Opinion

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

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

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

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.

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

Box 1. The neutral theory and comparative genomics

Neutral theory

The neutral theory of molecular evolution stipulates that the vast majority of alleles that 'fix' in the population and become substitutions are neutral alleles with no effect on the function of the site in which they occur [37,38]. The proportions of mutations that are deleterious and neutral in functional and non-functional categories are shown (Figure IA). Mutations that disrupt function are deleterious and are removed from the population over time such that they are neither seen as polymorphisms nor as substitutions. The rate of evolution under neutral theory, r, is given by: $r = (1–f)*r_0$, where f is the fraction of deleterious mutations and r_0 is the rate of mutation. Thus $r \leq r_0$, and cases where $r < r_0$ are indicative of selective constraint.

Comparative genomics

The insight of comparative genomics was to use the framework of neutral theory and apply it to multiple species alignments to find functional elements. Species can be related to each other in phylogenetic trees with the time between species represented by the branch lengths, which can be estimated via a maximum-likelihood methodology in which the tree that allows best explanation of the data relative to a pre-specified model of evolution is chosen [39]. The sum over all branch lengths represents the total time of the tree. Tests for functional constraint acting on sites use a 'neutral' reference to control for linked selection, biased gene conversion, mutation rate, and more. Under neutral theory, functional sites should evolve more slowly ('be conserved') than the neutral reference, and therefore the total length of the tree at functional sites, t, should be smaller than the length of the tree at a neutral reference, t_0 (Figure IB). Those sites/elements/regions where t/t_0 ($\sim r/r_0$) is inferred to be less than 1 are inferred to be under selection and functional. Note however that few useful references are themselves truly neutral – e.g., synonymous sites [23,24] – and any selective constraint in the reference will make tests for purifying selection more conservative.

Moderate to strong purifying selection

In the selection model for Figure IC, A (black line) or C (blue line) is the optimal base pair, whereas C/A↔T↔G mutations are neutral among each other and less fit compared to the optimal. $\pi_{A+T} = 0.8$ implies that mutation is biased such that in the absence of selection the equilibrium frequency in a DNA sequence of A = T = 0.4 and the frequency of G = C = 0.2. Sequences of 100 kb in length were simulated at each selection coefficient over the 32 mammalian species tree (human to sloth) (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz44way), which has a total neutral divergence of around 4.75 substitutions/site [40]. The sites were assumed to be the independent, which may not always be a valid assumption to make (see [41]). Plotted in Figure IC is the median number of substitutions/site for each region estimated using genomic evolutionary rate profiling (GERP) [4]. This figure shows that when selection is allowed to vary over a large range, and is consistent across a tree, we quickly lose ability to distinguish weak from moderate from strong selection as even fairly weak selection ($|4N_e s|$ of about 3–5) generates complete conservation of the region regardless of the direction of mutation *vis à vis* selection.

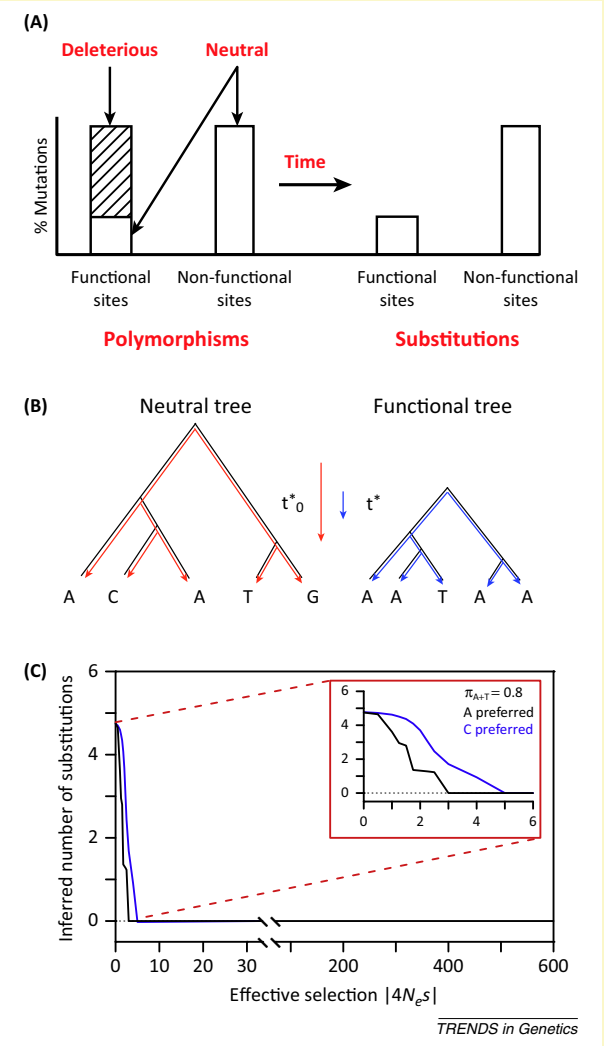

Figure I. **(A)** The spectrum of mutations versus substitution in functional versus non-functional sites. **(B)** Conservation over a species tree for functional versus non-functional sites. **(C)** Selection versus conservation over the mammalian phylogeny.

words the number of sites at which polymorphism was not observed in the sample at all [23,24].

Assuming mutation–selection balance – and this may not be a fair assumption [25] – one can calculate an analytical form for the SFS given the mutation rate, population size, and DFE (see [16,23,24] for details). We can use this to calculate the likelihood (λ) of observed spectra for a neutral reference, SFS_{ref}, and test set of sites, SFS_{test}.

$$\lambda_{full}(SFS_{test}, SFS_{ref}|\theta, \gamma, f) = \lambda_{test}(SFS_{test}|\theta, \gamma, f) \times \lambda_{ref}(SFS_{ref}|\theta) \quad (1)$$

To infer the DFE on the test set of sites, the SFS of the neutral reference anchors the effective mutation rate, θ, which sets the expected neutral amplitude and shape of the SFS. Example spectra for different strengths of selection are displayed in Box 2: Figure IB showing how selection skews the shape of the SFS from the neutral expectation. An appropriately chosen neutral reference can also help control for shared deviations from the assumptions of the SFS model, such as demography and linked selection affecting both the neutral and test set of sites (for more, see [16,23,24]). The likelihood of the parameters of the DFE (γ, the selection coefficients and f, the distribution of sites) is maximized using the SFS of the test set of sites. As there are an infinite number of possible distributions of selection coefficients, the problem is generally simplified by assuming a particular form of a distribution: often a

Box 2. Population genetics and the SFS

Density of polymorphisms

Unlike divergence between species where even weak selective constraint can cause complete conservation and inability to distinguish selective forces (see Box 1: Moderate to strong purifying selection), the density of polymorphisms within a species drops gradually as selection strength increases (Figure IA). Even so, sites under very strong purifying selection contribute little to observed polymorphism, making it difficult to detect SNPs at such sites unless a very deep sample of the population is taken. Change in the density of polymorphism relative to a neutral expectation is one hallmark of the action of purifying selection in polymorphism data.

Shape of the SFS

Selection changes the expected frequency of alleles in the population. Shown in Figure IB are three representative folded spectra (i.e., where allele frequencies range from 0 to 0.5 and no attempt is made to polarize alleles into derived and ancestral) evolving under different strengths of purifying selection: neutral, moderate ($4N_es = 10$), and strong ($4N_es = 200$) selection. To highlight the effect on shape in rare frequency classes: 17% of neutral mutations are singletons, they occur only once in this sample of 201 individuals from the population, versus 30% for moderately deleterious alleles and 72% for strongly deleterious alleles.

Statistical power of SFS methods

Figure IC represents the power to detect the action of purifying selection and determine its strength. The y-axis is the power (in units of Δlog-likelihood) to distinguish between the true parameters ($4N_es$, $p = 0.3$) and the null hypotheses $H_{0,\ neutral} = (0,0)$ and $H_{0,\ lethal} = (-\infty, p')$ where p' is estimated by maximum-likelihood. The dotted grey line represents 5% significance for the χ^2 test [$-2\Delta\log(L) = 3.84$]. Thus, there is good power to detect selection from neutrality, which increases as selection increases and is insensitive to the depth of sampling in the population (unbroken lines). Differentiating purifying selection of some finite strength from infinitely strong selection corresponding to lethality, or equivalently from reduction in mutation rate in the test region relative to the neutral reference, has much less power, decreases as the strength of selection increases, and is very sensitive to the sampling depth (broken lines). This is because the power to distinguish finite selection from lethality comes only from the shape of the SFS, and the stronger the selection, the deeper the sampling needed to reveal the skew of the allele frequency spectrum towards ultra-rare variants. Such excess of ultra-rare variants is expected under the model of finite strong selection, but is not expected under the models of infinitely strong selection or mutation rate reduction.

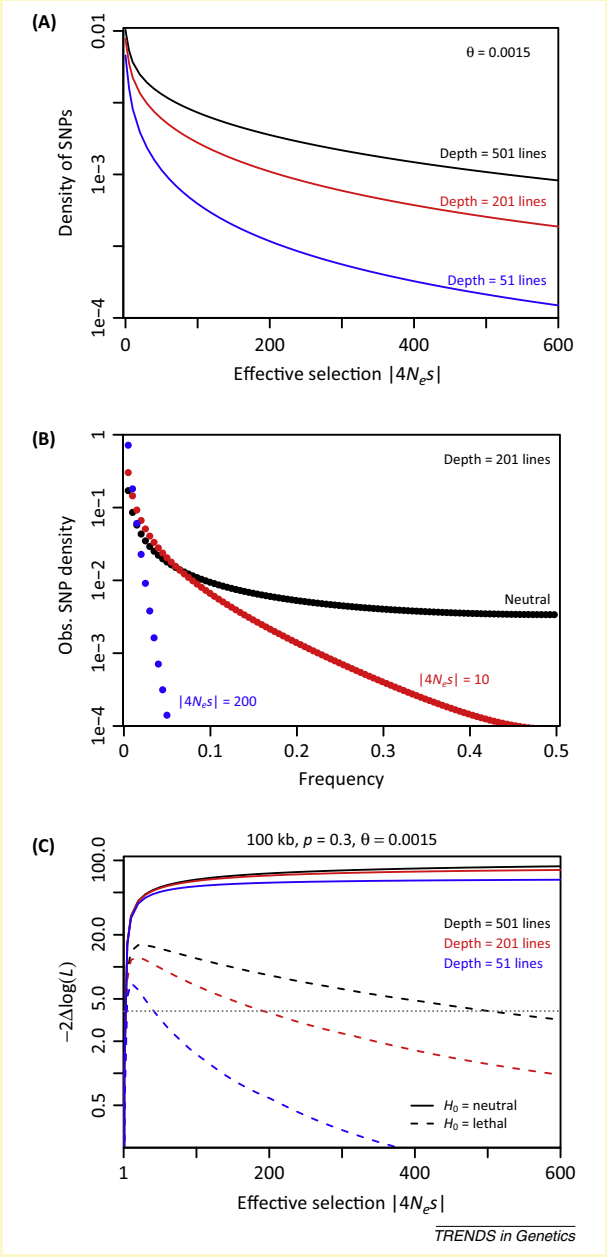

Figure I. **(A)** Amplitude of site frequency spectrum (SFS): selection versus the density of polymorphisms for different sample depths. θ = 0.0015. **(B)** Shape of SFS: the fraction of observed polymorphisms over the frequency of the minor allele in the population for different strengths of selection in a sample of 201 individuals. **(C)** Power to detect and resolve different strengths of selection in 100 kb of independent sites where 30% of the sites are functional ($p = 0.3$) and with a human-like θ value of 0.0015. Abbreviations: Obs., observed; SNP, single-nucleotide polymorphism.

gamma, lognormal, or categorical distribution of sites over selection coefficients is used. The gamma and log-normal distributions are popular choices because they can take a large number of shapes defined by only two parameters, whereas the categorical distribution does not assume a shape at all, but requires a parameter for the percentage of sites in each defined category of selection strength. Note that the choice of the form of DFE can greatly affect its biological interpretation and should be done with care [22,26].

To exemplify the power of the SFS to detect selection and resolve its strength, we use a categorical distribution where 30% of sites are evolving under a given selection coefficient and 70% are neutral (Box 2, Figure IC). To simplify matters, here we assume all non-neutral mutations are deleterious – other models incorporating beneficial and adaptive mutations are possible [17,27]. The unbroken and broken lines display the power to distinguish selection of particular intensity from neutrality (i.e., to detect functionality) and from infinitely strong selection respectively.

Amplitude and shape combined drive the power to detect selection from neutrality and weaker selection coefficients. This power increases quickly with the strength of selection. It turns out to be harder to distinguish strong but finite selection from infinitely strong selection driven by

complete lethality because both are very effective at removing polymorphisms from the sample (i.e., they both decrease the amplitude of the SFS) and generate an otherwise neutral-looking SFS in small sample polymorphism datasets. This distinction between strong but finite selection and infinitely strong selection is very important because infinitely strong selection is indistinguishable in its effect from a reduction in mutation rate in the tested region. Given that there can exist the possibility that mutation rate varies systematically between the set of test sites and the reference, this creates a problem of inference. By contrast, finitely strong selection is a clear indication of functionality and can never be confused with mutation rate variation. However, very few sites under strong selection have polymorphisms and those SNPs they do have are at very low frequency in the population (Box 2, Figure IA,B). Thus, one needs to sample the population deeply to capture the signal of strong but finite purifying selection: an excess of ultra-rare alleles relative to the expectation from infinitely strong selection or mutation rate variation (Box 2: Power to detect and resolve the strength of selection). Datasets that have such deep sampling of a single species are indeed becoming available [28–30].

Unfortunately, even in deeply sampled populations and neutral loci, few sites are polymorphic, which limits the ability to call individual elements as functional. SFS methods thus need a large number of sites for their analysis – we used 100 kb. Deeper sampling furnishes diminishing returns in terms of the amount of polymorphism added (Box 2, Figure IA). For instance, take a human-like θ value of 0.0015. Assuming even a sample depth of 10001 chromosomes sequenced from the population, there would still only be ~1.5 SNPs every 100 bp of neutral sequence. It should be noted that this assumes a constant population size, whereas humans have recently had a rapidly expanding population [31]. In a rapidly expanding population there will be a greater proportion of rare alleles such that sequencing deeply will net more polymorphisms than expected under the above [31]. Nevertheless, the expected gain of power to detect and resolve selection at a small number of sites from sampling a single species deeper and deeper is not as great as deeply sampling the variation in that element in multiple species.

Adding polymorphisms from closely related species where a specific element has maintained its functionality should greatly increase the power to detect that element. Because the frequency spectra of the different species are independent, the likelihoods of the models can be multiplied by their fit to the polymorphism data in each species. To simplify the problem for illustrative purposes, let us assume that each species has the same parameters (θ, sample depth) and that the element/set of sites is present and functional in each species. The log-likelihood of the multi-species SFS model is simply the number of species multiplied by the log-likelihood of the single-species model. Essentially, by adding the polymorphism of five species, one is looking at fivefold the number of sites all with the same expected SFS pattern: in other words, a 200 bp element becomes, in effect, a 1000 bp element.

In Figure 1 we show the expected power gains in terms of the number of sites needed to detect and resolve

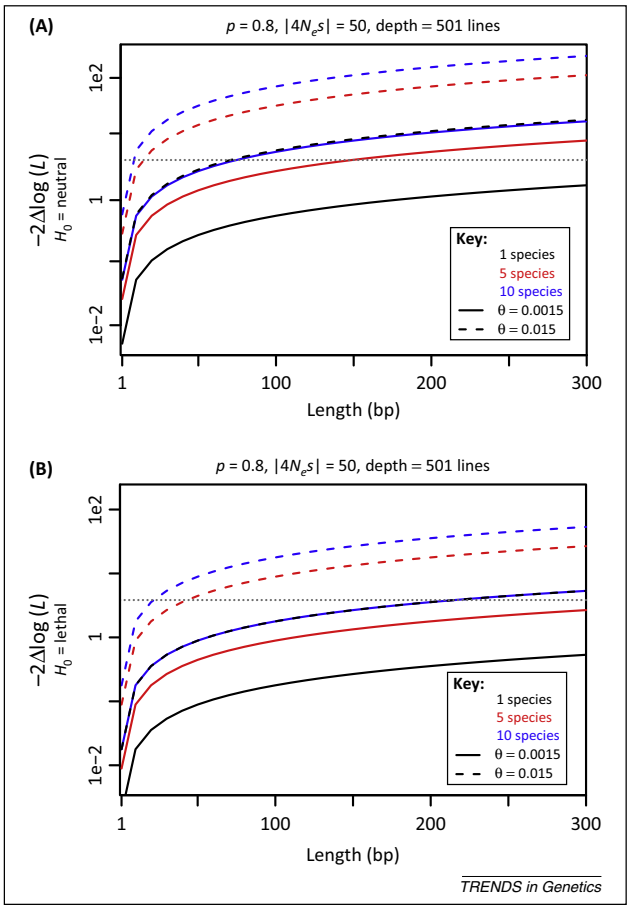

Figure 1. Comparative population genomics. The power to detect purifying selection on either an individual functional element or any small collection of sites is improved greatly by the addition of multiple species. We model a functional element in which 80% of the sites are functional and maintained by a selective force of $4N_es = -50$ across 1, 5, or 10 species (for details of $4N_es$ see Glossary: Effective selection). Each species has been sequenced to a depth of 501 individuals from a population. (A) The *y* axis is the power (in units of Δlog-likelihood) to distinguish between the true parameters ($4N_es = -50$, $p = 0.8$) and the null hypotheses $H_{0,\ neutral} = (0,0)$. (B) The *y* axis is the power (in units of Δlog-likelihood) to distinguish between the true parameters ($4N_es$, $p = 0.8$) and the null hypotheses $H_{0,\ lethal} = (-\infty, p')$. The dotted grey line represents 5% significance for the χ^2 test [$-2\Delta\log(L) = 3.84$]. We can see that as we increase the number of species, especially when those species have a higher level of polymorphism, we gain substantial power to detect shorter functional elements. When θ is on the order of 1%, and we go from a single population genomic dataset to a dataset from 10 species, we move from being able to detect purifying selection acting on an 80 bp element to being able to detect it acting on an 8 bp element. Similarly, panel (B) shows that, to distinguish finite selection from infinitely strong selection or mutation rate variation, 220 bp are needed if data from only a single species are available, versus 22 bp if polymorphisms from 10 species are used. Note that the increase in power (in log-likelihood space) is proportionate to the number of polymorphisms, such that increasing number of species, length, and θ all cause a proportional increase in power.

selection by having polymorphism data in multiple species. Having spectra from 10 closely related *Drosophila* species would allow for the detection of small functional elements, ~8 bp compared to ~80 bp in only one species, and 22 bp would be enough to distinguish moderate from strong and lethal selection. Because of the lower θ in species with human-like levels of polymorphism, only long elements or small groups of sites (totaling ~80 bp) whose function has been conserved across the 10 species can be distinguished from neutral sequences; ~220 bp would be required to distinguish moderate from lethal selection.

The above exposition demonstrates what power can be gained from multi-species polymorphism datasets. Although identifying functional elements conserved over 10 species is typically not difficult, these 10 species can be arbitrarily close in phylogenetic distance (although if they are too close one must then model incomplete lineage assortment and the general non-independence of polymorphism between species). Further, discovered elements can be categorized by the average strength of purifying selection maintaining their functionality – a more meaningful metric of their functional significance than their longevity across the phylogenetic tree. One not only gains an increase in resolution over analyzing the polymorphism from a single species but also profits from a more biologically relevant assignation of functional importance than that which divergence data alone are capable of delivering.

Comparative population genomics: combining polymorphism and divergence data

Methods using divergence have the power to detect the action of selection operating on short elements, or even individual sites, by leveraging the efficiency of selection over long evolutionary history. However, this efficiency comes at the cost that the rate of evolution provides limited information about the true functional significance of the element because even moderate or weak selection will lead to complete conservation. Divergence methods thus favor the detection of persistent functional elements, and not necessarily those that are particularly important in a focal species.

Polymorphism data allow the calculation of the distribution of selection coefficients in large sets of sites using recent evolutionary history, but our ability to detect selection acting on a small number of sites or a single functional element is limited by the low density of polymorphisms within a single species. One can partially ameliorate these issues by analyzing the polymorphism from multiple related species, but even greater power may be available by combining polymorphism and divergence data into one overarching framework.

Contrasting polymorphism with divergence data is not a new concept, particularly in tests for adaptation [25]. Comparisons of inferences of constraint made from polymorphism versus divergence data often show good congruence between the two: the conservation of sites along the tree is correlated with both lower SNP density and rarer-frequency SNPs within a single species, indicating an enrichment for constraint in conserved sites [23,32,33].

We applaud recent efforts to leverage combined polymorphism and divergence data to estimate the fraction of deleterious mutations with greater efficacy than either alone [34–36]. Beyond the increased potential for simply detecting functional elements, particularly exciting is that the combination of multi-species polymorphism and divergence in a unified framework allows better modeling of the evolutionary history and functional importance of known elements. The long-term evolutionary history of a group of sites can inform which sites are likely to be under selection, thereby increasing the ability of polymorphism methods to detect and resolve the strength of selection and, in turn, providing more information for the modeling of the long-term evolutionary history.

Comparative population genomics represents an exciting new prospect for detecting functional elements, describing their functional importance, and imputing their evolutionary history. For application to the human genome, it will require large new sequencing projects to sequence many related species to sufficient depth, as outlined in Box 2 and Figure 1. Shallow polymorphism datasets are already available for some of our closest relatives, and more are expected to come on-line [36], but more depth will allow the estimation of strong values of selection coefficients. By utilizing polymorphisms within many species, as well the divergence between them, we will have increased power to build more accurate DFEs on individual genomic elements and small groups of sites. We will thereby not only gain a more accurate picture of the percent of functionality in the human genome, but also a more detailed picture about the distribution of functional importance across it.

References

1 Hardison, R.C. (2003) Comparative genomics. *PLoS Biol.* 1, e58
2 Siepel, A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050
3 Stark, A. *et al.* (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* 450, 219–232
4 Davydov, E.V. *et al.* (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP. *PLoS Comp. Biol.* 6, e1001025
5 Pollard, K.S. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20, 110–121
6 Lindblad-Toh, K. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478, 476–482
7 Graur, D. *et al.* (2013) On the immortality of television sets: 'function' in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* 5, 578–590
8 Khatun, J. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74
9 Eddy, S.R. (2013) The ENCODE project: missteps overshadowing a success. *Curr. Biol.* 23, R259–R261
10 Doolittle, W.F. (2013) Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. U.S.A.* 110, 5294–5300
11 Ward, L.D. and Kellis, M. (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337, 1675–1678
12 Siepel, A. *et al.* (2006) New methods for detecting lineage-specific selection. In *Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006)*. pp. 190–205
13 Hall, S.S. (2012) Journey to the genetic interior. *Sci. Am.* 307, 80–84
14 Andolfatto, P. (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437, 1149–1152
15 Bergman, C.M. and Kreitman, M. (2001) Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* 11, 1335–1345
16 Boyko, A.R. *et al.* (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4, e1000083
17 Fay, J.C. *et al.* (2001) Positive and negative selection on the human genome. *Genetics* 158, 1227–1234
18 Hartl, D.L. *et al.* (1994) Selection intensity for codon bias. *Genetics* 138, 227–234
19 Lawrie, D.S. *et al.* (2011) Faster than neutral evolution of constrained sequences: the complex interplay of mutational biases and weak selection. *Genome Biol. Evol.* 3, 383
20 Halpern, A.L. and Bruno, W.J. (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.* 15, 910–917
21 Tamuri, A.U. *et al.* (2012) Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation–selection models. *Genetics* 190, 1101–1115

22 Eyre-Walker, A. and Keightley, P.D. (2007) The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8, 610–618
23 Lawrie, D.S. *et al.* (2013) Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genet.* 9, e1003527
24 Keightley, P.D. and Halligan, D.L. (2011) Inference of site frequency spectra from high-throughput sequence data: quantification of selection on nonsynonymous and synonymous sites in humans. *Genetics* 188, 931–940
25 Messer, P.W. and Petrov, D.A. (2013) Frequent adaptation and the McDonald–Kreitman test. *Proc. Natl. Acad. Sci. U.S.A.* 110, 8615–8620
26 Kousathanas, A. and Keightley, P.D. (2013) A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics* 193, 1197–1208
27 Clemente, F. and Vogl, C. (2012) Evidence for complex selection on fourfold degenerate sites in *Drosophila melanogaster*. *J. Evol. Biol.* 25, 2582–2595
28 Mackay, T.F.C. *et al.* (2012) The *Drosophila melanogaster* genetic reference panel. *Nature* 482, 173–178
29 Weigel, D. and Mott, R. (2009) The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol.* 10, 107
30 The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073
31 Keinan, A. and Clark, A.G. (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336, 740–743
32 Goode, D.L. *et al.* (2010) Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. *Genome Res.* 20, 301–310
33 Cooper, G.M. *et al.* (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods* 7, 250–251
34 Wilson, D.J. *et al.* (2011) A population genetics–phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet.* 7, e1002395
35 Gronau, I. *et al.* (2013) Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. *Mol. Biol. Evol.* 30, 1159–1171
36 De Maio, N. *et al.* (2013) Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Mol. Biol. Evol.* 30, 2249–2262
37 Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* 217, 624
38 King, J.L. and Jukes, T.H. (1969) Non-Darwinian evolution. *Science* 164, 788–798
39 Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376
40 Meyer, L.R. *et al.* (2013) The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Res.* 41, D64–D69
41 Siepel, A. and Haussler, D. (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* 21, 468–488