PERSPECTIVE

Epistasis Analysis Goes Genome-Wide

Jianzhi Zhang*

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, United States of America

* jianzhi@umich.edu

Epistasis, a term coined by William Bateson in 1909 [1], refers to the interdependence of mutations in their phenotypic effects. Let the phenotypic value of a trait relative to that of the wild type be f_A and f_B for mutants A and B, respectively, and let the phenotypic value of the corresponding double mutant be f_{AB}. Although variation exists, epistasis is usually defined by $\varepsilon = f_{AB} - f_A f_B$ and is said to be positive when $\varepsilon > 0$ and negative when $\varepsilon < 0$.

Life would have been much simpler, and perhaps even boring, if epistasis were completely absent. In reality, however, epistasis abounds, rendering biology full of surprises and complexity. For instance, a commonly encountered type of epistasis is synthetic lethality, where simultaneously deleting two genes from the genome of a normal organism is lethal despite the fact that deleting each of them separately is viable. Using the notation introduced above, we can describe synthetic lethality by $f_A > 0$, $f_B > 0$, and $f_{AB} = 0$; consequently $\varepsilon < 0$. A simple mechanistic explanation of synthetic lethal epistasis is that the two genes investigated are functionally similar and hence redundant. Clearly, studying epistasis helps us to understand the functional relationship between genes, which is critical to uncovering the inner workings of biological systems. Epistasis can explain why hybrids between species are typically inviable or infertile [2] and is believed to underlie the intriguing phenomenon that some human disease-causing mutations are fixed in other species with no apparent detriment [3]. Furthermore, epistasis is assumed in many evolutionary theories. For example, the mutational deterministic hypothesis of the evolution of sexual reproduction [4] and the hypothesis of reduction in mutational load by truncation selection against deleterious mutations [5] both depend on overall negative epistasis. Thus, verifying these hypotheses requires confirming the prevalence of negative epistasis.

Epistasis is typically detected by demonstrating the inequality between f_{AB} and $f_A f_B$ or some consequences of this inequality. The advent of next-generation sequencing and other genomic technologies is quickly enlarging the scale of epistasis studies. Of special significance is the recent completion of the yeast genetic interaction map, which includes nearly all 36 million epistasis values for pairs of ~6,000 yeast genes estimated from the growth rates of single- and double-gene–deletion mutants [6]. Although this map provides unprecedented data of epistasis between null mutations of different genes, it offers no information on the epistasis between mutations at different sites within a gene or that between non-null mutations in different genes.

Complementing the above coarse-grained epistasis map are nucleotide-resolution epistasis maps of individual genes or segments of genes [7–10]. For instance, Li et al. synthesized a yeast tRNA gene with error, creating all possible single-point mutation variants of the gene as well as tens of thousands of variants with multiple mutations [7]. They then used a high-throughput method to measure the fitness of yeast strains, each carrying a variant tRNA gene at the place of the endogenous gene, and estimated epistasis between mutations. Interestingly, negative epistasis was found to be more prevalent than positive epistasis [7]. In principle, such a map can be constructed for every gene in the yeast genome to acquire a general picture of epistasis.

OPEN ACCESS

Citation: Zhang J (2017) Epistasis Analysis Goes Genome-Wide. PLoS Genet 13(2): e1006558. doi:10.1371/journal.pgen.1006558

Editor: Takashi Gojobori, National Institute of Genetics, JAPAN

Published: February 16, 2017

Funding: This work is supported by grants from the United States National Institutes of Health. The funder had no role in the preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

Even with both the coarse- and fine-grained epistasis maps, we still do not have epistasis data between non-null mutations of different genes, which constitute the largest part of a complete nucleotide-resolution epistasis map of a genome. Take yeast as an example: this largest part contains potentially 10^{14} interactions. Obviously, determining the complete epistasis map is an enormous challenge.

In this issue of *PLOS Genetics*, Skwark and colleagues [11] harnessed population genomic data to approach this challenge. Specifically, they developed a computational method termed genomeDCA to detect epistasis using genotype and allele frequencies estimated from genome sequences of thousands of individuals of the same species. Their method is a modification of direct-coupling analysis (DCA), a statistical method in structural biology for predicting direct residue contacts within and between proteins [12,13]. What is the similarity between physical residue contacts and epistasis that allows the borrowing of DCA for identifying epistasis? Well, if two sites contact in protein structure, the amino acid that sits at one site likely impacts what amino acids can occupy the other site. One can imagine the scenario in which amino acid A at site 1 interacts well with amino acid B at site 2, but when A is mutated to A′ or when B is mutated to B′, the interaction is disrupted. However, when both residues are mutated, the interaction may be restored, resulting in $\varepsilon = f_{A'B'} - f_{A'B} f_{AB'} > 0$, where f is the strength of interaction relative to that between A and B. If fitness increases with the interaction strength, one would frequently observe the genotype of AB or A′B′ at the two sites but rarely encounter AB′ or A′B when many species are examined. Thus, protein sequences from many species provide information about residue contacts as well as epistasis. This idea forms the basis of DCA in structural biology, although the actual statistical analysis is more sophisticated due to a number of confounding factors such as the phylogenetic nonindependence of protein sequences.

GenomeDCA is similar to DCA but is applied to genome sequences of a large number of conspecifics. When there is no population structure and when all individuals are recombining freely, linkage disequilibrium between two nucleotide sites should approach zero unless the relative fitness of the double mutant does not equal the multiplication of those of the corresponding single mutants (Fig 1). Hence, detection of linkage disequilibrium between two sites indicates epistasis. Of course, sites that are close in chromosomal location may be at linkage disequilibrium due to limited recombination. Thus, genomeDCA should be applied to sites that are sufficiently far apart on the same chromosome or located on different chromosomes. The main advantage of genomeDCA over traditional epistasis analyses is its genomic scale and its ability to test epistasis for many mutation pairs from one large set of genome sequences. The chief disadvantage is that, given the limited number of conspecifics sampled, only a tiny fraction of all possible mutations is observed, which dictates the number of mutation pairs for which epistasis can be evaluated by genomeDCA. In other words, one may not be able to test epistasis between a predetermined pair of mutations, because these mutations may not be present in the genome sequences or may not be sufficiently common to guarantee statistical power. This situation applies to many deleterious mutations, which are either not present or have low frequencies even in reasonably large samples. Nevertheless, depending on the purpose of the epistasis analysis, one could argue that it is the epistasis between observed mutations that is most relevant. In this regard, genomeDCA detects the most relevant epistasis. Another limitation of genomeDCA is that it detects only fitness epistasis, whereas traditional methods can detect epistasis in other traits.

Skwark et al. applied genomeDCA to two large population genomic datasets: 3,156 genomes of *Streptococcus pneumoniae* isolates and 3,442 genomes of *Streptococcus pyogenes* isolates. Their results are biologically interesting. For example, they detected over 5,000 epistatic interactions in the former dataset, over three quarters of which are between sites in three genes (*pbp2x*, *pbp1a*, and *pbp2b*) that confer resistance to beta-lactam antibiotics. With the

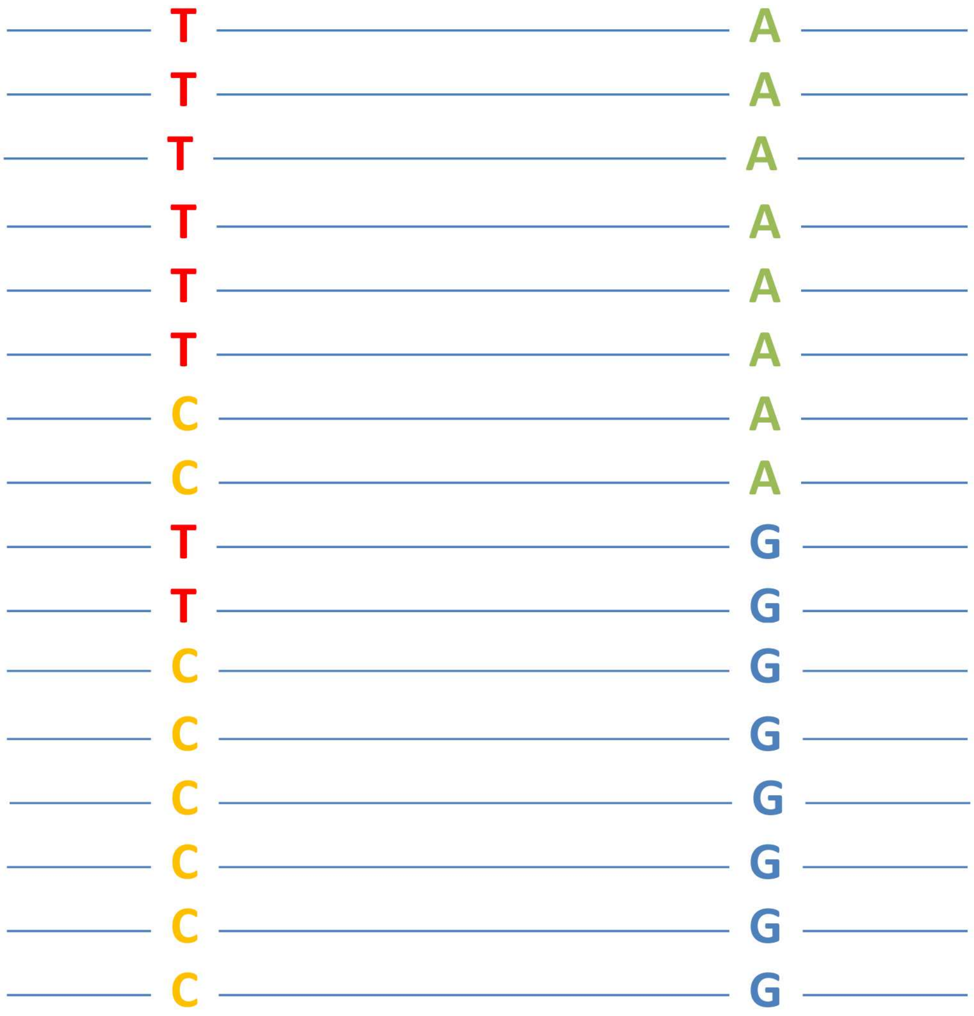

Fig 1. A hypothetical example of epistasis between two nucleotide sites that is detectable by genomeDCA. Each line represents a (haploid) genome sequence from an individual, where two freely recombining nucleotide positions are shown. Each position has two states in the population. In this example, the genotype frequencies of TA and CG are both 6/16, exceeding the expectation (0.5 × 0.5 = 4/16) from allele frequencies under no epistasis. Should this discrepancy be statistically significant, epistasis between the two sites is detected (genotypes TA and CG are fitter than TG and CA).

doi:10.1371/journal.pgen.1006558.g001

precipitous drop of the cost of DNA sequencing and rapid accumulation of population genomic data of humans, human pathogens, and genetic model organisms, genomeDCA promises to offer a cost-effective survey of epistasis at the genomic scale. As epistasis data accrue, one can start looking for general patterns and underlying mechanisms of epistasis [14], which will ultimately aid our understanding of the organizing principles of biological systems.

References

1. Bateson W (1909) Mendel's Principles of Heredity. Cambridge: Cambridge University Press.
2. Orr HA (1996) Dobzhansky, Bateson, and the genetics of speciation. Genetics 144: 1331–1335. PMID: 8978022
3. Xu J, Zhang J (2014) Why human disease-associated residues appear as the wild-type in other species: genome-scale structural evidence for the compensation hypothesis. Mol Biol Evol 31: 1787–1792. doi: 10.1093/molbev/msu130 PMID: 24723421

4. Kondrashov AS (1988) Deleterious mutations and the evolution of sexual reproduction. Nature 336: 435–440. doi: 10.1038/336435a0 PMID: 3057385

5. Crow JF, Kimura M (1979) Efficiency of truncation selection. Proc Natl Acad Sci U S A 76: 396–399. PMID: 16592610

6. Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, et al. (2016) A global genetic interaction network maps a wiring diagram of cellular function. Science 353. doi: 10.1126/science.aaf1420 PMID: 27708008

7. Li C, Qian W, Maclean CJ, Zhang J (2016) The fitness landscape of a tRNA gene. Science 352: 837–840. doi: 10.1126/science.aae0568 PMID: 27080104

8. Puchta O, Cseke B, Czaja H, Tollervey D, Sanguinetti G, et al. (2016) Network of epistatic interactions within a yeast snoRNA. Science 352: 840–844. doi: 10.1126/science.aaf0965 PMID: 27080103

9. Bank C, Hietpas RT, Jensen JD, Bolon DN (2015) A systematic survey of an intragenic epistatic landscape. Mol Biol Evol 32: 229–238. doi: 10.1093/molbev/msu301 PMID: 25371431

10. Olson CA, Wu NC, Sun R (2014) A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Curr Biol 24: 2643–2651. doi: 10.1016/j.cub.2014.09.072 PMID: 25455030

11. Skwark MJ, Croucher NJ, Puranen S, Chewapreecha C, Pesonen M, et al. (2017) Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. PLoS Genet in press.

12. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A 106: 67–72. doi: 10.1073/pnas.0805923106 PMID: 19116270

13. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A 108: E1293–1301. doi: 10.1073/pnas.1111471108 PMID: 22106262

14. He X, Qian W, Wang Z, Li Y, Zhang J (2010) Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. Nat Genet 42: 272–276. doi: 10.1038/ng.524 PMID: 20101242