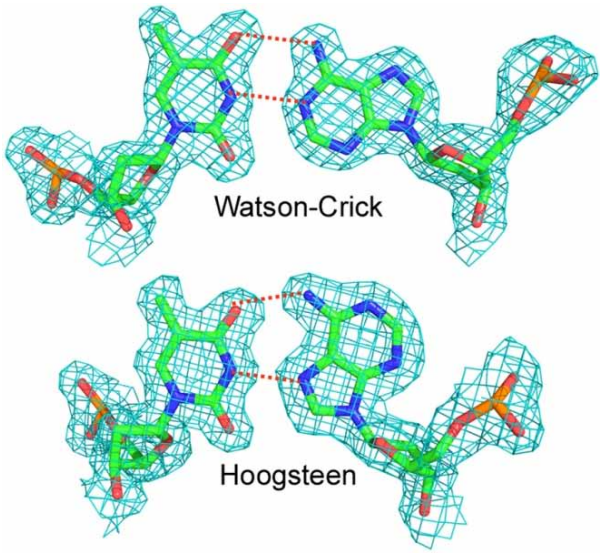


Figure 4. Electron density map (2Fo-Fc) at 1σ level shown for an A:T base pair of a p53-DNA binding site in Watson-Crick5 and Hoogsteen geometry[1].

References:

1. Kitayner, M. *et al.* Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. Nat. Struct. Mol. Biol. 17 (2010).

2. Joshi, R. *et al.* Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell 131, 530-43 (2007).

3. Rohs, R. *et al.* The role of DNA shape in protein-DNA recognition. Nature 461, 1248-53 (2009).

4. Rohs, R. *et al.* Origins of Specificity in Protein-DNA Recognition. Annu. Rev. Biochem. 79 (2010).

GENSPACE: COMMUNITY-DRIVEN KNOWLEDGE SHARING IN GEWORKBENCH

GAIL KAISER LAB

geWorkbench, the bioinformatics platform of the MAGNet Center (http://www.geworkbench.org), is frequently expanded with 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. Also, most people don't like reading long articles on how to use a tool. Using genSpace, we aim to provide context-sensitive information in geWorkbench directly, describing how best to use the vast array of tools that are available. We answer questions such as:

- What do I do first?
- Which tools work well together?
- Where does this tool fit in a typical workflow?

genSpace logs, aggregates, and data mines geWorkbench users' activities to answer these questions. Users have the option of having their data logged anonymously or with their user names, or of opting out of this data collection entirely.

One of the main goals of genSpace is to provide collaborative filtering and knowledge-sharing features to geWorkbench users through recommendations presented via social networking metaphors such as "people like you …". The genSpace module for recommendations was first included in geWorkbench v1.7.0, which was released on 17 July 2009. Some of the features included are described below.

geWorkbench users can search for workflows (sequences of tools) that include or start with a particular tool. Figure 5 shows an example of a query for "All Workflows" that include SMLR Classifier. genSpace also provides "Real-time Workflow Suggestions" as a user is using geWorkbench. Figure 6 shows such an example. The user has just run SOM Analysis and the genSpace module provides recommendations of the superflows that include SOM Analysis and the possible next steps.

All of these recommendations are derived using data mining of past usage history and collaborative filtering techniques.

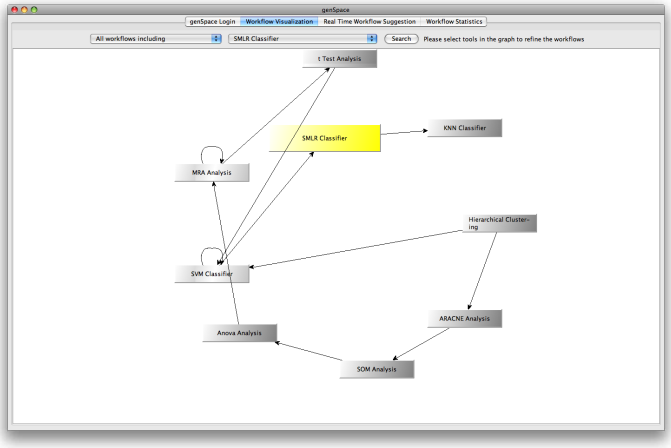


Figure 5. Workflow Visualization

The genSpace recommendations module requires a collection of user data in order to generate useful suggestions - the "cold start" problem. For this reason, The the genSpace module for recording geWorkbench users' activities was first included in geWorkbench v1.6.3, which was released on 8 January 2009, about six months prior to the release of the recommendations module with geWorkbench v1.7.0.

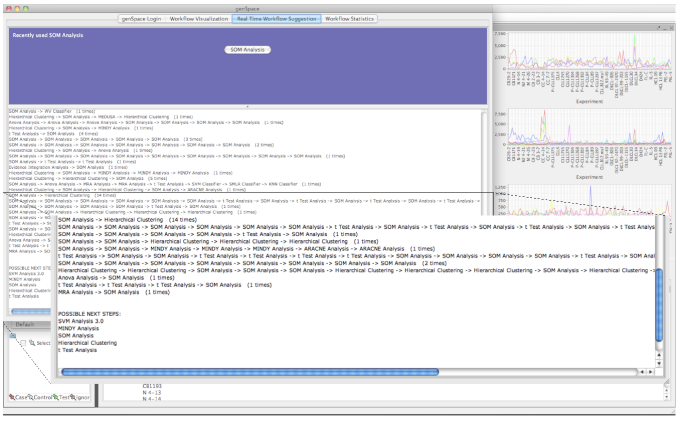


Figure 6. Real-time Workflow Suggestions

As of 25 March 2010, we have collected over 6500 user logs from roughly 150 distinct users, including users from Germany, Switzerland, Brazil, and the U.K.

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

SKYLINESEARCH: SEMANTIC RANKING AND VISUALIZATION OF PUBMED SEARCH RESULTS

KENNETH ROSS LAB

SkylineSearch is a semantic search and visualization engine recently released by the group of Kenneth Ross, and available at skyline.cs.columbia.edu. The engine aims to enhance the user experience during one of the most common and often daunting tasks, scientific literature search. Many life sciences researchers search PubMed as part of their daily activities. With the number of articles in PubMed growing from year to year, and with many queries returning thousands of high-quality matches, there is a clear need for relevance ranking of results. Such ranking is not currently available in PubMed.

Users specify queries as combinations of terms from the MeSH ontology. SkylineSearch implements several novel ranking functions that use articles' MeSH annotations to determine the relevance of each article to a user's query. A two-dimensional visualization, termed a skyline, plots article relevance against its publication date. The y coordinate shows the computed relevance (higher is better), while the x coordinate increases with the age of the publication. This visualization allows the user to identify articles of interest and save them for future reference, optionally tagging them with descriptive keywords. Figure 7 presents a screenshot of the SkylineSearch system, in which the user executed the MeSH query "Autoimmune Diseases AND Pregnancy".

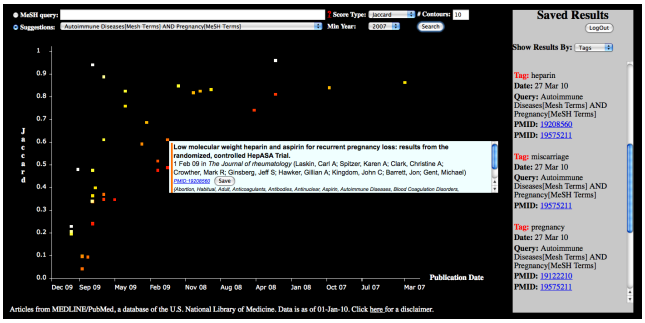


Figure 7. SkylineSearch system, displaying results of the MeSH query "Autoimmune Diseases AND Pregnancy". Publication date is plotted on the x-axis, while query relevance is plotted on the y-axis. Selecting one of the results in the visualization will display details about the article in a pop-up. Saved and tagged results are presented in the Saved Results area, on the right-hand side.

The ranking functions used by SkylineSearch, as well as ways of efficiently computing them on the scale of PubMed and MeSH, are described in Stoyanovich et al., 2010.

References

1. J. Stoyanovich, W. Mee, and K. A. Ross. Proceedings of the 2010 International Conference on Data Engineering, March 2010. http://www.cs.columbia.edu/~kar/pubsk/ICDE10_conf_full_040.pdf "Semantic Ranking and Result Visualization for Life Sciences Publications."

TOWARD DETECTING AND PREDICTING EPILEPTIC SEIZURES

DAVID WALTZ LAB

Epilepsy affects approximately 50 million people worldwide. Anti-epileptic drugs are able to control seizures in about 2/3 of patients, but patients who are not helped by drugs, cannot drive, swim or climb a ladder, and for those with frequent seizures, epilepsy can dominate their lives. Surgical excision of the epileptogenic brain region is the only other option today, though only 60% of patients become seizure-free after surgery.

David Waltz of MAGNet and CCLS (the Center for Computational Learning Systems), along with CCLS researchers Haimonti Dutta, Ansaf Salleb-Aouissi, Hatim Diab, Phil Gross, and students Karthik M Ramasamy, Stanley German, Shen Wang and Huascar Fiorletta, have been collaborating with Catherine Schevon and Ronald Emerson of the Comprehensive Epilepsy Center at Columbia University Medical Center on improving options for such patients. The joint project aims to locate markers that could indicate impending seizures. Today, before surgery, implanted electrodes record EEG signals during both habitual seizures and interictal, or quiescent periods, gathering roughly 7 TeraBytes of data for each patient. This data is used to plan surgery. Near term project goals focus on detecting seizures in this data – large seizures are easy to detect, but patients typically have many small seizures whose initiation sites are difficult to pin-point. To establish the mechanism of seizure development long-term fine-scale recordings such as HFOs (High Frequency Oscillations) are being studied. These very brief 200-500 Hz brain oscillations, are common in epilepsy patients, but essentially unknown in non-epileptics.

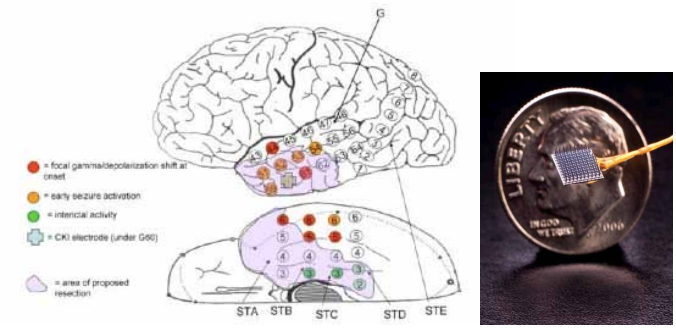
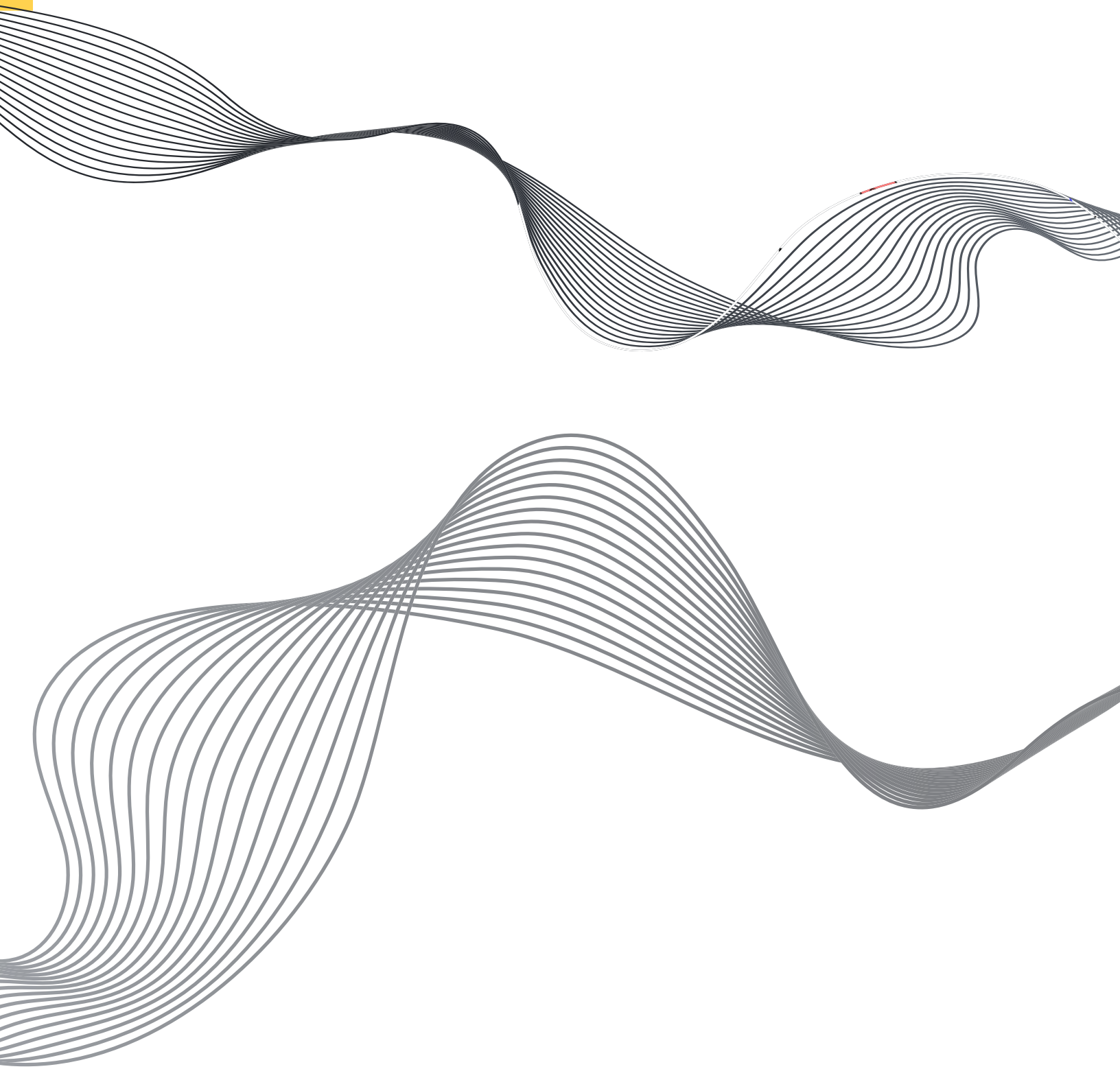


Figure 7: (Left) An implantation with traditional subdural grid and strip electrodes in a patient with medically refractory temporal lobe epilepsy.
(Right) The Neuroport Micro Electrode Array (Blackrock Microsystems, Salt Lake City, UT) measuring 4 mm x 4 mm with 1 mm microelectrodes arranged in a 10 X 10 grid used for collection of iEEG data.

The analysis of this data using advanced machine learning techniques is likely to provide insights regarding mechanisms of seizure initiation, and in the long run could lead to devices that could warn a patient of an impending seizure, apply drugs or electrical signals that would interrupt and prevent seizures. Further information about projects in the EWarn group at CCLS is available at : http://www.ccls.columbia.edu/research/epilepsy-early-warning



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

SPRING 2010
Issue No. 3