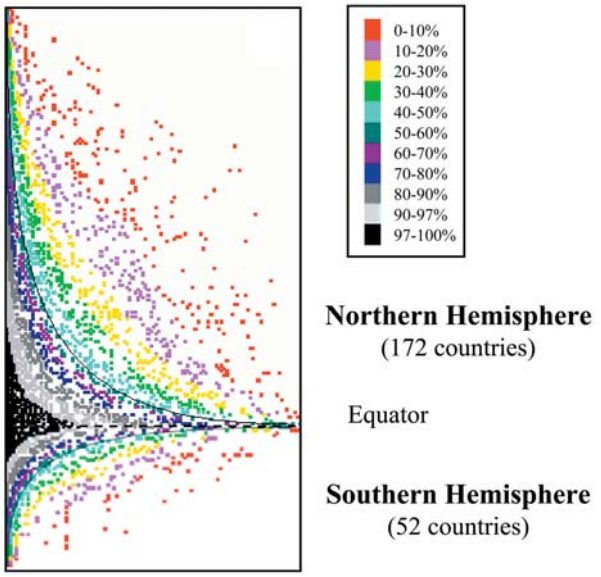
Synopses of Research Articles

Ecology Drives the Global Distribution of Human Diseases

It's no surprise that the Amazonian rainforest contains far more species than, say, the Siberian tundra. Over 50% of the world's species live in tropical rainforests, which cover just 6% to 7% of the earth's terrestrial surface. That the number of marine and terrestrial species declines with distance from the equator is a well-documented phenomenon called the latitudinal species diversity gradient. What's proven challenging, however, is figuring out what drives this pattern. Over 30 hypotheses have been proposed in the past two decades, but only four have garnered serious attention. These four focus on variables relating to area and energy factors, geographic constraints, and habitat diversity. Understanding the factors—both contemporary and ancient—responsible for the diversity gradient could help answer one of the fundamental questions in evolutionary ecology: what regulates species diversity? But teasing out the likely mechanisms behind this diversity has practical implications as well: mounting evidence suggests that ecological and climatic conditions influence the emergence, spread, and recurrence of infectious diseases. Global climate change is likely to aggravate climate-sensitive diseases in unpredictable ways.

Increasingly, public health programs aimed at preventing and controlling disease outbreaks are considering aspects of the ecology of infectious diseases—how hosts, vectors, and parasites interact with each other and their environment. The hope is that by understanding how ecological factors impact the global distribution of parasitic and infectious diseases, public health officials can predict and contain future outbreaks. Even though parasitic and infectious organisms account for a major fraction of the biological diversity on the planet, few studies have analyzed the factors affecting the spatial distribution of these organisms or attempted to quantify their contribution to biodiversity. In this issue, Vanina Guernier, Michael Hochberg, and Jean-François Guégan address the influence of ecological factors on the biological diversity and distribution of parasitic and infectious diseases and find that climatic factors are the most important determinant of the global distribution of human pathogens.

The current understanding of human disease and availability of complete datasets on many parasitic and infectious diseases, the researchers explain, present a unique opportunity to explore the relationship between parasitic and infectious disease species richness (defined in their study as total number of pathogens within a given country's borders) and latitude. This information, in turn, can help identify potential factors that affect diversity gradients. After compiling epidemiological data on 332 different human pathogens across 224 countries, Guernier et al. used sophisticated statistical modeling methods to identify and characterize the influence of a number of potential contributing factors on species richness. After adjusting the model to control for cofactors that might influence the relationship between latitude and species richness indirectly rather than directly (cofactors such as the size of countries and demographic, economic, and environmental variables), the researchers



The number of pathogen species increases towards the equator

confirmed that, on average (seven times out of ten), tropical areas harbor a larger number of pathogen species than more temperate areas. In other words, the species richness of human pathogens follows the same pattern seen in other species.

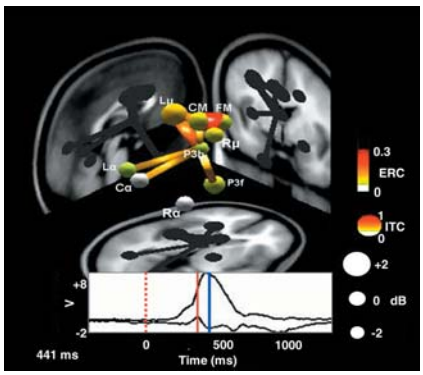
These results, Guernier et al. argue, suggest that the latitudinal species diversity gradient "might be generated in large part by biotic interactions." This in turn indicates that current estimates of species diversity, which ignore parasites and infectious organisms, are "substantially underestimated." The authors went on to explore groupings of individual pathogen species within larger parasitic and infectious disease communities along the gradient and found that species present at northern latitudes are a subset of those present in equatorial areas, rather than a different set of species (a phenomenon called "nestedness"). Since nestedness is strongly associated with latitude, which is typically used as a proxy for a range of climatic factors, the researchers investigated the relationship between various climatic variables and pathogen diversity. The climatic variable most strongly correlated with diversity was the maximum range of precipitation of a region.

The finding that climatic factors are largely responsible for the spatial distribution of human pathogens has important implications for predicting and managing future infectious disease outbreaks. These results counter the conventional assumption that socioeconomic conditions are the most important factor in controlling disease, indicating that global climate change could have far more significant effects on global patterns of disease, with diseases once relegated to the tropics migrating to temperate zones, for example. Identifying the links between ecology and disease, however, could lay the foundation for effective preventive strategies.

Guernier V, Hochberg ME, Guégan J-F (2004) Ecology drives the worldwide distribution of human diseases. DOI: 10.1371/journal.pbio.0020141

Deconstructing Brain Waves: Background, Cue, and Response

Light waves from an awaited signal—a white circle—arrive at the subject's eye; within a fraction of a second, the subject's thumb presses a button. Between eye and thumb lies the central nervous system, its feats of perception, integration, and response largely opaque to scientific scrutiny. Imaging techniques like magnetic resonance imaging can detail brain anatomy but can only broadly show changes in activity levels occurring over seconds—indirect echoes of brain function. Electrodes stuck to the scalp record coordinated neuronal symphonies, and wires inserted among neurons can capture the single-cell firing patterns of the individual instruments of the neural orchestra. But how these electrical signals map to information processing within and across neural circuits remains blurry. A new analysis sharpens the focus by separating individual brain wave patterns, measured from multiple sites across the scalp, into nine distinct process classes, each centered in an anatomically relevant brain area and producing predictable patterns as human subjects receive visual cues and produce responses.



Schematic representation of the source and strength of task-related EEG signals. Click here to view animation.

Scalp electroencephalograms (EEGs) are dominated by waves of synchronized neuronal activity at specific frequencies. Decades of research have associated wave patterns recorded at different scalp regions with different states of alertness—attending, drowsy, sleeping, or comatose; eyes open or closed—and gross abnormalities, such as seizure, brain damage, and tumor. In order to separate EEG responses to specific events from background, state-related activity, researchers repeat an experiment like the button-press exercise tens or hundreds of times and average the EEG across trials. By averaging out background activity, this technique reveals a characteristic waveform, called an event-related potential (ERP). It differs by electrode location, but often contains a large positive wave that peaks 300 milliseconds or more after an awaited visual cue.

In the current paper, Scott Makeig et al. argue that ERP averaging removes important information about ongoing processes and their interactions with event-related responses. Instead of averaging multiple recordings from each of 31 electrode sites, the authors applied an algorithm that seeks independent signal sources contributing to the individual tracings. The researchers measured signal source activities by the frequency and phase of wave patterns and source locations by comparing signal strength and polarity at different electrodes. Altogether, the researchers identified nine classes of maximally independent sources, each having similar locations and activities across subjects. The results dovetail neatly with prior anatomical and functional observations.

This analysis demonstrates that average waveforms identified in ERP studies probably sum multiple, separate processes from several brain regions. In particular, the large positive ERP seen 300 milliseconds or more after a visual cue reflects different waveforms from frontal, parietal, and occipital cortex—areas involved in task planning, spatial relationships and movement, and visual processing, respectively. In addition, this study showed a two-cycle burst of activity in the 4–8 (theta) frequency band after button presses—another common ERP feature. The theta activity was coordinated across several signal sources, and localized to areas associated with planning and motor control. Notably, the planning component seemed to lead the motor signal. Suppression or resynchronization of several EEG processes followed the visual cue or button press. The authors theorize that such coordination might influence the speed or impact of communication between brain areas and help retune attention after significant events.

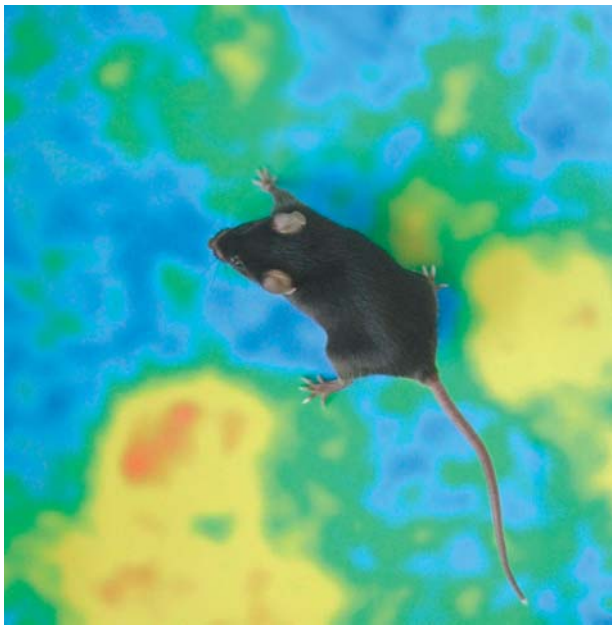
Using this approach in more subjects, and under differing conditions, could provide an unprecedented glimpse of how the brain translates perception and planning into action. The results suggest that EEG data contain an untapped richness of information that could give researchers and clinicians a new window into thought in action.

Makeig S, Delorme A, Westerfield M, Jung T-P, Townsend J, et al. (2004) Electroencephalographic brain dynamics following manually responded visual targets. DOI: 10.1371/journal.pbio.0020176

A Method for Studying Calcium Dynamics in Transgenic Mice

Calcium makes up just 2% of the human body, and 99% of it is sequestered in bones and teeth. The remainder exists within and around cells, influencing a variety of cellular processes, from transcription and cell survival to nerve signaling and muscle contraction. Calcium inhabits the intra- and extracellular space as a gradient, with extracellular concentrations some 1,000 times greater than those inside the cell. These gradients are maintained by calcium pumps. Calcium signaling operates mostly through voltage- and ligand-gated calcium channels (ligands are signaling molecules), both in the plasma membrane of the cell and in the membranes enclosing intracellular organelles. Signaling is typically initiated by an influx of calcium across the plasma membrane or by the release of calcium from an organelle, such as the endoplasmic reticulum. In neuron-to-neuron signaling, calcium signaling helps convert electrical signals into chemical signals in the form of neurotransmitters. The arrival of an electrical signal at a nerve terminal opens the many calcium channels in the nerve terminal, admitting a stream of calcium ions. The increased intracellular calcium concentration in turn releases the resident neurotransmitters accumulated in the nerve terminal, converting the electrical signal into a chemical signal. As neurotransmitters bind to their receptors in the next target neuron, they change the cell's membrane potential, prompting the neuron to generate an electrical signal, thereby converting the chemical signal back into an electrical one and completing the signaling circuit. Since calcium dynamics mediates most neuronal information flow, it can be used as a general measure of neural activity.

Through an elaborate network of electrical activity, the brain encodes, combines, and interprets signals to process information about the world. Simultaneous measurement of this activity in multiple brain locations has provided valuable insight into how neural networks function. But since electrical recordings can't pinpoint activity in the fine branches of individual neurons or pick up biochemical (nonelectrical) signals, researchers are increasingly turning to approaches that measure calcium concentrations as a proxy for neuron



Calcium-indicator-expressing mouse on a Ca^{2+} activity odor map (Image by Rolf Sprengel)

activity. Now Mazahir Hasan et al. have created mice engineered to stably express two different kinds of fluorescent calcium indicator proteins (FCIPs) in the brain (the fluorescence produced by these proteins can be seen when the brain is viewed with a two-photon microscope). Because the indicators are incorporated into the mouse genomes, this approach offers the possibility of targeting specific cells (by using promoters that specify which cells the genes should be activated in), allowing researchers to map the activity of select neuron populations.

Fluorescent proteins can be incorporated into a gene of interest to help researchers track that gene's protein in living tissue. FCIPs report calcium concentrations by changing fluorescence when they bind to calcium. Introducing calcium indicators into neural tissues was largely impractical, often failing to target specific cell types, until these genetically engineered indicators were developed in the late 1990s, allowing the desired specificity. While FCIPs have been used to good effect in worms, fruitflies, and zebrafish—and just recently in mouse muscle—they had not been stably and functionally expressed in the mammalian brain until now.

To deliver the FCIPs to the mouse brain, Hasan et al. used a regulatory promoter called Ptet (in the tetracycline system), which offers the possibility of targeting the expression of the FCIPs in different neural populations. To test the functionality of the proteins, they used fluorescence microscopy to analyze neurons from mouse lines and found high levels of FCIP expression. The real test, however, was whether the FCIPs could fulfill their promise as a probe for calcium activity. When the authors electrically stimulated brain slices from a mouse line expressing moderate to high levels of FCIP (electrical stimulation is known to increase intracellular calcium concentration), fluorescence increased rapidly following the stimulus. Significant changes in FCIP fluorescence were also observed when live mice responded to odor stimulation. That "fast and robust" FCIP signals were detected in live animals responding to sensory stimulation, the authors argue, proves the promise of FCIPs as a reporter on the activity of select neural populations in living systems. And since these indicator proteins retain stable functional expression over time (8- to 12-week-old mice continued to express the proteins), they could help researchers track neuronal activity over extended periods.

While a variety of bugs remain to be worked out with FCIPs—it's unclear, for example, why only the Ptet promoter generates high levels of FCIP expression in the brain and why not all neurons in a given population express the proteins—Hasan et al. demonstrate that the tetracycline system supports stable expression of the calcium indicators. The FCIP approach avoids the complications of invasive techniques like surgically administering dyes and produces a more interpretable signal, since the cell populations are already known. Because FCIPS can be used in living animals, they can reveal where and when neurons are firing. And because FCIP mice can be crossed with mice containing mutations in genes important for neural function, this method could reveal how specific genes contribute to the construction of neural networks.

Hasan MT, Friedrich RW, Euler T, Larkum ME, Giese G, et al. (2004) Functional fluorescent Ca^{2+} indicator proteins in transgenic mice under TET control. DOI: 10.1371/journal.pbio.0020163

Information Transport across a Membrane

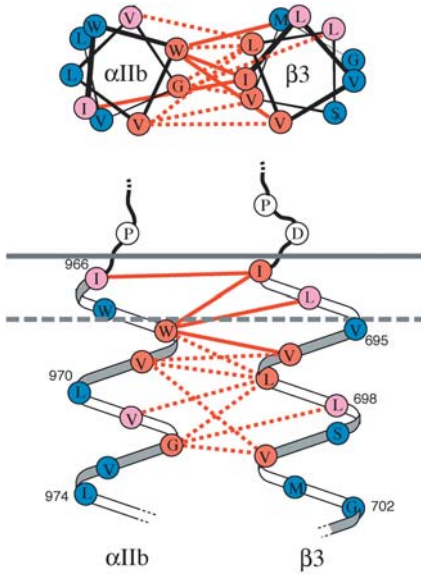
From a biochemical perspective, a living cell is a collection of molecules jam-packed into a confined space by a flexible barrier, called the plasma membrane. A diverse array of proteins embedded in the plasma membrane act as conduits between the cell interior and its external environment, conveying nutrients, metabolites, and information. The life of a cell—as well as that of any multicellular organism—depends on a cell's ability to communicate with its neighbors, both near and far. One way cells do this is with transmembrane receptors outfitted with both extracellular and intracellular domains that mediate information flow between the cell's external and internal environment. One class of transmembrane receptors, called integrin receptors, specializes in interacting with and binding to other cells and the extracellular matrix, a complex of molecules surrounding cells that provides structural support. By integrating various components of the extracellular matrix, integrins (also known as adhesion receptors), play an important role in such diverse processes as cell differentiation, programmed cell death, wound healing, and metastasis.

Integrins can be regulated by signals within the cell to bind to their ligands with either low or high affinity. While a multitude of integrin ligands have been identified and the general mechanics of both the extracellular and intracellular domains of these receptors are known, exactly how a signal crosses the receptor's transmembrane segment to regulate affinity has remained obscure. Now, Bing-Hao Luo, Timothy Springer, and Junichi Takagi have taken a mutational approach to shed light on the inner workings of the transmembrane segment and to explain how it transmits information.

Much of what we know about the function of integrins has come from studying the crystal structures and models obtained from structural analysis. These analyses have generated information not only about the structure and composition of the extracellular and intracellular domains of integrins, but also about the conformational changes that accompany signaling events. Integrins contain a large extracellular

domain, a transmembrane segment, and a relatively short intracellular "tail." Integrins are heterodimers—molecules that contain two subunits composed of different amino acids—made up of an α chain and a β chain. Tight association of the two subunits is associated with an inactive, or low-affinity, state of the extracellular ligand-binding domain. Separation of the intracellular subunits is associated with a dramatic conformational change and activation of the extracellular domain, changing a bent structure with a downward-pointing ligand-binding site into an extended one with an outwardly stretched ligand-binding site. This mechanism differs from most transmembrane signaling molecules, which usually achieve activation through association with their target molecules.

To investigate how the transmembrane segment mediates these changes, Luo, Springer, and Takagi systematically replaced amino acids in both the α and β transmembrane domains of the heterodimer with cysteines, creating the potential for binding interactions through a chemical reaction, disulfide bond formation, between the two subunits. By analyzing 120 possible cysteine pairs, the researchers not only confirmed the structure of the transmembrane region as helical but also mapped the proximal amino acid residues between the helices. To understand how the helical transmembrane domains transmit signals, the team introduced activating mutations in the amino acids of the α subunit cytoplasmic tail. Using this approach, they observed the loss of the contact between the subunits, indicating a separation of the transmembrane helices. Furthermore, when disulfide bond formation occurred, linking the transmembrane segments together, activation was suppressed. While previous models had proposed various modes of subunit movements, including hinge- and piston-like models, these results strongly support the notion that lateral separation of the subunits is the driving force behind the signal. As many diseases arise from defects in integrin adhesion, understanding the conformation and mechanism of integrin activation could suggest promising avenues for drug development aimed at correcting such defects.



Association between integrin α and β subunit transmembrane domains

Luo B-H, Springer TA, Takagi J (2004) A specific interface between integrin transmembrane helices and affinity for ligand. DOI: 10.1371/journal.pbio.0020153

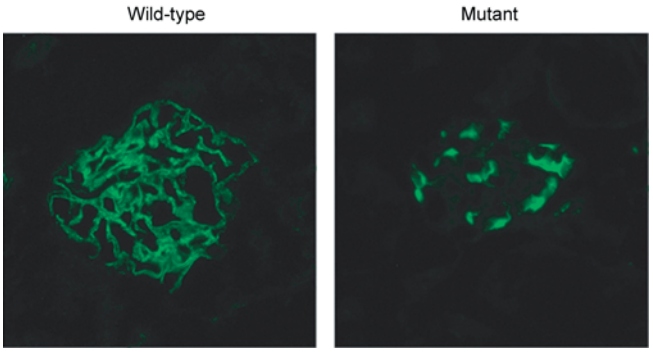
A Protein's Role in Progressive Renal Disease

Focal segmental glomerulosclerosis (FSGS) made a brief media splash last year when a kidney transplant forced NBA superstar Alonso Mourning into early retirement. Mourning's condition elicited a flood of calls from fans offering their kidneys, but most people with kidney disease are not so lucky. Some 56,000 patients await transplants; many have waited over five years. FSGS, which underlies about 25% of the 60,000 kidney-related deaths each year, causes inflammation and irregular scarring in the glomeruli, clusters of blood vessels in the kidney that filter toxins from the blood. These lesions, which allow protein and blood to escape into the urine, cause progressive kidney failure. FSGS commonly occurs as an outgrowth of various primary disorders, including obesity, HIV infection, diabetes, and hypertension. Though it's not clear what causes FSGS, this form of renal pathology is becoming more common. By using the genes underlying inherited forms of FSGS as probes, scientists hope to uncover the mechanisms that unleash the disease and to find ways to stem the damage.

Mutations in the *ACTN4* gene, which encodes a protein called α-actinin-4, cause an inherited form of FSGS. The protein normally remodels actin filaments, the primary structural component of muscle and cytoskeleton. Having a single mutated copy of the gene can cause FSGS in humans, though it is unclear how. In this issue of *PLoS Biology*, Martin Pollak and his colleagues at Brigham and Women's Hospital at Harvard Medical School use a three-pronged approach to figure out how the defective protein wreaks renal havoc and how these physiological changes lead to FSGS. Using biochemical analysis, cell-based studies, and a newly developed "knockin" mouse model, the researchers report that FSGS-related mutations cause α-actinin-4 to engage in various aberrant behaviors that ultimately rob the protein of its function and poison cells.

In previous experiments, Pollak's team had discovered that some families with the inherited form of FSGS carried mutations in *ACTN4*. In these individuals, the disease appeared to strike podocytes (glomeruli epithelial cells) first. While engineering mice designed to carry mutations in this gene, the researchers created mice that lacked detectable *Actn4* expression. These "knockout" mice developed severely damaged podocytes and progressive glomerular disease. In the current experiments, the researchers returned to their "knockin" mice, which carry two copies of the mutation found in the families with inherited FSGS. They also generated "normal" mice and mice harboring one normal and one mutant copy of the gene.

In the biochemical experiments, the researchers investigated the mutant protein's binding behavior. Typically, two α-actinin-4 proteins form a twosome without incident, but here the mutants behaved badly, assuming improper structural conformations and forming aggregates rather than pairs. Next, Pollak and



Mutant α-actinin-4 is mislocalizes and aggregates in renal glomeruli

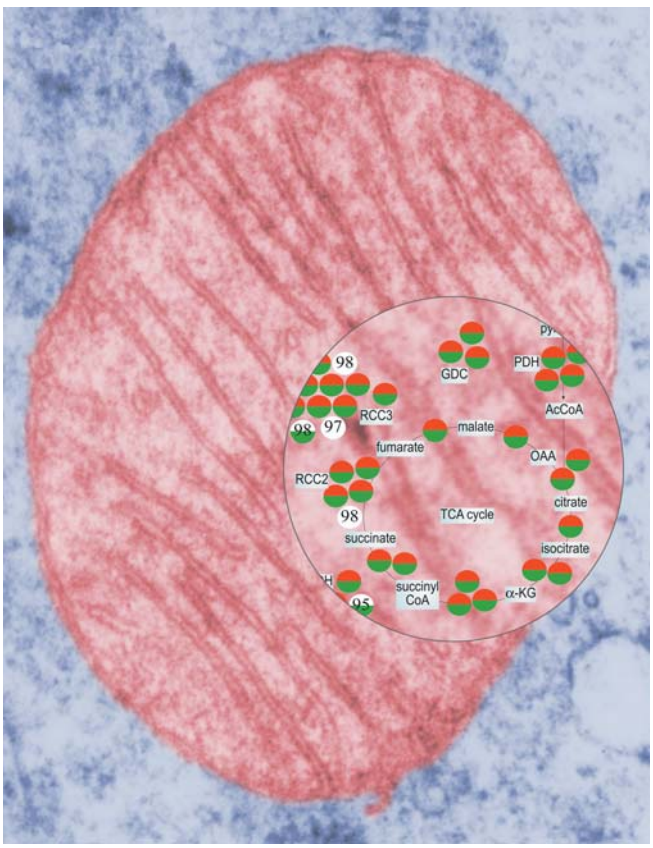
colleagues introduced the genes with the *Actn4* mutation into podocytes, using a variety of methods, to see where in the cells the expressed proteins turned up. They also introduced fluorescently labeled mutant and normal *Actn4* genes into podocytes that were grown from the three mouse types: normal proteins were diffused throughout the cytoplasm in each cell type, but the mutant proteins showed an uneven distribution. Analysis of various tissues taken from the knockin mice revealed normal levels of mRNA transcripts—indicating normal gene transcription—but "markedly reduced" α-actinin-4 protein levels. The mutant proteins, it turns out, were manufactured normally but were degraded far more quickly than normal proteins. Electron microscopy showed that podocytes in the kidneys of the knockin mice had structural defects, while the mice themselves had significantly higher levels of protein in their urine than mice with one or two normal copies of the gene did.

The finding that FSGS-associated *Actn4* mutations produce α-actinin-4 aggregates with significantly reduced life span, the authors explain, suggests two possible mechanisms of initiating disease: aggregation and the toxic affects of aggregation could injure podocytes, or loss of α-actinin-4 function caused by rapid degradation of the protein could produce injury. Pollak and colleagues argue that both factors likely play a role: *Actn4* mutations lead to both reduced α-actinin-4 activity and protein aggregation, and the loss of protein function and the toxic effects of protein aggregation produce glomerular injury. As inherited renal disease typically emerges later in life, it may be that α-actinin-4 aggregation causes incremental but cumulative podocyte damage over time. So what does this mean for patients with progressive renal disease? While these findings may not translate into clinical applications anytime soon, they do suggest that therapies aimed at repairing the structure or expression of these essential cytoskeletal proteins might return the renegade proteins to the fold.

Yao J, Le TC, Kos CH, Henderson JM, Allen PG, et al. (2004) α-actinin-4-mediated FSGS: An inherited kidney disease caused by an aggregated and rapidly degraded cytoskeletal protein. DOI: 10.1371/journal.pbio.0020167

Combining Measures to Characterize Subcellular Machinery

Understanding how the cell functions—or breaks down—implies an understanding of the assembly lines, transportation systems, and powerhouses that keep it running. Global approaches are needed to identify the numerous proteins essential to each cellular machine. But which techniques are best? Lars Steinmetz and colleagues applied and evaluated a variety of methods to define mitochondrial proteins, and report that sets of complementary approaches are needed to characterize a cellular subsystem.



Zooming in on mitochondria (Image by Peter Seibel, design by Shayna Roosevelt)

Yeast mitochondria make an ideal subject for study. About two-thirds of the estimated 700 mitochondrial proteins have been identified to date, leaving fertile ground for new finds. The researchers previously compiled a list of 477 proteins with compelling evidence for mitochondrial involvement. This list provides a well-defined reference set against which to test protein-finding methods. Moreover, the well-studied and accessible yeast genome is well-suited for exploration and genetic manipulation. Since mitochondria are very similar among all eurkaryotes (organisms whose cells have nuclei), the results will prove relevant across species.

Steinmetz and colleagues triangulated results from multiple techniques to identify new candidate mitochondrial proteins. They compared the reference protein list to their new data from protein, mRNA, and gene knockout studies, and to 19 published datasets from other researchers, to evaluate the success of different techniques at finding known mitochondrial proteins. Then they combined evidence across studies to identify a set of proteins that likely characterizes most of the mitochondrial machinery.

The researchers first identified proteins from yeast mitochondria using a technique called liquid chromatography mass spectrometry, which separates the proteins by water insolubility (also called hydrophobicity), then identifies each by the mass and molecular charge of its constituents. By comparing this approach with others, the authors show that this proteomic technique alone is by no means comprehensive, nor error-free. Mass spectrometry is biased toward finding more abundant proteins, and the purified mitochondria can contain contaminants from elsewhere in the cell. To address these issues, the authors compared their protein data with a protein study from another group, the reference protein

set, and a recent subcellular localization study. Potential mitochondrial proteins identified by more than one protein approach were more likely to localize to mitochondria in the localization study than were proteins identified by only one approach. This finding suggests that, compared to either method alone, a combination of protein and localization measures can more robustly identify proteins residing in mitochondria.

But since mitochondria, as the cell's power plants, are integrated into other cellular machinery, the authors argue, methods targeting proteins that are physically located to mitochondria should be complemented with functional approaches. Proteins with mitochondrial roles, regardless of concentration or location in the cell, are better identified by approaches that associate mRNA expression or gene deletion—which removes proteins or renders them inoperable—with changes in mitochondrial function.

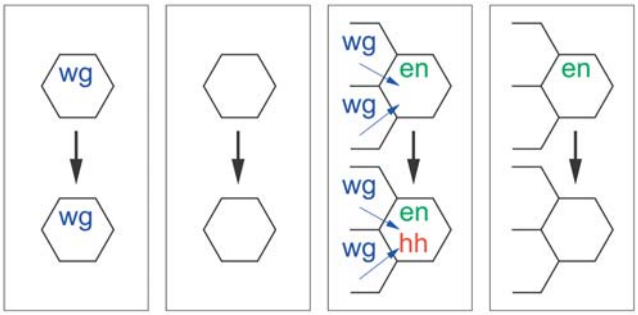
By comparing results from multiple methods against the reference protein list, the researchers evaluated the likelihood that each protein was mitochondrial. They compiled a list of 691 top candidates. This multi-technique analysis easily outperformed any single study in terms of its ability to identify proteins in the reference set, and of the proportion of known versus unconfirmed proteins located. As mitochondria are well-conserved across species, the results provide a candidate gene list for finding human counterparts that might be associated with mitochondrial disorders.

Future studies can use this analysis to evaluate which research methods are likely to be most informative in other cell systems. This paper demonstrates the power of combining techniques with differing strengths in order to zero in on proteins that might elude any single approach, resulting in a more complete parts list for specific cellular machinery.

Prokisch H, Scharfe C, Camp DG II, Xiao W, David L, et al. (2004) Integrative analysis of the mitochondrial proteome in yeast. DOI: 10.1371/journal.pbio.0020160

Computation Approach Shows Robustness of the Striped Pattern of Fruitfly Embryos

Since the days of ancient Greece, mathematics has been used to describe the world in the hopes of identifying underlying laws of nature. Physicists have long relied on mathematics to understand the behavior and interaction of particles too small to observe directly. Since it's not always possible to determine the behavioral properties of a single atom or electron, physicists characterize the behavior of these particles in terms of probability and the law of averages. Likewise, it's not always easy to tell how a single protein contributes to the behavior of a cell or organism. Faced with increasingly immense datasets—from genomes, proteomes, gene expression networks, cell signaling pathways, and more—biologists are turning to the tools of higher mathematics. High-throughput technologies like genome sequencers and microarrays generate a global picture of genomic or cellular activity, but such datasets have a high noise-to-signal ratio—the details are often subject to multiple interpretations. One way computational methods can help separate the signal from the noise is by determining the likelihood of a given set of interactions and presenting a range of possible network behaviors. When sufficient information about a biological pathway is available, experimental evidence can enhance modeling approaches to help refine the nature and role of putative network behaviors. (To learn more about computational biology, see the essay "A Calculus of Purpose," by Arthur Lander, also in this issue of *PLoS Biology*.)



Cell behaviors for segment polarity patterning

One model system ripe for computational analysis is the fruitfly *Drosophila melanogaster*, genetically the best-understood multicellular organism. *Drosophila* development proceeds through a complex series of both sequential and simultaneous events. An elaborate network of genetic interactions transforms a single-cell *Drosophila* egg into a multicellular embryo with 14 discrete segments. These segments are the result of a series of hierarchical decisions, as one set of genes induces the characteristic expression pattern of another set: "gap" genes direct the striped expression pattern of "pair rule" genes, which induce expression of segment polarity genes, whose messenger RNA (mRNA) and protein products produce the characteristic 14-segment polarity pattern. While the molecules and pathways that generate the segment polarity pattern are well known, little is known about the quantitative nature of their interactions: in what concentrations do the components (for example, mRNAs and proteins) exist and what parameters (for example, binding constants, transcription rates, and gene product life spans) govern their interactions?

Four years ago a group of researchers led by George von Dassow developed a model of the genetic interactions that define segment polarity, called the segment polarity network. The model used a parameter set of 48 numerical values for each computer simulation of the segment polarity pattern. Since quantitative information about the network was unavailable, the group used random values for each of the parameters, repeating the simulation for nearly 250,000 different random parameter sets. The model proved remarkably robust—the network output was largely insensitive to variation in parameter values, with a surprisingly large fraction of random parameter sets generating the desired segment polarity pattern. That so many random variables could produce the pattern means either that almost any set of parameter values can work or that only a few of the parameters are important. Now, Nicholas Ingolia reveals the mechanism accounting for this robustness and bolsters the model with recent experimental evidence.

To investigate the reason for the original model's robustness, Ingolia asked whether the parameters of the model could be deconstructed into the properties of individual

cells. It's known, for example, that the stable expression of two genes, called *wingless (wg)* and *engrailed (en)*, within specific cells of a "prepattern" laid down early in embryogenesis is converted into the segment polarity pattern by an intercellular signaling network. *Wg* and *en* operate through positive feedback loops that activate their own expression, a process that is destined to end up with individual cells in one of two stable states of gene expression (an outcome called bistability). Since each stable state is intrinsically robust—that is, resistant to changing parameters—Ingolia hypothesized that the parameters that generate the robustness of the segment polarity pattern in von Dassow's model are those that produce this bistability.

Using computational methods to simulate the behavior of individual cells, Ingolia shows that individual cells in the original model adopt three different stable states of *wg* and *en* expression. The overall pattern of the model, as well as its insensitivity to parameter variation, Ingolia concludes, emerges from the stable expression states of single cells. Parameters that do not produce bistability within single cells, Ingolia found, almost never generate the correct pattern, while those that do produce bistability are much more likely than randomly chosen parameters to generate the striped segment pattern. When Ingolia added new experimental variables to the model—the signaling protein produced by the *sloppy-paired* gene and its interactions with *en*—he could reduce the fraction of parameter sets that satisfied the bistability requirement but nonetheless failed to produce the segment polarity pattern, refining the model to reflect the realities of the cell.

Such computational approaches are allowing biologists to gain valuable insights into the real-world properties and behavior of staggeringly complex biological networks. It's been over 2,000 years since Pythagoras proposed that the laws of heaven and earth reflect a numerical harmony rooted in mathematical laws. Whether that notion holds for biology, bit by bit the tools of higher mathematics are peeling back the layers of complexity to identify underlying properties of living systems.

Ignolia NT (2004) Topology and robustness in the *Drosophila* segment polarity network. DOI: 10.1371/journal.pbio.0020123

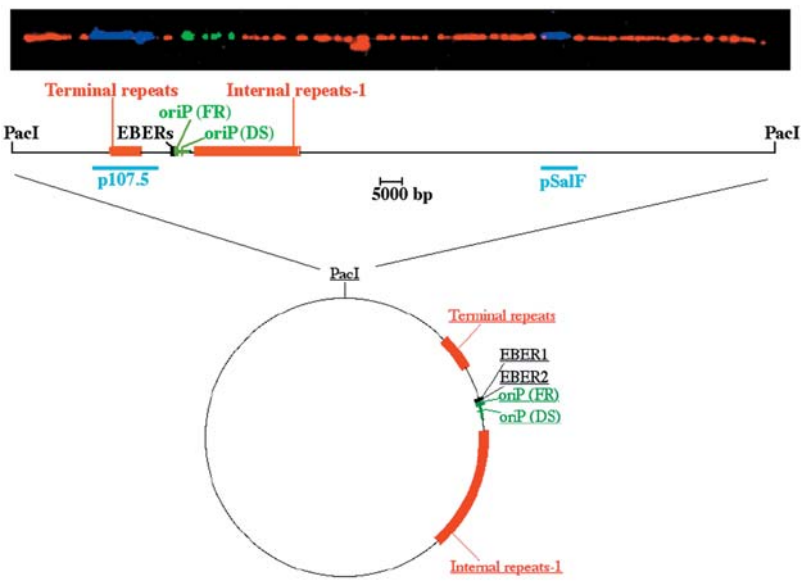
Initiation of DNA Replication: The Genomic Context

Every time a cell divides, it must first duplicate its entire genome. Barring the occasional error, the daughter cells inherit identical copies of the parent cell's genome. With a typical human cell containing almost 9 feet of DNA made of 3 billion base pairs crammed into a nucleus about 5 microns (.0002 inches) in diameter, that's no small feat. To accomplish the job, cells engage specialized teams of protein machines, each performing different tasks during the various stages of DNA replication: initiation, duplication, quality control, and repair. Much of what we know about the molecular mechanisms of DNA replication comes from studies of bacteria. In the bacterial genome, which consists of several million base pairs, replication begins at a single site, spanning about 100 base pairs. The regulation and mechanisms of replication, even in the compact bacterial genome, are so complex that 51 years after Watson and Crick reported that the structure of DNA "immediately suggests" a mechanism for its replication, biologists are still working out the details and regulation of that mechanism.

Before duplication, aptly named initiator proteins bind to DNA at replication initiation sites and break the bonds holding the complementary base pairs together, separating the double helix locally into single strands and creating two Y-shaped junctions at either end called replication forks. At each replication fork, a complex of proteins continues the business of unzipping the DNA and using the exposed single strands as templates to generate complementary daughter strands. What controls when and how individual initiation sites are activated in mammalian cells has remained obscure. Is initiation restricted to specific sites? Do specific DNA sequences control initiation events locally? Examining individual molecules of fluorescently labeled replicating DNA, Paolo Norio and Carl Schildkraut report that initiation events are not controlled by individual initiation sites but occur throughout the genome. And the activation of these sites appears to depend on what's happening at the genomic level.

Using a novel technique called single molecule analysis of replicated DNA (SMARD), Norio and Schildkraut use the Epstein Barr virus (EBV) in human B cells as a model system for studying DNA replication. During the latent stage of infection, the EBV genome exists as an episome—a circular piece of extrachromosomal DNA. It replicates only once per cell cycle, during the DNA synthesis stage, and uses its host's replication machinery to do so. Using nucleotide analogs that can be detected by immunofluorescence (since the analogs attract antibodies that are fluorescently labeled), the researchers can determine the position, direction, and density of the replication forks, and then determine how replication starts, progresses, and terminates.

Norio and Schildkraut studied replication using two strains of the EBV virus grown in human B cells, their natural target. Previous studies, which had largely focused on the



Linearized EBV episome imaged by fluorescent microscopy and aligned with the corresponding genomic map

activity of individual initiation sites, had suggested that different EBV strains vary in how initiation sites are activated and that specific initiation sites or regions likely regulate replication. Looking at larger genomic regions, Norio and Schildkraut found something different: not only do initiation sites occur throughout the genomes, but their activity "differs dramatically" in the two EBV strains and even within a strain. Differences were seen in the order of initiation site activation, in the direction of replication fork movement, and in the speed of duplication in different parts of the genome. While the two largely similar viral genomes do show some genetic differences, the authors dismiss the idea that these local differences could explain the observed variations in replication control. It's more likely, they conclude, that epigenetic modifications (such as changes in chromatin structure) produce the differences in the order and frequency of activation of initiation sites across genomic regions.

It seems that initiation events are not restricted to specific genomic areas, and experimentally induced loss of individual initiation sites does not significantly affect EBV genome replication (because other sites take up the slack). This redundancy provides flexibility in determining which sites are activated. Since the EBV genome uses human replication machinery to duplicate its genome, these findings likely apply to DNA replication in mammalian cells as well. The very survival of the cell—and the health of the organism it inhabits—depends upon the faithful replication of the genome. Using processes that operate at the genomic level may afford cells the means to manage an unwieldy genome, and perhaps, more importantly, guarantee their genes safe passage to the next generation.

Norio P, Schildkraut CL (2004) Plasticity of DNA replication initiation in Epstein-Barr virus episomes. DOI: 10.1371/journal.pbio.0020152

Turning Down the Volume: Why Some Genes Tolerate Less Noise

All organisms have evolved complex mechanisms designed to exquisitely regulate the expression of appropriate genes at their correct levels. Natural random variation in the processes of regulation and expression, however, limits the precision with which protein production can be controlled. This subtle variation, or "noise," in the expression of genes has been studied with increasing interest. Though much progress has been made in understanding the amount of noise that exists and the cellular processes that underlie it, the physiological impact of noise, and whether it is biologically relevant or can just be ignored, has been less clear.

To answer this question, Hunter Fraser et al. asked whether noise in gene expression exerts an equal effect on all genes in the genome. Is noise in gene expression irrelevant to the fitness and well-being of cells, or do cells need to minimize noise in the expression of some or all genes? If noise in gene expression has a negative impact on cells, they reasoned, that impact should vary from gene to gene depending on the gene's function. There should be selection to minimize noise for those genes most crucial to cell survival and function. Thus, a genome-wide analysis of noise in gene expression, they predicted, would show that genes for which "noisy" expression would be most harmful would display less of it.

The researchers examined this question in the budding yeast *Saccharomyces cerevisiae* because of the vast quantity and variety of genomic data available for this organism. Previous research has shown that the noise that exists in the expression of a gene is directly related to the rates of transcription and translation. Using data available from previous genome-wide studies, the authors were able to estimate these rates, and therefore the noise, for nearly every gene in the yeast genome.

After estimating the amount of noise in the expression level for nearly every gene, the authors examined two subsets of genes that they hypothesized would be particularly affected by noise. First they looked at "essential" genes, reasoning that since total lack of expression of these genes results in death, even small variations in expression resulting from noise would often exert a negative impact. Previous research had identified all the essential genes in yeast by deleting each gene individually and assessing the fitness of the resulting mutant. Here, the authors compared the levels of noise in this pool of essential genes to that of nonessential genes. They found that essential genes usually display less noise than nonessential genes, lending support to their hypothesis.

"Shhhhhhhhhhhhhh!" (Photo by Bryan Zeitler and Jennifer Zeitler)

They similarly examined genes encoding proteins involved in forming multiprotein complexes. Because these complexes are built of proteins in specific ratios, over- or underexpression of one component will hinder the accurate assembly of productive complexes. So a high degree of noise could interfere with the coordinated expression necessary for proteins involved in these complexes. The authors again used data from previous research to choose members of this group: they relied on two studies which had identified a large number of multiprotein complexes in yeast. Using this group of genes for comparison, the authors found that, like essential genes, genes encoding proteins involved in multiprotein complexes generally display less noise than other genes.

This study draws a simple but fundamental conclusion about noise in eukaryotic gene expression—noise has physiological consequences. Importantly, the fact that noise is minimized in those gene groups for which noisy expression would be most harmful suggests that factors contributing to noise are subject to natural selection. This study also demonstrates the power of using the growing number of genome-scale datasets in this type of analysis. Researchers will undoubtedly continue to mine the available data to draw biological conclusions not anticipated by the original authors.

Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB (2004) Noise minimization in eukaryotic gene expression. DOI: 10.1371/journal.pbio.0020137

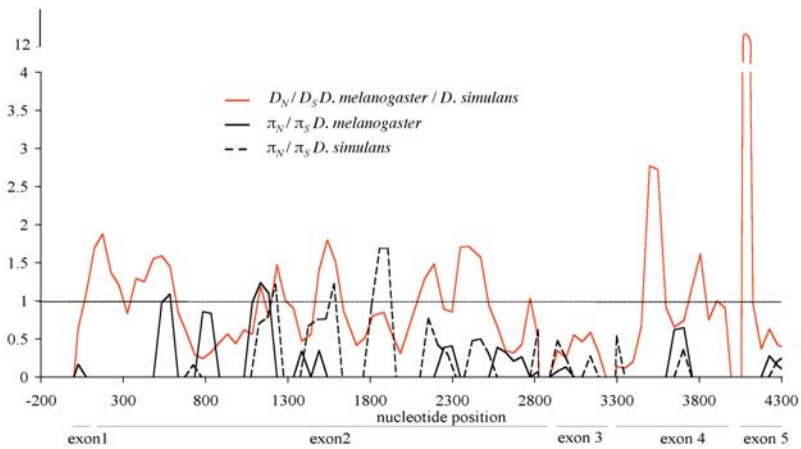
A Gene Responsible for Hybrid Incompatibility in *Drosophila*

Nearly 150 years after Darwin published *On the Origin of Species*, biologists are still debating how new species emerge from old—and even the definition of species itself. Darwin demurred from offering a hard and fast definition, suggesting that such a thing was "undiscoverable." One of the more enduring definitions characterizes organisms as distinct reproductive units and species as groups of individuals that can interbreed and produce viable, fertile offspring. The lack of genetic exchange between species, called reproductive isolation, lies at the heart of this definition. Environmental changes can create physical barriers between populations that preclude mating between the populations. Reproductive isolation can also involve changes at the genetic level, when molecular barriers prevent two recently diverged populations from producing viable or fertile offspring. Such factors limit gene flow between diverging species and allow the emergence of genetically novel yet sound populations—that is, new species.

At the heart of reproductive isolation is a phenomenon called hybrid incompatibility, in which closely related species are capable of mating but produce inviable or sterile offspring. The classic example of hybrid incompatibility is the male donkey–female horse cross, which yields a sterile mule, but many other cases have been documented among mammals, and thousands of plant crosses produce infertile offspring. Much has been learned about the genetic architecture of hybrid incompatibility by studying the offspring of closely related, or "sibling," fruitfly species in the lab. Sibling species are morphologically very similar, or even indistinguishable, but typically do not interbreed in nature. In the lab, their offspring are either sterile or inviable, a fate that varies depending on the gender of the offspring and species of the parents. To elucidate the molecular mechanisms of reproductive isolation, biologists must first identify candidate hybrid incompatibility genes. Species- or lineage-specific functional divergence is an essential trait of these genes. (That is, the genes evolve different functions after the species diverge from their common ancestor.) While several such candidate genes have been identified in the fruitfly *Drosophila melanogaster*, none has been shown to display this functional divergence. Now, working with *D. melanogaster* and its sibling species *D. simulans* and *D. mauritiana*, Daniel Barbash, Philip Awadalla, and Aaron Tarone establish the functional divergence of a candidate hybrid compatibility gene and confirm its status as a true speciation gene.

Since the 1930s, investigations of reproductive isolation have been guided by the Dobzhansky-Muller model, which attributes hybrid incompatibility to the interactions between two or more genes that have evolved independently in two isolated populations. These independently evolving genes diverge functionally, and the interactions of these functionally divergent genes in a hybrid individual are responsible for the defective phenotypes observed (either inviability or sterility). If this is the case, the alleles, or versions, of the gene causing hybrid incompatibility should have distinct phenotypes in the two species. A corollary of the model says that the diverged allele *(A)* and not the ancestral allele *(a)* causes the incompatibility phenotype, which means that experimental manipulations of *A* but not *a* should affect the hybrid incompatibility phenotype.

Barbash et al. tested the model's predictions by genetically manipulating the alleles of the *Hybrid male rescue (Hmr)* gene from each sibling species and observing the mutations' effects on the flies' hybrid offspring. In previous experiments the researchers had shown that loss-of-function mutations in the *D. melanogaster Hmr* gene "rescue" hybrid individuals from the hybrid incompatibility phenotype (male inviability) typically observed in the offspring of crosses between *D. melanogaster* and its sibling species, and that increased *Hmr* activity suppresses rescue and kills hybrids. If *D. melanogaster Hmr* has functionally diverged between the species, then transgenes containing *Hmr* from sibling species should not cause the hybrid incompatibility phenotype caused by the *D. melanogaster Hmr*. The researchers tested this hypothesis by introducing transgenic *Hmr* genes from sibling species into *D. melanogaster*. In all cases, the hybrid male offspring of *D. melanogaster/D. mauritiana* and *D. melanogaster/D. simulans* crosses "were at least as viable as their brothers without the transgene."



Multiple regions of *Hmr* show evidence for divergence driven by positive selection

To examine this divergence at the genomic level, Barbash et al. compared the divergence of 250 genes in *D. melanogaster* and *D. simulans* and found that the *Hmr* gene was among the most rapidly evolving genes. By examining the frequency of mutations that have accumulated between *D. melanogaster* and sibling species relative to the number of mutations accumulated within species, the authors show that the mutations between species were by and large not neutral and that they occurred after *D. melanogaster* diverged from its sibling species, indicating that the gene has been under positive natural selection. Barbash et al. have not only identified a bona fide speciation gene by demonstrating its functional divergence, they've also created a platform for investigating the mechanisms through which such genes cause hybrid incompatibility and lay the groundwork for speciation.

Barbash DA, Awadalla P, Tarone AM (2004) Functional divergence caused by ancient positive selection of a *Drosophila* hybrid incompatibility locus. DOI: 10.1371/journal.pbio.0020142

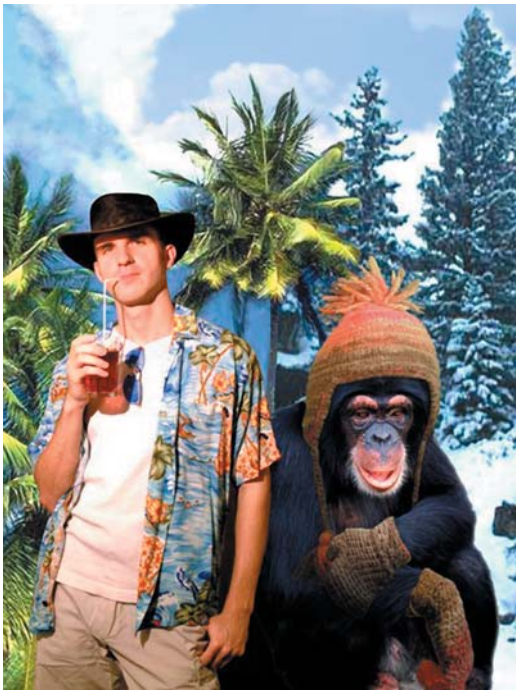
A DNA Recombination "Hotspot" in Humans Is Missing in Chimps

When Francis Collins and Craig Venter reported the draft sequence of the human genome in 2001, Collins described the so-called book of life as more of a life sciences encyclopedia. In it, we can find our evolutionary history written in the fossil record of our DNA, a parts manual listing the genes and proteins needed to build and operate a human being, and a medical text, gleaned from the genetic variants linked to human disease. Unfortunately, he added, the texts are written in a language "we don't entirely know how to read yet." Since then, biologists have made great progress in extracting meaning from the human genome. Humans are 99.9% alike genetically, and that 0.1% makes all the difference in terms of appearance, personality, and susceptibility to disease. That 0.1% promises to shed light on the evolutionary forces that control genetic variation as well as the genetic origins of human disease.

Very small genetic variations—including differences of a single DNA base, called single nucleotide polymorphisms, or SNPs—occur through random mutations. Individuals have two of each chromosome (one from the mother and one from the father), and the combination of SNPs found together on one chromosome can change through the random shuffling of genetic material between the two chromosomes when sperm and egg cells are produced during meiosis. By studying the location and frequency of this reassortment in the genome, biologists hope to understand how recombination affects the overall pattern of SNP variation and how these patterns relate to human disease. An international collaborative effort called the HapMap Project aims to identify the most common SNP associations within chromosomes, known as haplotypes, and then determine which haplotypes are associated with disease. This approach relies on what's known as "linkage disequilibrium"—the nonrandom association of alleles (gene variants) at different locations on a chromosome—to facilitate their search for candidate disease genes. Adjacent SNPs show strong linkage disequilibrium, which means that researchers can select a limited number of SNPs as markers for a haplotype and test their association with disease rather than testing each SNP.

Patterns of linkage disequilibrium depend on the rate of recombination—higher recombination rates typically cause less linkage. Demographic factors and chance also affect levels of linkage disequilibrium; while both vary across populations, it has been thought that recombination rates do not. Recombination appears to favor specific genomic regions, termed hotspots, but the observation that the recombinant chromosomes are not passed down in equal proportions suggests that recombination hotspots may be short-lived, appearing as transient blips on the evolutionary radar. Exploring this possibility, Susan Ptak et al. compared a well-studied recombination hotspot in humans, called TAP2, with a similar region in our closest evolutionary cousins, chimpanzees, to see whether they are similarly endowed.

Since recombination occurs relatively rarely, researchers have relied largely on indirect methods to determine regional recombination rates. Though recent advances have made sperm analysis in humans more practical (though still technically challenging), such techniques are less feasible with chimps because collecting large amounts of sperm from individual males might compromise their success in mating competition or reduce the genetic diversity of endangered chimp populations. Here, the researchers used an indirect approach to estimate recombination rates from the patterns of linkage distribution, which "reflect the rate and distribution of recombination events in the ancestors of the sample." They focused on chimps from a single subspecies because the reported high level of genetic differentiation between subspecies could skew estimates of recombination rate variation. Analysis of the TAP2 region revealed 47 SNPs in the human and 57 in the chimp, with an overall lower level of linkage disequilibrium in humans: strong linkage was seen only in adjacent pairs of SNPs in humans, but was found in both adjacent and more distant pairs in the chimps. Using a statistical approach to characterize recombination rate variation between the two species, Ptak et al. found "extremely strong support" for rate variation in humans but found strong evidence against such variation in chimps.



The TAP2 region harbors a recombination hotspot in humans. What about in chimpanzees: hot or cold?

Humans and chimps diverged from a common ancestor five to six million years ago and differ at only 1.2% of base pairs on average. That the recombination hotspot does not exist in both species suggests that hotspots are not stable and can evolve fairly quickly. If recombination rates within a small genomic area—at the level of a few thousand bases—can change in such a short time frame between such closely related species, Ptak et al. reason, they may do so within species, too. Such a prospect has important implications for the HapMap Project and disease association studies that rely on linkage disequilibrium. While haplotypes offer a shortcut for identifying candidate disease genes based on typing a certain number of markers, the number of markers required depends on the strength of linkage disequilibrium. If recombination rates differ across human populations, as these results suggest, then the strength of linkage disequilibrium will too—which means that association studies might need to adjust the number of markers needed to flag candidate disease genes in different populations.

Ptak SE, Roeder AD, Stephens M, Gilad Y, Pääbo S, et al. (2004) Absence of the TAP2 human recombination hotspot in chimpanzees. DOI: 10.1371/journal.pbio.0020155

Annotation Marathon Validates 21,037 Human Genes

Scientists at the annotation "marathon" of 41,118 cDNA clones

The announcement of the human genome sequence three years ago was widely hailed as one of the great scientific achievements in modern history, and with good reason. Determining the structure and nature of the genetic code promises to provide valuable insights into human evolution and the molecular basis of disease. But sequencing the genome is just the first step toward this decidedly worthy goal—the monumental task of ascribing biological meaning to those sequences has just begun. And while researchers know a great deal about some of the 30,000 or so genes in the human genome, they have yet to ascribe function to the majority of them. Takashi Gojobori and a large international team of collaborators have now taken a big step toward narrowing this knowledge gap.

Deciphering the human genome presents such a daunting challenge in part because it's so huge, making it difficult to distinguish genetic signal from noise. Simpler organisms have much more compact genomes. In the case of brewer's yeast, for example, genes that encode proteins account for about 70% of the genome. In contrast, only about 1% to 2% of the human genome codes for proteins. That translates to about one gene for every 2,000 bases for yeast compared to about one gene for every 150,000 bases for humans.

The low density of human genes makes identifying them difficult enough, but this process is further complicated by how genes are organized in the human genome. The functional parts are broken up into smaller segments called exons, which are separated in the genome by intervening sequences called introns. This configuration also occurs in simpler organisms, but since the number and size of introns is relatively small in simpler organisms, it's easier to tell what's a gene and what isn't. In humans, the introns are extremely long, as are the gaps between the genes, and the exons are tiny in comparison; plus, it takes many more of these short, scattered exons to make one gene.

One approach to this problem is to use computer algorithms that scan the genome sequence looking for segments of DNA sequences that could potentially encode proteins. Gojobori and colleagues, however, used a different approach. They analyzed the sequences of 41,118 full-length cDNAs available from six sequencing centers around the world. These cDNAs are stretches of DNA that represent genes that have already been expressed and used by the cell for protein production. Since all the exons have been spliced together and the introns removed, these cDNAs correspond to the functional versions of these genes, allowing researchers to work backward, looking for the sequences in the genome.

In order to process the 41,118 cDNAs, the researchers used a combination of computer algorithms and expert human analysis. To tackle such an enormous project, 158 genome scientists, representing 67 institutions from 12 countries, gathered in Japan in the summer of 2002. Over the course of a ten-day annotation marathon, the scientists validated, mapped, and annotated the cDNAs. As things stand, the team has been able to assemble the cDNAs into over 20,000 strong candidates for human genes.

From just the initial analysis of the data generated by this group, several valuable findings about the human genome have emerged: there are over 5,000 candidates for new genes, including an exciting group of several hundred that do not appear to encode proteins; up to 4% of the genome appears not to be represented in the current human genome sequence; and several thousand DNA sequence variants have been uncovered that will be useful for disease mapping studies.

But perhaps most important of all, the data from this study have been collected and assembled into a large searchable database called H-Invitational Database, which is linked to other functional databases around the world. This will be an invaluable resource for geneticists, and will serve as a starting point for further analyses. Future research on the human genome will be aimed at expanding the list of known genes and analyzing the properties of these genes. This study not only moves us closer to a complete functional description of the human genome, it also builds on the traditions of international cooperation and large-scale collaboration that played such an important part in deciphering the sequence itself.

Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, et al. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. DOI: 10.1371/journal.pbio.0020162