

Predicting the Genetic Loci of Past Evolution

VIRGINIE COURTIER-ORGOGOZO

ARNAUD MARTIN

Introduction

Evolution reveals itself by the changes in observable characteristics of biological populations over successive generations. Here we focus on the DNA mutations underlying phenotypic changes that have occurred during natural evolution of populations or species, as well as through domestication and experimental evolution. The search for the mutations responsible for evolutionary changes started with iconic case studies such as the ABO blood group gene (F.I. YAMAMOTO *et al.* 1990) or hemoglobin for drepanocytosis and malaria resistance (V.M. INGRAM 1957). As instances of genes causing phenotypic changes between populations and species started to accumulate, some researchers noticed that the mutations causing evolution did not appear to be randomly distributed across the genomes. Intriguing cases of repeated evolution at the genetic level were reported, with recurrent genetic changes involved in the evolution of similar phenotypes in distant taxa (A.H. PATTERSON *et al.* 1908; N. GOMPEL and B. PRUD'HOMME 2009; A. KOPP 2009; P.A. CHRISTIN *et al.* 2010; A.E. LOBKOVSKY and E.V. KOONIN 2012; G.L. CONTE *et al.* 2012; A. MARTIN and V. ORGOGOZO 2013; D.L. STERN 2013; T. LENSER and G. THEISSEN 2013). Furthermore, certain types of phenotypic changes seemed to be preferentially associated with certain broad categories of mutations (S.B. CARROLL 2000; E.H. DAVIDSON 2006; G.A. WRAY 2007; B. PRUD'HOMME; N. GOMPEL and S.B. CARROLL 2007; S.B. CARROLL 2008; D.L. STERN and V. ORGOGOZO 2008; V.J. LYNCH and G. WAGNER 2008; G. WAGNER and V.J. LYNCH 2008; M.A. STREISFELD and M.D. RAUSHER 2011; D.L. STERN 2011). For example, morphological evolution in animals was suggested to preferentially involve cis-regulatory mutations rather than coding changes. Please note that “preferentially” refers here to the final consequence of selection and of other population genetics processes and does not necessarily mean that mutations occur in a non-random fashion. Genetic variations occur throughout the genome, some of them being gradually

eliminated while others are maintained, thus allowing enrichments in certain types of mutations when one looks at the result of selection and of other population genetics processes over multiple generations.

Observation of consistent patterns in the genetic loci of evolution has had two main consequences on evolutionary biology research. First, it prompted the elaboration of various explanatory hypotheses. Second, it led some biologists to propose that the genetic loci of evolution are partly predictable, in the sense that for a given phenotypic change that occurred in the past the underlying mutations can be guessed with reasonable confidence. What is predicted here is the genetic causes of evolutionary events that occurred in the past (V. ORGOGOZO 2015; V. ORGOGOZO, B. MORIZOT, A. MARTIN 2015), and not the mutations that will occur in the future (for such cases see LOBKOVSKY and KOONIN 2012; R.E. LENSKI 2017; M. LÄSSING, V. MUSTONEN and A.M. WALCZAK 2017).

In this paper we examine the predictions regarding the genetic loci of past evolution: what kinds of predictions are they? What are they based on? How accurate are they?

Nature of the Predictions

Published predictions about the genetic loci of evolution do not arise from complex mechanistic models. They simply derive from the observation of repeated cases of genetic evolution and the identification of the context and the hypothetical causes that lead to repetitions. For example, the loss of larval trichomes in *Drosophila sechellia* was found to be caused by multiple mutations in five distinct enhancers of the *svb* gene, with each enhancer regulating trichome development in a specific body part (N. FRANKEL *et al.* 2010). Such recurrent evolution at the same locus, together with the special “hub” position of *svb* in the gene network for trichome development, suggested that *svb* was a hotspot gene for trichome evolution in flies (A.P. MCGREGOR *et al.* 2007; STERN and ORGOGOZO 2008; STERN 2011). The independent loss of trichomes in *Drosophila exoana* was thus inferred to involve cis-regulatory mutations in *svb*, as in *D. sechellia*, and indeed this was found to be true (N. FRANKEL, S. WANG, D.L. STERN 2012). Agricultural pests and mosquitoes have repeatedly evolved resistance to pyrethroid insecticides such as DDT via coding mutations in the *para voltage-gated sodium channel* gene (*syn. para, vgsC*) (D.M. SODERLUND 2008). Weston *et al.* predicted that pyrethroids may also affect non-pest organisms that populate sprayed areas (D.P. WESTON *et*

al. 2013). Not only they found that pyrethroid resistance had evolved in populations of freshwater crustaceans exposed to agricultural run-off, but they also uncovered typical mutations in *para* gene, mirroring the evolutionary mechanisms previously observed in the targeted pests. Thus, a specific evolutionary pressure left a predictable genetic signature in the environment that can now be detected. In principle, comparable predictions could be done in other situations where human activity has chemically modified the environment.

Predictions rely on the assumption that the set of already known loci of evolution on which predictions are based are identified via an unbiased approach. Linkage mapping studies and association studies screen the entire genome sequence and are thus supposed to be unbiased in their detection of the genetic loci. However, one should keep in mind that once a genomic region is narrowed down to a few candidate genes through linkage mapping or association mapping, knowledge from past studies might favor for validation tests the candidate genes that are already known to be involved in a similar phenotypic change. As a result, even genome-wide mapping studies carry some bias towards already known genes.

Predictions are usually formulated along the following lines: “For a given phenotypic change, it is predicted that the causal mutations are such and such”. Formulations can also be relatively more complex. For evolution of red flowers in *Penstemon barbatus*, Wessinger and Rausher not only predicted the causal genes (*F35'H* or *F3'H*) and the fact that the mutations should create loss-of-function alleles but also provided gene-specific details about the expected mutations: “when it involves elimination of F3'5'H activity, functional inactivation or deletion of this gene tends to occur; however, when it involves elimination of F3'H activity, tissue-specific regulatory substitutions occur and the gene is not functionally inactivated” (C.A. WESSINGER and M.D. RAUSHER 2014).

Predicting the genetic loci of evolution is a retrodiction

Predictions are usually inferences about the future, based on current knowledge about the past. When the causal temporality is reversed, some authors prefer to use the term retrodiction. Retrodicting can be defined as making an inference about an event that occurred in the past (A. LOVE, personal communication; J. FETZER 2017). If this past event is already known, then retrodiction is the act of providing a *post hoc* explanation for it. For

example, Darwin retrodicted why species similar to those found on oceanic islands are usually found on the nearest mainland (C. DARWIN 1859). In other cases, retrodiction can produce facts that are yet unknown and that we should observe. For example, Darwin retrodicted from the existence of orchids with very long nectaries that we should discover moths with equally long proboscis (DARWIN 1859). Predictions are not always explanatory: they do not necessarily rely on a model or on causal explanations. For example, a prehistorical astronomer without a heliocentric theory who accumulates observations such as “the sun always rises above that hill” would predict that the sun will rise above the hill again, simply by noticing the repetitions. In our case, published predictions about the genetic loci of evolution do not arise from complex mechanistic models, they instead derive from the observation of repeated cases of genetic evolution. In this sense, evolutionary geneticists resemble prehistorical astronomers without a heliocentric theory. Using repetitions among the known loci of evolution to make predictions about the past is thus a retrodiction. Moreover, predictions about past genetic loci concern a genetic difference that exists today between two living taxa (a property of the present state). Predictions about the genetic loci of past evolution are thus retrodictions about the past and the present.

Predictions at various genetic levels

Predictions can be made at the gene level, as for *svb* or the *para* sodium channel gene, but also at higher and lower genetic levels: at the level of a nucleotide, of part of a coding region, of a specific enhancer, of a group of genes, and also for broader categories of genetic changes (Table 1).

Table 1. Repetitions can be detected and predicted at various genetic levels, from specific nucleotides to general classes of mutations

Genetic Level	Example of prediction or of repeated evolution
---------------	--

<p>Nucleotide</p>	<p>Resistance to cyclodiene insecticides was successfully predicted to be associated with amino acid substitutions at a single residue (A302) within the gamma-aminobutyric acid (GABA) receptor sub-unit named Rdl in the cat flea <i>Ctenocephalides felis</i> for 8 of the 9 laboratory strains that were tested (C. BASS <i>et al.</i> 2004).</p>
<p>Part of a coding region</p>	<p>Various amino acid substitutions in the DIII and DIV pore loops of the sodium channel Nav1.4 explain tetrodotoxin resistance in newts, snakes, pufferfishes (G. TOLEDO <i>et al.</i> 2016) and to red tide toxin-resistance in clams (V.M. BRICELJ <i>et al.</i> 2005).</p>
<p>Enhancer</p>	<p>Pelvic loss in sticklebacks originating from 13 different lakes is due to various deletions with distinct breakpoints of the same <i>pitx1</i> enhancer region in 9 cases and to other yet unknown genetic change(s) in 4 cases (Y.F. CHAN <i>et al.</i> 2010).</p>
<p>Gene</p>	<p>The gene <i>svb</i> was successfully predicted to cause an evolutionary loss of trichomes in <i>Drosophila ezoana</i> (N. FRANKEL, S. WANG, D.L. STERN 2012).</p> <p>The gene <i>WntA</i> was successfully predicted to cause wing color pattern variation across many <i>Heliconius</i> species and populations (A. MARTIN <i>et al.</i> 2012), and beyond (J.R. GALLANT <i>et al.</i> 2014).</p>
<p>Gene Family, Gene Paralogs</p>	<p>Interspecific changes in anthocyanin pigment intensity in flowers was successfully predicted to involve preferentially mutations in transcription factor genes of the MYB family (M.A. STREISFELD and M.D. RAUSHER 2011).</p>
<p>Genetic Pathway</p>	<p>It was successfully predicted that flower color evolution in <i>Penstemon barbatus</i> was caused by the inactivation of one of the candidate anthocyanin pathway genes, <i>F3'H</i>, <i>F3'5'H</i> and <i>DFR</i> (C.A. WESSINGER and M.D. RAUSHER 2014).</p>

Gene Ontology	Animal morphological evolution is predicted to involve preferentially signaling ligand encoding genes, i.e. genes associated with the gene ontology (GO) terms “extracellular region” and “receptor binding” (A. MARTIN V. and ORGOGOZO 2017).
Broader Class of Mutation	Cis-regulatory mutations are the predominant mechanism underlying the evolution of morphology in animals (S.B. CARROLL 2008; D.L. STERN, V. ORGOGOZO 2008).

Some predictions carry more information than others. Predicting the impact of a projectile within 1 m is better than within 1 km. In a first approximation, the *inference gain* can be estimated by the inverse of the probability of the predicted outcome according to the null model. In most cases, the null model considers that mutations can occur with equal probability at all nucleotide positions within a genome (ORGOGOZO 2015), or with equal probability in all genes of a genome. The prediction that trichome pattern evolution in *Drosophila* is likely to involve the *svb* gene (MCGREGOR *et al.* 2007) has an inference gain of approximately 17,500 (*svb* is one gene out of the total of 17,737 genes present in the genome of *D. melanogaster* (FIBase FB2017_04)). In contrast, the prediction that morphological evolution is likely to involve signaling ligand encoding genes (A. MARTIN and V. ORGOGOZO 2017) has a lower inference gain, estimated to be about 180 in the stickleback fish, as the proportion of genes associated with gene ontology (GO) terms “extracellular region” and “receptor binding” in the stickleback *G. aculeatus* (BROADS1) is 115/20,787 (MARTIN and ORGOGOZO 2017). Predictions at the level of the gene or at the nucleotide level have a higher inference gain than predictions about broader classes of mutations.

The causes of genetic repetition

In general, authors put forward several arguments for the over-representation of certain genetic loci for phenotypic traits of interests and suggest that their combination explains repeated evolution. Arguments for the importance of cis-regulatory mutations versus coding changes are not reviewed

here and can be found in CARROLL 2000, CARROLL 2008, DAVIDSON 2006, WRAY 2007, STERN and ORGOGOZO 2008, D.L. STERN and V. ORGOGOZO 2009, B.-Y. LIAO, M.-P. WENG and J. ZHANG 2010, STERN 2011.

Several non-exclusive hypotheses have been proposed to explain why some genetic changes repeatedly drive certain phenotypic changes or adaptations during evolution (Table 2).

Table 2. Non-exhaustive list of arguments that have been proposed to explain hotspot genes

Argument	Example
Mutation bias	<p><i>Pitx-1</i> is located within a fragile chromosome region, which exhibits a higher rate of breakage and is thus prone to deletions caused by DNA break repair by non-homologous end joining (NHEJ) (Y.F. CHAN <i>et al.</i> 2010).</p> <p>Primate chimeric genes involved in retroviral defense have evolved multiple times through the transposition of <i>CypA</i>, encoding a protein with high affinity for viral proteins, with a <i>Trim5alpha</i>, encoding a protein involved in virus degradation. Such an event seems intuitively unlikely but it did occur several times, probably because <i>CypA</i> is prone to retrotransposition and formation of hybrid genes by exon shuffling (R. MALFAVON-BORJA. 2013; Z. ZHANG. 2003).</p>
Large target size	<p>The <i>Mt1R</i> coding region contains a large number of sites which, when mutated, yield a viable phenotypic change (N. GOMPEL and B. PRUD'HOMME 2009).</p> <p><i>Oca2</i> is a large gene: in humans it contains 24 exons spanning 345 kb (N. GOMPEL and B. PRUD'HOMME 2009).</p>

<p>Intragenic modularity</p>	<p>In the mouse <i>Peromyscus</i> the <i>Agouti</i> gene harbors a large cis-regulatory region decomposed into several modules, allowing many nucleotide sites to be tweaked to generate precise and specific pigmentation changes in various parts of the body (C.R. LINNEN. 2009).</p>
<p>Genetic potentiality</p>	<p>The reduced content of amylose, a characteristics of cereals with sticky texture, can only arise via mutations within the <i>Waxy</i> gene, which encodes a granule-bound starch synthase (T. LENSER and G. THEISSEN 2013).</p> <p>No other mutations than coding mutations in opsin genes can lead to change in light wavelength sensitivity (S.C. MORRIS 2003).</p>
<p>Architecture of Gene Networks</p>	<p>The <i>svb</i> gene sits at a “hub” position in the gene network for trichome development (A.P. MCGREGOR. 2007).</p>
<p>Optimal Pleiotropy</p>	<p>Interspecific changes in anthocyanin pigment intensity in flowers is more likely to involve transcription factor genes of the MYB family than basic helix-loop-helix (bHLH) or WD40 repeat family members because MYB genes usually have more tissue-specific functions and thus fewer pleiotropic effects (M.A. STREISFELD and M.D. RAUSHER 2011; A. KOPP 2009).</p>
<p>Mutational Path</p>	<p>A mutation in tuberculosis strains providing resistance to a first antibiotic acted as a preliminary condition and strong predictor for the secondary evolution of multidrug resistance on five continents (A. MANSON 2017).</p>

Hemiplasy: Standing Variation, Introgression, and Horizontal Gene Transfer	<p>The non-shattering allele <i>sh4</i> in rice was already present at low frequency within wild populations, so that independent selection has probably driven the same allele to fixation in distinct varieties (T. LENSER and G. THEISSEN 2013).</p> <p>Adaptation to high altitude in dogs evolved via an allele of <i>EPAS1</i> that was already present in highland wolves, by interbreeding (Z. FAN, D. ORTEGA-DEL VECCHYO, R.K. WAYNE 2017).</p>
--	--

A first category of explanation can be attributed to mutational bias: certain mutations are more likely to occur than others. For instance, a region that is prone to structural variation or transposon insertion is likely to undergo repeated rearrangements, thus facilitating certain gene-to-phenotype changes (Y.F. CHAN *et al.* 2010). Adaptation to high-altitude in *Andean* house wrens and hummingbirds has been associated with single point mutations in the βA -*globin* gene and these mutations appear to lie within CpG sites, which are known to be chemically unstable and highly mutable (S.C. GALEN *et al.* 2015; A. STOLTZFUS and D.M. McCANDLISH 2015; M. LYNCH and B. WALSH 2007). A recent study of various high-altitude species (C. NATARAJAN *et al.* 2016) discovered other mutations in βA -*globin* that increase affinity for oxygen, indicating that other genetic paths are theoretically possible. In *Andean* house wrens the spontaneous mutation rate appears to have biased evolution towards certain genetic paths. Spontaneous mutation rates are higher for transitions (A \leftrightarrow G or C \leftrightarrow T) than for transversions (LYNCH and WALSH 2007). Using this property as a test case for investigating the role of mutation biases, Stoltzfus and McCandlish compiled from existing literature a list of putatively adaptive amino-acid changes that have evolved in parallel in natural or experimental contexts. They found a fourfold excess of transitions over transversions (A. STOLTZFUS and D.M. McCANDLISH 2017), suggesting that the repeatability of adaptive coding changes is at least partly explained by biases in the mutations. This is consistent with a “first come, first served” model where even if a number of possible mutational paths to adaptation exist, the ones that are more likely to emerge in the first place are more accessible to selection, and are thus repeatedly observed when the environmental challenge is replicated.

Two other categories of explanations (named “Intragenic Modularity” and “Large Target Size” in Table 2) are related to the fact that compared to other loci that may provide mutational paths to the considered phenotypic change, a given gene may be favored because of its properties at the DNA sequence level: a large intergenic region that can provide relatively more possibilities for a modular control, or a large coding region allowing many amino-acid sites to be tweaked.

A third category of explanation deals with gene function itself rather than just the physical properties of the stretch of DNA hosting their information. Kopp used the term “Optimal Pleiotropy” to propose that only certain genes may host the potential for tuning a given phenotype without deleterious effects. In this concept, the emphasis is on the capacity of certain genes to yield specific effects, whether they simply have a limited number of roles, or on the contrary, a large number of roles but sufficient modularity to allow genetic uncoupling of these roles (A. KOPP 2009). Similarly, the “Architecture of Gene Networks” may highlight “hub” or “input/output” genes that are more likely to coordinate a cascade of changes and thus, drive effects accessible to selection (McGREGOR *et al.* 2007; STERN and ORGOGOZO 2008).

A fourth mechanism deals with “Permissive Mutations” that may relax the constraints preventing a given change and thus open a new valley in the adaptive landscape. This phenomenon was reviewed recently (J.F. STORZ 2016) and has been mostly studied in the framework of protein evolution (R.D. TARVIN *et al.* 2017; J.D. BLOOM, L.I. GONG, D. BALTIMORE 2010; A. MANSON *et al.* 2017; T.N. STARR, L.K. PICTON, J.W. THORNTON 2015). The predictability is of the form “if change A has happened, then change B is likely as well because we observed that A is often followed by B”. In principle, such phenomenon could also apply to genome-wide epistasis, with mutational paths being contingent on allelic states present at distant loci on the genome, as has been observed in “Evolve and Resequence” experiments (J.B. ANDERSON *et al.* 2010; A. LONG *et al.* 2015).

Finally, another explanation underlines the higher capacity of certain derived alleles to circulate among the branches of a species phylogeny, due to various mechanisms such as incomplete lineage sorting or standing genetic variation (*e.g.* C.T. MILLER *et al.* 2007), genetic introgressions in intermixing populations or closely related species (*e.g.* J. ENCISO-ROMERO *et al.* 2017; Z. FAN, D. ORTEGA-DEL VECCHYO, R.K. WAYNE 2017; L.T. DUNNING *et al.* 2017), or even horizontal gene transfers between distant branches of the

tree of life (e.g. F.W. Li *et al.* 2014; J.A. METCALF *et al.* 2014; J. ROPARS *et al.* 2015). These cases, named “collateral evolution” or “hemiplasies” (reviewed in STERN 2013; MARTIN and ORGOGOZO 2013) contrast with *stricto sensu* occurrences of genetic parallelism (R.W. SCOTLAND 2011; STERN 2013; J.F. STORZ 2016) because the alleles are identical-by-descent rather than identical-by-state. The derived phenotypes may be “convergent”, i.e. show a pattern of homoplasy due to a discontinuity on a phylogeny, but the genotypes are not, hence representing a so-called “hemiplasy” (J.C. AVISE and T.J. ROBINSON 2008). As our understanding of gene flow is rapidly improving in the phylogenomic era, it is likely that the *a priori* observation of connectivity between two lineages will increase our capacity to make predictions on causal mechanisms of gene-to-phenotype change (R.W. WALLBANK *et al.* 2016).

In summary, predictable genetic changes may be 1) under the influence of a mutational bias (site-specific rates of mutation); 2) in a large locus, more prone to change due to a sequence size parameter (narrow mutational target size, independently of genetic function – large target size, intragenic modularity); 3) in a gene that is inherently poised to tweak a given trait due to properties of its molecular function or regulatory interactions (narrow mutational target size, due to the structure of the genotype-phenotype map; genetic potentiality; architecture of gene networks; optimal pleiotropy); 4) contingent upon the pre-existence of other changes in the genome (mutational path); 5) a by-product of allele sorting or transfer processes (hemiplasy).

Prediction accuracy

Good predictions are not only the ones with high inference gain, but also the ones with high *accuracy*. As a matter of fact, the prediction that tomorrow will be a sunny day is better with 95% accuracy than with 50% accuracy. Accuracy corresponds to the ratio of correct predictions to the total number of cases evaluated. How accurate are the predictions about the genetic loci of evolution? Some appear to be 100% accurate whereas others are not. Pigmentation evolution in flies was predicted to involve mostly cis-regulatory mutations (CARROLL 2008). A recent review compiles 32 cases of pigmentation evolution in various *Drosophila* species: all of them are caused by cis-regulatory mutations and none are caused by coding changes (J.H. MASSEY and P.J. WITTKOPP 2016). If we suppose that existing approaches are not biased

towards cis-regulatory changes, then it means that so far the prediction for cis-regulatory evolution is 100% accurate. Resistance to trimethoprim in *Escherichia coli* was correctly predicted to be associated with genes encoding dihydrofolate reductase enzymes in 308 of the 320 tested resistant strains, thus giving an accuracy of 96% (A. BROLUND *et al.* 2010). An examination of 192 worldwide populations of *Arabidopsis thaliana* exhibiting natural variation in flowering time found that approximately 70% of the early-flowering strains carried deleterious mutations in the hotspot gene *FRIGIDA* (C. SHINDO *et al.* 2005). The prediction that variation in flowering time in *A. thaliana* should involve deleterious mutations in *FRIGIDA* has thus an estimated accuracy of 70%. In tetrapods, the *MC1R* receptor and its antagonist *Agouti* together account for 54% of the 206 published cases of pigmentation variation (MARTIN and ORGOGOZO 2017). This percentage is inflated by a “caveman effect”, which is a type of sampling bias: these two genes might be called “pigmentation hotspot genes” because there are precisely the loci researchers look at first when digging for genetic changes driving pigment variation. Nonetheless, the 54% value gives a maximum estimate of the accuracy of the prediction for a commonly studied trait.

The prediction that genes encoding signaling ligands should be responsible for morphological evolution is one of the less accurate predictions that have been presented. According to experimental data, about 20% of the cases where an animal morphological difference has been mapped to a gene involve a signaling ligand gene (80/391, MARTIN and ORGOGOZO 2017, www.gephebase.org). In sticklebacks, 14 genome-wide QTL studies ended with the identification of the causal gene and 4 of them identified a secreted ligand gene, thus giving an estimated accuracy of 28%. The ligand prediction is still better than the null model of each gene having an equal probability of being responsible for the phenotypic change, as ligand genes represent less than 5% of the total number of genes in a genome (MARTIN and ORGOGOZO 2017). A more accurate formulation of the ligand prediction is thus that for animal morphological evolution signaling ligand genes are over-represented compared to their proportion in genomes.

Interestingly, the accuracy of actual predictions does not appear to be determined by the strength of the arguments that have been proposed to substantiate them. A wealth of arguments have been proposed for the importance of cis-regulatory mutations in morphological evolution (CARROLL 2000; B. PRUD'HOMME, N. GOMPEL, S.B. CARROLL 2007; WRAY 2007) and

yet this prediction does not appear to be accurate for plant morphological evolution (M.A. STREISFELD and M.D. RAUSHER 2011).

Inference gain varies with phenotypes

Certain phenotypic traits are associated with predictions of high inference gain while other phenotypic changes call for low inference gain predictions. Predictions about metabolic activity or resistance to particular molecules appear to have more inference gain than predictions about morphological differences. For example, evolution of C4 photosynthesis can be associated with a few specific amino acid changes in the *PEPC* gene (P.A. CHRISTIN *et al.* 2007) and antifolate resistance in *Plasmodium falciparum* with only 6 mutations in the dihydrofolate reductase (*DHFR*) locus (M.S. COSTANZO and D.L. HARTL 2011).

In contrast, pigmentation pattern evolution can be caused by mutations in at least 10 genes in *Drosophila* flies and 13 genes in vertebrates (Table 3).

Table 3. Non-exhaustive list of genes responsible for evolution of pigmentation pattern in natural populations in Vertebrates and *Drosophila* flies. See www.gephebase.org for details

Vertebrates	<i>Drosophila</i> flies
<i>Agouti</i>	<i>bab1, bab2</i>
<i>Atrn</i>	<i>Dat</i>
<i>BCO2</i>	<i>Dll</i>
<i>CYP2J2</i>	<i>ebony</i>
<i>EIF2F2</i>	<i>omb</i>
<i>HERC2</i>	<i>pdm3</i>
<i>KITL</i>	<i>tan</i>
<i>MC1R</i>	<i>yellow</i>
<i>MGRN1</i>	<i>wg</i>
<i>Oca2</i>	
<i>SLC45A2</i>	
<i>Taqpep</i>	
<i>TPCN2</i>	

Importantly, additional data about the phenotypic change of interest can help narrow down the number of candidate genetic loci. Coat-darkening phenotypes in natural populations of vertebrates have been associated to only two of these pigmentation genes, the *Agouti* signaling protein (*Agouti*) and melanocortin-1 receptor (*Mc1r*). So these are the best candidates genes for natural coat-darkening phenotypes. Of note, spontaneous coat-darkening phenotypes in mice have also been associated with mutations in two other genes, *attractin* (*Atrn*), and *mahogunin* (*Mgrn*) (E.P. KINGSLEY *et al.* 2009). Moreover, knowing whether the coat-darkening phenotype is dominant or recessive reduces the number of candidate mutations further: a gain-of-function in *Mc1R* is inferred for dominant phenotypes and a loss-of-function in *Agouti* for recessive traits (E. EIZIRIK *et al.* 2003). The better characterized the phenotypic change, the more inference gain one can have. For example, pelvic-reduced sticklebacks were found to exhibit a left-right directional asymmetry of pelvic bones, as in *Pitx1*-null mice, thus strengthening the prediction that the underlying gene should be *Pitx1* (M.D. SHAPIRO *et al.* 2004, Y.F. CHAN *et al.* 2010).

Rockman and others have argued in favor of a polygenic model behind complex traits (R.A. FISHER 1918; M.V. ROCKMANN 2012; E.A. BOYLE, Y.I. LI and J.K. PRITCHARD 2017), suggesting that the genetic loci underlying certain complex traits may not be predictable. We do not doubt the generality of this view, but it is useful to stress that traits assumed to be multigenic are sometimes found to be oligogenic (S. MAKVANDI-NEJAD *et al.* 2012). Focusing on large-effect loci can facilitate the discovery and identification of some of the genetic causes of phenotypic variation. This said, it remains critical to look for alternative mutations and avoid the pitfall of ascertainment bias. The power of contemporary association mapping in naturally variable populations and of genome-wide genotyping for QTL analysis shall facilitate our escape from the low-hanging large-effect loci and help us draw a balanced view of genetic predictability in the next decades.

Does prediction accuracy vary with time scale and taxonomic range?

As the taxon of interest becomes more distant from the set of species from which genetic knowledge is available and from which predictions are elaborated, predictions may be less likely. For example, table 3 shows that

genes involved in pigmentation evolution in vertebrates cannot be used as candidate genes for pigmentation evolution in flies. C4 photosynthesis evolved many times independently in grasses and sedges through mutations in phosphoenolpyruvate carboxylase (PEPC) via a limited number of amino acid positions and the distribution of sites that have repeatedly mutated differ significantly between grasses and sedges, indicating that the genetic basis of C4 photosynthesis evolution is slightly different between taxa (CHRISTIN *et al.* 2007). A meta-analysis of ~25 cases (G.L. CONTE *et al.* 2012) suggested that probability of gene reuse declines with divergence time between the two taxa under consideration. However, this trend was not reproduced with a larger dataset (118 cases, Fig. 6 in J.R. GALLANT *et al.* 2014). Therefore, we cannot conclude from current data that independent evolution of the same phenotype is more likely to involve mutations in the same genetic locus when taxa are closely related than when they are distantly related.

Conclusion

Even though predictions about the loci of past evolution do not rely on advanced theoretical models, they have proved relatively accurate so far. Predicting the mutations of the past can help not only to understand the mechanisms of evolution, but also to genetically-engineer domesticated species and to infer the mutations that will occur in pathogenic microorganisms.

Acknowledgements

We thank S. Caianiello, S. Campanella, D. Ceccarelli, G. Frezza and E. Gagliasso for organizing the 2017 conference and for encouraging the writing of this review. We also thank A. Stoltzfus, G. Conte B. Morizot and A. Love for stimulating discussions and A. Matamoro-Vidal for comments on the manuscript. The research leading to this paper has received funding from the European Research Council under the European Community's Seventh Framework Program to VCO (FP7/2007-2013 Grant Agreement #337579) and from the John Templeton Foundation to AM and VCO (JTF award #43903).

References

- ANDERSON J.B., FUNT J., THOMPSON D.A., PRABHU S., SOCHA A., SIRJUSINGH C. *et al.* 2010, *Determinants of Divergent Adaptation and Dobzhansky-Muller Interaction in Experimental Yeast Populations*, in «Current Biology», 20(15), pp. 1383-1388.
- AVISE J.C., ROBINSON T.J. 2008, *Hemiplasy: A New Term in the Lexicon of Phylogenetics*, in «Systematic Biology», 57(3), pp. 503-507.
- BASS C., SCHROEDER I., TURBERG A., FIELD L.M., WILLIAMSON M.S. 2004, *Identification of the Rdl Mutation in Laboratory and Field Strains of the Cat Flea, Ctenocephalides felis (Siphonaptera: Pulicidae)*, in «Pest Management Science», 60(12), pp. 1157-1162.
- BLOOM J.D., GONG L.I., BALTIMORE D. 2010, *Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance*, in «Science», 328(5983), pp. 1272-1275.
- BOYLE E.A., LI Y.I. and PRITCHARD J.K. 2017, *An Expanded View of Complex Traits: From Polygenic to Omnigenic*, in «Cell», 169(7), pp. 1177-1186.
- BROLUND A., SUNDQVIST M., KAHLMEYER G., GRAPE M. 2010, *Molecular Characterisation Of Trimethoprim Resistance in Escherichia Coli and Klebsiella Pneumoniae During a Two Year Intervention On Trimethoprim Use*, in «PLoS One», 5(2):e9233.
- BRICELJ V.M., CONNELL L., KONOKI K., MACQUARRIE S.P. 2005, *Sodium Channel Mutation Leading To Saxitoxin Resistance in Clams Increases Risk of PSP*, in «Nature», 434(7034), pp. 763.
- CARROLL S.B. 2000, *Endless Forms: The Evolution of Gene Regulation and Morphological Diversity*, in «Cell», 101(6), pp. 577-580.
- CARROLL S.B. 2008, *Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution*, in «Cell», 134(1), pp. 25-36.
- CHAN Y.F., MARKS M.E., JONES F.C., VILLARREAL G., SHAPIRO M.D., BRADY S.D. *et al.* 2010, *Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer*, in «Science», 327(5963), pp. 302-305.
- CHRISTIN P.A., SALAMIN N., SAVOLAINEN V., DUVAL M.R., BESNARD G. 2007, *C4 Photosynthesis Evolved in Grasses via Parallel Adaptive Genetic Changes*, in «Current Biology», 17(14), pp. 1241-1247.
- CHRISTIN P.A., WEINREICH D.M., BESNARD G. 2010, *Causes and Evolutionary Significance of Genetic Convergence*, in «Trends in Genetics», 26(9), pp. 400-405.
- CONTE G.L., ARNEGARD M.E., PEICHEL C.L., SCHLUTER D. 2012, *The Probability of Genetic Parallelism and Convergence in Natural Populations*, in «Proceedings of the Royal Society of London B», 279(1749), pp. 5039-5047.

- COSTANZO M.S., HARTL D.L. 2011, *The Evolutionary Landscape of Antifolate Resistance in Plasmodium Falciparum*, in «Journal of genetics», 90(2), pp. 187-190.
- DARWIN C. 1859, *On the Origin of Species by Means of Natural Selection*, Murray, London.
- DAVIDSON E.H. 2006, *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*, Academic Press, San Diego.
- DUNNING L.T., LUNDGREN M.R., MORENO-VILLENA J.J., NAMAGANDA M., EDWARDS E.J., NOSIL P. et al. 2017, *Introgression and Repeated Co-Option Facilitated tRecurrent Emergence of C4 Photosynthesis Among Close Relatives*, in «Evolution», 71(6), pp. 1541-1555.
- ENCISO-ROMERO J., PARDO-DÍAZ C., MARTIN S.H., ARIAS C.F., LINARES M., McMILLAN W.O. et al. 2017, *Evolution of Novel Mimicry Rings Facilitated by Adaptive Introgression in Tropical Butterflies*, in «Molecular Ecology». doi: 10.1111/mec.14277. Epub ahead of print.
- EIZIRIK E., YUHKI N., JOHNSON W.E., MENOTTI-RAYMOND M., HANNAH S.S., O'BRIEN S.J. 2003, *Molecular Genetics and Evolution of Melanism in The Cat Family*, in «Current Biology», 13(5), pp. 448-453.
- FAN Z., ORTEGA-DEL VECCHYO D., WAYNE R.K. 2017, *EPAS1 Variants in High Altitude Tibetan Wolves Were Selectively Introgressed Into Highland Dogs*, in «PeerJ», 5:e3522.
- FETZER J. 2007, *Carl Hempel*, The Stanford Encyclopedia of Philosophy (Fall 2017 Edition), ed. E.N. Zalta, <https://plato.stanford.edu/archives/fall2017/entries/hempel/>.
- FISHER R.A. 1918, *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*, in «Transactions of the Royal Society Edinburgh», 52, pp. 399-433.
- FRANKEL N., DAVIS G.K., VARGAS D., WANG S., PAYRE F., STERN D.L. 2010, *Phenotypic Robustness Conferred by Apparently Redundant Transcriptional Enhancers*, in «Nature», 466(7305), pp. 490-3.
- FRANKEL N., WANG S., STERN D.L. 2012, *Conserved Regulatory Architecture Underlies Parallel Genetic Changes and Convergent Phenotypic Evolution*, in «Proceedings of the National Academy of Sciences», 109(51), pp. 20975-20979.
- GALEN S.C., NATARAJAN C., MORIYAMA H., WEBER R.E., FAGO A., BENHAM P.M. et al. 2015, *Contribution of a Mutational Hot Spot To Hemoglobin Adaptation in High-Altitude Andean House Wrens*, in «Proceedings of the National Academy of Sciences», 112(45), pp. 13958-13963.
- GALLANT J.R., IMHOFF V.E., MARTIN A., SAVAGE W.K., CHAMBERLAIN N.L., POTE B.L. et al. 2014, *Ancient Homology Underlies Adaptive Mimetic Diversity Across Butterflies*, in «Nature Communications», 5:4817.

- GOMPEL N., PRUD'HOMME B. 2009, *The Causes of Repeated Genetic Evolution*, in «Developmental Biology», 332(1), pp. 36-47.
- HOEKSTRA H.E., COYNE J.A. 2007, *The Locus Of Evolution: Evo Devo and the Genetics of Adaptation*, in «Evolution», 61(5), pp. 995-1016.
- INGRAM V.M. 1957, *Gene Mutations in Human Haemoglobin: The Chemical Difference Between Normal and Sickle Cell Haemoglobin*, in «Nature», 180(4581), pp. 326-328.
- KOPP A. 2009, *Metamodels and Phylogenetic Replication: A Systematic Approach to the Evolution of Developmental Pathways*, in «Evolution», 63(11), pp. 2771-2789.
- LÄSSIG M., MUSTONEN V., WALCZAK A.M. 2017, *Predicting Evolution*, in «Nature Ecology and Evolution», 1:0077.
- LENSER T., THEISSEN G. 2013, *Molecular Mechanisms Involved in Convergent Crop Domestication*, in «Trends in Plant Science», 18(12), pp. 704-714.
- LENSKI R.E. 2017, *Convergence and Divergence in a Long-Term Experiment with Bacteria*, in «The American Naturalist», 190(S1), pp. S57-S68.
- LI F.W., VILLARREAL J.C., KELLY S., ROTHFELS C.J., MELKONIAN M., FRANGEDAKIS E. et al. 2014, *Horizontal Transfer of an Adaptive Chimeric Photoreceptor From Bryophytes To Ferns*, in «Proceedings of the National Academy of Sciences», 111(18), pp. 6672-6677.
- LIAO B.-Y., WENG M.-P., ZHANG J. 2010, *Contrasting Genetic Paths to Morphological and Physiological Evolution*, in «Proceedings of the National Academy of Sciences», 107, pp. 7353-7358.
- LINNEN C.R., KINGSLEY E.P., JENSEN J.D., HOEKSTRA H.E. 2009, *On the Origin and Spread of an Adaptive Allele in Deer Mice*, in «Science», 325(5944), pp. 1095-1098.
- LOBKOVSKY A.E., KOONIN E.V. 2012, *Replaying the Tape of Life: Quantification of the Predictability of Evolution*, in «Frontiers in Genetics», 3:246.
- LONG A., LITI G., LUPTAK A., TENAILLON O. 2015, *Elucidating the Molecular Architecture of Adaptation via Evolve and Resequencing Experiments*, in «Nature Reviews. Genetics», 16(10):567.
- LYNCH V.J., WAGNER G.P. 2008, *Resurrecting the Role of Transcription Factor Change in Developmental Evolution*, in «Evolution», 62(9), pp. 2131-2154.
- LYNCH M. and WALSH B. 2007, *The Origins of Genome Architecture*, Sinauer Associates, Sunderland.
- MANSON A.L., COHEN K.A., ABEEL T., DESJARDINS C.A., ARMSTRONG D.T., Barry III C.E. et al. 2017, *Genomic analysis of globally diverse Mycobacterium Tuberculosis Strains Provides Insights Into Emergence and Spread of Multidrug Resistance*, in «Nature genetics», 49(3):395.

- MAKVANDI-NEJAD S., HOFFMAN G.E., ALLEN J.J., CHU E., GU E., CHANDLER A.M. *et al.* 2012, *Four Loci Explain 83% of Size Variation in the Horse*, in «PLoS One», 7(7):e39929.
- MALFAVON-BORJA R., WU L.I., EMERMAN M., MALIK H.S. 2013, *Birth, Decay, and Reconstruction of an Ancient TRIMCyp Gene Fusion in Primate Genomes*, in «Proceedings of the National Academy of Sciences», 110(7), pp. E583-E592.
- MARTIN A., PAPA R., NADEAU N.J., HILL R.I., COUNTERMAN B.A., HALDER G. *et al.* 2012, *Diversification of Complex Butterfly Wing Patterns by Repeated Regulatory Evolution of a Wnt Ligand*, in «Proceedings of the National Academy of Sciences», 109(31), pp. 12632-12637.
- MARTIN A., ORGOGOZO V. 2013, *The Loci of Repeated Evolution: A Catalog of Genetic Hotspots of Phenotypic Variation*, in «Evolution», 67(5), pp. 1235-1250.
- MASSEY J.H., WITTKOPP P.J. 2016, *The Genetic Basis of Pigmentation Differences Within and Between Drosophila Species*, in «Current Topics in Developmental Biology», 119, pp. 27-61.
- MCGREGOR A.P., ORGOGOZO V., DELON I., ZANET J., SRINIVASAN D.G., PAYRE F., STERN D.L. 2007, *Morphological Evolution Through Multiple Cis-Regulatory Mutations at a Single Gene*, in «Nature», 448(7153), pp. 587-90.
- METCALF J.A., FUNKHOUSER-JONES L.J., BRILEYA K., REYSENBACH A.L., BORDENSTEIN S.R. 2014, *Antibacterial Gene Transfer Across the Tree Of Life*, in «eLife», 3:e04266.
- MILLER C.T., BELEZA S., POLLEN A.A., SCHLUTER D., KITTLES R.A., SHRIVER M.D., KINGSLEY D.M. 2007, *cis-Regulatory Changes in Kit Ligand Expression and Parallel Evolution of Pigmentation in Sticklebacks and Humans*, in «Cell», 131(6), pp. 1179-1189.
- MORRIS S.C. 2003, *Life's Solution: Inevitable Humans in a Lonely Universe*, Cambridge University Press, Cambridge (MA).
- NATARAJAN C., HOFFMANN F.G., WEBER R.E., FAGO A., WITT C.C., STORZ J.F. 2016, *Predictable Convergence in Hemoglobin Function has Unpredictable Molecular Underpinnings*, in «Science», 354(6310), pp. 336-339.
- OMELCHENKO M.V., GALPERIN M.Y., WOLF Y.I., KOONIN E.V. 2010, *Non-Homologous Isofunctional Enzymes: A Systematic Analysis of Alternative Solutions in Enzyme Evolution*, in «Biology Direct», 5(1):31.
- ORGOGOZO V. 2015, *Replaying the Tape of Life in the Twenty-First Century*, in «Interface Focus», 5(6): 20150057.
- ORGOGOZO V., MORIZOT B., MARTIN A. 2015, *The Differential View Of Genotype-Phenotype Relationships*, in «Frontiers in Genetics», 6:179.

- PATERSON A.H., LIN Y.R., LI Z., SCHERTZ K.F., DOEBLEY J.F., PINSON S.R. *et al.* 1995, *Convergent Domestication of Cereal Crops by Independent Mutations at Corresponding Genetic Loci*, in «Science», pp. 1714-1718.
- PRUD'HOMME B., GOMPEL N., CARROLL S.B. 2007, *Emerging Principles of Regulatory Evolution*, in «Proceedings of the National Academy of Sciences», 104(suppl 1), pp. 8605-8612.
- ROCKMAN M.V. 2012, *The QTN Program and the Alleles That Matter For Evolution: All That's Gold Does Not Glitter*, in «Evolution», 66(1), pp. 1-17.
- ROPARS J., DE LA VEGA R. C.R., LÓPEZ-VILLAVICENCIO M., GOUZY J., SALLET E., DUMAS É. *et al.* 2015, *Adaptive Horizontal Gene Transfers Between Multiple Cheese-Associated Fungi*, in «Current Biology», 25(19), pp. 2562-2569.
- SHAPIRO M.D., MARKS M.E., PEICHEL C.L., BLACKMAN B.K. 2004, *Genetic and Developmental Basis of Evolutionary Pelvic Reduction in Threespine Sticklebacks*, in «Nature», 428(6984), pp. 717-23.
- SHINDO C., ARANZANA M.J., LISTER C., BAXTER C., NICHOLLS C., NORDBORG M., DEAN C. 2005, *Genetic and Developmental Basis of Evolutionary Pelvic Reduction in Threespine Sticklebacks*, in «Plant Physiology», 138(2), pp. 1163-1173.
- SODERLUND D.M. 2008, *Genetic and Developmental Basis of Evolutionary Pelvic Reduction in Threespine Sticklebacks*, in «Pest Management Science», 64(6), pp. 610-616.
- STARR T.N., PICTON L.K., THORNTON J.W. 2015, *Genetic and Developmental Basis of Evolutionary Pelvic Reduction in Threespine Sticklebacks*, in «PLoS Genet», 11:e1005206.
- STERN D.L. 2011, *Evolution, Development, & the Predictable Genome*, Roberts and Co. Publishers.
- STERN D.L. 2013, *The Genetic Causes of Convergent Evolution*, in «Nature Reviews Genetics», 14(11), pp. 751-764.
- STERN D.L., ORGOGOZO V. 2008, *The Loci of Evolution: How Predictable is Genetic Evolution?*, in «Evolution», 62(9), pp. 2155-2177.
- STERN D.L., ORGOGOZO V. 2009, *Is Genetic Evolution Predictable?*, in «Science», 323(5915), pp. 746-751.
- STORZ J.F. 2016, *Causes of Molecular Convergence and Parallelism in Protein Evolution*, in «Nature Reviews Genetics», 17(4), pp. 239-50.
- STOLTZFUS A., McCANDLISH D.M. 2015, *Mutation-Biased Adaptation in Andean House Wrens*, in «Proceedings of the National Academy of Sciences», 112(45), pp. 13753-13754.
- STOLTZFUS A., McCANDLISH D.M. 2017, *Mutational Biases Influence Parallel Adaptation*, in «Molecular Biology and Evolution», 34(9), pp. 2163-2172.

- STREISFELD M.A., RAUSHER M.D. 2011, *Population Genetics, Pleiotropy, and The Preferential Fixation of Mutations During Adaptive Evolution*, in «Evolution», 65(3), pp. 629-642.
- TARVIN R.D., BORGHESI C.M., SACHS W., SANTOS J.C., LU Y., O'CONNELL L.A. et al. 2017, *Interacting Amino Acid Replacements Allow Poison Frogs To Evolve Epibatidine Resistance*, in «Science», 357(6357), pp. 1261-1266.
- TOLEDO G., HANIFIN C., GEFFENEY S., BRODIE E.D. 2016, *Convergent Evolution of Tetrodotoxin-Resistant Sodium Channels in Predators and Prey*, in «Current Topics in Membranes», 78, pp. 87-113.
- WAGNER G.P., LYNCH V.J. 2008, *The Gene Regulatory Logic of Transcription Factor Evolution*, in «Trends in Ecology and Evolution», 23(7), pp. 377-385.
- WESSINGER C.A., RAUSHER M.D. 2014, *Predictability and Irreversibility of Genetic Changes Associated With Flower Color Evolution in *Penstemon Barbatus**, in «Evolution», 68(4), pp. 1058-1070.
- WESTON D.P., POYNTON H.C., WELLBORN G.A., LYDY M.J., BLALOCK B.J., SEPULVEDA M.S. et al. 2013, *Multiple Origins of Pyrethroid Insecticide Resistance Across the Species Complex of a Nontarget Aquatic Crustacean, *Hyaella Azteca**, in «Proceedings of the National Academy of Sciences», 110(41), pp. 16532-16537.
- WALLBANK R.W., BAXTER S.W., PARDO-DIAZ C., HANLY J.J., MARTIN S.H., MALLET J. et al. 2016, *Evolutionary Novelty in a Butterfly Wing Pattern Through Enhancer Shuffling*, in «PLoS Biology», 14(1):e1002353.
- WRAY G.A. 2007, *The Evolutionary Significance of Cis-Regulatory Mutations*, in «Nature Reviews Genetics», 8(3), pp. 206-216.
- YAMAMOTO F.I., CLAUSEN H., WHITE T., MARKEN J., HAKOMORI S.I. 1990, *Molecular Genetic Basis of the Histo-Blood Group ABO System*, in «Nature», 345(6272), pp. 229-233.
- ZHANG Z., HARRISON P.M., LIU Y., GERSTEIN M. 2003, *Millions of Years of Evolution Preserved: A Comprehensive Catalog of the Processed Pseudogenes in the Human Genome*, in «Genome Research», 13(12), pp. 2541-2558.