Cell

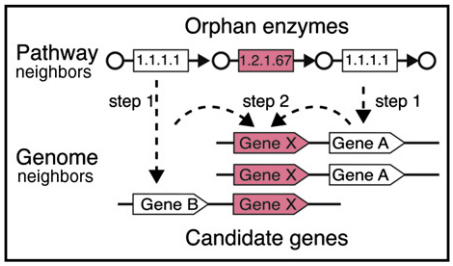
# Genomics: Think Global, Act Local

Long a slogan for environmentalists, "think global, act local" could be a new rallying cry for biologists. As genome-wide techniques advance and their costs drop, scientists are expanding into larger and larger territories—metagenomics, global proteomic approaches, and analyses of thousands of genomes. These massive data sets are opening up new possibilities for understanding some of the smallest details of the genome. Here, we look at four such cases—investigating the evolutionary role of insertions and deletions in the genome, connecting an orphan enzyme with its gene, mapping the fine details of chromatin structure, and characterizing global interactions between proteins and RNA—all of which depend on a combination of global thinking and local action.

## A Genome-wide Strategy to Create High-Payoff Mutations

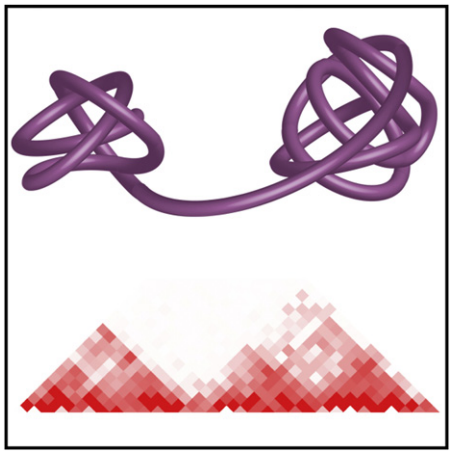
Genomic regions of structural variation, typically hundreds to thousands of base pair insertions and deletions, are frequently associated with disease through unknown mechanisms. These areas do not occur randomly throughout the genome but rather cluster together in ''hot spots'' of hypermutability. Previous studies in human cancer cells found that regions of hypermutability often overlapped with regions of DNA hypomethylation. Now, a study by Li et al. (2012) combines these two observations and demonstrates that methylation deserts (stretches of DNA hypomethylation) in the germline demarcate hot spots of structural variation, generating rapidly evolving regions of the genome. By analyzing the methylomes from four human sperm samples, they find significant overlap between hypomethylation and recent structural variants, which were identified through evolutionary and population genetic studies. Li et al. also examine samples from patients with various neurological disorders associated with structural variation and demonstrate that DNA hypomethylation is significantly associated with the etiological mutations. Li and colleagues postulate that the correlation between DNA hypomethylation and hypermutable regions represents an epigenetic mechanism that has evolved to select areas of the genome that may generate potentially beneficial mutations with a high pay-off; the regions susceptible to this process will necessarily also be high-risk regions, and thus many mutations will lead to disease, but some will lead to significant gains in fitness. For example, genomic regions related to brain development are likely candidates for this type of hypermutability. This hypothesis is supported by their finding that methylation deserts are associated with de novo mutations causing schizophrenia, bipolar disorder, and autism. To further test this hypothesis, however, the underlying mechanism for establishing these hot spots will need to be discovered.
*Li, J., et al. (2012). PLoS Genet. 8, e1002692.*

## Welcome to the Neighborhood: Identifying Genes for Orphan Enzymes




**Genomic neighbors of the pathway neighbors serve as possible candidate sequences for orphan enzymes. Image courtesy of Takuji Yamada.**

A sequenced genome may be the blueprint of an organism, but it often comes up short in identifying the tools used to realize that blueprint—enzymes. Orphan enzymes, which have known biochemical functions but unknown genes, often represent crucial intermediate steps in metabolic pathways. Identifying their genes is a necessary first step for metabolic engineering, drug design, and proteomic approaches. Yamada et al. (2012) now develop a computational approach to tackle this problem by using metabolic pathway information about orphan enzymes and prokaryotic genomic and metagenomic sequences. Their algorithm relies on the observation that genes involved in the same metabolic pathways are often physically close to one another in prokaryotic genomes. This generates genomic neighborhoods that can then be searched in more detail for the gene encoding the orphan enzyme. Yamada et al. validate their approach by expressing candidate sequences, purifying the proteins, and testing them with the appropriate enzymatic assays. Their method also predicts functions for protein domains of previously unknown functions. Finally, Yamada and colleagues assess how their predicted genes alter global metabolic models. They find that many orphan enzymes represent new reactions; this alters the predicted essentiality of certain genes in metabolic models. One drawback to this new approach is that it requires an orphan enzyme to have some known pathway neighbors a priori. As more data become available from genome-wide functional analyses, such as gene lethality screens, this limitation may be overcome through greater reliance on genomic neighborhood information.
*Yamada T., et al. (2012). Mol. Syst. Biol. 8, 581.*



Schematic illustrating topological domains (top) and the underlying Hi-C data represented as the frequency of chromatin interactions. Image courtesy of Jesse Dixon.

## Eminent Domain

New techniques to determine the 3D structure of the genome, such as chromosome conformation capture (3C) and Hi-C, have provided an increasingly detailed picture of chromatin structure. Dixon et al. (2012) now take this method to the next level with their comprehensive, high-resolution Hi-C analysis of mouse embryonic stem cells (ESC), human ESC, and human fibroblasts. In all three cell types, more than 90% of the genome exists in self-interacting regions of megabases, which they term topological domains. These domains are conserved between cell types and species, suggesting that this organization is a universal feature of metazoans. As expected, the boundaries of these domains are enriched with CTCF, a factor known to bind to insulator elements. However, the boundaries are also associated with other factors, such as housekeeping genes and proteins found at active promoters and gene bodies. These observations suggest that topological domains are generated, in part, by transcriptional activity. Interestingly, Dixon and colleagues also find that high levels of the heterochromatin mark H3K9me3 were prevalent in differentiated cells, forming large domains bounded by the topological domain boundaries, but they do not observe this in ESC, implying that the boundaries are established prior to the formation of heterochromatic regions and may serve as guides for later chromatin changes. The functional significance of this chromatin organization is still not well understood, and expanding this type of analysis to other cell types, including tumor cells, will help to determine the role of topological domains in both normal cellular processes and disease states.

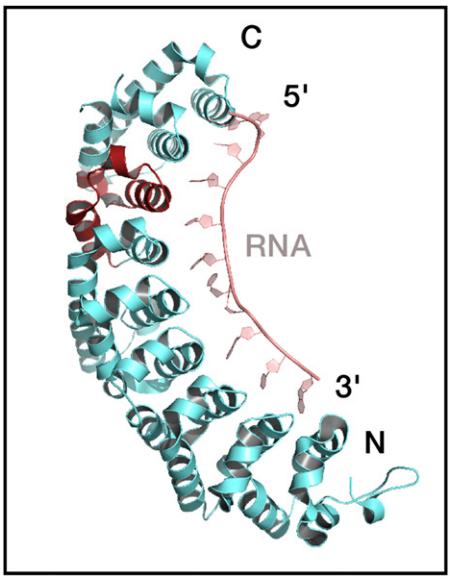
*Dixon J., et al. (2012). Nature 485, 376–380.*

## A Global View of RNA-Protein Interactions

Although new methods have been developed for large-scale characterization of RNA-protein interactions, many of these approaches are time-consuming and biased towards only the strongest interactions. Now, Campbell et al. (2012) introduce a new method, SEQRS, which speeds up and refines the global identification of protein-RNA interactions. SEQRS combines in vitro selection, high-throughput sequencing, and bioinformatic analysis to comprehensively characterize protein-RNA-binding events. Applying SEQRS, Campbell et al. find that the affinity of a protein for a particular RNA sequence can be modified by the presence of a second protein, as has been shown for DNA-protein interactions. They propose that some RNA-binding proteins can exert their regulatory influence without binding RNA at all but rather by modifying the interaction between a second protein and RNA. Campbell and colleagues also use SEQRS to identify new RNA regulatory elements and quantitatively assess the relative strengths of binding to these motifs. Finally, they apply their method to characterize possible off-target binding of RNA-binding proteins engineered to recognize a novel RNA sequence and observe that these designer proteins are highly specific, which is encouraging for the development of RNA-binding proteins as therapeutic molecules. Ultimately, the SEQRS results will need to be verified in vivo, but clearly this rapid, comprehensive approach will serve as an excellent starting point for deeper investigations into protein-RNA interactions.

*Campbell Z., et al., (2012). Cell Rep. 1, 570–581.*

**Rhiannon Macrae**




Global analysis of binding specificity for normal and mutant RNA-binding proteins using the SEQRS method. Image courtesy of Zachary Campbell and Marvin Wickens.