Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Plant Biology

# Developing maps of fitness consequences for plant genomes

Zoé Joly-Lopez[1], Jonathan M Flowers[1,2] and Michael D Purugganan[1,2]

Predicting the fitness consequences of mutations, and their concomitant impacts on molecular and cellular function as well as organismal phenotypes, is an important challenge in biology that has new relevance in an era when genomic data is readily available. The ability to construct genomewide maps of fitness consequences in plant genomes is a recent development that has profound implications for our ability to predict the fitness effects of mutations and discover functional elements. Here we highlight approaches to building fitness consequence maps to infer regions under selection. We emphasize computational methods applied primarily to the study of human disease that translate physical maps of within-species genome variation into maps of fitness effects of individual natural mutations. Maps of fitness consequences in plants, combined with traditional genetic approaches, could accelerate discovery of functional elements such as regulatory sequences in non-coding DNA and genetic polymorphisms associated with key traits, including agronomically-important traits such as yield and environmental stress responses.

**Addresses**
[1] Center for Genomics and Systems Biology, Department of Biology, 12 Waverly Place, New York University, New York, NY 10003, United States
[2] Center for Genomics and Systems Biology, NYU Abu Dhabi Research Institute, NYU Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates

Corresponding author: Purugganan, Michael D (mp132@nyu.edu)

**Current Opinion in Plant Biology** 2016, **30**:101–107

This review comes from a themed issue on **Genome studies and molecular genetics**

Edited by **Yves Van de Peer** and **J Chris Pires**

For a complete overview see the Issue and the Editorial

Available online 4th March 2016

**http://dx.doi.org/10.1016/j.pbi.2016.02.008**



One of the great challenges in biology is to determine the fitness consequences of individual polymorphisms across the genome. Over the last few years, high-throughput functional genomics and whole genome resequencing have enabled discovery of functional elements in non-coding DNA and comprehensive descriptions of single nucleotide polymorphisms (SNP) and other genetic variants in plant genomes. For example, genome-wide SNPs have been catalogued for many species including *Cicer* (chickpea) [1], *Zea* (maize) [2], *Oryza* (rice) [3], date palms [4], and Chlamydomonas [5], and the 3000 Rice Genomes Project recently reported more than 30 million polymorphisms in 3024 rice varieties [6••]. In the next few years such 'SNP atlases' will become available for many other crops and their wild relatives.

With whole genome sequences now widely available, evolutionary biologists are revisiting the long-standing challenge [7,8] of predicting the fitness effects of mutations. In principle, expanding these predictions to a genomewide scale would allow us to construct maps of fitness variation that describe the probability that a mutation will impact fitness and predict both the magnitude and sign (i.e. beneficial or deleterious) of their effect. In practice, estimating the fitness effect of mutations remains one of the great objectives in molecular evolution [8], but recent advances in diverse evolutionary and experimental approaches have improved the prospects of constructing maps of fitness consequences in plant genomes.

Maps of fitness consequences have potentially widespread applications. From an evolutionary perspective, they provide a basis for predicting whether a mutation is beneficial and improves a fitness-related trait or is deleterious and negatively impacts traits such as crop yield or resistance to disease that are targeted for improvement. From the perspective of molecular biology, maps of fitness consequences provide clues as to which positions in the genome impact a cellular function (Box 1). Since mutations that impact fitness must also affect function, identification of sites that affect fitness may assist with quantifying the fraction of the genome that is functional — a subject of recent controversy in human genetics [9] — and identifying polymorphisms that modify traits of interest [10,11].

Here we review recent advances in methods to construct genomewide maps of fitness variation. Although many of these approaches have primarily been applied to the study of human diseases, we focus on methods that are applicable to plants and highlight how they may profoundly improve efforts to discover functional elements in plant genomes.

## Genome-wide maps of fitness consequences

Regions of the genome that are conserved in evolution represent a special class of sites, where purifying selection

**Box 1 The link between fitness and function.**

Mutations that impact fitness must also impact a cellular function. This observation is the basis for applying principles of molecular evolution to assess the impact of a mutation's function indirectly by predicting its fitness effect. Such fitness effect predictions are made through methods based on population genetics theory that quantify the proportion of sites under selection or the strength of selection acting on collections of sites in the genome (Box 2).

What is the relationship between fitness and function? The fitness-function relationship is the extent to which a change in allele function will lead to a change in fitness. This relationship is often assumed to be linear. However limited experimental data indicate non-linear relationships that vary from locus to locus and depend on genetic background and the environment [65]. Rest *et al.* [66•] quantified the effect of changes in expression in *LCB2* in yeast on fitness and reported an 'S'-shaped, or sigmoidal, relationship. Hartl *et al.* [65] quantified this relationship for activity at β-*galactosidase* and fitness in *E. coli* and found it approximates a saturation curve as it does in other metabolic contexts. The exact nature of the relationship between fitness and function is sure to be complex and will remain unknown except for exceptional study systems.

has preserved a sequence over long periods of evolutionary history. Constraint-based, or conservation-based methods aim to identify these slow evolving sequences and the functional elements they encode using multiple sequence alignments from phylogenetically diverse species. In practice, such 'phylogenetic footprinting' methods assign scores to positions in the genome indicating the degree of conservation across species [12–14] (Box 2) thereby enabling the discovery of elements that, when mutated, are expected to impact fitness. Application of these approaches in plants benefits from 70 published genomes [15], which enable the localization of sequences that have been conserved over different evolutionary time scales and discovery of functional elements restricted to closely-related species (Figure 1).

**Box 2 Evolutionary genetic approaches to mapping fitness effects to the genome**

The development of maps of fitness consequences benefit from a number of complementary approaches.

- Constraint-based methods: These methods primarily use phylogenetic and homology-based inference to identify sites with low rates of substitution across a phylogeny. Sites are assigned scores that are typically interpreted in the context of fitness (e.g. neutral vs. deleterious) without the need to pre-classify sites into groups [12–14].
- SFS methods: A class of methods that use a histogram of allele frequencies, or SFS, to estimate the magnitude of fitness effects in a pre-defined class of sites relative to a neutral class based on allele frequency distributions. In principle, classes of sites subject to selection can be distinguished from those evolving neutrally and estimates of the distribution of fitness effects can be obtained [8].
- Comparative population genomic methods: This class of methods uses intra-specific diversity and between-species divergence data in pre-defined classes of sites to estimate the proportion of sites subject to selection [24••] or the magnitude and sign of the fitness effect [27]. Methods in this class require a neutral class of sites to be contrasted with the site class of interest.
- Effect class methods: These methods predict the impact of mutation on fitness or function by considering properties unique to each effect class. Most methods return a score indicating the likelihood that an individual mutation will impact function [67].

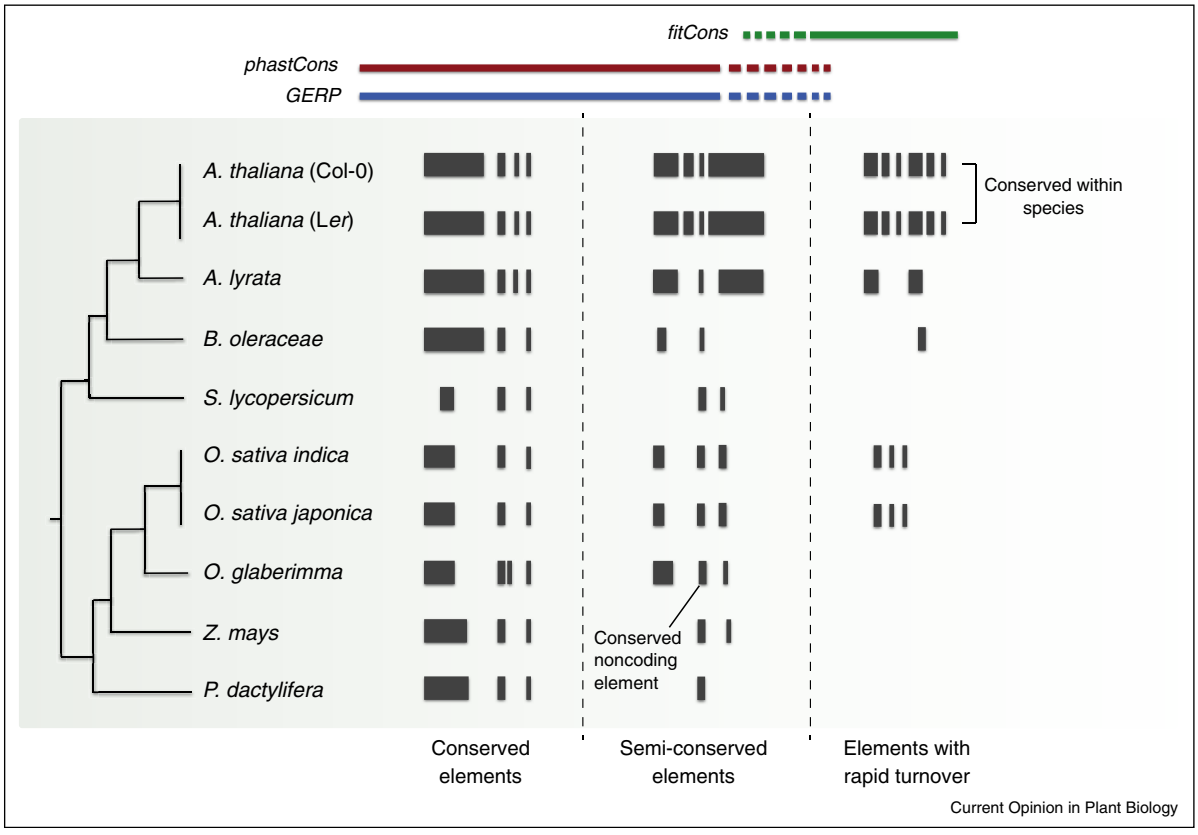
The power of such comparatives approaches is illustrated by a genome-wide, high-resolution atlas of >90 000 conserved noncoding sequences (CNSs) in the Brassicaceae family [16••]. In this study, whole genome sequences from nine closely-related crucifer species and intra-specific diversity data from two species were used to establish that CNSs sequences identified in multi-species alignments are under purifying selection in *Arabidopsis thaliana* and *Capsella grandiflora* populations [16••].

Constraint-based approaches rely on comparisons across multiple species to detect functional elements maintained over million-year timescales [12–14] and can be limited by a number of factors [17•] including low sensitivity to recent changes in constraints associated with either losses or gains in function in protein-coding genes or non-coding DNA. For example, characterizing recently evolved elements is limited by the fact that closely related species sequences are conserved owing to recent common ancestry and distinguishing between conservation owing to recent ancestry or evolutionary constraint is problematic [17•]. In plant genomes, CNSs, such as transcription factor binding sites (TFBSs), experience a more rapid evolutionary turnover compared with animal TFBSs [18••]. Thus, while constraint-based methods are a powerful means to detect elements associated with ancient conserved functions, they are limited in their ability to characterize elements conserved over shorter timescales, like plant TFBSs.

The site frequency spectrum (SFS) is a histogram of allele frequencies, which can be used to infer population demography, identify genomic regions subject to selection, and estimate the distribution of fitness effects of mutations [19]. The shape of the SFS is sensitive to the strength of selection acting on a class (e.g. nonsynonymous) of mutations; purifying selection, for example, shifts the site frequency spectrum towards lower frequencies relative to neutral mutations. SFS methods can thus estimate the distribution of fitness effects (i.e. the fraction of sites subject to different magnitudes of selection, Box 2).

In the context of the study of fitness consequences of genetic polymorphisms, SFS methods have been used to estimate the proportion of new mutations that are neutral, weakly or strongly deleterious from population data [20,21]. In practice, these methods incorporate the observed shape of the SFS (Box 2) for both a neutral and a selected class of mutation and apply numerical methods to estimate the proportion of mutations that are under selection by assuming a distribution of fitness effects and incorporating population parameters such as the mutation

**Figure 1**




Comparison of constraint-based approaches and fitCons to uncover functional elements conserved over different evolutionary time scales in plant genomes. The phylogeny shows representatives of diverged monocot and eudicot species as well as ecotypes for the species *O. sativa* and *A. thaliana*. The grey blocks represent regions of the genome with the following characteristics: (left) conserved syntenic blocks across multiple genomes, (middle) blocks showing stronger synteny between closely related species (*A. thaliana* and *A. lyrata*) but also having smaller elements (e.g. conserved non-coding sequences) that are conserved across more diverged species (semi-conserved elements), and (right) blocks that show intraspecies rather than interspecies conservation (rapid turnover). While constraint-based approaches (e.g. phastCons and GERP) are designed to uncover conserved and semi-conserved elements, fitCons integrates both divergence between relatively closely related species and population genomic data to enable the discovery of semi-conserved elements but also those exhibiting rapid turnover.

rate and effective population size [21]. Although these methods are limited by a strong dependence on prior site class definitions and the relative scarcity of intra-specific polymorphism in strongly selected site classes, they should enable characterization of deleterious mutations in plant genomes [22,23•] and may assist in quantifying the proportion of selected mutations.

Methods that combine population genomics and divergence data represent a powerful means of discovering sites in the genome with fitness consequences. Prominent among these methods is the fitCons method, which estimates the probability of fitness effects of mutations in classes of sites defined by a common function (e.g. a TFBS) by integrating intra-specific polymorphism and between-species divergence data with functional genomic information [24••]. Its foundation is a statistical method called Natural Selection from Interspersed Genomically Coherent Elements (INSIGHT) [25••], which is conceptually similar to population genetics methods that use patterns of polymorphism and divergence to identify departures from neutral expectations [26–29] (Box 2). The contrast between polymorphism and divergence is a powerful approach to inferring recent selection and the INSIGHT approach to pooling dispersed sites enables the discovery of noncoding elements that may have been subject to recent selection [25••]. In this respect, fitCons complements constraint-based methods by identifying functional elements that are recent in origin (Figure 1).

To generate a fitCons map, genomic regions are first partitioned into classes of sites that share similar functional attributes determined across multiple assays (e.g. RNA-seq, DNase-seq, ChIP-seq). To be successful, this will require generating highly informative genomic data types (e.g. non-redundant data sets) with high quality sequencing (e.g. depth, assembly). Sites within a class are assigned a fitCons score that reflects the probability of a

fitness consequence of mutations as inferred by INSIGHT. This approach has the advantage of being annotation-free and facilitates prediction of *cis* regulatory elements and measurement of the global influence of recent natural selection across the genome [24••]. With the increased availability of plant genomic and functional data, we believe that the approach described by Gulko *et al.* [24••] will facilitate assessment of the fitness effects of mutations, estimates of the proportion of plant genomes subject to selection, and discovery of functional elements in plant genomes.

A related approach is the Combined Annotation-Dependent Depletion (CADD) method that integrates information from diverse genome annotations into a single measure (*C* score) to identify putative deleterious, or pathogenic, variants [30•]. The approach relies on mutation-disease association databases such as ClinVar [31] to train a machine learning algorithm to predict fitness consequences. At present, this limits the application of CADD to humans, but highlights the need for development of comparable plant databases [30•,32].

## Characterizing fitness effects by functional effect class

Genomewide approaches to assess fitness consequences are complemented by methods that predict the effects of mutations in specific functional classes. Perhaps most well known in this class of methods are those that evaluate the effect of missense mutations on protein structure with the aim of identifying pathogenic effects (e.g. Sift [33], PolyPhen2 [34], and Provean [35]). Many of these methods make predictions by combining diverse sources of information including sequence conservation across phylogeny, protein structure, gene network topology (e.g. SuSpect [36]), and clinical information on known mutation-disease associations (e.g. GESPA [37]). Such methods may assist in identifying mutations for crop improvement. For example, Shihab *et al.* [38•] implemented the Functional Analysis Through Hidden Markov Models (FATHMM) method to prioritize mutations in starch pathways and storage proteins for improvement in wheat.

Other methods aim to assess fitness effects of mutations in other annotation classes including microRNAs (mrSNP [39]), non-coding RNAs (RNAsnp [40]), and splice sites (SNPlice [41]). Most of these predictive methods can be implemented in plants and it should become possible to pre-compute exhaustively the effects of all possible mutations in each of these effect classes in well-annotated plant genomes [42].

## Mapping the effects of single mutations

Fitness is manifested through specific organismal phenotypes; thus, another way to examine the genetic basis for fitness is via forward genetic approaches whose aim is to associate mutations with traits of interest. Genetic mapping techniques including quantitative trait loci (QTL) mapping and genome-wide association studies (GWAS) have been widely applied in plant genomics to map key phenotypes such as fitness-related traits relevant to crop improvement [11].

When coupled with map-based cloning of causal mutations, these methods complement genomewide maps of fitness consequences by linking specific mutations to traits of interest. For example, the seed protective structure called the awn, varies in structure and number among cereal crop species. Long awns are found in the wild rice *Oryza rufipogon*, while domesticated *Oryza sativa* rice have been selected to have short or no awns to facilitate harvesting and storage [43]. Mapping of phenotypes to specific mutations such as awn traits to naturally-occurring alleles in *Awn-1* (*An-1*) and *LONG AND BARBED AWN1* (*LABA1*) genes in rice [44•,45•] greatly enriches fitness consequence maps by both establishing the trait impacted by specific mutations and suggesting a mechanistic basis for the trait.

Characterization of selective sweeps provides another means of further enriching fitness consequence maps through the identification of regions associated with the fixation of adaptive mutations. Such sweep regions can be identified by various approaches, including local reduction of genetic diversity [46], extended haplotype homozygosity, or local skew in SFS [47]. Regions of positive selection associated with local adaptation can also be inferred by local elevation of sequence divergence or reductions in gene flow using Fst outlier, and other methods [48,49]. These techniques have been widely applied in plant evolutionary genomics [49] and are the basis for important conclusions related to crop domestication [50]. In practice, inferring selective sweeps is challenging due to the confounding effects of population demography as is characterization of the adaptive mutation responsible for a selective sweep which may range in size from $\sim 10^3$ to $10^6$ bps [51]. Nevertheless, when coupled with fine-mapping approaches, these methods provide a means of enriching maps of fitness consequences through the discovery of adaptive mutations.

## Conclusion and outlook

As whole genome resequencing and functional genomic datasets proliferate, the ability to distinguish among neutral, deleterious or adaptive variants in the form of fitness consequence maps will have increasing utility for evolutionary geneticists and plant biologists.

While such maps of fitness consequences will be useful, it should be noted that those generated by approaches such as fitCons are probabilistic in nature and predict fitness consequences for large groups of sites. These maps provide hypotheses for sites subject to selection and verifying the impacts of individual mutations of interest will

require experimental validation [24••,52,53]. Fortunately, new experimental approaches such as CRISPR/Cas and deep mutational scanning [54•], offer the possibility of systematically evaluating the impact of site-specific mutations [55]. The CRISPR/Cas system has been applied in multiple plant species including *Arabidopsis* [56•,57,58], rice [56•], wheat [59], maize [60], sorghum [58], tobacco [57,58], and citrus [61]. Other approaches involve developing and improving high-throughput phenotypic facilities to simultaneously monitor multiple traits [62] or mutations [54•].

Another area that will require attention centers on the need for appropriate repositories to visualize large-scale datasets and fitness maps to facilitate the discovery of functional elements. Phytozome [15] has recently expanded the ability to visualize SNP data, VISTA conservation tracks, and other whole genome plant datasets via JBrowse [63], and CoGE [64] allows users to customize their own instances of JBrowse and conduct comparative analysis that should improve the ability to discover conserved elements. Similar resources that facilitate visualization of information relevant to discovery of mutations with fitness consequences will be important for identifying candidate functional polymorphisms.

We envision that maps of fitness consequences can be seen as part of a tool kit designed to discover gene regions or mutations useful to quantitative geneticists, evolutionary plant biologists and crop breeders. For example, genome-wide maps of fitCons scores coupled with the location of putative selective sweeps or large QTL regions could help in fine mapping genomic regions with important alleles, and provide the basis of subsequent functional analyses. Also, future applications for plant genetics include targeting adaptive mutations (i.e. mutations formed in response to an environment in which the mutations were selected) by comparing multiple fitness consequence maps of a given crop variety under different environmental conditions (e.g. drought, high salinity). Linking adaptive mutations to QTL-mapped traits would also help distinguish mutations that have a deleterious effect from mutations that may be locally adaptive under certain conditions. Deleterious mutations could also be targeted and potentially bred out of a population [23•].

Finally, fitness consequence maps not only have applications in inferring function in plant genomes, but can also help in addressing long-standing problems in evolution including estimation of the proportion of the genome that is subject to selection and predicting the fitness consequence of mutations in regions subject to selection.

## Acknowledgements

We apologize to all colleagues whose relevant work we could not cite because of space limitation. We thank Adam Siepel, Adrian E. Platts, Evan Baugh, and Olivia Wilkins for fruitful discussions. We acknowledge support through grants from the Zegar Family Foundation, NYU Abu Dhabi Research Institute and the US NSF Plant Genome Research Program.

## References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Bajaj D, Das S, Badoni S, Kumar V, Singh M, Bansal KC, Tyagi AK, Parida SK: **Genome-wide high-throughput SNP discovery and genotyping for understanding natural (functional) allelic diversity and domestication patterns in wild chickpea**. *Sci Rep* 2015, **5**:12468.

2. Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, Meitinger T, Strom TM, Fries R, Pausch H *et al.*: **A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array**. *BMC Genomics* 2014, **15**:823.

3. Alexandrov N, Tai S, Wang W, Mansueto L, Palis K, Fuentes RR, Ulat VJ, Chebotarov D, Zhang G, Li Z *et al.*: **SNP-Seek database of SNPs derived from 3000 rice genomes**. *Nucleic Acids Res* 2015, **43**:D1023-D1027.

4. Hazzouri KM, Flowers JM, Visser HJ, Khierallah HSM, Rosas U, Pham GM, Meyer RS, Johansen CK, Patrick ZF, Masmoudi K *et al.*: **Whole genome re-sequencing of date palms yield insights into diversification of a fruit tree crop**. *Nat Commun* 2015, **6**:8824.

5. Flowers JM, Hazzouri KM, Pham GM, Rosas U, Bahmani T, Khraiwesh B, Nelson DR, Jijakli K, Abdrabu R, Harris EH *et al.*: **Whole-genome resequencing reveals extensive natural variation in the model green alga *Chlamydomonas reinhardtii***. *Plant Cell* 2015, **27**:2353-2369.

6. •• Li JY, Wang J, Zeigler RS: **The 3000 rice genomes project: new opportunities and challenges for future rice research**. *Gigascience* 2014, **3**:8.
This report highlights the contribution of the 3000 rice genome project. This joint international effort resequenced 3000 *O. sativa* sampled rice accessions from 89 countries, which produced millions of genomic reads. The findings of these sequencing efforts include the discovery of more than 18.9 million single nucleotide polymorphisms (SNPs).

7. Kimura M: *The Neutral Theory of Molecular Evolution*. Cambridge University Press; 1984.

8. Eyre-Walker A, Keightley PD: **The distribution of fitness effects of new mutations**. *Nat Rev Genet* 2007, **8**:610-618.

9. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E: **On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE**. *Genome Biol Evol* 2013, **5**:578-590.

10. Albert FW, Kruglyak L: **The role of regulatory variation in complex traits and disease**. *Nat Rev Genet* 2015, **16**:197-212.

11. Huang X, Han B: **Natural variations and genome-wide association studies in crop plants**. *Annu Rev Plant Biol* 2014, **65**:531-551.

12. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies**. *Genome Res* 2010, **20**:110-121.

13. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al.*: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res* 2005, **15**:1034-1050.

14. Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, Sidow A: **Distribution and intensity of constraint in mammalian genomic sequence**. *Genome Res* 2005, **15**:901-913.

15. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N *et al.*: **Phytozome: a comparative platform for green plant genomics**. *Nucleic Acids Res* 2012, **40**:D1178-D1186.

16. •• Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM *et al.*: **An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions**. *Nat Genet* 2013, **45**:891-898.

Haudry *et al.* use comparative genomics methods using nine closely related *Brassicaceae* species to identify thousands of conserved non-coding sequences. They additionally use population genomics between two of these species to validate that the CNSs detected remain under selection in recent evolutionary time.

17. • Lawrie DS, Petrov DA: **Comparative population genomics: power and principles for the inference of functionality**. *Trends Genet* 2014, **30**:133-139.
Interesting review that looks at comparative population genomics to discover functional elements. The authors argue that integrating different data source can be powerful to detect functional elements that are under different levels of selection.

18. •• Burgess DG, Xu J, Freeling M: **Advances in understanding cis regulation of the plant gene with an emphasis on comparative genomics**. *Curr Opin Plant Biol* 2015, **27**:141-147.
This review compares different approaches that have been used to uncover conserved non-coding regions in plants. In addition to describing their similarities and differences, this review compares the efficiency of the methods by testing them against two robust CNS datasets in plants.

19. Henn BM, Botigue LR, Bustamante CD, Clark AG, Gravel S: **Estimating the mutation load in human genomes**. *Nat Rev Genet* 2015, **16**:333-343.

20. Fay JC, Wyckoff GJ, Wu CI: **Positive and negative selection on the human genome**. *Genetics* 2001, **158**:1227-1234.

21. Eyre-Walker A, Woolfit M, Phelps T: **The distribution of fitness effects of new deleterious amino acid mutations in humans**. *Genetics* 2006, **173**:891-900.

22. Mezmouk S, Ross-Ibarra J: **The pattern and distribution of deleterious mutations in maize**. *G3 (Bethesda)* 2014, **4**:163-171.

23. • Renaut S, Rieseberg LH: **The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops**. *Mol Biol Evol* 2015, **32**:2273-2283.
This study looks at the effects of plant domestication and the potential reduction in the efficiency of natural selection at purging harmful mutations from the genome. Using different wild and domesticated accessions of the common sunflower, cardoon, and globe artichoke, Renaut and Rieseberg show that the distribution of different kinds of mutations in wild and domesticated sunflower families correlates with an excess of deleterious mutations as a consequence of domestication over the last 4000 years.

24. •• Gulko B, Hubisz MJ, Gronau I, Siepel A: **A method for calculating probabilities of fitness consequences for point mutations across the human genome**. *Nat Genet* 2015, **47**:276-283.
This paper describes an innovative and very promising approach that measure probabilities of mutational fitness consequences (fitCons) for nucleotide positions within different classes of elements in the genome, including noncoding regions. fitCons are calculated by combining functional genomic methods with the computational method INSIGHT described by Gronau *et al.*, 2013. The fact that the fitCons method does not rely on previous annotations or a priori trait selection makes it possible to be applied in plant genomes.

25. •• Gronau I, Arbiza L, Mohammed J, Siepel A: **Inference of natural selection from interspersed genomic elements based on polymorphism and divergence**. *Mol Biol Evol* 2013, **30**:1159-1171.
This paper describes the computational method INSIGHT, which is a method that uses a generative probabilistic model to contrast patterns of polymorphism and divergence in elements with those in flanking neutral sites. This method is particularly useful to infer selection on dispersed noncoding elements such as TFBS and noncoding RNAs.

26. McDonald JH, Kreitman M: **Adaptive protein evolution at the Adh locus in Drosophila**. *Nature* 1991, **351**:652-654.

27. Sawyer SA, Hartl DL: **Population genetics of polymorphism and divergence**. *Genetics* 1992, **132**:1161-1176.

28. Smith NG, Eyre-Walker A: **Adaptive protein evolution in Drosophila**. *Nature* 2002, **415**:1022-1024.

29. Andolfatto P: **Adaptive evolution of non-coding DNA in Drosophila**. *Nature* 2005, **437**:1149-1152.

30. • Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants**. *Nat Genet* 2014, **46**:310-315.
This paper describes the computational method CADD, which has been used in humans to generate a '*C* score', which is a measure of deleteriousness for each variant (SNP or short insertions–deletions). CADD integrates different types of annotations and therefore can be of clinical relevance in humans, through the use of the ClinVar database.

31. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR: **ClinVar: public archive of relationships among sequence variation and human phenotype**. *Nucleic Acids Res* 2014, **42**:D980-D985.

32. Seaver SM, Gerdes S, Frelin O, Lerma-Ortiz C, Bradbury LM, Zallot R, Hasnain G, Niehaus TD, El Yacoubi B, Pasternak S *et al.*: **High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource**. *Proc Natl Acad Sci U S A* 2014, **111**:9645-9650.

33. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm**. *Nat Protoc* 2009, **4**:1073-1081.

34. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**:248-249.

35. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP: **Predicting the functional effect of amino acid substitutions and indels**. *PLoS ONE* 2012, **7**:e46688.

36. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **SUSPECTS: enabling fast and effective prioritization of positional candidates**. *Bioinformatics* 2006, **22**:773-774.

37. Khurana JK, Reeder JE, Shrimpton AE, Thakar J: **GESPA: classifying nsSNPs to predict disease association**. *BMC Bioinformatics* 2015, **16**:228.

38. • Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR: **Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models**. *Hum Mutat* 2013, **34**:57-65.
This paper introduces an HMM-based method that predicts phenotypic effects of missense mutations and uses wheat as a case study. It is an interesting example where a method developed for humans was applied to plants.

39. Deveci M, Catalyurek UV, Toland AE: **mrSNP: software to detect SNP effects on microRNA binding**. *BMC Bioinformatics* 2014, **15**:73.

40. Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J: **RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs**. *Hum Mutat* 2013, **34**:546-556.

41. Mudvari P, Movassagh M, Kowsari K, Seyfi A, Kokkinaki M, Edwards NJ, Golestaneh N, Horvath A: **SNPlice: variants that modulate Intron retention from RNA-sequencing data**. *Bioinformatics* 2015, **31**:1191-1198.

42. Katsonis P, Lichtarge O: **A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness**. *Genome Res* 2014, **24**:2050-2058.

43. Kovach MJ, Sweeney MT, McCouch SR: **New insights into the history of rice domestication**. *Trends Genet* 2007, **23**:578-587.

44. • Luo J, Liu H, Zhou T, Gu B, Huang X, Shangguan Y, Zhu J, Li Y, Zhao Y, Wang Y *et al.*: **An-1 encodes a basic helix-loop-helix protein that regulates awn development, grain size, and grain number in rice**. *Plant Cell* 2013, **25**:3360-3376.
Characterization of *An-1*, which encodes a basic helix-loop-helix protein that regulates cell division in rice. *An-1* has been one of the major targets for artificial selection in cultivated rice because it has an effect on awn development and yield.

45. • Hua L, Wang DR, Tan L, Fu Y, Liu F, Xiao L, Zhu Z, Fu Q, Sun X, Gu P *et al.*: **LABA1, a domestication gene associated with long, barbed awns in wild rice**. *Plant Cell* 2015, **27**:1875-1888.
Exhaustive study that characterizes *LABA1*, a gene involved in lawn development and encoding a cytokinin-activating enzyme. *LABA1* lies in a previously identified selective sweep on chromosome 4 and sequence analysis between cultivated rice and wild rice shows a frameshift deletion

in the *LABA1* of cultivated rice, which may have been selected for in early rice domestication.

46. Smith JM, Haigh J: **The hitch-hiking effect of a favourable gene**. *Genet Res* 1974, **23**:23-35.

47. Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI: **Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora***. *PLoS Genet* 2014, **10**:e1004622.

48. Wright S: **Genetical structure of populations**. *Nature* 1950, **166**:247-249.

49. Holsinger KE, Weir BS: **Genetics in geographically structured populations: defining, estimating and interpreting F(ST)**. *Nat Rev Genet* 2009, **10**:639-650.

50. Wang M, Yu Y, Haberer G, Marri PR, Fan C, Goicoechea JL, Zuccolo A, Song X, Kudrna D, Ammiraju JS *et al.*: **The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication**. *Nat Genet* 2014, **46**:982-988.

51. Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C *et al.*: **Whole-genome sequencing of multiple *Arabidopsis thaliana* populations**. *Nat Genet* 2011, **43**:956-963.

52. Ciofani M, Madar A, Galan C, Sellars M, Mace K, Pauli F, Agarwal A, Huang W, Parkurst CN, Muratet M *et al.*: **A validated regulatory network for Th17 cell specification**. *Cell* 2012, **151**:289-303.

53. Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, Balakishnan B, Liang R, Zhang Y, Lyon S *et al.*: **Comparison of predicted and actual consequences of missense mutations**. *Proc Natl Acad Sci U S A* 2015, **112**:E5189-E5198.

54. • Fowler DM, Fields S: **Deep mutational scanning: a new style of protein science**. *Nat Methods* 2014, **11**:801-807.
Interesting method used to generate large-scale mutational data for nearly any protein using cells *in vitro* and *in vivo*. Not applied in plants but promising as a future method, coupled with rapidly developing technology and high-throughput sequencing.

55. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E: **A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity**. *Science* 2012, **337**:816-821.

56. • Feng Z, Zhang B, Ding W, Liu X, Yang DL, Wei P, Cao F, Zhu S, Zhang F, Mao Y *et al.*: **Efficient genome editing in plants using a CRISPR/Cas system**. *Cell Res* 2013, **23**:1229-1232.
First report that demonstrates the use of the CRISPR/Cas system in plants to target mutations and corrections. This paper shows evidence for both *Arabidopsis* and rice.

57. Li JF, Norville JE, Aach J, McCormack M, Zhang D, Bush J, Church GM, Sheen J: **Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9**. *Nat Biotechnol* 2013, **31**:688-691.

58. Jiang W, Zhou H, Bi H, Fromm M, Yang B, Weeks DP: **Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice**. *Nucleic Acids Res* 2013, **41**:e188.

59. Wang Y, Cheng X, Shan Q, Zhang Y, Liu J, Gao C, Qiu JL: **Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew**. *Nat Biotechnol* 2014, **32**:947-951.

60. Liang Z, Zhang K, Chen K, Gao C: **Targeted mutagenesis in Zea mays using TALENs and the CRISPR/Cas system**. *J Genet Genomics* 2014, **41**:63-68.

61. Bortesi L, Fischer R: **The CRISPR/Cas9 system for plant genome editing and beyond**. *Biotechnol Adv* 2015, **33**:41-52.

62. Yang W, Guo Z, Huang C, Duan L, Chen G, Jiang N, Fang W, Feng H, Xie W, Lian X *et al.*: **Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice**. *Nat Commun* 2014, **5**:5087.

63. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH: **JBrowse: a next-generation genome browser**. *Genome Res* 2009, **19**:1630-1638.

64. Lyons E, Freeling M: **How to usefully compare homologous plant genes and chromosomes as DNA sequences**. *Plant J* 2008, **53**:661-673.

65. Hartl DL, Dykhuizen DE, Dean AM: **Limits of adaptation: the evolution of selective neutrality**. *Genetics* 1985, **111**:655-674.

66. • Rest JS, Morales CM, Waldron JB, Opulente DA, Fisher J, Moon S, Bullaughey K, Carey LB, Dedousis D: **Nonlinear fitness consequences of variation in expression level of a eukaryotic gene**. *Mol Biol Evol* 2013, **30**:448-456.
Interesting study looking at fitness and function in yeast with response to abiotic stress. It shows that fitness and function are not linearly related.

67. Katsonis P, Koire A, Wilson SJ, Hsu TK, Lua RC, Wilkins AD, Lichtarge O: **Single nucleotide variations: biological impact and theoretical interpretation**. *Protein Sci* 2014, **23**:1650-1666.