CellPress

Comparative population genomics: power and principles for the inference of functionality

David S. Lawrie[1,2] and Dmitri A. Petrov[2]

[1] Department of Genetics, Stanford University, Stanford, CA, USA
[2] Department of Biology, Stanford University, Stanford, CA, USA

The availability of sequenced genomes from multiple related organisms allows the detection and localization of functional genomic elements based on the idea that such elements evolve more slowly than neutral sequences. Although such comparative genomics methods have proven useful in discovering functional elements and ascertaining levels of functional constraint in the genome as a whole, here we outline limitations intrinsic to this approach that cannot be overcome by sequencing more species. We argue that it is essential to supplement comparative genomics with ultra-deep sampling of populations from closely related species to enable substantially more powerful genomic scans for functional elements. The convergence of sequencing technology and population genetics theory has made such projects feasible and has exciting implications for functional genomics.

Sequence constraint: the key to searching for function in the genome

Comparative genomics uses the pattern of evolutionary conservation in aligned sequences between species to detect functional elements [1]. The rationale for this approach is that many mutations in functional sequences should be deleterious and thus weeded out of the population by purifying selection. This in turn should generate the canonical signature of sequence conservation between species: a lower rate of substitution at functional sites than that at neutrally evolving, non-functional sites.

Methods based on this principle have been successful in locating previously unidentified functional elements, illuminating the evolutionary history of known functional elements, and estimating the percentage of functional sites in a genome [2–6]. This final application has been the topic of recent controversy, particularly in relation to what percentage of the human genome is functional [7–10]. Methods couched in comparative genomics typically predict that ~5% of sites in the human genome are functional [6,7,11,12]. In stark contrast, experimental evidence from the Encyclopedia of DNA Elements (ENCODE) consortium indicates that anywhere from 20 to 80% of the human genome appears to participate in some sort of biochemical activity [8,13]. This difference likely indicates that not all biological activity is relevant to the biological function of the organism, and underscores the key advantage of

Glossary

Polymorphism: a new mutation in a population creates a 'polymorphism', a genetic variant that is present in some but not all individuals. In the case of a base-pair mutation, this is known as a single nucleotide polymorphism (SNP). A measure for the amount of expected polymorphism in a population is θ, the population-level mutation rate, which is equal to $4N_e\mu$, where N_e (the effective population size) is how many independent lineages exist in the current population, and μ is the per-site, per-lineage mutation rate. The expected number of neutral polymorphic sites, the density of polymorphism, seen in a sample of individuals from a population is determined by θ and by the number of individuals sequenced from the population, the sample depth.
Substitution: if a new mutation rises to 'fixation' in the population such that every member of the population shares that mutation, then it has become a fixed difference (substitution) between that population/species and another. The accumulation of fixed differences can be used as a proxy for the amount of time since the last common ancestor of two species.
Effective selection: the effective selection coefficient measures how much the trajectory of a mutation in the population is controlled by random genetic drift or by deterministic selection – the higher the absolute value of the coefficient, the more the probability that a mutation will become fixed is driven by selection. A neutral mutation has a coefficient of 0. For diploid organisms, the effective selection coefficient is four times the effective population size (N_e) multiplied by the selection coefficient (s): $4N_es$. The selection coefficient measures the fitness dis/advantage of one mutation relative to another. We define weak selection $|4N_es| < 5$, moderate as $5 < |4N_es| < 20$, and strong as $20 < |4N_es| < \infty$. Lethal mutations have effectively infinite selection acting against them. Other papers may use different classifications.
Confounding factors: many factors other than selection on the sites themselves can skew a site frequency spectrum (SFS) such as linked selection, mutation rate, biased gene conversion, and demography. Linked selection can be the effects from nearby adaptive mutations rising quickly to fixation, known as a selective sweep, or from purifying selection removing nearby deleterious alleles from the population linked to the site of study. Different sites have different mutation rates not only based on location in the genome but also on their (and that of their neighbor's) base-pair composition. Biased gene conversion is similar to natural selection mathematically, but is actually the result of a combination of mismatch repair that is biased in favor of some nucleotides compared to others and strand invasion during recombination that generates mismatched heteroduplexes when recombination occurs at a heterozygote site. Demography is the natural history of the population (e.g., population size changes, population substructure, migration, etc.) and can effect the expected SFS on a genome-wide scale.
Likelihood: hypothesis testing relies on the difference in maximum likelihood of two statistical models to explain the data: the null model is the hypotheses being tested against and the alternative model being tested for. Whether the null hypothesis is rejected depends on the difference in likelihoods between the two models and the chosen significance level.

Corresponding author: Lawrie, D.S. (dlawrie@stanford.edu).

0168-9525/$ – see front matter
 http://dx.doi.org/10.1016/j.tig.2014.02.002

evolutionary approaches for predicting functionality: by focusing on the fitness effects of mutations, they focus on functionality as it relates to the overall performance of the organism as noticed by natural selection.

However, divergence tests rely on a large number of related species for their power to detect individual functional elements, which implies that they have the greatest power to detect functional elements maintained over a long period of evolutionary history. Any relaxation of constraint or recently arisen functionality in the organism or small clade of interest will limit the power of such methods to detect functional elements. Indeed, evidence suggests that, as one considers more closely related species, estimates of the percentage of conserved sites increase even if the power to detect individual elements decreases [14,15]. Conservation as a signal of functional constraint thus suffers from some drawbacks that cannot be ameliorated by increasing the number of analyzed species.

Polymorphism within a species offers a more recent snapshot of the evolutionary history of a population. The obvious advantage is that selection need not be present over long evolutionary history to be discovered and, as expected, estimates of functionality in the human genome rise another 4% using polymorphism data [11]. Moreover, models using polymorphism data can deliver a more fine-grained picture of the functional importance of sites in the form of a detailed distribution of fitness effects (DFE) [16–18]. To calculate the DFE, mutations are binned according to their frequency within the sampled population. The resulting histogram, known as the site frequency spectrum (SFS), can be used to determine the fraction of sites evolving under a given strength of selection. The key shortcoming of this approach is a lack of resolution due to the usually low levels of polymorphism within a single species. Without enough polymorphism to provide statistical power, the DFE and therefore functionality can only be determined for large, coherent groups of sites subject to *a priori* similar selective pressures, such as all synonymous sites in a region, but unfortunately not for single sites.

In this Opinion paper we suggest that the development of new approaches combining comparative genomics with ultra-deep population sampling within multiple closely related species should provide much additional power and precision in the study of genomic functionality. We argue that such a unified approach will allow us to ameliorate the problems inherent in both divergence- and polymorphism-based methods.

Comparative genomics

The neutral theory of evolution (Box 1) stipulates that functional regions of the genome should evolve more slowly than neutral regions. For a given sequence alignment X between two species, 'A' and 'B', separated by time t_0 in neutral regions, one can infer the expected number of substitutions that occurred given a substitution model (see [19] for more details). If the inferred t is less than t_0 then the rate of evolution, r, for those sites is less than r_0 and the region is marked as conserved and under purifying selection.

This framework can be extended to multiple species over a phylogeny (Box 1: Comparative genomics). Such methods are known as 'phylogenetic footprinting' because the functionality of a genomic element should leave a 'footprint' of conservation on the evolutionary history of that element. More species add more power to differentiate functional from neutral elements by adding both more information content from the sequence alignment and by increasing the total branch length of the tree.

The logic of the neutral theory is generally interpreted to mean that natural selection should simply reduce the overall rate at which substitutions occur along the phylogenetic tree. However, one can also model selective constraint explicitly by assigning fitness parameters to each base pair and then calculating the probabilities of fixation for every possible substitution [20]. For instance, a coding site may favor A over C, G, or T in model of selective constraint. Because only A would encode the 'optimal' function in this example, mutations from A would be deleterious, mutations towards A would be beneficial, and all other mutations would be neutral. Mixing mutational biases with such selective forces can have complex effects on the inference of conservation when selection is weak (Box 1: Moderate to strong purifying selection) and can even lead to *prima facie* impossible situations where natural selection for constant functionality increases and not decreases the rate of evolution (see [19] for more).

Moreover, as tempting as it would be to estimate the strength of selection from divergence data alone, this cannot be done with much precision, especially for strong selection coefficients [19,21]. Examining the case where there is only one optimal base pair (Box 1: Moderate to strong purifying selection) shows the efficacy of purifying selection (constraint) over a tree: a small, linear increase in the strength of consistent purifying selection causes a large, exponential drop in the rate of evolution.

Weak to moderate constraint is thus capable of conserving sites over even large phylogenetic distances, and increasing the number of species/tree length results in only a limited increase in power to distinguish strong from moderate or weak purifying selection. Further, any substitution as the result of a transient relaxation of constraint will generate an estimate of constant weak selection over the tree. Meanwhile, attempting to carry out estimation of the strength of selection at individual branches comes at the expense of losing the power of phylogenetic footprinting over the full tree. Thus there are inherent difficulties with using divergence data to assess the importance of an element to the fitness of an organism.

Population genetics

Both the density of polymorphisms ('amplitude'; Box 2: Density of polymorphism) and the frequency distribution of observed SNPs ('shape'; Box 2: Shape of the SFS) contain information about the magnitude of selection operating on a group of sites. Many classic methods use the shape of the SFS to estimate the DFE [16,22]. These approaches can suffer from lack of power to detect strong selection, especially in shallow samples (see [22,23]; Box 2: Figure IC broken lines). More recent methods combine the information from the shape of the SFS with the expected change in polymorphism density by adding 'amplitude' information in the form the '0 frequency' class to the SFS, in other